ANÁLISE DE MEDIDAS DE ACURÁCIA EM PROCESSOS DE CLASSIFICAÇÃO BASEADAS NA MATRIZ DE CONFUSÃO

Josiane Aparecida Cardoso de Souza
Universidade Federal de Minas Gerais
Av. Pres. Antônio Carlos, 6627 - Pampulha, Belo Horizonte - MG, 31270-901
josisouzakta@hotmail.com

Marcelo Azevedo Costa
Universidade Federal de Minas Gerais
Av. Pres. Antônio Carlos, 6627 - Pampulha, Belo Horizonte - MG, 31270-901
azevedo@est.ufmg.br

Britaldo Silveira Soares Filho
Universidade Federal de Minas Gerais
Av. Pres. Antônio Carlos, 6627 - Pampulha, Belo Horizonte - MG, 31270-901
britaldo@csr.ufmg.br

Álvaro Lédo Ferreira
Universidade Federal de Minas Gerais
Av. Pres. Antônio Carlos, 6627 - Pampulha, Belo Horizonte - MG, 31270-901
alvaroledoferreira@gmail.com

RESUMO

O uso de dados coletados através do Sensoriamento Remoto como produtos emergentes institui a necessidade de entender e verificar sua acurácia, especialmente na análise da acurácia temática. Este artigo tem como objetivo analisar a variação de cinco índices de acurácia: Acurácia Total, Kappa de Cohen, Kappa Ponderado, Índice de Tau e Pi de Scott, aos quais são baseados na matriz de confusão. As matrizes foram criadas a partir de dois cenários: quando as probabilidades são iguais para cada célula da matriz e quando os elementos da diagonal possuem maior probabilidade de ocorrência. Os resultados mostraram que no primeiro cenário os índices de acurácia são reduzidos com o aumento do número de classes, devido à soma da diagonal ser baixa. E no segundo cenário os índices de acurácia apresentam valores melhores quando $k = 200$.

PALAVRAS CHAVE. Acurácia. Métodos de classificação. Matriz de confusão.

ABSTRACT

The use of data collected by Remote Sensing as emerging products establishes the necessity to understand and verify its accuracy, especially on thematic accuracy analysis. This paper has the objective of analyze the variation of 5 index of accuracy: Overall Accuracy, Cohen's Kappa, Weighted Kappa, Tau Coefficients and Scott's pi which are based on the confusion matrix. The matrix was created based on two scenarios: when the probability is equal for each cell of the matrix and when the elements of the diagonal have greater probability of occurrence. The results showed that in the first scenario the indices of accuracy reduce with the rise of the number of classes due to the sum of the diagonal being small. And in the second scenario the indices of accuracy show values better when $k = 200$.

KEYWORDS. Accuracy. Classifications Methods. Confusion Matrix.

1. Introdução

O Sensoriamento Remoto é considerado uma ferramenta importante no fornecimento de dados de uso e cobertura da Terra (Anderson et al, 1976; Turner, 2005; Whitehead, Hugenholtz, 2014; Haack, Mahabir, Kerkering, 2015; Costa et al, 2018). Estes dados são fontes de informações imprescindíveis para a tomada de decisões nas áreas científicas, atividades de manejo e conservação do solo (Lyons et al, 2018), dentre outras. Exemplos da aplicação do uso de sensoriamento remoto podem ser encontrados para estimar e mapear a área basal de povoamentos de eucaliptos (dos Reis et. al. (2018); avaliar quantitativamente a suscetibilidade de deslizamento de terra (Liu e Duan, 2018); analisar o nível de infestação de ervas daninhas em uma produção agrícola (Ottosen et al, 2019) e mediar emissões de gases de efeito estufa (Li et al, 2017), dentre outros.

Existem diversas maneiras de criar mapas a partir do sensoriamento remoto, sendo imprescindível analisar a sua acurácia, isto é, a sua capacidade de representar fidedignamente os dados reais. Para isso, a classificação de imagens é a abordagem mais comum para mensurar a acurácia (Lyons, 2018). A classificação de imagem é geralmente referida ao mapeamento temático e a rotulação de objetos e características de determinados grupos baseados em seus atributos (Sokal, 1974; Jensen, 2015). A avaliação da acurácia indica a qualidade do mapa que foi criado através dos dados de sensoriamento remoto, sendo dividida em (a) acurácia posicional e (b) acurácia temática (Congalton, 2009.p. 02).

A acurácia posicional corresponde à diferença da concordância locacional entre dados gerados por Sensoriamento Remoto e pontos terrestres conhecidos ou outra fonte considerada mais precisa que o mapa (Bolstad, 2005; Congalton, 2009.p.12). Seu cálculo é feito através da Raiz do Erro Quadrático Médio (REQM) (Bolstad, 2005). A acurácia temática avalia a concordância entre a classificação ou atributos dos recursos de um mapa gerados e a verdade em campo, representada através de uma matriz de confusão (Congalton, 2009.p. 32). Através da matriz de confusão podem ser aplicadas várias medidas para avaliar a acurácia dos dados, dentre elas tem-se a Acurácia Total, a estatística Kappa de Cohen, o Índice de Tau, dentre outras. Também é possível analisar a acurácia de cada classe individual, através de medidas como acurácia do produtor e do usuário, Kappa condicional, etc. (Couto, 2003).

