Projecto LANDAU - Metodologia geoestatística para a caracterização da incerteza espacial e actualização de cartas de ocupação do solo

Classificação uni-temporal da ocupação do solo através de dados de média resolução espacial - MERIS

Action 3.2 - Task 3

Relatório de execução

Joel Dinis, Pedro Rodrigues, Márcia Gonçalves, Rita Nicolau, Rui Reis

Abril 2012

ÍNDICE

1 Introdução 5

2 Métodos 5

2.1 Dados e área de aplicação 5

2.2 Nomenclatura 6

2.3 Algoritmos de classificação ensaiados 6

2.3.1 Amostragem destinada ao treino dos algoritmos de classificação 15

2.4 Validação dos mapas produzidos 15

2.4.1 Amostragem destinada à validação 16

3 Resultados 17

4 Conclusões 22

5 Referências Bibliográficas 22

ANEXO I – Matrizes de erro / confusão associadas à classificação da ocupação do solo, com base em dados MERIS 25

1 Introdução

O presente relatório descreve as actividades desenvolvidas no âmbito da *Action 3.2* - Task 3 do projecto LANDAU. As actividades em causa visaram a classificação automática da ocupação do solo em Portugal Continental a partir de dados de satélite de média resolução espacial (imagem MERIS), de acordo com a nomenclatura de ocupação do solo LANDAU.

Na produção de mapas de ocupação de solo a partir de dados MERIS testaram-se dez metodologias de classificação automática distintas, sendo objecto do presente estudo a avaliação do desempenho de tais metodologias.

2 Métodos

No presente ponto descrevem-se os dados, a nomenclatura e as metodologias empregues na classificação automática da ocupação do solo a partir de dados de satélite.

Grande maioria dos algoritmos vocacionados para classificação automática da ocupação do solo, carece de um processo de treino com vista à aprendizagem. Para o desenvolvimento deste processo é necessário proceder à recolha de amostras de treino. Findo o processo de treino do classificador, inicia-se a classificação da ocupação de solo propriamente dita. Os mapas gerados neste contexto são seguidamente submetidos a um processo de validação, que é desenvolvido com base em amostras (de teste ou de validação) que são necessariamente distintas das amostras utilizadas no treino dos classificadores.

2.1 Dados e área de aplicação

De entre as três escalas de análise previstas pelo projecto LANDAU (média, elevada e muito elevada resolução espacial), no presente trabalho utilizou-se a proporcionada por dados do sensor MERIS (relativos a Maio de 2005). A informação correspondente é a menos detalhada das três previstas (pixeis de 300 m) e consequentemente a mais restritiva para a classificação da ocupação do solo. Quando aplicada às três áreas de estudo previamente seleccionadas para o projecto, esta escala de análise impossibilitaria o reconhecimento e classificação de todas as classes de espaço de interesse para o projecto, tendo-se por isso optado por alargar o presente estudo a Portugal Continental.

A informação de referência utilizada para apoiar a recolha de amostras de treino dos

classificadores e de amostras de validação da cartografia produzida, foi a seguinte:

- Imagens aéreas orto-rectificadas, com uma resolução espacial de 50 cm e uma resolução espectral de 4 bandas (1995, 2005, 2007);
- Inventário Florestal – IF (2005);
- Cartografia CORINE Land Cover (2000, 2006).

2.2 Nomenclatura

A nomenclatura de ocupação do solo adoptada no presente trabalho foi anteriormente proposta no contexto do projecto, sendo por isso designada de LANDAU. A sua proposta emergiu da necessidade de obtenção de uma nomenclatura compatibilizada com o sistema de classificação Land Cover Classification System (LCCS), que fosse funcional às múltiplas escalas de análise utilizadas no projecto. Trata-se de uma nomenclatura que na sua versão mais desagregada inclui 15 classes de ocupação/uso do solo (vide Quadro 1).

Quadro 1 – Nomenclatura LANDAU

LANDAU – Nível 1	LANDAU – Nível 2
1 Artificial Areas (Territórios Artificializados)	1.1 Continuous Artificial Areas (Áreas Artificiais Contínuas)
	1.2 Discontinuous Artificial Areas (Áreas Artificiais Descontínuas)
2 Croplands (Áreas Agrícolas)	2.1 Irrigated Agriculture (Agricultura de Regadio)
	2.2 Non-irrigated Agriculture (Agricultura de Sequeiro)
	2.3 Rice Crops (Arrozais)
3 Natural and Semi-natural Vegetated Areas (Florestas e Meios Naturais e Semi-naturais)	3.1 Broadleaved Forest (Floresta de Folhosas)
	3.2 Coniferous Forest (Floresta de Resinosas)
	3.3 Mixed Forest (Floresta Mista)
	3.4 Grassland (Vegetação Herbácea)
	3.5 Shrubland (Matos)
	3.6 Baren to Sparsely Vegetated Areas (Vegetação Esparsa)
4 Bare Land (Solo Nu)	4 Bare Land (Solo Nu)
5 Burnt Areas (Áreas Ardidas)	5 Burnt Areas (Áreas Ardidas)
6 Wetlands (Zonas Húmidas)	6 Wetlands (Zonas Húmidas)
7 Water bodies (Corpos de Água)	7 Water bodies (Corpos de Água)

2.3 Algoritmos de classificação ensaiados

Os algoritmos classificação ensaiados no presente estudo foram os seguintes: Maximum Likelihood (ML), Linear Discriminant Classifier (LDC), Diagonal Quadratic Discriminant Classifier (DQDC), Minimum Distance (MD), K-Nearest Neighbours (KNN), Parzen Classifier (PARZEN), Classification and Regression Tree (CART), Support Vector Machine (SVM), Backpropagation Multilayer Perceptron (BMP) e Iterative Self-Organizing Data Analysis Technique (ISODATA).

O Quadro 2 subdivide os algoritmos referidos segundo as suas propriedades e modo operação de forma a possibilitar uma apresentação coerente dos mesmos.

Quadro 2 – Tipos de algoritmos

Tipo de Classificador	Algoritmo
Paramétricos	ML, LDC, DQDC, MD
Geométricos ou "Preguiçosos"	KNN, PARZEN
Baseados em regras de decisão	CART
Optimização não probabilística	SVM, BMP
Não assistidos	ISODATA

Classificadores paramétricos

Os classificadores paramétricos admitem a hipótese da normalidade dos dados, ou seja que o comportamento espectral de cada classe pode ser modelado por uma distribuição multivariada normal. Assim, estes algoritmos admitem a existência de um vector médio e de uma matriz variância-covariância para cada classe definida na fase de treino. Deste modo, é possível deduzir a equação da função de discriminação de um classificador paramétrico, recorrendo à função de densidade de probabilidade normal multivariada (Kuncheva, 2004):

$$g_i(x) = \log p_i - \frac{1}{2}\log\det(\Sigma_i) - \frac{1}{2}(x-\mu_i)^T\Sigma_i^T(x-\mu_i) \quad (1)$$

Onde x é o vector a ser classificado, μ_i é o vector médio da classe i, Σ_i a sua matriz de variância-covariância e p_i a probabilidade *à priori* associada à classe i. A função de discriminação apresentada na equação (1) define fronteiras de decisão quadráticas, i.e. hiperparaboloides de separabilidade, motivo pelo qual o classificador que a utiliza no processo de classificação é designado por *Quadratic Discriminant Classifier*, sendo mais usualmente conhecido por Classificador de Máxima Verosimilhança (adiante designado por ML).

