O Kappa está morto?
Uma discussão sobre índices baseados em matrizes de confusão

Camilo Daleles Rennó

DPI-GEO-REFERATA
São José dos Campos, 16 de setembro de 2015

Pontius X Congalton

Congalton, R. G.; Oderwald, R. G.; Mead, R. A. Assessing Landsat classification accuracy using discrete multivariate analysis statistical techniques. Photogrammetric Engineering and Remote Sensing, 49: 1671-1678, 1983

Congalton, R. G.; Green, K. Assessing the accuracy of Remote Sensed Data: principles and practices. 2 ed. Boca Raton, CRC Press. 2009

Pontius, R. G.; Millones, M. Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. International Journal of Remote Sensing, 32(15): 4407-4429, 2011 ***

Pontius, R. G.; Santacruz, A. Quantity, exchange, and shift components of difference in a square contingency table. International Journal of Remote Sensing, 35(21): 7543-7554, 2014

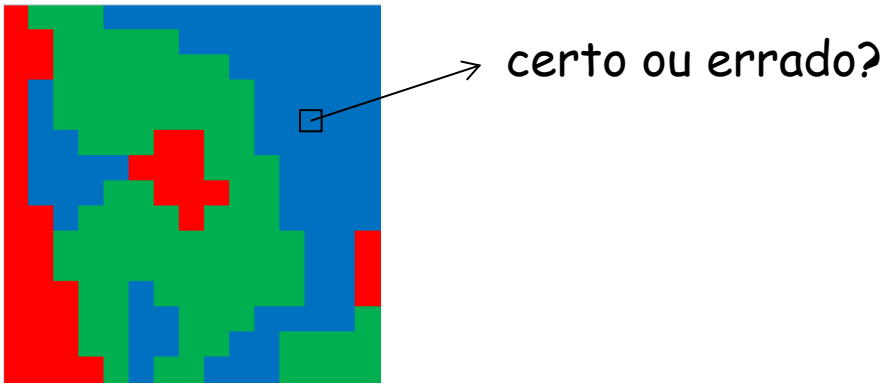
"We hope this article marks the end of the use of Kappa and the beginning of the use of these two components: quantity disagreement and allocation disagreement."

Avaliação dos Erros de Classificação

Numa classificação tradicional (rígida), considera-se que todo elemento (pixel ou polígono) está associado a uma única classe temática.

Resultado da classificação: imagem classificada ou mapa temático

O erro surge sempre que esta associação diverge da VERDADE.



mapa

A VERDADE, muitas vezes, representa apenas uma REFERÊNCIA (resultado ideal).

Avaliação dos Erros de Classificação

A referência pode ser obtida a partir de:

- dados pré-existentes (levantamentos, mapas, literatura, etc)
  - apesar de ter custo quase zero, as informações podem estar desatualizadas
  - erros pré-existentes são raramente conhecidos
  - pode haver diferentes definições para a mesma classe (diferença semântica)

- dados de campo
  - em geral, envolve um custo elevado (logística, localização precisa, equipes grandes, etc)
  - pode haver grande defasagem temporal entre a obtenção do dado usado na classificação e daquele utilizado na checagem (comum em ambientes dinâmicos)
  - a amostragem pode ser enviesada (somente pontos com fácil acesso são checados)

- fotointerpretação (geralmente usando imagens com resolução mais fina)
  - apesar do baixo custo, o resultado depende a experiência do fotointérprete (ideal: diferente de quem fez a classificação)
  - pode ser enviesado ao induzir um resultado positivo quando já se conhece o resultado da classificação (ideal: total independência entre os processos)
  - também pode ter problemas em ambientes dinâmicos (defasagem temporal)

Avaliação dos Erros de Classificação

Fatores importantes a serem considerados durante a avaliação dos erros:

- Unidade amostral utilizada na avaliação
- Dependência espacial (tamanho e distribuição das amostras)
- Representação da exatidão (acertos) e dos confusões (erros)