Este trabalho procura avaliar medidas de acurácia utilizando dados simulados. Procurou-se analisar a variação das medidas de acurácia em função do número de classes e do tamanho da amostra. Foram avaliadas as medidas Acurácia Total, estatística Kappa de Cohen, Kappa Ponderado, Índice de Tau e Pi de Scott.

2. Materiais e Métodos

2.1 Matriz de confusão

A matriz de confusão ou matriz de erro (Congalton, 1983; Stehman, Wickham, 2011; Khatami, Mountraki, Stehman, 2017) é uma ferramenta com arranjo quadrado que apresenta o grau de similaridade entre dados de referência (geralmente dispostos em colunas, derivados de fotos ou vídeos aéreos, medição de solo, etc) e dados de classificação (geralmente dispostos em linhas, gerados a partir de mapas ou dos dados de sensoriamento remoto) (Congalton, 1991). A representação da matriz de confusão é apresentada na Tabela 1.

Assume-se que n amostras estão dispostas dentre k^2 células, em que cada amostra é atribuída a uma das k classes no mapa, distribuídas de maneira independente a uma das mesmas k classes no conjunto de dados de referência. Deste modo, n_{ij} corresponde ao número de amostras distribuídas na classe i ($i = 1, 2, ..., k$) no conjunto de dados de classificação e classe j ($j = 1, 2, ..., k$) no conjunto de dados de referência. Já n_{i+} é o somatório do número de amostras denominadas como classe i do mapa e n_{+j} é o somatório do número de amostras da classe j nos dados de referência, respectivamente mostrados pelas equações 1 e 2.

$$n_{i+} = \sum_{j=1}^{k} n_{ij} \quad (1)$$

$$n_{+j} = \sum_{i=1}^{k} n_{ij} \quad (2)$$

A equação 1 corresponde à soma marginal de cada linha dos dados de classificação e a equação 2 corresponde à soma marginal de cada coluna dos dados de referência.

A diagonal principal da matriz de confusão representa a concordância entre os dados de classificação e os dados de referência sendo representada por n_{ii}, onde $i = j$ ($i = 1, 2, ..., k$) (Congalton, 2009). Portanto, a classificação correta é representada pelas amostras da diagonal principal da matriz de confusão e a classificação incorreta pelas amostras fora da diagonal principal (Couto, 2003; Stehman, 2006).

		Dados de referência (j = colunas)			
		1	2	k	Total das linhas n_{i+}
Dados de classificação (i = linhas)	1	n_{11}	n_{12}	n_{1k}	n_{1+}
	2	n_{21}	n_{22}	n_{2k}	n_{2+}
	k	n_{31}	n_{32}	n_{kk}	n_{k+}
Total das colunas n_{+j}		n_{+1}	n_{+2}	n_{+k}	n

Tabela 01: Adaptado de Congalton (2009, p.60)
Fonte: Congalton (2009)

Denotando p_{ij} como a proporção de amostras correspondente à *i*, *j*-ésima célula (n_{ij}), a equação 3 mostra a proporção de amostras que cada célula possui de uma combinação particular de classes de cobertura de terra classificada e referenciada (Salk et al, 2018).

$$p_{ij} = \frac{n_{ij}}{n} \quad (3)$$

Quando $i = j$ tem-se a proporção da diagonal principal da matriz:

$$p_{ii} = \frac{n_{ii}}{n} \quad (4)$$

Então a proporção marginal das linhas (p_{i+}) e colunas (p_{+j}) são definidas como:

$$p_{i+} = \sum_{j=1}^{k} p_{ij} = \frac{n_{i+}}{n} \quad (5)$$

$$p_{+j} = \sum_{i=1}^{k} p_{ij} = \frac{n_{+j}}{n} \quad (6)$$

com p_{i+} e p_{+j} definidas previamente como "concordância ao acaso".

Além de exibir resultados empíricos convenientes, a matriz de confusão facilita a análise estatística (Card, 1982), sendo utilizados diversos métodos de análises de acurácia destacando-se acurácia total, Kappa e Tau.

Este artigo analisa cinco medidas de acurácia: Acurácia Total (p_O), Kappa de Cohen ($\widehat{K}$), Kappa ponderado ($\widehat{K}_w$), Índice Tau (T) e Pi de Scott (π). A matriz de confusão foi construída baseando-se em dois cenários. No primeiro cenário, cada célula possui igual probabilidade (($P_{ij} =$

$1/k^2$) de que uma amostra caia em uma das k classes e no segundo, as células das diagonais possuem maior probabilidade de que as amostras caiam em uma das k classes do que nas células fora das diagonais ($P_{ii} = \lambda P_{ij}$, $\lambda > 1$). Em outras palavras, a probabilidade de as amostras caírem na diagonal (P_d) é λ vezes maior que caírem fora da diagonal (P_f).