Na prática, os vectores médios e as matrizes de variância-covariância não são conhecidos, pelo que é necessário estimá-los. Para tal, recorre-se aos respectivos estimadores de máxima verosimilhança:

$$\hat{\mu}_i = \frac{1}{N_i}\sum_{w\in i} x_w \quad (2)$$

$$\hat{\Sigma}_i = \frac{1}{N_i - 1}\sum_{w\in i}(x_w - \hat{\mu}_i)^T(x_w - \hat{\mu}_i) \quad (3)$$

Deste modo, o processo de classificação recorrendo à função de discriminação quadrática pode ser resumido na seguinte equação:

$$l(x) = argmax_w \left\{ \log p_w - \frac{1}{2} \log \det(\hat{\Sigma}_w) - \frac{1}{2} (x - \hat{\mu}_w)^T \hat{\Sigma}_w^T (x - \hat{\mu}_w) \right\} \quad (4)$$

onde o *label* (código numérico representativo da classe) a atribuir ao *pixel* x será igual ao da classe que maximize a função de discriminação quadrática.

Impondo a restrição de que todas as classes apresentem igual padrão de variabilidade (hipótese da homocedasticidade), então a equação (1) pode ser simplificada através da remoção do logaritmo da matriz de variância-covariância, por este não ser discriminativo (Kuncheva, 2004):

$$g_i(x) = \log p_i - \frac{1}{2} (x - \mu_i)^T \Sigma^{-1} (x - \mu_i) \quad (5)$$

A função de discriminação apresentada na equação (5) define fronteiras de decisão lineares, ou seja hiperplanos de separabilidade, pelo que o classificador que a utiliza é conhecido por *Linear Discriminant Classifier* (adiante designado por LDC).

Na prática, a matriz de variância-covariância é determinada através da média ponderada das matrizes de variância-covariância estimadas para cada classe, utilizado como ponderador a probabilidade *à priori* de cada classe:

$$\hat{\Sigma} = \sum_i p_i \Sigma_i \quad (6)$$

A regra de classificação é análoga à do ML:

$$l(x) = argmax_W \left\{ \log p_W - \frac{1}{2} (x - \hat{\mu}_W)^T \hat{\Sigma}^T (x - \hat{\mu}_W) \right\} \quad (7)$$

Finalmente, torna-se ainda possível impor uma restrição adicional a cada uma das funções de discriminação: a da independência entre dimensões. Esta restrição implica que cada par de dimensões distintas do espaço de classificação sejam independentes entre si, i.e.:

$$\sigma_{ij} = 0, se\ i \neq j \quad (8)$$

Assim, as matrizes de variância-covariância, quer as estimadas quer a ponderada, são matrizes diagonais. Por este motivo, o classificador que utiliza a regra de classificação quadrática, a par da imposição de independência entre dimensões é conhecido por *Diagonal Quadratic Discriminant Classifier* (adiante designado por DQDC).

De modo similar, o classificador que utiliza a regra de classificação linear a par da imposição da independência entre dimensões é conhecido por *Diagonal Linear Discriminant Classifier*, que para fins de classificação o torna equivalente ao classificador da mínima distância (*Minimum Distance Classifier,* adiante designado por MD).

Classificadores geométricos ou "preguiçosos"

Os classificadores designados de geométricos (Kuncheva, 2004) são também conhecidos como classificadores "preguiçosos" (*"Lazy classifiers"*; Hastie *et al*, 2009), por não realizarem nenhuma síntese da amostra de treino. Ao contrário dos restantes, estes classificadores não processam as unidades de treino para "aprenderem" a reconhecer os padrões das classes, mas utilizam-nas directamente comparando-as com o objecto (no presente estudo, o pixel) a ser classificado.

O classificador "preguiçoso" mais recorrente é o classificador dos K vizinhos mais próximos (*K-Nearest Neighbors* - adiante designado por KNN). A "proximidade" é definida segundo uma função de distância, sendo a distância euclidiana a mais frequentemente utilizada (Hastie *et al*, 2009). Demonstra-se (Kuncheva, 2004) que a probabilidade de um objecto *x* do espaço de classificação (pixel) pertencer a uma classe *w* da nomenclatura é:

$$p(w_i|x) \approx k_i w/k \tag{9}$$

onde k_w é o número de unidades de treino pertencentes à classe *w* nas *k* unidades mais próximas de *x*. Deste modo, a regra de classificação para o classificador KNN é dada por (Kuncheva, 2004):

$$l(x) = argmax_w \left\{\frac{k_w}{k}\right\} \tag{10}$$

Na prática a classificação de x é realizada em três passos: 1) calculo da distância de *x* a cada unidade de treino; 2) selecção das *k* unidades de treino mais próximas; 3) classificação de *x* por meio da atribuição do *label* mais frequente.

O segundo classificador "preguiçoso" ensaiado foi o classificador de Parzen (adiante designado por PARZEN). Este classificador recorre a uma função de discriminação de *kernel*, sendo a função *kernel* gaussiana a mais frequentemente utilizada. Demonstra-se (Kuncheva, 2004) que a função de discriminação do classificador de Parzen é dada por:

$$g_i(x) = \sum_{u \in i} \exp\left\{-\frac{(x-u)^T S^{-1}(x-u)}{2h^2}\right\} \quad (11)$$

onde *u* é uma unidade de treino, *h* é um parâmetro que define a altura do *kernel* e S é uma matriz semelhante à matriz de variância-covariância que define a forma do *kernel*. Na prática, a definição do parâmetro h e S é difícil (Kuncheva, 2004) pelo que o parâmetro h é normalmente definido empiricamente e o parâmetro S é definido pela matriz identidade. Deste modo, a regra de classificação adoptada pelo PARZEN pode ser definida do seguinte modo:

$$l(x) = \operatorname{argmax}_w \left\{\sum_{u \in i} \exp\left\{-\frac{(x-u)^T (x-u)}{2h^2}\right\}\right\} \quad (12)$$

Classificadores baseados em regras de decisão

Os classificadores baseados em regras de decisão (do inglês: rule based classifiers) apresentam uma estrutura em árvore, pelo que são usualmente conhecidos como classificadores em árvore ou árvores de classificação. Existem inúmeras implementações deste tipo de classificadores. No presente estudo foi implementado o algoritmo Classification And Regression Tree, abreviadamente designado de CART, tal como apresentado em Kuncheva (2004). Os classificadores baseados em árvores de decisão procuram encontrar formas de dividir sucessivamente o universo em vários subconjuntos, através do desenvolvimento de regras de decisão sucessivas (testes) sobre valores de atributos da variável de interesse. Este procedimento é realizado através de uma árvore composta por nós (que indicam testes), por ramos (ou arcos que indicam decisões resultantes dos testes) e por folhas ou nós terminais (que representam classes). As folhas são identificadas na sequência de um ou mais testes cujo resultado aponte para uma única classe ou para uma classe maioritária. A classificação consiste em seguir o caminho ditado pelos sucessivos testes colocados ao longo da árvore até que seja encontrada uma folha que indica a classe a atribuir ao pixel que pretendemos classificar. O algoritmo CART fundamenta-se numa partição binária recursiva, produzindo uma árvore que pode ser percorrida da sua raiz até às folhas respondendo apenas a questões simples do tipo sim/não.

Existem três passos fundamentais na construção de uma árvore de classificação: o cálculo de uma medida de impureza que se pretende minimizar, o cálculo do ponto de divisão da árvore (um nó) e a decisão de subdividir ou não um nó da árvore.