Unidade Amostral

A comparação entre um mapa temático e uma referência pode ser feita por:

- pontos simples (ou pixels de uma imagem ou células de uma grade)
- grupos de pontos (avaliação contextual)
- polígonos (ou objetos)
- grupos de polígonos

Importante: cada unidade amostral representa apenas uma verificação
a unidade amostral a ser avaliada não depende do classificador
(p. ex. classificador por região pode ser avaliado por pontos)

Dependência Espacial

A utilização de pontos amostrais (ou polígonos) próximos cria a falsa impressão de que o conjunto amostral é grande quando, na realidade, há muita redundância de informação

A presença de dependência espacial impacta principalmente a avaliação da incerteza (variância) das estimativas obtidas a partir da amostra

Para minimizar esses problemas, determina-se qual a distância (Lag em x e em y) a partir do qual os pontos escolhidos podem ser considerados independentes

Para isso, deve-se conhecer a estrutura de relação espacial dos dados utilizados (função de autocorrelação e/ou semi-variograma)

Pode ser negligenciado se o tamanho da amostra for muito menor do que o tamanho total do mapa a ser avaliado e se a escolha das amostras for totalmente aleatória

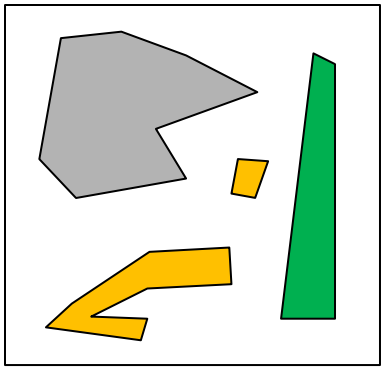
Atenção especial para pontos coletados em campo (alta concentração) e classificações de imagens segmentadas (poucos polígonos)

Representação da Exatidão

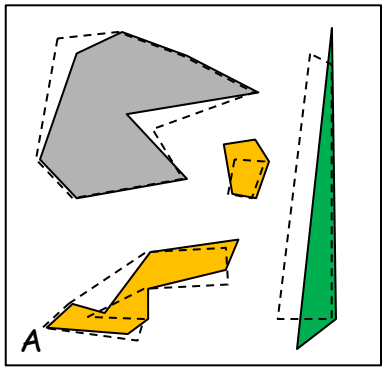
A avaliação da exatidão de uma classificação pode ser feita através de:

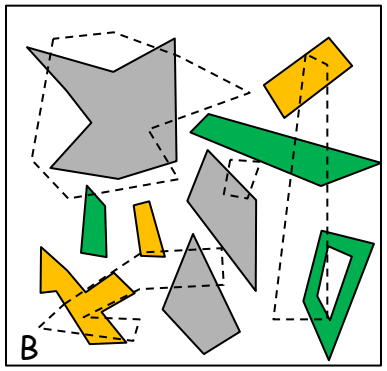
- Tabelas (matriz de confusão – acertos e confusões entre classes)
- Índices (global ou por classe)
- Mapas de incerteza (grades numéricas ou mapas temáticos)

Exatidão Temática

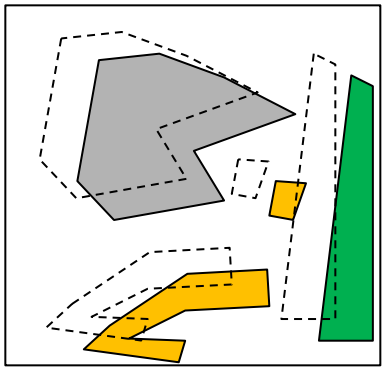


mapa de referência
(verdade)





problemas de
deslocamento
diferenças de
resolução
espacial



erros de comissão (inclusão) e
omissão (exclusão) se compensam

Qual dos mapas A ou B é é melhor com relação a referência?

com relação ao posicionamento: A

com relação à área total de cada classe: B

Matriz de Confusão

A partir dos pares de pontos (polígonos ou grupos) avaliados, constrói-se a

Matriz de Confusão

(matriz de erro)