Para cada cenário foram simuladas três classes de tamanhos k = 3, 5 e 10 e, para cada classe, simulou-se um número de amostras de 200 e 500. Foram geradas 1000 matrizes, uma matriz de confusão para cada análise com distribuição multinomial e 95% de confiança. As subseções 2.1.1 até 2.1.5 explicam as medidas de acurácias utilizadas nesse artigo, a seção 3 apresenta os resultados e a seção 4 as conclusões.

As simulações e análises foram realizadas utilizado o software *RStudio*, e os pacotes: *stats*, *irr* e *psych*.

2.1.1 Acurácia total

A acurácia total ou acurácia global, apresentada por Story e Congalton (1986), tem como objetivo analisar a concordância geral entre os dados de classificação e os de referência através da soma da diagonal principal dividida pelo total de amostras, representada por p_o (Eq. 7).

$$p_o = \frac{\sum_{i=1}^{k} n_{ii}}{n} \tag{7}$$

p_o também é chamada de porcentagem concordância por autores como Story, Congalton (1986), Naesset (1996), Ma e Redmond (1995), e de proporção observada por Koukolas, Blackburn (2001), Congalton (2009), e representada pelo somatório a proporção das amostras classificadas corretamente (p_{ii}) ((Eq.8).

$$p_O = \sum_{i=1}^{k} p_{ii} \tag{8}$$

2.1.2 Kappa de Cohen

A análise Kappa desenvolvida por Cohen (1960) e é uma técnica multivariada discreta utilizada em avaliação de acurácia (Bishop et al, 1975). Mede a proporção de concordância, eliminando-se a chance ao acaso (Cohen, 1960). O resultado gerado pela análise Kappa é a estatística KHAT ($\hat{K}$) e mede o quão bem os dados de classificação concordam com os dados de referência. Para o estimador $\hat{K}$ é necessário determinar a proporção da concordância atual ou observada p_O (Eq. 7) e a proporção de concordância casual p_c (Eq. 9), ao qual corresponde ao produto das proporções marginais das linhas e colunas dos dados de referência e de classificação.

p_c também é chamada de proporção esperada.

$$p_c = \sum_{i=1,j=1}^{k} p_{i+}p_{+j} \tag{9}$$

Assumindo um modelo de amostragem multinomial, a estimativa de máxima verossimilhança de $\hat{K}$ é dada por:

$$\hat{K} = \frac{p_o - p_c}{1 - p_c} \tag{10}$$

Vale ressaltar que o numerador dessa equação é a diferença entre os dados observados e os esperados (dados de classificação), sendo esta estatística similar à análise da família Qui- quadrado (χ^2).

Os resultados de $\hat{K}$ variam entre -1 e +1, onde -1 corresponde a nenhuma concordância, 0 relaciona-se a concordâncias iguais entre os dados e +1 corresponde a uma perfeita concordância. Autores como Landis e Koch (1977) caracterizaram o nível de concordância $\hat{K}$ em três intervalos: valor maior que 0,80 representa forte concordância, entre 0,40 e 0,80 concordância moderada e menor que 0,40 uma concordância fraca.

2.1.3 Kappa ponderado

O Kappa ponderado foi desenvolvido por Cohen (1968) e aplicado à área de sensoriamento remoto por Rosenfield e Fitzpatrick-Lins (1986). É apropriado quando algumas discordâncias (células não diagonais) são mais importantes que outras, de modo que o pesquisador possa atribuir diferentes pesos (Rosenfield, 1986).

O peso atribuído para a i,j-ésima célula na matriz de confusão é denotado por w_{ij}. Isto significa que a proporção p_{ij} na a i,j-ésima célula é ponderada por w_{ij}. Segundo Fleiss (1969), w_{ij} é restrito ao intervalo $0 \leq w_{ij} \leq 1$ para $i \neq j$, e os pesos que representam a máxima concordância são iguais a 1 ($w_{ij} = 1$). A nova equação para a proporção observada é

$$p_o^* = \sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} p_{ij} \tag{11}$$

E a nova equação para a proporção esperada é:

$$p_c^* = \sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} p_{i+} p_{+j} \tag{12}$$

Assim, o Kappa ponderado é obtido como:

$$\widehat{K}_w = \frac{p_o^* - p_c^*}{1 - p_c^*} \tag{13}$$

2.1.4 Índice Tau

O coeficiente *Tau* foi desenvolvido por Klecka (1980) e introduzido no sensoriamento remoto por Ma e Redmond (1995). Similar ao Kappa de Cohen, seu objetivo é analisar o nível de concordância entre os dados da matriz de confusão. Couto (2003) afirma que *Tau* mede o progresso da acurácia dos dados de classificação em função do ajuste aleatório de amostras para as classes.