No presente estudo, o cálculo da impureza baseou-se na medida de Gini (Kuncheva, 2004). De acordo com a medida de Gini, a impureza (*G*) num nó será dada por:

$$G = 1 - \sum_j P_j^2 \tag{13}$$

onde *P* representa a proporção de elementos da classe j no nó. A impureza varia entre zero e $1 - {}^{1}/_{c}$, onde c representa o número de classes da nomenclatura. O valor mínimo de impureza será zero se todas a unidades de treino existentes num nó pertencerem à mesma classe. O valor máximo ocorre quando existe igual número de unidades de treino do nó.

O cálculo do ponto de divisão da árvore foi realizado resolvendo o seguinte problema de optimização:

$$(\alpha, \beta) = \ argmax_{(\alpha,v)} \{G - \ w_L(u,v)g_L(u,v) - w_R(u,v)g_R(u,v)\} \tag{14}$$

onde α é a dimensão do ponto de divisão, β é o valor do ponto de divisão, *G* é o valor de impureza global no nó, w_L é a proporção de unidades de treino à esquerda do ponto de corte do espaço de classificação, w_R é a proporção de unidades de treino à direita do ponto de corte do espaço de classificação, g_L é o valor de impureza à esquerda do ponto de corte do espaço de classificação e g_R é o valor de impureza à direita do ponto de corte do espaço de classificação.

No presente estudo, a decisão de subdividir ou não um nó da árvore fundamentou-se no resultado de um teste de hipóteses baseado na distribuição Qui-quadrado. A estatística deste teste corresponde ao valor médio de duas variáveis (χ_L^2 e χ_R^2), que contabilizam as unidades de treino de classes da nomenclatura que se encontram à esquerda e à direita de cada nó. A equação (15) explicita a contabilização efectuada à esquerda de cada nó (χ_L^2), que é equivalente à efectuada à direita de cada nó (χ_R^2) (Kuncheva, 2004) :

$$\chi_L^2 = \sum_{i=1}^{c} \frac{(NN_{Li} - N_L N_i)^2}{NN_L N_i} \tag{15}$$

onde c é o número de classes na nomenclatura, N é o número de unidades de treino no nó, N_L é o número de unidades de treino à esquerda do nó, N_{Li} é o número de unidades de treino à esquerda do nó pertencentes à classe i, e N_i é o número de unidades de treino da classe i no nó.

A decisão de subdividir ou não um nó da árvore pode ser tomada por comparação da estatística acima mencionada com o valor tabelado (numa distribuição Qui-quadrado com c–1 graus de liberdade), ou alternativamente, por comparação da estatística com um valor fornecido pelo utilizador, valor este que deverá variar entre zero e dez (Kuncheva, 2004). Nesta gama de variação, valores próximos de zero tendem a criar árvores de pequena extensão (curtas) e valores próximos de dez tendem a criar árvores longas. O nó será subdivido em dois se o valor da estatística for superior ao valor tabelado (ou ao fornecido pelo utilizador). Caso contrário, o nó é qualificado como uma folha (nó terminal) sendo-lhe atribuído o *label* da classe mais frequente.

Uma vez construída a árvore é usual proceder à sua podagem (*prunning*) para reduzir os possíveis efeitos de sobre-ajustamento (do inglês: *over-fitting*; Hastie et al., 2009) e assim aumentar o seu poder de generalização. Contudo, esse procedimento é computacionalmente exigente e difícil de optimizar (Kuncheva, 2004). Na implementação adoptada neste estudo a podagem foi realizada durante a construção da árvore (*pre-prunning*), ajustando para tal o valor que controla a decisão de divisão da árvore.

Classificadores de optimização não probabilística

Nesta secção foram incluídos dois tipos de classificadores: o *Suport Vector Machine* (adiante designado por SVM) e o *Backpropagation Multilayer Perceptron* (adiante designado por BMP). Estes dois classificadores foram incluídos na mesma secção porque ambos tentam identificar, de forma não probabilística, fronteiras de separabilidade óptima.

O algoritmo SVM baseia-se na utilização de uma superfície de decisão para separar as classes de espaço, maximizando a sua separação. Tal superfície é um hiper-plano óptimo. Os pontos que confinam com a fronteira deste hiper-plano são designados de vectores de suporte (*support vectors*), constituindo objectivo da fase de treino a sua identificação. Usualmente as superfícies de separação são lineares, porém é possível definir superfícies não lineares. A natureza geométrica destes superfícies é definida pela função kernel utilizada. No presente estudo, utilizou-se a função gaussiana radial:

$$K(x, y) = \exp(-\alpha \|x - y\|^2) \tag{16}$$

onde o *x* e *y* são vectores do espaço de classificação e o parâmetro α é um escalar estabelecido pelo utilizador. O procedimento para a definição das superfícies de separação é complexo e baseia-se no algoritmo de Plat (1998). O processo de treino

do SVM é computacionalmente exigente pelo que nem sempre é possível trabalhar com dados de treino muito volumosos. No presente estudo, foi necessário definir uma subamostra de treino. Contudo, uma vez treinado, o SVM é rápido no processo de classificação.

Como descrito por Fonseca (1994), "no seu exemplo mais típico, uma rede neuronal artificial é formada por três tipos de elementos a que se dá o nome de neurónios: os neurónios da camada de entrada que introduzem os dados na rede, os neurónios das camadas internas e os neurónios da camada de saída que produzem o resultado. Cada ligação entre neurónios possui um peso, tendo cada neurónio, em geral, um nível de patamar que influência directamente a sua saída. O valor de excitação de cada neurónio é calculado efectuando a soma das suas várias entradas pesadas pelos respectivos coeficientes (pesos) ao que é somado o seu valor de patamar.
A resposta global da rede é ditada pelos valores apresentados pelos neurónios da camada de saída. Normalmente associa-se a cada neurónio da camada de saída uma dada classe e o objectivo é que a rede, na presença de um dado exemplo, active a saída correspondente à classe a que esse exemplo pertence. A aprendizagem das redes neuronais consiste então em ajustar os pesos e os valores de patamar de cada um dos neurónios da rede, de forma a que esta tenha a capacidade de classificar correctamente todos os exemplos presentes no conjunto de treino.

Segundo o método de classificação BMP, os exemplos do conjunto de treino são mostrados um a um à rede, sendo calculada a diferença entre a saída obtida e a desejada. Esse valor de erro é então propagado da camada de saída para a camada de entrada sendo efectuado um ajuste nos valores dos pesos e dos patamares dos neurónios das diversas camadas tendente à diminuição do erro obtido usando para tal um método de gradiente." (Fonseca, 1994)

O neurónio é no fundo uma função com a seguinte expressão analítica:

$$v = \phi\left(w + \sum_{i=0}^{n} w_i x_i\right) \tag{17}$$

onde ϕ é a função da activação (neste estudo a função sigmoide), o vector $[w, w_0, \quad w_1, \ldots, w_n]$ é um vector de pesos, *x* é o vector de *input* e *v* é valor de activação. Atendendo a que alguns estudos demonstram que três camadas são suficientes para criar um BMP suficientemente preciso para a classificação (Schowengerdt, 1997), O BMP adoptado neste estudo é composto por três camadas: a

camada inicial ou de entrada, constituída por tantas entradas quanto o número de dimensões em consideração, uma camada interna formada por 20 neurónios e uma camada de saída constituída por tantos neurónios quantas as classes. O número de neurónios na camada interior foi definido experimentalmente. O *label* a ser atribuído durante o processo de classificação é definido pelo índice do neurónio de máxima activação. Assim, por exemplo, se o neurónio nº 1 na camada de saída for o neurónio com valor máximo de activação, então o objecto recebe o label da primeira classe da nomenclatura.

