Revista Tecnologia e Sociedade
ISSN: 1984-3526

https://periodicos.utfpr.edu.br/rts

Identifying factors impacting the overall accuracy in image classification problems: a statistical approach

Wesley Bertoli
Departamento Acadêmico de Estatística, Universidade Tecnológica Federal do Paraná – Curitiba, PR
ORCID: 0000- 0002-4671-1268
wbsilva@utfpr.edu.br

José Marcato Junior
Faculdade de Engenharias, Arquitetura e Urbanismo e Geografia, Universidade Federal de Mato Grosso do Sul – Campo Grande, MS
ORCID: 0000-0002-9096-6866
jose.marcato@ufms.br

Lucas Yuri Dutra de Oliveira
Faculdade de Engenharias, Arquitetura e Urbanismo e Geografia, Universidade Federal de Mato Grosso do Sul – Campo Grande, MS
ORCID: 0000-0002-5958-9193
lucas.oliveira@ufms.br

ABSTRACT

Image classification is a subject of pattern recognition that can be applied in several areas. Obtaining highly-accurate classification involves choosing optimal set-ups from which images will be classified. In this process, controllable variables can affect the overall classification accuracy, such as the image's spatial resolution and the classification method. In this sense, we have designed a factorial experiment where the classification accuracy of an image (from Curitiba, Paraná, Brazil) was obtained from three satellites and three classification methods. The Kruskal-Wallis test was applied to evaluate if the variability across factor levels supports the hypothesis that the experimental factors' effects are statistically significant. Then, we evaluated which factor levels differed from each other using post-hoc tests. Our findings suggest that the image's spatial resolution and the interaction between Satellite and Classification Method are determinants in obtaining accurate image classifications in a geographical context.

KEYWORDS: Factorial design, image classification, Kruskal-Wallis test, overall classification accuracy, spatial resolution.

INTRODUCTION

Anthropogenic actions may cause the most diverse effects on the environment, requiring monitoring to mitigate the impacts. Monitoring land use becomes essential, as the activity carried out can generate impacts such as deforestation, in addition to causing physical and chemical changes in the soil (Araújo *et al*., 2004) and also in the water. The primary way to perform this monitoring of land use is through remote sensing (Vasconcelos and Novo, 2004). Compared with traditional monitoring techniques, remote sensing has advantages as it requires less time and costs to apply (Abdelmalik, 2018), especially when the area of interest is extensive. Several space missions provide us with Earth observation orbital data, such as Landsat, Cbers, Sentinel, Planet, and RapidEye, which, when combined with the correct geoprocessing techniques, provide us with reliable results in the most diverse applications.

An important method for monitoring land use is image classification, which evaluates each pixel and classifies it based on previously established parameters. Some factors can interfere with the classifier's performance, such as the spatial resolution of the image, the classification technique, and the number of samples per class. To perform image classification, we can use traditional remote sensing techniques, such as Pereira and Guimarães (2018) who used Minimum Distance (MD), Maximum Likelihood (ML), and Spectral Angle Mapping (SAM) methods to map the land use and occupation, concluding that the methods present an excellent performance providing reliable results.

There is also the possibility of using machine learning techniques, which present accurate results like traditional techniques. Tian *et al*. (2016) applied the Random Forest method for identifying wetlands in arid regions of China, while Noi and Kappas (2018) applied the Random Forest *k*-Nearest Neighbors and Support Vector Machine methods in the process of classifying land use and occupation ground.

This work aims to analyze data of a factorial experiment, which was conducted to identify which factors significantly influence the image classification process, considering the spatial resolution of the image and classification methods as the qualitative variables to be investigated. We chose the MD, ML, and SAM methods because they are traditional remote sensing techniques and provide equally accurate results compared to more complex techniques.