Ma e Redmond (1995) relatam que a diferença crítica entre os *Kappa* e *Tau* está nas probabilidades, enquanto o *Kappa* é baseado na probabilidade a *posteriori*, *Tau* usa a probabilidade a *priori*, ou seja, o *Kappa* depende da matriz de confusão e o *Tau* pode ser calculado antes mesmo da construção da matriz de confusão. De acordo com Klecka (1980) o coeficiente de *Tau* pode ser calculado através da seguinte equação:

$$T = \frac{p_o - p_r}{1 - p_r} \tag{14}$$

Para probabilidades a *prori* de associação iguais entre os grupos:

$$T_e = \bar{\tau} = \frac{p_o - (1/k)}{1 - (1/k)} \tag{15}$$

T_e representa o coeficiente Tau para casos onde as classificações são baseadas em probabilidades iguais de pertencer à classe. Ma e Redmond (1995) afirmam que T_e aproxima da acurácia total (p_o) à medida que o número de classes aumenta.

2.1.5 Pi de Scott

O índice Pi de Scott (Scott's π), desenvolvido por Scott (1955), corrige a concordância esperada através do número de classes e a frequência com que são utilizadas (Scott, 1955). Sua equação é similar ao Kappa (Eq. 16), diferenciando-se no cálculo da concordância esperada obtida através das proporções de cada classe em toda a amostra de unidades do conteúdo (Craig, 1981; Riff, Lacy e Fico, 2014, p.132).

$$\pi = \frac{p_o - p_s}{1 - p_s} \tag{16}$$

A concordância esperada depende do número de classes e da frequência com que cada um deles são utilizados em um dado teste (Scott, 1955; (Riffe, Lacy e Fico, 2014, pag. 130). Assumindo que as proporções de concordância são iguais entre as classificações e as unidades nas classes de toda população são conhecidas (Craig, 1981). O valor de p_s é dado por

$$p_s = \sum_{i=1}^{k} \left(\frac{p_{i+} + p_{+i}}{2} \right)^2 \quad (17)$$

O índice pode variar de 0 a 1, independentemente do número de classes, permitindo uma comparação com a acurácia total (Craig, 1981). Quanto mais próximo de um, maior é a concordância dos classificadores. A concordância esperada mínima ou incerteza máxima acontece quando todas as classes são utilizadas com igual frequência, e aumenta quando há qualquer desvio de distribuição uniforme de frequências entre as classes (Scott, 1955).

Pi de Scott é útil, principalmente, em situações em que a avaliação de confiabilidade é fundamentada na concordância de todo ou qualquer tamanho de subconjuntos, independentemente dos classificadores (Craig, 1981). Pi de Scott pode ser considerada uma medida aproximada quando a confiabilidade da classificação excede o esperado (acaso).

3. Resultados

Em um primeiro cenário foram avaliadas as medidas de acurácia com igual probabilidade de que uma amostra caia em uma das *k* classes, variando as classes *k* = 3, 5 e 10 e mantendo o número de amostras primeiramente igual a 200 e posteriormente igual a 500.

Quando *n* = 200, o índice de acurácia total (p_o) reduz com o aumento do número de classes (Figura 1) devido a redução da soma da diagonal da matriz de confusão, uma vez que este índice utiliza somente a diagonal da matriz de confusão (ignorando os valores fora da diagonal). Ao aumentar o número de classes, os valores mínimos e máximos da soma da diagonal principal
reduziram (Figura 2).

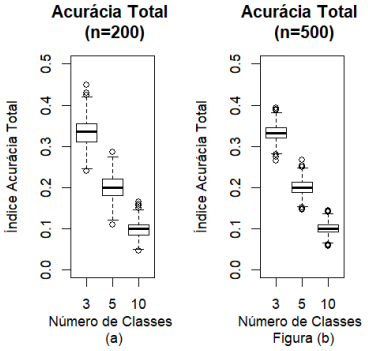


Figura 1: *boxplot* do Índice de Acurácia Total para *n* = 200

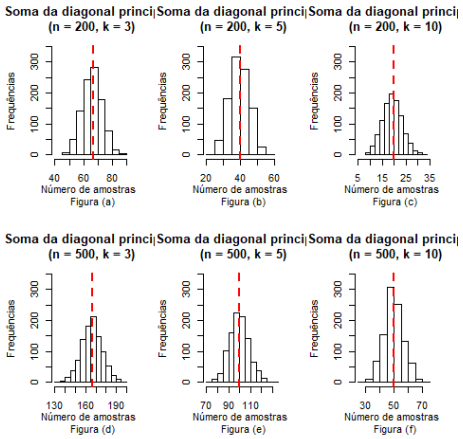


Figura 2: histogramas referentes à frequência da soma da diagonal principal para as classes *k* = 3 (Figura (a)), *k* = 5 (Figura (b)) e *k* = 10 ((Figura (c)) para n = 200 e classes *k* = 3 (Figura (d)), *k* = 5 (Figura (e)) e *k* = 10 ((Figura (f)) para n = 500.