O processo de classificação no BMP realiza-se do seguinte modo (Kuncheva, 2004): Os pesos das ligações entre os neurónios são aleatoriamente inicializados. Seguidamente, a primeira unidade de treino é "empurrada" para o BMP e são calculados os valores de activação na camada interior e na camada de saída. Posteriormente procede-se à avaliação do erro e ao calculo da correcção a aplicar a cada peso, quer na camada interior quer na camada de saída. O processo repete-se para a próxima unidade de treino. O procedimento descrito replica-se o número de vezes consideradas necessárias para baixar o erro global até um determinado limiar (*threshold)*.

Classificadores não assistidos

O classificador ISODATA (*Iteractive Self-Organizing Data Analysis Technique*) é um classificador não assistido baseado noutro classificador não assistido, o K-Means (Mathler, 2004). O ISODATA foi desenvolvido para tornar o processo de criação de grupos homogéneos (*clusters*) menos dependente do utilizador. Na sua implementação no software ArcGIS da ESRI, o ISODATA é aplicado para produzir um ficheiro de assinaturas espectrais para cada *cluster*. Seguidamente, o utilizador corre o classificador da máxima verosimilhança para produzir um mapa de *clusters* (dendograma). Através da análise do dendrograma o utilizador identifica as classes de interesse.

As regras utilizadas pelo operador para classificar os *clusters* em classes de informação foram as seguintes: i) se um conjunto de clusters (entre 2 a 4) se encontram próximos, então estes são isolados dos restantes e a sua ocupação maioritária é avaliada; ii) *clusters* com área de ocupação reduzida e/ou fragmentada, podem ser reclassificadas por contexto. O número de *clusters* definidos à partida foram definidos experimentalmente.

2.3.1 Amostragem destinada ao treino dos algoritmos de classificação

Para o treino dos algoritmos de classificação foi recolhida uma amostra de 11957 observações (pontos). No caso do algoritmo de classificação SVM, esta amostra foi reduzida para 1195 pontos por motivos de eficiência computacional. A amostra de treino foi deterministicamente recolhida sobre a Cartografia CORINE Land Cover (2006), na sequência da sua reclassificação em classes de nível 2 da nomenclatura LANDAU. Inicialmente a amostra de treino só incluía polígonos representativos das classes LANDAU, mas estes foram posteriormente convertidos em pontos devido ao facto de muitos classificadores não processarem informação poligonal.

O Quadro 3 disponibiliza informação sobre o número de pontos amostrados por classes de nível 2 da nomenclatura LANDAU, bem como sobre a representação de cada classe, em termos de área, no território estudado (Portugal Continental).

Quadro 3 – Nº de pontos da amostra de treino por classes da Nomenclatura LANDAU

Nomenclatura LANDAU nível 2	Nº de pontos amostrados	Representatividade da classe no Continente (%)
1.1 Áreas Artificiais Contínuas	472	0.7
1.2 Áreas Artificiais Descontínuas	472	2.6
2.1 Agricultura de Regadio	684	2.4
2.2 Agricultura de Sequeiro	2304	29.0
2.3 Arrozais	450	0.6
3.1 Floresta de Folhosas	1060	11.3
3.2 Floresta de Resinosas	851	6.0
3.3 Floresta Mista	839	5.3
3.4 Vegetação Herbácea	979	9.4
3.5 Matos	2065	29.0
3.6 Vegetação Esparsa	574	1.4
4 Solo Nu	118	0.4
5 Áreas Ardidas	459	0.4
6 Zonas Húmidas	450	0.3
7 Corpos de Água	180	1.4

2.4 Validação dos mapas produzidos

Para comparação dos mapas produzidos e análise da sua utilidade para aplicações específicas importa conhecer a exactidão na classificação da ocupação do solo produzida por cada algoritmo, ou seja a exactidão temática de cada mapa. Neste sentido, na presente secção descreve-se o método de amostragem utilizado para validar os mapas produzidos, a regra de concordância adoptada na validação da classificação, bem como as medidas de exactidão temática, derivadas da matriz de erro/confusão, que retratam o desempenho na classificação da ocupação do solo de

cada algoritmo testado.

Sendo inviável avaliar o erro cometido por cada algoritmo de classificação em todas as posições do domínio espacial, a validação da classificação da ocupação do solo produzida por cada algoritmo é usualmente realizada a partir de uma amostra representativa das classes de ocupação/uso do solo presentes no domínio espacial. Esta amostra deverá ser preferencialmente recolhida de modo aleatório por forma a possibilitar generalizações sobre a área de interesse.

2.4.1 Amostragem destinada à validação

Para validação dos mapas de ocupação do solo foi recolhida uma amostra de 750 observações (pontos), equitativamente distribuídas pelas 15 classes de ocupação/uso do solo do nível 2 da nomenclatura LANDAU. A dimensão da amostra foi determinada considerando o risco do produtor (< 15%) e o risco do utilizador (< 5%), mas também o esforço operacional de concretização.

Através da consulta da informação de referência listada em 2.1 foi identificada a classe de espaço que com maior probabilidade descreve cada ponto amostrado, bem como uma segunda classe de espaço que poderá igualmente descrever a ocupação/uso do solo no mesmo ponto, destinando-se esta última a situações em que há ambiguidade na atribuição da primeira classe de espaço.

A matriz de erro/confusão (ou tabela de contingência) foi obtida por confronto da classificação de cada observação da amostra de validação com a classificação atribuída pelo algoritmo para a mesma posição do domínio espacial. Admitiu-se a existência de concordância na classificação se uma das duas classes de espaço atribuídas à observação amostrada, a partir da consulta de informação de referência, coincidir com a classe de espaço indicada pelo algoritmo de classificação.

As medidas de avaliação do desempenho na classificação estimadas a partir da matriz de erro/confusão foram a exactidão global, a exactidão do produtor e a exactidão do utilizador, cujo cálculo se explicita de seguida.

Assumindo a existência de k classes de ocupação/uso do solo, em cada célula da matriz de erro/confusão descreve-se o número (n_{ij}) de observações (pontos) da amostra de validação que foram classificados: através do algoritmo na classe de ocupação/uso do solo i (i = 1, 2, …, k) e através de informação de referência na classe de ocupação/uso do solo j (j = 1, 2, …, k).

Se a amostra de validação for com posta por n observações (pontos), represente-se

por:

n_{i+} as observações (pontos) que foram classificadas na classe de ocupação/uso do solo i através do algoritmo, sendo $n_{i+} = \sum_{j=1}^{k} n_{ij}$

n_{+j} as observações (pontos) que foram classificadas na classe de ocupação/uso do solo j com base na informação de referência, sendo $n_{+j} = \sum_{i=1}^{k} n_{ij}$

Uma vez que a exactidão global avalia a proporção de observações (pontos) correctamente classificadas na amostra de validação, a estimativa desta medida ($\hat{P}$) será dada por:

$$\hat{P} = \frac{\sum_{i=1}^{k} n_{ii}}{n} \quad (18)$$

A exactidão do produtor exprime a proporção entre o número de observações (pontos) correctamente classificados numa classe e o número total de observações da amostra de validação que efectivamente pertencem a essa classe. A estimativa da exactidão do produtor ($\hat{P}_p$) corresponde a:

$$\hat{P}_p = \frac{n_{jj}}{n_{+j}} \quad (19)$$

A exactidão do utilizador exprime a proporção de observações da amostra de validação que efectivamente pertence à classe a que foi atribuída. A estimativa da exactidão do utilizador ($\hat{P}_u$) corresponde a:

$$\hat{P}_u = \frac{n_{ii}}{n_{i+}} \quad (20)$$

3 Resultados

No presente capítulo descrevem-se os principais resultados decorrentes da classificação da ocupação do solo obtida através dos dez algoritmos ensaiados. No Anexo I disponibilizam-se as matrizes de erro/confusão associadas aos classificadores empregues. No Quadro 4 compararam-se os algoritmos testados face às medidas de exactidão auferidas na classificação das diferentes classes de uso do solo da nomenclatura LANDAU (nível 2).