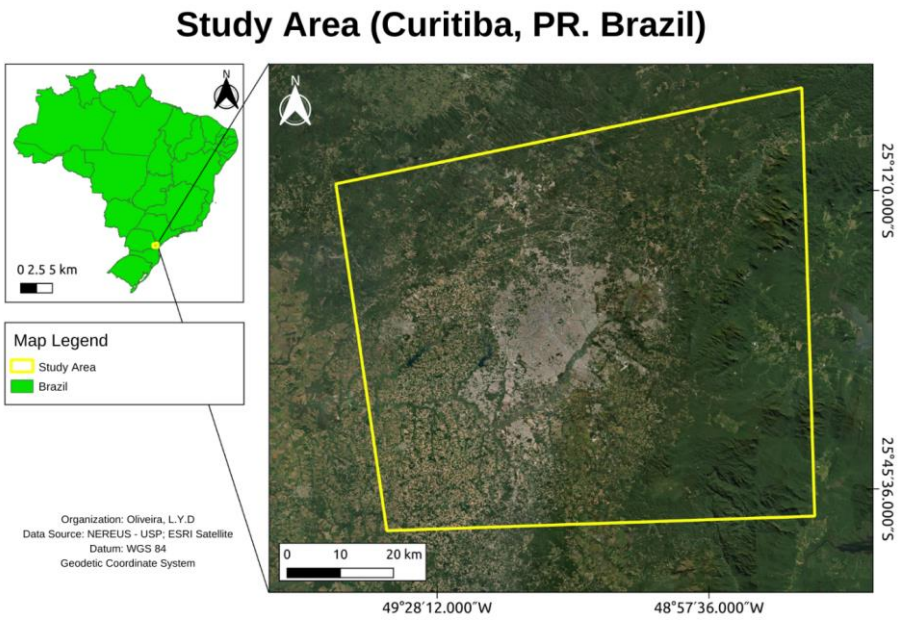
This paper is organized as follows. Section 2 illustrates the study area, presents the image classification techniques, and describes the adopted statistical methods for the collected data analysis. In Section 3, we present and discuss the obtained results. General comments and concluding remarks are addressed in Section 4.

MATERIAL AND METHODS

Study Area

We have selected the city of Curitiba (capital of Paraná state, Brazil) as the study area for this work. The images were cut out (Figure 1), so working with the same area in each classification was possible.

Figure 1 - Geographical representation of the study area.



Source: The Autors

To evaluate controllable factors' influence (and their interactions) in the classification process, we have considered the two variables: spatial resolution of images and classification method, both segmented by the number of samples per class. Three orbital images of the study area were selected (from the same period): Landsat 8, Cbers-4, and Sentinel- 2, which were made available free of charge by the United States Geological Survey (USGS), Instituto Nacional de Pesquisas Espaciais (INPE), and planet.com, respectively. Each image has a different spatial resolution: 30m for Landsat 8, 20m for Cbers-4, and 10m for Sentinel-2.

The Semi-Automatic Classification Plugin (SCP), available in the open-source QGIS software, was used to perform the classification. The SCP calculates the spectral signature of the classes based on the samples of each class (Congedo, 2016), performing the classification for the entire area. We have applied the MD, ML, and SAM classification methods. Each method is briefly explained in the following, as defined by Congedo (2016), adapted from Richards (2013).

Minimum Distance

The MD method is based on the Euclidean distance between spectral signatures, that is

$$d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2},$$

where **x** and **y** are the spectral signature vectors of an image pixel and a training area, respectively. Besides, n denotes the number of image bands. In this method, each pixel in the image is associated with the closest spectral signature according to the discriminant function

$$\boldsymbol{x} \in \mathcal{C}_k \iff d(\boldsymbol{x}, \boldsymbol{y}_k) < d(\boldsymbol{x}, \boldsymbol{y}_j) \quad \forall\, k \neq j,$$

where C_k *(k = 1, 2, 3)* is the *k*-th land cover class and $\boldsymbol{y}_k$ is the associated spectral signature vector.