Referência

Classificação	1	2	⋯	c	Total
1	x_{11}	x_{12}	⋯	x_{1c}	x_{1+}
2	x_{21}	x_{22}	⋯	x_{2c}	x_{2+}
⋮	⋮	⋮	⋱	⋮	⋮
c	x_{c1}	x_{c2}	⋯	x_{cc}	x_{c+}
Total	x_{+1}	x_{+2}	⋯	x_{+c}	n

x_{ij}: número de pontos da classe j (referência), classificados na classe i (classificação)

x_{kk}: número total de pontos corretamente classificados da classe k

x_{+j}: número total de pontos avaliados da classe j na referência

x_{i+}: número total de pontos avaliados da classe i na classificação

Matriz de Confusão

Observações importantes:

- considera que as classes são excludentes (cada ponto pertence a apenas uma classe);
- todos os pontos avaliados devem pertencer a alguma classe, ou seja, o classificador não pode considerar a classe "não classificado";
- a utilização de classes muito semelhantes (representando subtipos) pode induzir a um excesso de erros (ou confusões) que certamente prejudicarão a avaliação global da classificação;
- a interpretação dos resultados está diretamente dependente da unidade amostral adotada;
- índices globais podem estar enviesados para a amostra avaliada

Avaliação da Exatidão

Classificação \ Referência	A	B	C	D	Total
A	13	8	0	0	21
B	8	10	0	3	21
C	0	5	27	4	36
D	0	0	0	32	32
Total	21	23	27	39	110

Exatidão Total (ou Global) $= \dfrac{\sum_{k=1}^{c} x_{kk}}{n}$

- mínimo = 0
- máximo = 1 (ou 100%)

$$= \frac{13 + 10 + 27 + 32}{110} = 74{,}5\%$$

Avaliação da Exatidão

Classificação \ Referência	A	B	C	D	Total
A	13	8	0	0	21
B	8	10	0	3	21
C	0	5	27	4	36
D	0	0	0	32	32
Total	21	23	27	39	110

O que significa uma exatidão de 74,5% ?

- Se uma amostra, dentre as 110, fosse escolhida ao acaso, a probabilidade desta estar corretamente classificada seria de 74,5%

- Se um ponto (ou polígono) fosse escolhido ao acaso no mapa, a probabilidade desta estar corretamente classificada seria de 74,5% se a amostragem representar as reais proporções de cada classe.

Avaliação da Exatidão

Classificação \ Referência	A	B	C	D	Total
A	13	8	0	0	21
B	8	10	0	3	21
C	0	5	27	4	36
D	0	0	0	32	32
Total	21	23	27	39	110

Pontius e Millones (2011) sugerem sempre utilizar a matriz não enviesada (ajustada para as proporções reais de cada classe)

	Proporção		
A	25%	1%	1%
B	25%	2%	95%
C	25%	95%	3%
D	25%	2%	1%
exatidão	71,9%	98,1%	45,7%

Avaliação da Exatidão

Ponto de vista do Produtor (Referência)

Quanto da classe k foi "vista" pelo classificador?

Classificação \ Referência	A	B	C	D	Total
A	13	8	0	0	21
B	8	10	0	3	21
C	0	5	27	4	36
D	0	0	0	32	32
Total	21	23	27	39	110

$$\text{Exatidão do Produtor da classe } k = \frac{x_{kk}}{x_{+k}}$$

$$\text{Exatidão do Produtor da classe B} = \frac{10}{23} = 43{,}5\%$$

$$\text{Erro de omissão da classe B} = \frac{8+5+0}{23} = 56{,}5\%$$

Avaliação da Exatidão

Ponto de vista do Consumidor (Classificação)

Quanto do que foi classificado como k é realmente da classe k?