Quadro 4 – Medidas de exactidão temática dos algoritmos de classificação testados

$\hat{P}_p$ = Exactidão do Produtor $\hat{P}_u$ = Exactidão do Utilizador	ML		LDC		DQDC		MD		KNN		PARZEN		CART		SVM		BMP		ISODATA	
	$\hat{P}_p$	$\hat{P}_u$	$\hat{P}_p$	$\hat{P}_u$	$\hat{P}_p$	$\hat{P}_u$	$\hat{P}_p$	$\hat{P}_u$	$\hat{P}_p$	$\hat{P}_u$	$\hat{P}_p$	$\hat{P}_u$	$\hat{P}_p$	$\hat{P}_u$	$\hat{P}_p$	$\hat{P}_u$	$\hat{P}_p$	$\hat{P}_u$	$\hat{P}_p$	$\hat{P}_u$
1.1 Áreas Artificiais Contínuas	59%	74%	56%	61%	40%	47%	28%	47%	61%	35%	35%	55%	46%	64%	28%	52%	42%	60%	18%	64%
1.2 Áreas Artificiais Descontínuas	78%	61%	58%	55%	42%	34%	36%	25%	52%	80%	73%	44%	70%	44%	52%	49%	62%	50%	0%	0%
2.1 Agricultura de Regadio	88%	78%	88%	98%	78%	89%	76%	93%	89%	76%	71%	88%	92%	87%	79%	82%	84%	81%	28%	93%
2.2 Agricultura de Sequeiro	69%	90%	78%	90%	65%	87%	47%	77%	74%	60%	52%	81%	63%	79%	82%	33%	69%	81%	89%	34%
2.3 Arrozais	83%	85%	71%	85%	60%	69%	53%	65%	78%	51%	57%	67%	55%	82%	37%	68%	57%	88%	41%	38%
3.1 Floresta de Folhosas	49%	84%	82%	88%	54%	71%	54%	69%	72%	47%	47%	67%	59%	68%	46%	69%	74%	81%	47%	77%
3.2 Floresta de Resinosas	73%	92%	92%	71%	89%	64%	85%	59%	62%	75%	88%	63%	70%	69%	86%	68%	83%	80%	84%	68%
3.3 Floresta Mista	77%	52%	80%	69%	45%	38%	45%	37%	46%	63%	38%	33%	60%	48%	11%	50%	66%	63%	19%	39%
3.4 Vegetação Herbácea	73%	85%	92%	63%	67%	43%	84%	46%	52%	60%	75%	59%	60%	55%	51%	34%	77%	69%	26%	11%
3.5 Matos	83%	61%	81%	71%	43%	68%	44%	62%	38%	89%	77%	58%	83%	42%	98%	35%	94%	37%	79%	57%
3.6 Vegetação Esparsa	82%	93%	90%	96%	66%	56%	64%	52%	55%	46%	66%	49%	44%	65%	0%	0%	54%	82%	0%	0%
4 Solo Nu	90%	82%	64%	91%	64%	74%	56%	76%	77%	46%	48%	71%	54%	90%	32%	89%	26%	87%	56%	82%
5 Áreas Ardidas	75%	100%	88%	85%	78%	74%	86%	67%	83%	65%	65%	79%	82%	95%	57%	78%	75%	100%	96%	43%
6 Zonas Húmidas	94%	67%	92%	90%	67%	60%	65%	74%	59%	65%	65%	55%	73%	75%	47%	55%	86%	68%	12%	46%
7 Corpos de Água	76%	95%	72%	97%	72%	95%	58%	100%	90%	52%	50%	89%	72%	90%	42%	91%	64%	82%	58%	78%
$\hat{P}$ = Exactidão Global	77%		79%		62%		59%		61%		61%		66%		51%		68%		45%	

Dos resultados apresentados constata-se que os algoritmos com melhor desempenho global na classificação do uso do solo foram o LDC ($\hat{P}$ =79%) e o ML ($\hat{P}$ =77%), sendo o ISODATA e o SVM os algoritmos que tiveram pior desempenho neste âmbito.

Uma análise por classes de nível 2 da nomenclatura LANDAU possibilita destacar que:

- As Áreas Artificiais Contínuas (1.1) foram classificadas com maior exactidão pelo algoritmo ML, quer do ponto de vista do produtor ($\hat{P}_p$=59%), quer do utilizador ($\hat{P}_u$ =74%);
- As Áreas Artificiais Descontínuas (1.2) foram melhor classificadas do ponto de vista do produtor pelo algoritmo KNN ($\hat{P}_p$=80%), embora através do algoritmo ML também se tenha obtido uma exactidão aceitável ($\hat{P}_p$=78%) na sua classificação. Do ponto de vista do utilizador, a exactidão na classificação destas áreas foi maior quando se empregou o algoritmo ML ($\hat{P}_u$ =61%).
- As áreas de Agricultura de Regadio (2.1) foram classificadas com maior exactidão do produtor pelo algoritmo CART ($\hat{P}_p$=92%). Estas áreas também foram classificadas com boa exactidão do produtor pelos algoritmos ML e LDC ($\hat{P}_p$=88% para ambos). A máxima exactidão do utilizador obtida na classificação destas áreas ($\hat{P}_u$ =98%) foi proporcionada pelo algoritmo LDC.
- Não obstante o razoável desempenho do algoritmo LDC na classificação da Agricultura de Sequeiro (2.2) ($\hat{P}_p$=78%), foi o algoritmo ISODATA o que permitiu classificar este tipo de agricultura com maior exactidão do produtor ($\hat{P}_p$=89%). No que diz respeito à máxima exactidão do utilizador obtida na classificação destas áreas, os algoritmos LDC e o ML revelaram um desempenho similar ($\hat{P}_u$ =90%).
- O algoritmo ML foi o que possibilitou classificar os Arrozais (2.3) com maior exactidão do produtor ($\hat{P}_p$=83%). A máxima exactidão do utilizador obtida na classificação destas áreas coube ao algoritmo BMP ($\hat{P}_u$ =88%).
- Na classificação da Floresta de Folhosas (3.1) o algoritmo LDC foi o que revelou melhor desempenho, apresentando não só a maior exactidão do produtor ($\hat{P}_p$=82%) mas também a maior exactidão do utilizador ($\hat{P}_u$ =88%).