Maximum Likelihood

The ML method is based on the Bayes' Theorem, which approximates the classes' probability distribution and then estimates which class each pixel belongs to. The Multivariate Normal distribution is typically adopted in this context (Richards, 2013). For this method, a sufficient number of pixels in each training sample is required so that it is possible to estimate the data covariance matrix. Following the Multivariate Normal distribution, the pixel vector's density is given by

$$f_k(\boldsymbol{x}) = \log p(\mathcal{C}_k) - \frac{1}{2} \log |\boldsymbol{\Sigma}_k| - \frac{1}{2} (\boldsymbol{x} - \boldsymbol{y}_k)^\intercal \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x} - \boldsymbol{y}_k),$$

where $p(C_k)$ is the probability that the correct class is C_k and $\boldsymbol{\Sigma}_k$ is the data covariance matrix in class C_k. For this method, the discriminant function is given by

$$\boldsymbol{x} \in \mathcal{C}_k \iff f_j(\boldsymbol{x}) < f_k(\boldsymbol{x}) \quad \forall\, k \neq j.$$

The ML method is one of the most used methods in supervised classifications (Congedo, 2016), even presenting a slower processing time than the MD method in most cases.

Spectral Angle Mapping

Finally, we have the SAM method, widely used for evaluating hyperspectral data. This method determines the spectral similarity between two spectra (signatures images and pixel training samples) by calculating their angle. Kruse *et al.* (1993) define the spectral angle as

$$\theta(\boldsymbol{x}, \boldsymbol{y}) = \cos^{-1}\left[\frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}}\right].$$

In this method, we classify the pixels as belonging to the class having the lowest angle, that is

$$\boldsymbol{x} \in \mathcal{C}_k \iff \theta(\boldsymbol{x}, \boldsymbol{y}_k) < \theta(\boldsymbol{x}, \boldsymbol{y}_j) \quad \forall\, k \neq j.$$

Data

The working dataset was built by matching the Satellite and Method levels for each number of samples per class, which yielded 27 classifications. The classes adopted in the classification process were: water bodies, vegetations, and urbanized areas. The adopted number of samples per class was 1, 5, and 10, being the greater the number of samples, the greater the number of pixels selected as training for the classifier. The Overall Classification Accuracy (OCA) for each factor level and replicated (samples) combination is presented in Table 1. The variable Satellite was codified as 1 for Landsat 8, 2 for Cbers-4, and 3 for Sentinel-2. As for the variable Methods, the adopted codes were 1 for MD, 2 for ML, and 3 for SAM.

Table 1 - Overall classification accuracy for each factor level and replicated combination

Satellite	Method	Sample	OCA (%)
1	1	1	11.56
1	2	1	9.51
1	3	1	11.54
1	1	5	11.04
1	2	5	0.26
1	3	5	11.28
1	1	10	11.09
1	2	10	0.44
1	3	10	11.27
2	1	1	100.00
2	2	1	94.94
2	3	1	13.77
2	1	5	99.94
2	2	5	89.45
2	3	5	13.63
2	1	10	99.94
2	2	10	95.49
2	3	10	13.70
3	1	1	32.32
3	2	1	30.57
3	3	1	31.95
3	1	5	29.87
3	2	5	29.55
3	3	5	29.60
3	1	10	29.57
3	2	10	29.93
3	3	10	29.57