Para um número de amostras igual a 200, as classes apresentaram níveis médios diferentes, as classes 3 e 5 apresentaram uma distribuição simétrica e a classe 10 apresentou uma distribuição assimétrica à esquerda (Figura 1 (a)). Para *n* = 500, a acurácia total também reduziu com o aumento do número de amostras e as classes 3 e 5 permaneceram distribuídas simetricamente, já a classe 10 apresentou uma assimetria à direita. O aumento das amostras de 200 para 500 fez com que os índices se concentrassem mais próximos da mediana e da média, como esperado.

O Kappa de Cohen sempre apresenta valor menor que a acurácia total porque, diferentemente do Índice de Acurácia Total, este considera todos os valores fora da diagonal da matriz de confusão. A Figura 3 (a), mostra que para *n* = 200 os intervalos variaram de -0.15 a 0.15, e as três classes apresentam distribuição simétrica.

Na equação do Kappa de Cohen são considerados os valores fora da diagonal principal (p_c), portanto coeficientes negativos ocorrem quando p_c é maior que p_o (proporção da soma da diagonal), mostrando que quanto menor p_c, menor será o valor de Kappa de Cohen (Figura 3 a). Para *n* = 500, houve um aumento no Índice de acurácia do Kappa de Cohen, o intervalo variou entre -0.1049 e 0.0474 (Figura 3 (b)). O comportamento dos dados, assim como n = 200, também apresentou distribuição simétrica para cada classe, e se concentra cada vez mais próximo de zero à medida que o número de classes aumenta.

O Kappa Ponderado não sofreu muitas alterações com o aumento de classes para amostras iguais a 200 e 500. A distribuição em ambos os casos foi simétrica. E, assim como o Kappa de Cohen, quando se aumenta o número de amostras, mantendo o número de classes, o índice de acurácia aumenta.

O coeficiente Pi de Scott resulta em valores muito próximos ao índice de Kappa de Cohen, ficando no mesmo intervalo entre -0.15 e 0.15. E os coeficientes negativos ocorrem pelo mesmo motivo que Kappa, senão aqui a proporção esperada corrigida (p_s) é maior que a proporção observada. As médias e seus desvios-padrões para as classes 3, 5 e 10 foram respectivamente - 0.0033±0.0517, -0.0035±0.0354, -0.0035±0.0227, sendo muito próximas umas das outras indicando uma baixa variabilidade dos dados. As classes 3 e 5 apresentaram para os dois casos uma distribuição simétrica, já para *k* = 10, no primeiro caso (*n*=200), os dados tiveram uma distribuição negativamente assimétrica e, no segundo caso (*n*=500), a distribuição foi assimétrica positiva.

O cálculo do Índice Tau subtrai da proporção observada a proporção relacionada ao número de classes (1/*k*), resultando em valores menores que Scott e Kappa, porém com intervalos muito próximos ao coeficiente de Pi de Scott e de Kappa (-0.15 e 0.15). A distribuição dos dados foi similar à distribuição da estatística Pi de Scott em ambos os casos para *n* = 200 e 500.