- Na classificação de Florestas de Resinosas (3.2) os melhores desempenhos em termos de exactidão do produtor e do utilizador foram facultados, respectivamente pelos algoritmos LDC e ML ($\hat{P}_p$=92%; $\hat{P}_u$=92%).
- O algoritmo LDC foi dos testados, o que permitiu classificar a Floresta Mista (3.3) com maior exactidão do produtor ($\hat{P}_p$=80%) e maior exactidão do utilizador ($\hat{P}_u$=69%).
- Na classificação da Vegetação Herbácea (3.4), a maior exactidão do produtor foi proporcionada pelo algoritmo LDC ($\hat{P}_p$=92%) e a maior exactidão do utilizador foi alcançada através do algoritmo ML ($\hat{P}_u$=85%).
- Apesar dos Matos (3.5) terem sido classificados com uma exactidão do produtor razoável pelos algoritmos ML e LDC (respectivamente, $\hat{P}_p$=83% e $\hat{P}_p$=81%), a maior exactidão do produtor para estas áreas foi obtida pelo algoritmo SVM ($\hat{P}_p$=98%). O algoritmo LDC permitiu alcançar a máxima exactidão do utilizador na classificação dos Matos ($\hat{P}_u$=71%).
- No que concerne à classificação da Vegetação Esparsa (3.6), quer a maior exactidão do produtor quer a maior exactidão do utilizador foram obtidas através do algoritmo LDC ($\hat{P}_p$=90%; $\hat{P}_u$=96%).
- Na classificação do Solo Nu (4) o ML foi o algoritmo que permitiu obter uma exactidão do produtor máxima ($\hat{P}_p$=90%). A maior exactidão do utilizador na classificação desta classe foi obtida pelo algoritmo LDC ($\hat{P}_u$=91%).
- A maior exactidão do produtor alcançada na classificação das Áreas Ardidas (5) resultou da aplicação do algoritmo ISODATA ($\hat{P}_p$=96%), tendo o desempenho dos algoritmos LDC e ML sido um pouco inferior (respectivamente, $\hat{P}_p$=88% e $\hat{P}_p$=75%). Estas áreas foram classificadas com máxima exactidão do utilizador ($\hat{P}_u$=100%) pelo algoritmo ML.
- Na classificação das Zonas Húmidas (6), a maior exactidão do produtor foi assegurada pelo algoritmo ML ($\hat{P}_p$=94%) e a maior exactidão do utilizador foi assegurada pelo algoritmo LDC ($\hat{P}_u$=90%).
- A maior exactidão do produtor alcançada na classificação dos Corpos de Água

(7) resultou da aplicação do algoritmo ML ($\hat{P}_p$=76%). A maior exactidão do utilizador na classificação desta classe coube ao algoritmo MD ($\hat{P}_u$=100%), não obstante a excelente exactidão do utilizador proporcionada pelo algoritmo LDC ($\hat{P}_u$=97%).

Os algoritmos com melhor desempenho global na classificação do uso do solo (LDC e ML) produziram maior erro na classificação das Áreas Urbanas (1), originando no caso das Áreas Urbanas Contínuas (1.1) exactidões do produtor e do utilizador que não excedem, respectivamente, 59% e 74%. Para as Áreas Urbanas Descontínuas (1.2), obtiveram-se exactidões do produtor ligeiramente superiores ($\hat{P}_p$≤78%), mas exactidões do utilizador mais reduzidas ($\hat{P}_u$≤61%). De um modo geral, a maior confusão verificada nas classes de urbano deve-se ao facto do tecido urbano do Continente não ser suficientemente denso para criar um sinal limpo, uma vez que se tratam de classes com uma elevada diversidade de elementos de superfície (telhados, alcatrão, jardins, etc). De facto, o tecido urbano continuo apresenta-se com um sinal pontilhado devido à diversidade de materiais e às zonas de pontos verdes. A dimensão do pixel também foi relevante para a pior classificação desta áreas, mas provavelmente mais relevante na classe de urbano descontinuo. Sendo esta classe caracterizada por pequenas estruturas urbanizadas, num pixel com 9 ha de área o sinal é dissolvido.

Na classificação das Áreas Agrícolas (2) os algoritmos citados denotaram exactidões do produtor iguais ou superiores a 69% e exactidões do utilizador iguais ou superiores a 78%.

A aplicação destes dois métodos na classificação de Solo Nu (4) possibilitou obter exactidões que oscilaram entre 64%-90% ($\hat{P}_p$) e 82%-91% ($\hat{P}_u$). Na classificação de Áreas Ardidas (5) obtiveram-se exactidões variáveis entre 75%-88% ($\hat{P}_p$) e 85%-100% ($\hat{P}_u$). Na classificação de Corpos de Água (7) as exactidões foram de 72%-76% ($\hat{P}_p$) e 95%-97% ($\hat{P}_u$).

Na classificação das Florestas e Meios Naturais e Semi-naturais (3) o LDC, em particular, revelou um bom desempenho conduzindo a exactidões do produtor iguais ou superiores a 80%. A exactidão do utilizador alcançada na classificação das Florestas e Meios Naturais e Semi-naturais denotou grande variabilidade, oscilando entre 52% e

96%. Neste grande grupo, os algoritmos tiveram pior desempenho na classificação da Floresta Mista (3.3). Esta situação pode ser justificada pelo facto da floresta mista se confundir com as florestas de folhosas e de resinosas, numa relação directa com a respectiva densidade de ocupação. Ou seja, se uma área interpretada como floresta mista apresentar uma proporção relevante de folhosas, o sinal espectral desse pixel irá aproximar-se do da classe folhosa; de modo semelhante para a classe de resinosas.

Na classificação das Zonas Húmidas (6), a exactidão do utilizador obtida através dos dois algoritmos variou entre 67% e 90% e a exactidão do produtor foi igual ou superior a 92%.

4 Conclusões

Face aos resultados apresentados, os algoritmos que revelaram melhor desempenho na classificação do uso do solo foram o LDC (Linear Discriminant Classifier) e o ML (Maximum Likelihood). A superioridade do LDC sobre os restantes classificadores é consistente com estudos anteriores aplicados a Portugal Continental, como o de Carrão e colaboradores (2008 e 2010). O LDC tem sido descrito na literatura (Kuncheva, 2004; Hastie *et al.*, 2009) como um classificador robusto, computacionalmente pouco exigente, que conduz a bons resultados, mesmo quando as hipóteses se desviam da realidade. Hastie e colaboradores (2009, p.111) salientam igualmente que no projecto STATLOG (Mitchie *et al.*, 1994), que envolveu a comparação de múltiplos classificadores em diferentes aplicações (entre as quais a identificação de grupos de cromossomas em dados genéticos), o LDC e o ML destacaram-se entre os três algoritmos com melhor desempenho na classificação de diversos conjuntos de dados. Particularizando, o LDC figurou entre os três primeiros na classificação de sete conjuntos de dados e o ML figurou entre os três primeiros na classificação de quatro conjuntos de dados. O presente estudo comprova que, à escala de uma imagens MERIS, o LDC apresenta melhores resultados do que os restantes algoritmos de classificação.

5 Referências Bibliográficas

Carrão, H., Araújo A., Caetano M., 2008. Land cover classification in Portugal with intra-annual time series of MERIS images, *2nd MERIS/(A)ATSR User Workshop*, 22 - 26 September 2008, Rome, Italy, unpaginated CD-ROM.

Carrão H., Araújo A, Gonçalves P., Caetano M., 2010. Multitemporal MERIS Images for land cover mapping at national scale : the case study of Portugal. International Journal

of Remote Sensing. Vol 31, No 8, April 2010, 2063-2082.

Congalton R G, Green K. Assessing the accuracy of remotely sensed data - Principles and practices. CRC Press, Taylor & Francis Group, 2nd ed. 2009.

Costa L M M, Zeilhofer P, Rodrigues W S. Avaliação do Classificador SVM (Support Vector Machine) no Mapeamento de Queimadas no Pantanal Mato-Grossense. Anais do III Simpósio Brasileiro de Ciências Geodésicas e Tecnologias da Geoinformação. Recife, 2010: 1-5.

Fonseca J M M R. Indução de Árvores de Decisão: HistClass - Proposta de um algoritmo não paramétrico. Dissertação de Mestrado. Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa. 1994.

Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: Data mining, inference and prediction. Springer, 2nd edition, 2009.

Kuncheva L. Combining Patterns Classifiers: Methods and algorithms. Wiley Interscience Publication, 2004.

Loureiro M L D S. Exploração das Características Espectrais de Imagens IKONOS para Caracterização da Ocupação do Solo: Comparação de Classificadores. Dissertação de Mestrado. Instituto Superior Técnico, Universidade Técnica de Lisboa. 2008.