Source: The Authors

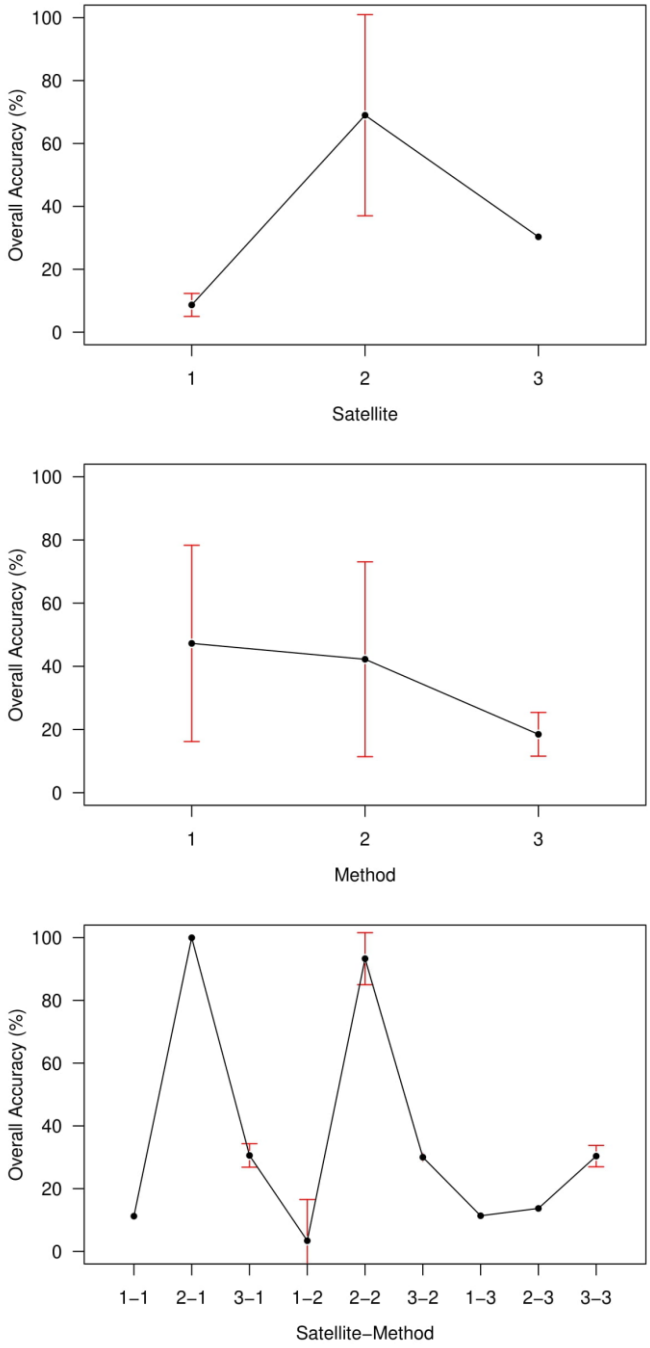
Statistical Methodology

We characterize the experimental outcome of interest (OCA) as a continuous random variable measured under conditions defined by factors (categorical variables with nominal levels). In this context, our primary hypothesis is that OCA values behave differently across Satellite/Method factors' levels and interactions between them. For instance, from Table 1, one may suggest that OCA varies substantially across satellites and that the SAM method provides low average accuracy.

Figure 2 depicts the OCA average values within Satellite/Method factors levels. The red lines correspond to the 95% confidence intervals for the grouped mean estimators. Noticeably, Satellite Cbers-4 provides higher classification accuracy, mainly when MD and ML methods are used. Besides, the rightest panel

also indicates that the OCA varies significantly on the interactions between Satellite and Method levels. These sample results suggest that further investigation should be conducted to evaluate whether the factors' effect is statistically significant for explaining the variability of the response variable.

Figure 2 - Grouped means of the overall classification accuracy.



Source: The Autors

The standard statistical approach to analyzing data from experimental designs is the parametric analysis of variance (ANOVA), from which one can investigate the

influence of factors on the average values of the response variable by comparing means across different groups (Montgomery, 2017). In our case, the developed experiment consists of a factorial design whose data can be described by the linear regression model

$$y_{ijk} = \mu_{ij} + \epsilon_{ijk},$$

with

$$\mu_{ij} = \mu + \beta_i + \gamma_j + (\beta\gamma)_{ij},$$

where μ is the overall mean effect, β_i (i = 1, 2, 3) is the effect of the *i*-th Satellite, γ_j (j = 1, 2, 3) is the effect of the j-th Method, $(\beta\gamma)_{ij}$ is the effect of the interaction between the i-th Satellite and the *j*-th Method, and ϵ_{ijk} is the random error. In this formulation, the observed responses are taken at each level of factors Satellite and Method in each one of the *k* replicates (*k* = 1, 2, 3) from samples per class.

Both factors (Satellite and Method) are of equal interest in our two-factor experiment. Specifically, using the ANOVA framework through the means model (2.1), we are interested in testing hypotheses about the effect of Satellites and Methods levels and the effect of the interactions between them. Formally, we want to test

$$\begin{cases} H_0\text{: } \beta_i = 0 \text{ for all } i \\ H_1\text{: At least one } \beta_i \neq 0, \end{cases} \quad \text{and} \quad \begin{cases} H_0\text{: } \gamma_j = 0 \text{ for all } j \\ H_1\text{: At least one } \gamma_j \neq 0, \end{cases} \quad \text{and} \quad \begin{cases} H_0\text{: } (\beta\gamma)_{ij} = 0 \text{ for all } i, j \\ H_1\text{: At least one } (\beta\gamma)_{ij} \neq 0. \end{cases}$$