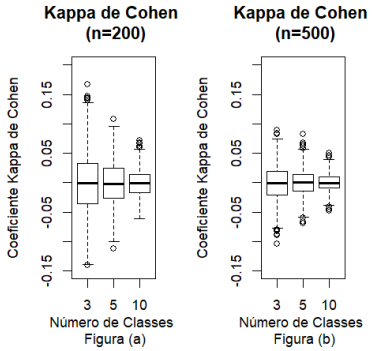


Figura 3: *boxplot* do coeficiente Kappa de Cohen para n = 200

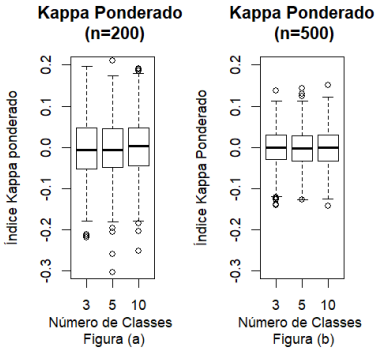


Figura 4: *boxplot* do coeficiente Kappa Ponderado

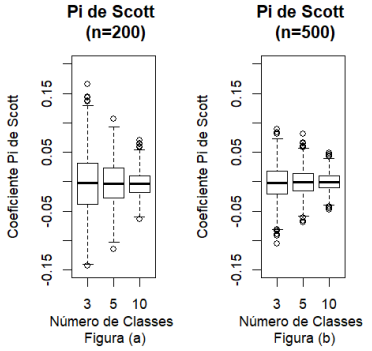


Figura 5: *boxplot* do Coeficiente Pi de Scott

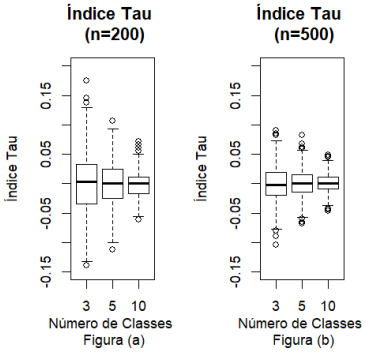


Figura 06: *boxplot* do Índice Tau

Em todos os casos, um valor pequeno da soma da diagonal influenciou o resultado dos índices, produzindo um baixo valor de proporção observada. Para os índices Kappa de Cohen, Kappa Ponderado e Pi de Scott ocorreu um o aumento das proporções esperadas dos índices, gerando índices com valores negativos.

No segundo cenário foram analisadas as medidas de acurácia com probabilidade maior para cada célula na diagonal da matriz, variando as classes k = 3, 5 e 10 e mantendo o número de amostras primeiramente igual a 200 e posteriormente igual a 500. Com a probabilidade de as amostras caírem na diagonal maior que fora da diagonal, os índices de acurácia total para ambos os números de amostras foram maiores quando comparados com o primeiro cenário (Figura 7). Para n = 200, os valores da acurácia se mantêm muito próximos para os três números de classes e os *outliers* diminuem quando se aumenta o número de classes. Para n = 500, os índices das matrizes com 3 e 5 classes foram próximos, já com 10 classes a acurácia reduziu e o número de *outliers* aumentou. A distribuição foi simétrica para ambos os tamanhos de amostras, em que no primeiro (n = 200) o número de outliers reduziu com o aumento da quantidade de classes e para n = 500 aconteceu o inverso.

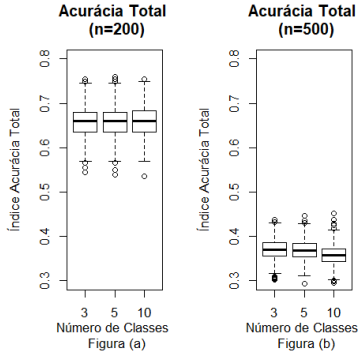


Figura 07: *boxplot* do Índice de acurácia total

A soma da diagonal principal para amostras de tamanho *n* = 200 não variou muito para as três classes, permanecendo próximo de 131, o que justifica os índices de acurácia total serem próximos. Já para a amostra *n* = 500, a soma da diagonal nas três classes ficou próxima de 180.

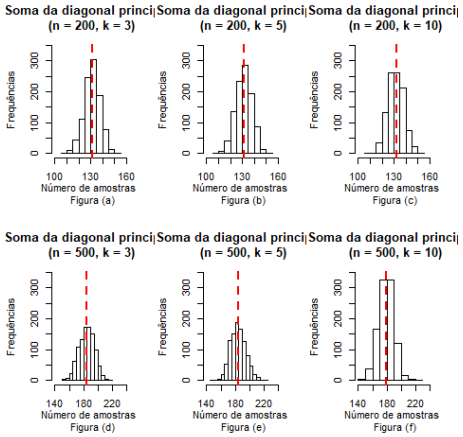


Figura 8: histogramas referentes à frequência da soma da diagonal principal para as classes *k* = 3 (Figura (a)), *k* = 5 (Figura (b)) e *k* = 10 ((Figura (c)) para n = 200 e classes *k* = 3 (Figura (d)), *k* = 5 (Figura (e)) e *k* = 10 ((Figura (f)) para n = 500

O coeficiente de Kappa aumentou quando as classes aumentaram, atingindo um valor máximo de 0.7265 para n = 200 e 0.3912 para n = 500. Porém, para o número de classes igual a 3, os coeficientes mínimos foram negativos (Figura 9). O cenário muda quando se aumenta o número de classes para 5 e 10, ocorrendo um aumento dos coeficientes. Quando n=500, os dados ficam mais concentrados em torno da média do que os dados quando n=200.

O Kappa ponderado teve seus valores ampliados e, assim como Kappa de Cohen, os valores mínimos foram negativos quando k=3, tanto em n=200 quanto n=500 (Figura 10). E como ocorrido no cenário 1, neste cenário 2 os índices Pi de Scott e Tau apresentaram comportamentos similares ao Kappa de Cohen (Figuras 11 e 12), além dos intervalos muito próximo deste.

Estes resultados são intuitivos, uma vez que os índices Kappa de Cohen, Kappa Ponderado e Pi de Scott dependem das proporções marginais, não sendo possível a validação dos dados neste trabalho.