Mathler P. Computer processing of remotely sensed images: An introduction. 3rd edition. Wiley, 2004.

Mitchie M, Spiegelhater D J, Taylor C C. Machine learning, neural and statistical classification. Ellis Horwood series in artificial intelligence, 1994.

Plat, J. C. Sequential Minimal Optimization: a fast algorithm for training support vector machines. 1998. URL: http://www.bradblock.com/Sequential_Minimal_Optimization_A_Fast_Algorithm_for_Training_Support_Vector_Machine.pdf (Acedido a 18 de Abril 2012).

Schowengerdt R. A. Remote sensing: models and methods for image processing. 2nd edition. Academic Press, 1997.

Souza B F S, Teixeira A S, Silva A T F. Classificação de Bioma Caatinga usando Support Vector Machine (SVM). Anais do XIV Simpósio Brasileiro de Sensoriamento Remoto. Natal, 2009: 7917-7924.

ANEXO I – Matrizes de erro / confusão associadas à classificação da ocupação do solo, com base em dados MERIS

Quadro I.1 – Matriz de erro / confusão associada à classificação da ocupação do solo produzida pelo algoritmo Maximum Likelihood – ML

		Informação de Referência															Total	$\hat{P}_u$ (%)
		4	**5**	**6**	**7**	**1.1**	**1.2**	**2.1**	**2.2**	**2.3**	**3.1**	**3.2**	**3.3**	**3.4**	**3.5**	**3.6**		
Classificação	**4**	45			1	4			5								55	82%
	5		38														38	100%
	6		9	46	9					1	3	1					69	67%
	7			2	38												40	95%
	1.1					29	10										39	74%
	1.2	5				16	40		3	2							66	61%
	2.1							45		2	8	1	1	1			58	78%
	2.2								36					4			40	90%
	2.3							6		40			1				47	85%
	3.1										26		5				31	84%
	3.2											36	3				39	92%
	3.3				1						15	11	37		7		71	52%
	3.4								4	2				33			39	85%
	3.5		3	1	1		1		4	1	1		1	7	45	9	74	61%
	3.6		1												2	41	44	93%
Total		50	51	49	50	49	51	51	52	48	53	49	48	45	54	50	750	
$\hat{P}_p$ (%)		90%	75%	94%	76%	59%	78%	88%	69%	83%	49%	73%	77%	73%	83%	82%		$\hat{P}$ =77%

Quadro I.2 – Matriz de erro / confusão associada à classificação da ocupação do solo produzida pelo algoritmo Linear Discriminant Classifier – LDC

		Informação de Referência															Total	$\hat{P}_u$ (%)
		4	**5**	**6**	**7**	**1.1**	**1.2**	**2.1**	**2.2**	**2.3**	**3.1**	**3.2**	**3.3**	**3.4**	**3.5**	**3.6**		
Classificação	**4**	32				3											35	91%
	5		45	2	6												53	85%
	6		1	45	4												50	90%
	7	1			36												37	97%
	1.1	2			1	28	15										46	61%
	1.2	4				19	29		1								53	55%
	2.1							44		1							45	98%
	2.2	1							36					2		1	40	90%
	2.3			1				5		35							41	85%
	3.1										46	1	5				52	88%
	3.2		2		2						1	44	3		10		62	71%
	3.3							1		3	9	3	36				52	69%
	3.4	7					3		8	9				46			73	63%
	3.5	3	1	1	1		3		1	1			1	2	44	4	62	71%
	3.6		2													47	49	96%
Total		50	51	49	50	50	50	50	46	49	56	48	45	50	54	52	750	
$\hat{P}_p$ (%)		64%	88%	92%	72%	56%	58%	88%	78%	71%	82%	92%	80%	92%	81%	90%		$\hat{P}$ =79%

Quadro I.3 – Matriz de erro / confusão associada à classificação da ocupação do solo produzida pelo algoritmo

Diagonal Quadratic Discriminant Classifier - DQDC

		Informação de Referência															Total	$\hat{P}_u$ (%)
		4	**5**	**6**	**7**	**1.1**	**1.2**	**2.1**	**2.2**	**2.3**	**3.1**	**3.2**	**3.3**	**3.4**	**3.5**	**3.6**		
Classificação	**4**	32			1	8			2								43	74%
	5		40	11	3												54	74%
	6		6	33	7										9		55	60%
	7		2		36												38	95%
	1.1	2				20	12		1					7		1	43	47%
	1.2	15				14	21		8	1				2			61	34%
	2.1							39		4	1						44	89%
	2.2					1	2		33					2			38	87%
	2.3							11		29	1					1	42	69%
	3.1										30	4	8				42	71%
	3.2		2								6	47	12		7		74	64%
	3.3									2	15	2	18		5	5	47	38%
	3.4	1			1	6	14		6	6				33		9	76	43%
	3.5		1	4	2	1								2	23	1	34	68%
	3.6			1			1		1	6	3		2	3	9	33	59	56%
Total		50	51	49	50	50	50	50	51	48	56	53	40	49	53	50	750	
$\hat{P}_p$ (%)		64%	78%	67%	72%	40%	42%	78%	65%	60%	54%	89%	45%	67%	43%	66%		$\hat{P}$ =62%

Quadro I.4 – Matriz de erro / confusão associada à classificação da ocupação do solo produzida pelo algoritmo Minimum Distance – MD

		Informação de Referência															Total	$\hat{P}_u$ (%)
		4	**5**	**6**	**7**	**1.1**	**1.2**	**2.1**	**2.2**	**2.3**	**3.1**	**3.2**	**3.3**	**3.4**	**3.5**	**3.6**		
Classificação	**4**	28				8			1								37	76%
	5		44	11	11												66	67%
	6		3	32	5										3		43	74%
	7				29												29	100%
	1.1	2			2	14	9			1				1		1	30	47%
	1.2	19				17	18		16					2			72	25%
	2.1							38		3							41	93%
	2.2						4		23	1				2			30	77%
	2.3	1						11		26	2						40	65%
	3.1							1			29	4	8				42	69%
	3.2		2		2						2	46	13		13		78	59%
	3.3									1	18	4	19		3	6	51	37%
	3.4					10	18		8	6				42		8	92	46%
	3.5		2	5	1	1								2	23	3	37	62%
	3.6			1			1		1	11	3		2	1	10	32	62	52%
Total		50	51	49	50	50	50	50	49	49	54	54	42	50	52	50	750	
$\hat{P}_p$ (%)		56%	86%	65%	58%	28%	36%	76%	47%	53%	54%	85%	45%	84%	44%	64%		$\hat{P}$ =59%

Quadro I.5 – Matriz de erro / confusão associada à classificação da ocupação do solo produzida pelo algoritmo K Nearest Neighbours – KNN

		Informação de Referência															Total	$\hat{P}_u$ (%)
		4	**5**	**6**	**7**	**1.1**	**1.2**	**2.1**	**2.2**	**2.3**	**3.1**	**3.2**	**3.3**	**3.4**	**3.5**	**3.6**		
Classificação	**4**	23			2	4			1								30	77%
	5		33	1	6												40	83%
	6	4	5	32	11	1									1		54	59%
	7	1		2	26												29	90%
	1.1	5	1			17	5										28	61%
	1.2	8			1	21	41	1	5							2	79	52%
	2.1							39		5							44	89%
	2.2	2			1	1			32					7			43	74%
	2.3	1						6		25							32	78%
	3.1		1							4	28	1	4			1	39	72%
	3.2		3		2						7	36	6		3	1	58	62%
	3.3	1						3		2	15	7	26			3	57	46%
	3.4	1		1		2		2	9	8			1	27		1	52	52%
	3.5	4	8	11	1	2	4		4	5	6	4		8	47	19	123	38%
	3.6			2		1	1		2		4		4	3	2	23	42	55%
Total		50	51	49	50	49	51	51	53	49	60	48	41	45	53	50	750	
$\hat{P}_p$ (%)		46%	65%	65%	52%	35%	80%	76%	60%	51%	47%	75%	63%	60%	89%	46%		$\hat{P}$ =61%