Applying the parametric ANOVA with fixed-effects requires careful checking for some assumptions, which rely on the distributional behavior of the errors: they should be independent and normally distributed with zero mean and constant variance (homoscedastic) among the groups. In this sense, researchers must adopt standard proce- dures for model-fit evaluation, which are typically based on finding specific patterns using graphical tools (scatter plots, histograms, Normal probability plots), beyond performing formal hypothesis tests for normality (Shapiro-Wilk, Kolmogorov-Smirnov) and heteroscedasticity (Bartlett, Levene) of the estimated residuals.

One should be aware that violating the ANOVA assumptions could lead to misleading model-based inferences. In this case, researchers often resort to the Kruskal-Wallis (KW) Rank Sum test, widely known as the nonparametric one-way ANOVA. The KW test extends the Wilcoxon Rank Sum test to three or more independent samples, and it has more statistical power than the parametric one-way ANOVA in case of nonnormality of the residuals.

The KW test can be introduced in the context where independent random samples are taken from *k* populations, and the interest is to test whether the populations' medians are statistically different. Let δ_i be the median of the i-th population (*i* = 1, . . . , *k*). Thus, one can test the equality of the medians by testing the hypothesis

$$\begin{cases} H_0\text{: } \delta_1 = \delta_2 = \cdots = \delta_k = 0 \\ H_1\text{: At least one } \delta_i \neq 0, \end{cases}$$

which is equivalent to test $\delta_1 = \delta_2 = \cdots = \delta_k = a$, with $a \neq 0$. In our experiment, we

have to individually compare $k = 3$ medians of every single factor and $k = 9$ medians from interactions between factors' levels. The formal procedure to perform a KW test is well-described in Ramachandran and Tsokos (2018).

RESULTS AND DISCUSSION

This subsection is dedicated to presenting and discussing the main results obtained after analyzing the data described in Subsection 2.2. All computations were performed using the R environment (R Development Core Team, 2020), and we have adopted a significance level of 5% to draw conclusions. Our first attempt was based on fitting the linear regression model (2.1) and then decomposing the variance of the response variable using the parametric ANOVA methodology. The obtained results are presented in Table 2.

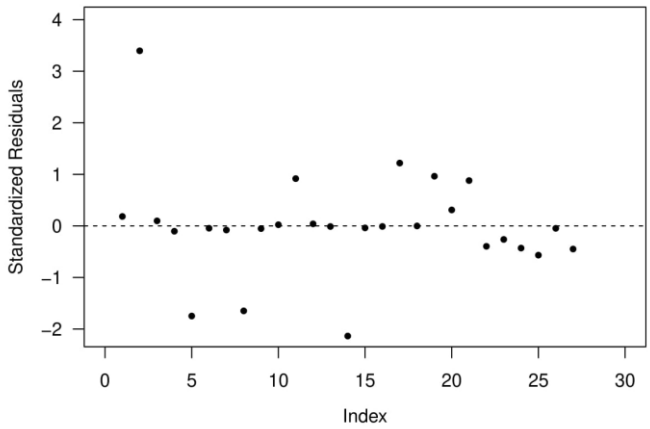
Table 2 - Parametric analysis of variance for the overall classification accuracy.

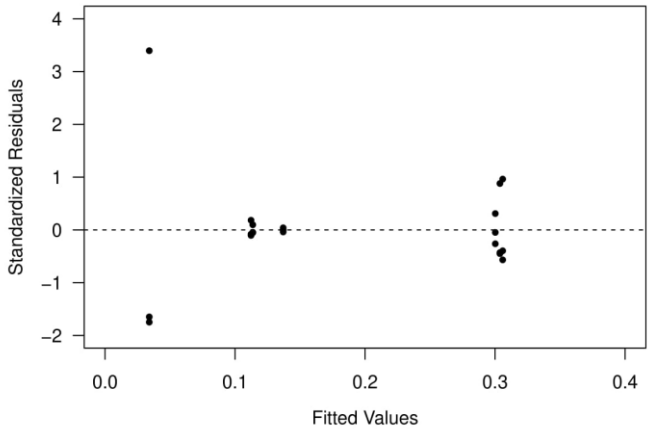
Source of Variation	Degree of Freedom	Sum of Squares	Mean Square	F value	p-value
Satellite	2	1.6807	0.8403	1733.11	< 0.001
Method	2	0.4254	0.2127	438.64	< 0.001
Satellite × Method	4	0.9691	0.2423	499.69	< 0.001
Residuals	18	0.0087	0.0005	-	-