Classificação \ Referência	A	B	C	D	Total
A	13	8	0	0	21
B	8	10	0	3	21
C	0	5	27	4	36
D	0	0	0	32	32
Total	21	23	27	39	110

Exatidão do Consumidor da classe k = $\frac{x_{kk}}{x_{k+}}$

Exatidão do Consumidor da classe B = $\frac{10}{21} = 47{,}6\%$

Erro de comissão (inclusão) da classe B = $\frac{8+0+3}{21} = 52{,}4\%$

Avaliação da Exatidão

Classificação \ Referência	A	B	C	D	Total
A	13	8	0	0	21
B	8	10	0	3	21
C	0	5	27	4	36
D	0	0	0	32	32
Total	21	23	27	39	110

Classe	exatidão produtor	erro omissão	exatidão consumidor	erro inclusão
A	61,9%	38,1%	61,9%	38,1%
B	43,5%	56,5%	47,6%	52,4%
C	100,0%	0,0%	75,0%	25,0%
D	82,1%	17,9%	100,0%	0,0%

Avaliação da Exatidão

Classificação \ Referência	A	B	C	D	Total
A	13	8	0	0	21
B	8	10	0	3	21
C	0	5	27	4	36
D	0	0	0	32	32
Total	21	23	27	39	110

Exatidão Total = 74,5%

E se a classificação fosse realizada de modo totalmente aleatório?

Avaliação da Exatidão

$$\frac{21 * 21}{110}$$

Classificação \ Referência	A	B	C	D	Total
A	4,01	4,39	5,15	7,45	21
B	4,01	4,39	5,15	7,45	21
C	6,87	7,53	8,84	12,76	36
D	6,11	6,69	7,85	11,35	32
Total	21	23	27	39	110

$$\text{Exatidão Total} = \frac{4{,}01 + 4{,}39 + 8{,}84 + 11{,}35}{110} = 26{,}0\%$$

Ou seja, 26,0% do acerto pode ter sido conseguido de modo casual !!!

Medida de Concordância Kappa

Classificação \ Referência	A	B	C	D	Total
A	13	8	0	0	21
B	8	10	0	3	21
C	0	5	27	4	36
D	0	0	0	32	32
Total	21	23	27	39	110

Índice Kappa (K) – medida de concordância

$$\hat{\kappa} = \frac{\theta_1 - \theta_2}{1 - \theta_2}$$

$$\theta_1 = \frac{\sum_{k=1}^{c} x_{kk}}{n}$$

exatidão total (observada)

$$\theta_2 = \frac{\sum_{k=1}^{c} x_{k+}\, x_{+k}}{n^2}$$

exatidão total (se classificação fosse aleatória)

K
- mínimo =< 0 $(\theta_1 < \theta_2)$
- máximo = 1 $(\theta_1 = 1)$

Índice Kappa

Classificação \ Referência	A	B	C	D	Total
A	13	8	0	0	21
B	8	10	0	3	21
C	0	5	27	4	36
D	0	0	0	32	32
Total	21	23	27	39	110

Índice Kappa (K) – medida de concordância

$$\theta_1 = \frac{\sum_{k=1}^{c} x_{kk}}{n} = 0{,}745$$

$$\theta_2 = \frac{\sum_{k=1}^{c} x_{k+}\, x_{+k}}{n^2} = 0{,}260$$

$$\hat{\kappa} = \frac{\theta_1 - \theta_2}{1 - \theta_2} = \frac{0{,}745 - 0{,}260}{1 - 0{,}260} = 0{,}6561$$

Se a classificação fosse totalmente aleatória, qual seria o valor esperado para o kappa? zero ⇨ Teste de hipótese