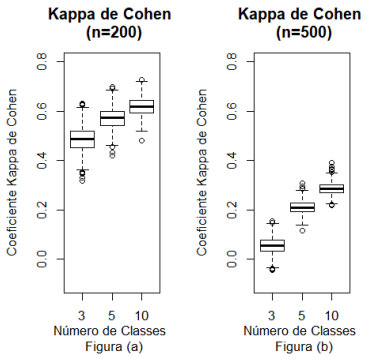


Figura 9: *boxplot* do Coeficiente de Kappa de Cohen

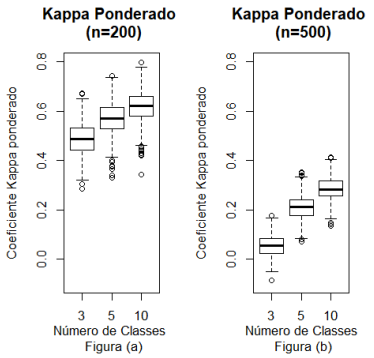


Figura 10: *boxplot* do Índice do Kappa ponderado

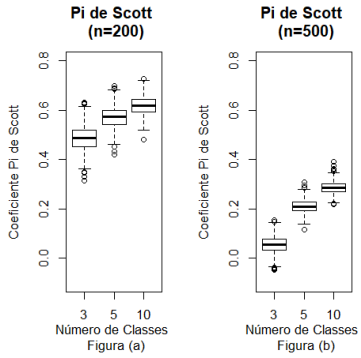


Figura 11: *boxplot* do Índice de Pi de Scott

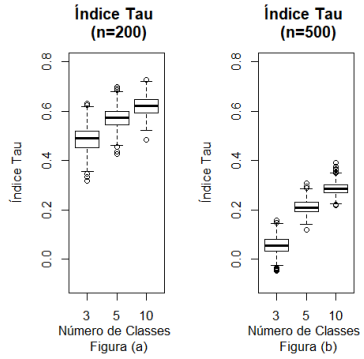


Figura 12: *boxplot* do Índice de Tau

4. Conclusão

No primeiro cenário foi observado que, mantendo as amostras fixas e aumentando as classes, os índices de acurácia diminuem. Já aumentando as amostras e mantendo a classe, os índices de acurácia aumentam e se concentram mais próximos da média. No segundo cenário, os índices de acurácia apresentaram valores melhores quando k = 200, não necessariamente significando que um maior número de amostras leva a um maior índice de acurácia.

Valores reduzidos dos índices de acurácia estão ligados ao número de acertos, ou seja, quanto menor o valor da diagonal principal, menor será o valor dos índices. Apesar dos índices (exceto Acurácia Total) incluírem a proporção esperada em seus cálculos, um pequeno valor da proporção observada acarretará a proporção esperada maior, resultando até mesmo em uma acurácia negativa. Este resultado é intuitivo, uma vez que os índices Kappa de Cohen, Kappa Ponderado e Pi de Scott dependem das proporções marginais, não sendo possível a validação deste.

Referências

Anderson, J.R., Hardy, J.T., Roach, J.T. and Witmer, R.E. (1976). A Land Use and Land Cover Classification System for Use with Remote Sensor Data. *Geological Survey Professional Paper 964*. 28 p.

Card, D. H. (1982) Using known map category marginal frequencies to improve estimates of thematic map accuracy. *Photogrammetric Engineering and Remote Sensing*, 48, pp. 431-439.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20 (1), pp. 37-46. doi: . https://doi.org/10.1177/001316446002000104

Congalton, R. G. and Mead, R. A. (1983). A quantitative method to test for consistency and correctness in photo interpretation. *Photogrammetric Engineering and Remote Sensing*, 49: 69–74.

Congalton, R.G.; Oderwald, R.; Mead, R. Assessing Landsat classification accuracy using discrete multivariate analysis statistical techniques. Photogramm. Eng. Remote Sens. 1983, 49, 1671–1678.

Congalton, R.G. (1991) A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sensing Environment*, 37, 35–46. doi: https://doi.org/10.1016/0034-4257(91)90048- B

Costa, H., Almeida, D., Vala, F., Marcelino, F. and Caetano, M. (2018) Land Cover Mapping from Remotely Sensed and Auxiliary Data for Harmonized Official Statistics. *ISPRS*

Internatinal Journal Geo-Information. 7, 157. doi:10.3390/ijgi7040157

Couto, P. (2003) Assessing the accuracy of spatial simulation models. Ecological Modelling, 167, 181–198. doi:10.1016/S0304-3800(03)00176-5

Craig RT. (1981) Generalization of Scott's index of intercoder agreement. *Public Opinion Quarterly.*45:260–4.