Source: The Autors

These results allow us to conclude that the main effects of Satellite and Method are highly significant (p-values < 0.001). Notably, the same conclusion holds for the interaction between Satellites and Methods. In the following, one should check whether the fitted model met the assumptions of the parametric ANOVA, so the obtained results can be considered statistically valid. Figure 3 illustrates the behavior of the estimated standardized residuals across observations (left panel) and fitted values (right panel). These scatter plots suggest that the errors may not be homoscedastic on the levels of at least one of the factors, which was confirmed by Levene's test of equality of residual variances across Methods levels (p-value ≈ 0.0033).

Figure 3 - Scatter plots of the standardized residuals.

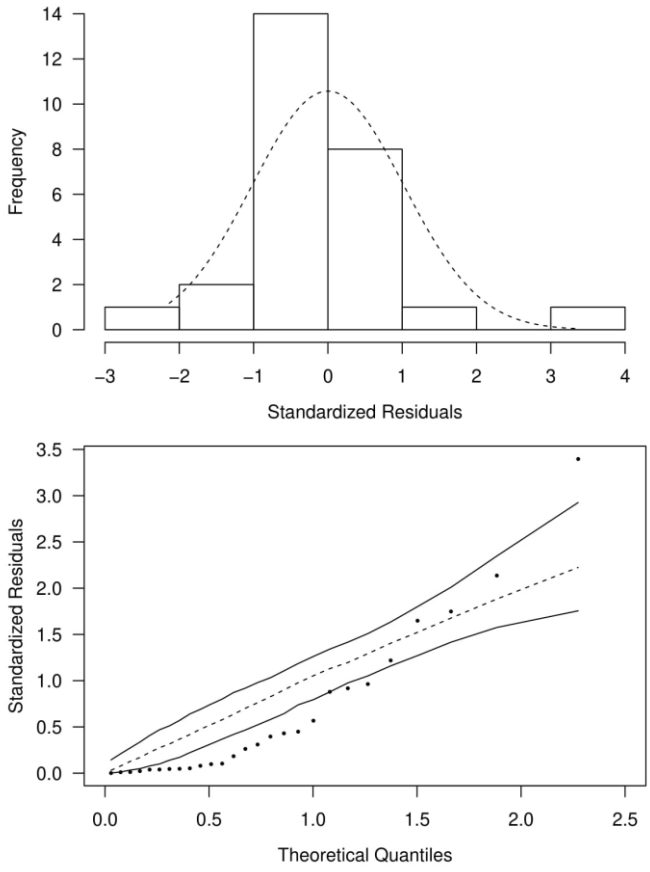




Souce: The Authors

Figure 4 depicts additional evidence that the obtained fit is not appropriate. The normality assumption for the residuals can be easily refuted by the behavior of its frequency distribution (left panel) and by the Shapiro-Wilk normality test (*p*-value ≈ 0.0011). Besides, the Half-Normal probability plot indicates a poor fit since most of the estimated standardized residuals are lying outside the simulated envelope (right panel).

Figure 4 - Frequency distribution and Half-Normal plot with simulated envelope for the standardized residuals



Source: The Authors

After model-fit checking, we conclude that our inferences cannot be derived from the parametric ANOVA. In this sense, we flexible the distributional assumptions to use a more appropriate statistical methodology: the KW test. As described in Subsection 2.3, this test is based on evaluating whether group medians (of a single factor) are significantly different. Therefore, we present in Table 3 the KW Statistic values and the respective *p*-values, assuming that, under H_0, these values are drawn from a Chi-squared distribution with $k - 1$ degrees of freedom.