Índice Kappa

$$\hat{\kappa} = \frac{\theta_1 - \theta_2}{1 - \theta_2} \qquad \theta_1 = \frac{\sum_{k=1}^{c} x_{kk}}{n} \qquad \theta_2 = \frac{\sum_{k=1}^{c} x_{k+}\, x_{+k}}{n^2}$$

$$Var(\hat{\kappa}) = \frac{1}{n}\left[\frac{\theta_1(1-\theta_1)}{(1-\theta_2)^2} + \frac{2(1-\theta_1)(2\theta_1\theta_2 - \theta_3)}{(1-\theta_2)^3} + \frac{(1-\theta_1)^2(\theta_4 - 4\theta_2^2)}{(1-\theta_2)^4}\right]$$

$$\theta_3 = \sum_{k=1}^{c} x_{kk}\left(x_{k+} + x_{+k}\right) \Big/ n^2$$

$$\theta_4 = \sum_{i=1}^{c}\sum_{j=1}^{c} x_{ij}\left(x_{i+} + x_{+j}\right)^2 \Big/ n^3$$

Pressupondo amostras independentes e TLC válido (amostra grande):

$$Z = \frac{\hat{\kappa} - \kappa}{\sqrt{\mathrm{Var}(\hat{\kappa})}} \sim N(0,1)$$

Outros Índices

Kappa condicional

Avalia a concordância para uma determinada classe

Kappa ponderado

Cada célula ij da matriz de confusão pode receber um peso ($0 \leq w_{ij} \leq 1$) permitindo que certos erros sejam mais importantes que outros

Tau

O índice Kappa pressupõe que ambas as probabilidades marginais (classificação e referência) sejam conhecidas antes mesmo da classificação

Para o índice Tau, utiliza-se as probabilidades *a priori* de cada classe (p_k) ao invés de estimá-las pelas proporções marginais obtidas após a classificação

Assim, este índice pode ser obtido por:

$$\hat{\tau} = \frac{\theta_1 - \theta_2'}{1 - \theta_2'} \qquad \theta_1 = \frac{\sum_{k=1}^{c} x_{kk}}{n} \qquad \theta_2' = \frac{\sum_{k=1}^{c} p_k x_{+k}}{n}$$

Na ausência de informação, utiliza-se o mesmo valor para todos p_k (classes equiprováveis, ou seja, $p_k = 1/c$)

Críticas/Sugestão de Pontius

- É sempre melhor (mais útil) focar na discordância e tentar explicar os erros do que focar na concordância e se preocupar como a aleatoriedade explica parte do acerto observado – o que é a essência do Kappa;

- Não conhece nenhum artigo onde os autores mudaram a conclusão quando compararam a exatidão total e o Kappa. Em geral, os autores apenas apresentam simultaneamente a exatidão total e o Kappa junto ao mapa avaliado;

- A avaliação da significância do Kappa é feita pressupondo-se que parte da exatidão total observada é casual (Kappa = 0?). Isso é quase sempre inútil;

- Sugerem particionar os erros em diferentes componentes:

    - Quantity (quantidade, grandeza)
    - Allocation (alocação, atribuição):
        - Exchange (troca, permuta)
        - Shift (mudança, deslocamento)

- Estes índices podem ser empregados na análise de trajetórias (mapas obtidos em 2 datas)

Quantity

	Referência				
	A	B	C	D	Total
A	13	8	0	0	21
B	8	10	0	3	21
C	0	5	27	4	36
D	0	0	0	32	32
Total	21	23	27	39	110

Observe que para a classe A, os erros de omissão são compensados pelos erros de inclusão.

Neste caso, a área estimada pela classificação é a mesma da referência, ou seja, a classificação poderia ser utilizada para estimar a **quantidade** da classe A presente na referência.

Quantity

	Referência				
	A	B	C	D	Total
A	13	8	0	0	21
B	8	10	0	3	21
C	0	5	27	4	36
D	0	0	0	32	32
Total	21	23	27	39	110

Já para classe B, há 2 amostras a mais na referência do que na classificação, o que representa um erro de 1,8% em relação ao total.

Classe	Quantity
A	0,0%
B	1,8%
C	8,2%
D	6,4%
Total	8,2%

Para todas as classes, Quantity = 8,2%

Allocation

	Referência				
	A	B	C	D	Total
A	13	8	0	0	21
B	8	10	0	3	21
C	0	5	27	4	36
D	0	0	0	32	32
Total	21	23	27	39	110

Em toda a matriz observa-se um total de 28 amostras erradas, o que representa 25,5%, ou seja, 1 – exatidão.