Dos Reis, A.A., Carvalho, M.C., de Mello, J.M., Gomide, L.R., Ferraz Filho, A.C. and Acerbi Junior, F.W. (2018) Spatial prediction of basal area and volume in Eucalyptus stands using Landsat TM data: An assessment of prediction methods. *New Zealand Journal of Forestry Science.* 48, 1. doi: 10.1186/s40490-017-0108-0

Fleiss J.L., Cohen J., Everitt B. (1969) Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, 72(5):323-327. doi:10.1037/h0028106

Haack, B., Mahabir, R. and Kerkering, J. (2015) Remote sensing-derived national land cover land use maps: a comparison for Malawi. *Geocarto International*, v. 30, n.3, 270–292 p., http://dx.doi.org/10.1080/10106049.2014.952355

Hughes, M. A., and Garrett, D. E. (1990). Intercoder reliability estimation approaches in marketing: a generalizability theory framework for quantitative data. Journal of Marketing Research, 27(2), 185–195. https://doi.org/10.2307/3172845

Khatami, R., G. Mountrakis, and S. V. Stehman. (2017). Mapping Per-Pixel Predicted Accuracy of Classified Remote Sensing Images. *Remote Sensing of Environment* 191: 156–167. doi:10.1016/j.rse.2017.01.025.

Klecka, W. R. (1980). Discriminant Analysis, Sage Publications Inc., Beverley Hills, First edition

Koukoulas, S. and Blackburn, G. A., (2001). Introducing New Indices for Accuracy Evaluation of Classified Images Representing Semi-Natural Woodland Environments. *Photogrammetric Engineering and Remote Sensing*, 67(4), pp. 499–510.

Landis, J.R. and Koch, G. G. (1977) The measurement of observer agreement for categorical data.
Biometrics, 33, pp. 159-174

Li, W., P. Ciais, S. Peng, Y. Chao, Y. Wang, M. Thurner, S. S. Saatchi, et al. (2017). Land-Use and Land-Cover Change Carbon Emissions between 1901 and 2012 Constrained by Biomass Observations. *Biogeosciences* 14, 22, 5053. doi:10.5194/bg-14-5053-2017.

Liu, J. and Duan, Z. (2018). Quantitative assessment of landslide susceptibility comparing statistical index, index of entropy, and weights of evidence in the Shangnan area, China. *Entropy*. 20, 868.doi:10.3390/e20110868

Lyons, M.B., Keith, D.A., Phinn, S.R., Mason, T.J. and Elith, J. (2018) A comparison of resampling methods for remote sensing classification and accuracy assessment. *Remote Sensing Environment. 208*, 145–153. https://doi.org/10.1016/j.rse.2018.02.026

Ma, Z. and Redmond, R. L. (1995) Tau coefficients for accuracy assessment of classification of remote sensing data. *Photogrammetric Engineering and Remote Sensing*, 61, pp. 435-439

Naesset, E. (1996) Use of the weighted Kappa coefficient in classification error assessment of thematic maps. *International Journal Geography Information System*, 10, pp. 591-604. doi: https://doi.org/10.1080/02693799608902099

Ottosen, Thor-Bjørn, Lommen, S. T.E., Skj\øth, C.A. (2019). Remote sensing of cropping practice in Northern Italy using time-series from Sentinel-2. *Computers and Electronics in Agriculture* 157, 232–238. https://doi.org/10.1016/j.compag.2018.12.031

Riffe, D., Lacy, S. and Fico, F. G. (2014). Analyzing Media Messages: Using Quantitative Content Analysis in Research. Lawrence Erlbaum Associates, London, third edition.

Rosenfield, G. and Fitzpatrick-Lins, K. (1986). A coefficient of agreement as a measure of thematic classification accuracy. Photogrammetric Engineering and Remote Sensing, 2 (52), pp. 223-227

Salk, C., Fritz, S., See, L., Dresel, D. and McCallum, I. (2018) An exploration of some pitfalls of thematic map assessment using the new map tools resource. *Remote Sensing*., 10, pp. 1-14. doi:10.3390/rs10030376

Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, pp. 321-325

Sokal, R.R. (1974). Classification: Purposes, principles, progress, prospects. *Science*, 185: 1115– 1123.

Stehman S.V. (2006). Design, analysis, and inference for studies comparing thematic accuracy of classified remotely sensed data: A special case of map comparison. *Journal of Geographical Systems*, 8, pp. 209-226

Stehman, S.V. and Wickham, J.D. (2011). Pixels, blocks of pixels, and polygons: Choosing a spatial unit for thematic accuracy assessment. *Remote Sensing of Environment*, 115, pp. 3044-3055. doi:10.1016/j.rse.2011.06.007

Story, M. and Congalton, R.G (1986). Accuracy Assessment: A User's Perspective. *Photogrammetric Engineering and Remote Sensing*, *52*, 397–399.
Turner, M.G (2005). Landscape Ecology: What Is the State of the Science? *Annual Review of Ecology, Evolution, and Systematics*. 36, 319–344.doi: 10.1146/annurev.ecolsys.36.102003.152614

Whitehead, K.; Hugenholtz, C.H (2014). Remote sensing of the environment with small unmanned aircraft systems (UASs), part 1: A review of progress and challenges. *Journal of Unmanned Vehicle Systems*, 02, 69–85. doi:10.1139/juvs-2014-0007.