Table 3 - Kruskal-Wallis tests for the overall classification accuracy

Source of Variation	Degree of Freedom	KW Statistic	*p*-value
Satellite	2	18.0000	0.0001
Method	2	1.6332	0.4419
Satellite × Method	8	24.6770	0.0018

Source: The Authors

The KW test results illustrate the importance of checking the parametric ANOVA assumptions. The factor Method was first assumed to have a significant effect on the OCA but not significant when a more suitable methodology was applied. However, the effect of Satellites remains highly significant, which reassures the same conclusion for the interaction between the main factors considered in our experiment.

Table 4 - Exact *p*-values of post-hoc tests for pairwise (Satellite-Method) comparisons.

Satellite	Method	Satellite							
		1	2	3	1	2	3	1	2
		Method							
		1	1	1	2	2	2	3	3
2	1	0.0524	-	-	-	-	-	-	-
3	1	0.7160	0.9202	-	-	-	-	-	-
1	2	0.9995	**0.0066**	0.3028	-	-	-	-	-
2	2	0.1768	0.9999	0.9943	**0.0327**	-	-	-	-
3	2	0.7793	0.8823	1.0000	0.3648	0.9879	-	-	-
1	3	1.0000	0.0813	0.8082	0.9976	0.2472	0.8599	-	-
2	3	0.9976	0.3331	0.9879	0.9025	0.6473	0.9943	0.9995	-
3	3	0.7485	0.9025	1.0000	0.3331	0.9916	1.0000	0.8352	0.9916

Source: The Authors

After identifying significant factors (and interactions) using a KW test procedure, one may be interested in performing *post-hoc* tests for pairwise comparisons to evaluate which factors levels combinations are statistically different from each other. For this purpose, we have adopted the distribution-free Nemenyi's rank comparison test (Nemenyi, 1963; Sachs, 1997) using the **kwAllPairsNemenyiTest** function from the **PMCMRplus** package. We have found Satel- lite Landsat 8 statistically different from the others (*p*-values < 0.005), and no differences were identified between Satellites Cbers-4 and Sentinel-2. The results regarding comparisons across interactions between Satellites and Meth-

ods are presented in Table 4. One can notice that classifications using the spatial resolutions provided by Landsat 8 and Cbers-4 are substantially different among MD and ML methods.

CONCLUDING REMARKS

Beyond using an extensive database on the availability of accurate orbital images and applying traditional or more complex techniques, geomatics researchers typically resort to image classification tools for monitoring land use and occupation. An accurate classification result strongly depends on the initial configuration of the adopted platform. In this sense, this work proposed to evaluate the interaction of two qualitative variables (spatial resolution of the image and classification method) in the image classification process, aiming to understand the influence of those variables (and their interactions). For that, we have performed a factorial experiment combining images of different spatial resolutions (Landsat 8: 30m, Cbers-4: 20m, and Sentinel-2: 10m) with different classification techniques (Minimum Distance, Maximum Likelihood, and Spectral Angle Mapping). After data collection, we assessed that the classical parametric ANOVA assumptions were not met, which led us to apply the Kruskal-Wallis test to derive our inferences and draw appropriate conclusions. The obtained results highlighted the importance of using a statistical method that flexible general parametric assumptions on our data as we may overestimate the effect of a specific factor on the overall classification accuracy. Finally, the main conclusions of the present study are that although the spatial resolution has slightly higher importance, both investigated variables are equally crucial to obtaining a reliable and accurate result in the image classification.

Identificando fatores que afetam a precisão geral em problemas de classificação de imagens: uma abordagem estatística

RESUMO

A classificação de imagens é um assunto de reconhecimento de padrões que pode ser aplicado em diversas áreas. A obtenção de uma classificação altamente precisa envolve a escolha de configurações ideais a partir das quais as imagens serão classificadas. Nesse processo, variáveis controláveis podem afetar a precisão geral da classificação, como a resolução espacial da imagem e o método de classificação. Nesse sentido, delineamos um experimento fatorial onde a precisão da classificação de uma imagem (de Curitiba, Paraná, Brasil) foi obtida a partir de três satélites e três métodos de classificação. O teste de Kruskal-Wallis foi aplicado para avaliar se a variabilidade entre os níveis dos fatores sustenta a hipótese de que os efeitos dos fatores experimentais são estatisticamente significativos. Em seguida, avaliamos quais níveis dos fatores diferiam entre si, usando testes post-hoc. Nossos resultados sugerem que a resolução espacial da imagem, que varia entre os satélites escolhidos para o estudo, e o método de classificação são determinantes na obtenção de classificações precisas de imagens em um contexto geográfico.