Como o erro devido a quantidade foi de 8,2%, então o restante 17,3% é devido a problemas de atribuição.

Mas qual tipo de erro de atribuição? Permuta ou mudança?

Allocation: Exchange or Shift?

	Referência				
	A	B	C	D	Total
A	13	8	0	0	21
B	8	10	0	3	21
C	0	5	27	4	36
D	0	0	0	32	32
Total	21	23	27	39	110

Observe que para a classe A, todos os erros foram devidos à permuta com a classe B, ou seja, amostras classificadas como B foram compensadas pelas amostras da classe B erroneamente classificadas como A. Dessa forma, para a classe A, todo o erro de atribuição foi do tipo exchange totalizando 16 amostras (8 + 8) ou 14,5% do total.

Allocation: Exchange or Shift?

	Referência				
	A	B	C	D	Total
A	13	8	0	0	21
B	8	10	0	3	21
C	0	5	27	4	36
D	0	0	0	32	32
Total	21	23	27	39	110

Já a classe B, 8 amostras erradas foram permutadas com a classe A (Exchange = 14,5%), mas 5 foram erroneamente classificadas como C e 3 foram erroneamente alocadas para a classe B sendo verdadeiramente da classe D.

Nesse caso, houveram 2 erros de quantidade (Quantity = 1,8%), restando 6 amostras erradas por mudança (Shift = 5,5%).

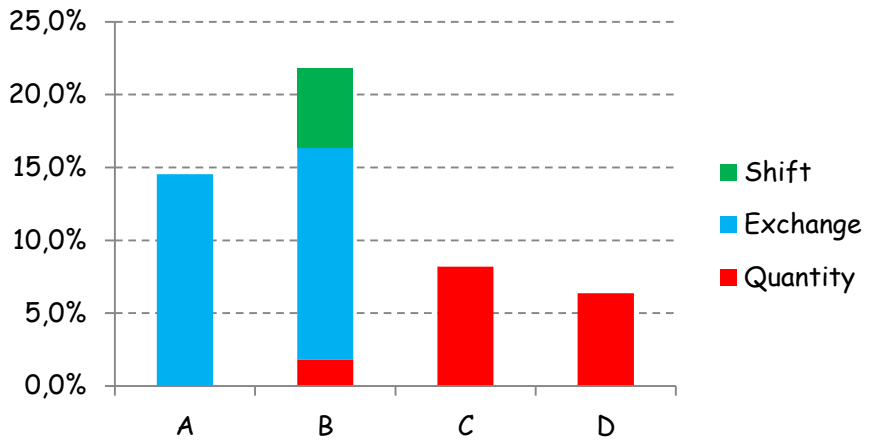
Avaliação Final

Referência

	A	B	C	D	Total
A	13	8	0	0	21
B	8	10	0	3	21
C	0	5	27	4	36
D	0	0	0	32	32
Total	21	23	27	39	110

Classe	Quantity	Allocation
A	0,0%	14,5%
B	1,8%	20,0%
C	8,2%	0,0%
D	6,4%	0,0%
Total	8,2%	17,3%

Exchange	Shift
14,5%	0,0%
14,5%	5,5%
0,0%	0,0%
0,0%	0,0%
14,5%	2,7%



Conclusões

- A Matriz de Confusão representa a parte mais importante na avaliação deste que represente corretamente os acertos e erros entre o mapa e a referência
- Ainda não há uma avaliação da significância dos índices propostos por Pontius, o que dificulta a comparação de resultados entre diferentes métodos por exemplo
- Em geral, as avaliações não consideram o componente espacial
- Os erros deveriam apontar para melhoria no método de classificação
- O uso de técnicas de reamostragem poderiam ser melhor aproveitadas para avaliações das incertezas, principalmente quando se dispõe de uma referência "completa" (por que confiar nos resultados de uma única amostra?)