Quadro I.6 – Matriz de erro / confusão associada à classificação da ocupação do solo produzida pelo algoritmo PARZEN

Classificação	Informação de Referência															Total	$\hat{P}_u$ (%)
	4	**5**	**6**	**7**	**1.1**	**1.2**	**2.1**	**2.2**	**2.3**	**3.1**	**3.2**	**3.3**	**3.4**	**3.5**	**3.6**		
4	24			2	6			1					1			34	71%
5		33	3	6												42	79%
6	4	8	32	12	1									1		58	55%
7	1		2	25												28	89%
1.1	4	1	1		17	7		1								31	55%
1.2	11		1	1	21	37	2	8	1				1		2	85	44%
2.1							36		4	1						41	88%
2.2	1			1	1			25					3			31	81%
2.3	1						11		28			2				42	67%
3.1		1							2	28	2	7			2	42	67%
3.2		3		2						6	44	10		5		70	63%
3.3	1	1					1		2	17	4	15		1	3	45	33%
3.4	1			1	1		1	10	9			1	38		2	64	59%
3.5	1	4	8		1	2		1					4	40	8	69	58%
3.6	1		2		1	5		2	3	7		5	4	5	33	68	49%
Total	50	51	49	50	49	51	51	48	49	59	50	40	51	52	50	750	
$\hat{P}_p$ (%)	48%	65%	65%	50%	35%	73%	71%	52%	57%	47%	88%	38%	75%	77%	66%		$\hat{P}$ =61%

Quadro I.7 – Matriz de erro / confusão associada à classificação da ocupação do solo produzida pelo algoritmo

Classification and Regression Tree - CART

		Informação de Referência															Total	$\hat{P}_u$ (%)
		4	**5**	**6**	**7**	**1.1**	**1.2**	**2.1**	**2.2**	**2.3**	**3.1**	**3.2**	**3.3**	**3.4**	**3.5**	**3.6**		
Classificação	**4**	27				3											30	90%
	5		42	2													44	95%
	6		2	36	7		1							1	1		48	75%
	7		1	3	36												40	90%
	1.1	1	1		1	23	7		3								36	64%
	1.2	13				22	35		3					4		3	80	44%
	2.1	1						46		3	2			1			53	87%
	2.2	3				1	1		33					4			42	79%
	2.3							4		27	1					1	33	82%
	3.1				1					3	30	2	7			1	44	68%
	3.2				4						2	35	7		3		51	69%
	3.3									3	15	8	29		1	4	60	48%
	3.4	4					1		11	6				28		1	51	55%
	3.5	1	4	6	1	1	4		2	5	1	5	4	8	44	18	104	42%
	3.6		1	2			1			2			1	1	4	22	34	65%
Total		50	51	49	50	50	50	50	52	49	51	50	48	47	53	50	750	
$\hat{P}_p$ (%)		54%	82%	73%	72%	46%	70%	92%	63%	55%	59%	70%	60%	60%	83%	44%		$\hat{P}$ =66%

Quadro I.8 – Matriz de erro / confusão associada à classificação da ocupação do solo produzida pelo algoritmo Support Vector Machine – SVM

		Informação de Referência															Total	$\hat{P}_u$ (%)
		4	**5**	**6**	**7**	**1.1**	**1.2**	**2.1**	**2.2**	**2.3**	**3.1**	**3.2**	**3.3**	**3.4**	**3.5**	**3.6**		
Classificação	**4**	16				1				1							18	89%
	5		29	8													37	78%
	6		8	23	11												42	55%
	7	1		1	21												23	91%
	1.1	3	1	1		14	7									1	27	52%
	1.2	8				17	26		1		1						53	49%
	2.1	1					1	42		6	1						51	82%
	2.2	13		1	11	8	5	1	49	7	22	1	8	16		8	150	33%
	2.3	1						6		17			1				25	68%
	3.1		1								29	2	10				42	69%
	3.2		4		1						5	44	10		1		65	68%
	3.3										3	1	4				8	50%
	3.4	2		1		2	4	4	7	14			1	20		4	59	34%
	3.5	5	8	14	6	8	7		3	1	2	3	1	3	52	37	150	35%
	3.6															0	0	0%
Total		50	51	49	50	50	50	53	60	46	63	51	35	39	53	50	750	
$\hat{P}_p$ (%)		32%	57%	47%	42%	28%	52%	79%	82%	37%	46%	86%	11%	51%	98%	0%		$\hat{P}$ =51%

Quadro I.9 – Matriz de erro / confusão associada à classificação da ocupação do solo produzida pelo algoritmo Backpropagation Multilayer Perceptron - BMP

Classificação	Informação de Referência															Total	$\hat{P}_u$ (%)
	4	**5**	**6**	**7**	**1.1**	**1.2**	**2.1**	**2.2**	**2.3**	**3.1**	**3.2**	**3.3**	**3.4**	**3.5**	**3.6**		
4	13				2											15	87%
5		38														38	100%
6	3	1	42	15	1											62	68%
7	6		1	32												39	82%
1.1	5				21	8									1	35	60%
1.2	9				22	31										62	50%
2.1	2						43	1	6			1				53	81%
2.2	1				1			34					6			42	81%
2.3							4		28							32	88%
3.1										42	2	6			2	52	81%
3.2		4		1							40	5				50	80%
3.3		1							1	12	2	29			1	46	63%
3.4	3			1		1	2	3	6				37	1		54	69%
3.5	8	6	6	1	3	10	2	11	8	1	4	2	5	51	19	137	37%
3.6		1								2		1		2	27	33	82%
Total	50	51	49	50	50	50	51	49	49	57	48	44	48	54	50	750	
$\hat{P}_p$ (%)	26%	75%	86%	64%	42%	62%	84%	69%	57%	74%	83%	66%	77%	94%	54%		$\hat{P}$ =68%

Quadro I.10 – Matriz de erro / confusão associada à classificação da ocupação do solo produzida pelo algoritmo

Iterative Self-Organizing Data Analysis Technique - ISODATA

Classificação \ Informação de Referência	4	5	6	7	1.1	1.2	2.1	2.2	2.3	3.1	3.2	3.3	3.4	3.5	3.6	Total	$\hat{P}_u$ (%)
4	28				5			1								34	82%
5		49	38	11							5	1		11		115	43%
6			6	6	1											13	46%
7				29			7			1						37	78%
1.1				1	9	1		2					1			14	64%
1.2						0	2		7							9	0%
2.1							14					1				15	93%
2.2	18				28	31		54					23		3	157	34%
2.3					2		27	1	20	1			1			52	38%
3.1										27	2	6				35	77%
3.2				2						7	46	13				68	68%
3.3										11		7				18	39%
3.4	3		1		4	7		2	20	9		4	10		33	93	11%
3.5		2	4	1					1	2	2	5		41	14	72	57%
3.6	1				1	11		1	1				3		0	18	0%
Total	50	51	49	50	50	50	50	61	49	58	55	37	38	52	50	750	
$\hat{P}_p$ (%)	56%	96%	12%	58%	18%	0%	28%	89%	41%	47%	84%	19%	26%	79%	0%		$\hat{P}$ =45%