PALAVRAS-CHAVE: Delineamento fatorial, classificação de imagens, teste de Kruskal-Wallis, precisão geral da classificação, resolução espacial.

ACKNOWLEDGEMENTS

This work is partially supported by CAPES (Finance Code 001).

REFERENCES

Abdelmalik, K. W. **Role of statistical remote sensing for Inland water quality parameters prediction.** The Egyptian Journal of Remote Sensing and Space Science, v. 21, n. 2, p. 193–200, 2018.

Araújo, E. A.; Lani, J. L.; Amaral, E. F.; Guerra, A. **Uso da terra e propriedades físicas e químicas do argissolo amarelo distrófico na Amazônia Ocidental.** Revista Brasileira de Ciência do Solo, v. 28, p. 307–315, 2004.

Congedo, L. **Semi-automatic classification plugin documentation**. R**elease 6.0.1.1.** Technical Report, 2016.

Kruse, F. A.; Lefkoff, A. B.; Boardman, J. W.; Heidebrecht, K. B.; Shapiro, A. T.; Barloon, P. .J.; Goetz, A. F.H. T**he spectral image processing system (SIPS) – interactive visualization and analysis of imaging spectrometer data.** Remote Sensing of Environment, v. 44, n. 2–3, p. 145–163, 1993.

Montgomery, D. C. **Design and analysis of experiments.** John Wiley & Sons, 2017.

Nemenyi, P. **Distribution-free multiple comparisons.** Ph.D. thesis, Princeton University, 1963.

Noi, P. T.; Kappas, M. **Comparison of random forest, k-nearest neighbor, and support vector machine classifiers for land cover classification using Sentinel-2 imagery.** Sensors, v. 18, n. 1, p. 18, 2018.

Pereira, L. F.; Guimarães, R. M. F. **Mapeamento multicategórico do uso/cobertura da terra em escalas detalhadas usando Semi-automatic Classification Plugin.** Journal of Environmental Analysis and Progress, v. 3, n. 4, p. 379–385, 2018.

QGIS Development Team, 2021. **QGIS Geographic Information System.** Open Source Geospatial Foundation Project.

R Development Core Team. **R: A language and environment for statistical computing.** Vienna, Austria: R Foundation for Statistical Computing, 2020.

Ramachandran, K. M.; Tsokos, C. P. **Mathematical statistics with applications.** Elsevier Academic Press, 2009.

Richards, J. A. **Remote sensing digital image analysis: An introduction**. Berlin, Germany: Springer, 2013.

Sachs, L. **Angewandte statistik**. Berlin, Germany: Springer, 1997.

Tian, S.; Zhang, X.; Tian J.; Sun, Q. **Random forest classification of wetland landcovers from multi-sensor data in the arid region of Xinjiang, China.** Remote Sensing, v. 8, n. 11, p. 954, 2016.

Vasconcelos, C. H.; Novo, E. M. L. M. **Mapeamento do uso e cobertura da terra a partir da segmentação e classificação de imagens-fração solo, sombra e vegetação derivadas do modelo linear de mistura aplicado a dados do sensor TM/Landsat5, na região do reservatório de Tucuruí-PA.** Acta Amazônica, v. 34, p. 487–493, 2004.

Recebido: 11/05/2022
Aprovado: 18/08/2022
DOI: 10.3895/rts.v18n54.15480

Como citar: BERTOLI, W. et al. Identifying factors impacting the overall accuracy in image classification problems: a statistical approach. **Rev. Tecnol. Soc**., Curitiba, v. 18, n. 54, p. 261-274, out./dez., 2022. Disponível em: https://periodicos.utfpr.edu.br/rts/article/view/15480. Acesso em: XXX.

Correspondência: