Mercator, Fortaleza, v.21, e21034, 2022. ISSN:1984-2201

# ENVIRONMENTAL FRAGILITY BY MACHINE LEARNING ALGORITHMS

https://doi.org/10.4215/rm2022.e21034

Cristiano Marcelo Pereira de Souza [a]* - Lucas Augusto Pereira Silva [b] - Gustavo Vieira Veloso [c]
Marcos Esdras Leite [d] - Elpídio Inácio Fernandes Filho [e]

(a) PhD in Soils and Plant Nutrition. Professor at State University of Montes Claros, Montes Claros (MG), Brazil.
**ORCID:** http://orcid.org/0000-0001-7692-1613. **LATTES:** http://lattes.cnpq.br/3795848668647630.
(b) PhD student in Geography. Federal University of Uberlândia, Uberlândia (MG), Brazil.
**ORCID:** http://orcid.org/0000-0001-5504-9029. **LATTES:** http://lattes.cnpq.br/4284728074543770.
(c) PhD in Soil Science and Plant Nutrition. Federal University of Viçosa, Viçosa (MG), Brazil.
**ORCID:** http://orcid.org/0000-0002-9451-2714. **LATTES:** https://lattes.cnpq.br/5446388671333942.
(d) PhD in Geography. Professor at State University of Montes Claros, Montes Claros (MG), Brazil.
**ORCID:** http://orcid.org/0000-0002-9020-6445. **LATTES:** http://lattes.cnpq.br/0392398629237265.
(e) PhD in Agronomy. Professor at the Federal University of Viçosa, Viçosa (MG), Brazil.
**ORCID:** http://orcid.org/0000-0002-9484-1411. **LATTES:** http://lattes.cnpq.br/9848935150180973.

**Article history:**
Received 25 August, 2022
Accepted 10 December, 2022
Published 15 December, 2022

**(*) CORRESPONDING AUTHOR**
**Address:** Unimontes. Avenida Dr. Rubens Braga, S/N, CEP: 39401-089, Vila Mauriceia, Montes Claros (MG), Brazil. Phone: (+55 38) 3229-8000.
**E-mail:** cmpsgeografia@gmail.com

## Abstract

The advancement of predictive models by Machine Learning Algorithms (ML) associated with environmental data enables the improvement of models of environmental fragility, which are essential tools for decision-making. This study aimed to derive a prediction of environmental fragility by testing ML associated with environmental covariates in the state of Minas Gerais. We use physical-environmental variables (soil, geology, climate, relief) with a weight of fragility for the attributes and calculation of the average to obtain a model of Potential Environmental Fragility (PEF). Subsequently, we extracted the PEF values to a 4,800-point grid, which was used to generate a new prediction by ML called PEFML. This prediction was based on testing five algorithms and a set of 105 environmental covariates. The results indicated that the best-performing PEFML prediction was the Random Forest model (R2 0.59 and RMSE 0.47), indicating a predominance of the low environmental fragility level. The PEF and PEFML models have strong correlations (0.7 Pearson); however, PEFML has stronger correlations with other environmental data. Therefore, the PEFML prediction is a robust model that captures information from covariates and has coherent spatial patterns.

**Keywords:** Environmental Fragility Model; Spatial Prediction; Random Forest; Environmental Planning.

## Resumo / Resumen

### FRAGILIDADE AMBIENTAL USANDO ALGORITMOS DE APRENDIZADO DE MÁQUINA

O avanço de modelos preditivos por Algoritmos de Aprendizado de Máquina (ML) associados à dados ambientais possibilita aprimoramento de modelos de fragilidade ambiental, os quais são importantes ferramentas para tomada de decisão. O objetivo desse estudo foi derivar uma predição de fragilidade ambiental, testando ML associados a covariáveis ambientais no estado de Minas Gerais. Utilizamos variáveis físico-ambientais (solo, geologia, clima, relevo) com peso de fragilidade para os atributos e cálculo da média obtendo o modelo de Fragilidade Ambiental Potencial (PEF). Posteriormente, extraímos os valores de PEF para uma grade de 4.800 pontos e usadas para gerar uma nova predição por ML, denominada PEFML. A predição foi com teste de cinco algoritmos e conjunto de 105 covariáveis ambientais. Comparamos os dois modelos de fragilidade ambiental (PEF e PEFML), inclusive com outros dados de riscos/vulnerabilidade/fragilidade. Os resultados indicaram que a predição de PEFML de melhor desempenho foi o modelo Random Forest (R2 0.59 e RMSE 0.47), indicando predomínio do nível fragilidade baixa. Os modelos de fragilidade PEF e PEFML têm forte correlação (0.7 Pearson), porém, PEFML possui correlações mais fortes com outros dados ambientais. Portanto, a predição de PEFML é um modelo robusto que capta informações de covariáveis e possui padrões espaciais coerentes

**Palavras-chave:** Modelos de fragilidade ambiental; Predição espacial; Random Forest; Planejamento Ambiental.

### FRAGILIDAD AMBIENTAL MEDIANTE ALGORITMOS DE APRENDIZAJE AUTOMÁTICO

El avance de los modelos predictivos mediante Machine Learning Algorithms (ML) asociados a datos ambientales permite mejorar los modelos de fragilidad ambiental, que son herramientas fundamentales para la toma de decisiones. Este estudio tuvo como objetivo derivar una predicción de la fragilidad ambiental mediante la prueba de ML asociado con covariables ambientales en el estado de Minas Gerais. Se utilizaron variables físico-ambientales (suelo, geología, clima, relieve) con peso de fragilidad para los atributos y cálculo de la media para obtener un modelo de Fragilidad Ambiental Potencial (PEF). Posteriormente, extrajimos los valores de PEF a una cuadrícula de 4800 puntos, que se utilizó para generar una nueva predicción de ML, llamada PEFML. Esta predicción se basó en la prueba de cinco algoritmos y un conjunto de 105 covariables ambientales. Los resultados indicaron que la predicción PEFML con mejor desempeño fue el modelo Random Forest (R2 0.59 y RMSE 0.47), indicando un predominio del bajo nivel de fragilidad ambiental. Los modelos PEF y PEFML muestran fuertes correlaciones (0,7 Pearson); sin embargo, PEFML tiene correlaciones más fuertes con otros datos ambientales. Por lo tanto, la predicción PEFML es un modelo robusto que captura información de covariables y tiene patrones espaciales coherentes.

**Palabras-clave:** Modelos de fragilidad ambiental; predicción espacial; Random Forest; Planificación ambiental.

# INTRODUCTION

Spatial modeling of environmental fragility is a crucial tool for territorial management and conservation of natural resources. In Brazil, a model of environmental fragility widely used is the proposal by Ross (1994) which is based on the theory of landscape ecodynamics (Tricart, 1977). The studies apply this model in two categories: (i) potential environmental fragility (PEF), based on slope, climate, and soil data; (ii) emergent environmental fragility, which adds land use data and demonstrates vulnerable areas associated with anthropic action. Studies have constantly improved the models of environmental fragility and methodological adaptations are proposed; an example is the incorporation of a geological component (Crepani et al., 2001; Spröl e Ross, 2006; Franco et al., 2011; Cruz et al., 2017; Campos et al., 2019; Costa et al., 2020). Furthermore, the new methodological improvements meet the future trend of increasingly using quantitative methods and robust analyzes in geoscience studies (Murray et al., 2009; Padarian et al., 2020). In this scenario, studies have applied new modifications in the modeling of environmental fragility, for example, fuzzy logic, neural networks, multicriteria analysis, and Bayesian networks (Spörl et al., 2011; Campos et al., 2019; Costa et al., 2020; Amorim et al., 2021).

In the scope of robust methods of analysis, there has been a rapid growth in the use of artificial intelligence, especially with Machine Learning algorithms (ML) in modeling studies in various fields of geosciences (Bergen et al., 2019; Gomes et al., 2019; Souza et al., 2020; Silva et al., 2023), but still little applied to studies of environmental fragility. The advantage of ML is to learn complex patterns for predicting spatiotemporal data using multiple data sources, e.g., environmental covariates (Kuhn e Johnson, 2013; Bergen et al., 2019; Padarian et al., 2020). Specifically on this aspect, some ML algorithms may have advantage over ot her predictive models, such as the kriging interpolation, which has geostatistics in its formulation and requires that the entered data have spatial dependence, and sometimes this spatial distribution pattern is non-existent in environmental data (Wang et al., 2020; Souza et al., 2022). Furthermore, regarding the covariates, studies have already suggested that the insertion of new data (covariates) to assist in modeling environmental fragility is potentially promising, since environmental vulnerability can be related to several physical-environmental factors (Cruz et al., 2017; Amorim et al., 2021); nevertheless, the approach with ML has not yet been adequately tested to predict models of environmental fragility.

Currently, with advances in geoinformation, there is a vast supply of information that works as covariates in modeling using ML, which can work to improve models of environmental fragility. For example, digital elevation model (DEM) generates several covariates linked to geomorphology (Sena et al., 2020); manipulating spectral bands from satellite images provides vegetation indices (Dias et al., 2021); climate models designed for the globe are updated with some frequency (Hijmans et al., 2005); and categorical data on environmental information are often made available in the form of a geographic database (Heineck et al., 2003; UFV et al., 2010). Studies in several areas show clear benefits of accuracy in modeling with increments of covariates (Hijmans et al., 2005; Gomes et al., 2019; Souza et al., 2022). Spatial prediction by ML is especially necessary when a given study area presents a large amount of spatial data with diverse physical environmental aspects (geodiversity), where studies based on qualitative analysis are not sufficient or time-consuming, or simpler prediction methods do not handle the data well (Bergen et al., 2019; Padarian et al., 2020; Souza et al., 2022). In Brazil, a area with geodiversity is Minas Gerais state, with varied geotectonic contexts (Machado e Silva, 2010; Costa, 2021), conditioning several geomorphological aspects, such as surfaces flattened by dissection, mountain ranges belonging to orogenic contexts, and pedodiversity with the most weathered soils in the world (Ker, 1997; Silva et al., 2018). In addition, it has a complex geoecological framework (that is, the presence of Cerrado, Atlantic Forest, and Caatinga) resulting from paleoclimatic events (Ab'sáber, 1970). Therefore, the progress of environmental fragility studies is to connect various environmental aspects, considering environmental heterogeneity, and ML has a fundamental role in this effort. This study aims to test machine learning algorithms to predict a new Potential Environmental Fragility (PEF) model, demonstrating which covariates are potentially explanatory for the levels of fragility in the state of Minas Gerais.

# MATERIALS AND METHODS

## STUDY AREA

The state of Minas Gerais is located in Southeastern Brazil between -23°0’ to -14°0’ S and -51°0’ to -40°0’ W (Figure 1). Based on the Köppen climate type, there types of climate in the region: Cwb (humid temperate climate with dry winter and moderately hot summer), Cwa (humid temperate climate with dry winter and hot summer), Aw (Savannah tropical climate with dry winter season), and As (Semi Tropical Climate Wet). The geological framework is marked by four major provinces: (i) the São Francisco Province, with crystalline rocks, often covered by metasedimentary sediments (Neoproterozoic); (ii) the Mantiqueira Province has massifs and hills developed in granitic/granitoid and metamorphic rocks (Proterozoic); (iii) the Tocantins Province comprises the granitic/granitoid and schist (Proterozoic) folded bands at the edge of the São Francisco Craton; and (iv) the Paraná Sedimentary Basin with mafic rocks covered by Cretaceous sandstones, forming extensive plateaus (Ab'sáber, 1970; Heineck et al., 2003). The state is marked by the presence of three important biomes – Atlantic Forest, Cerrado and Caatinga, in addition to transition zones called ecotones (Ab'sáber, 1970).

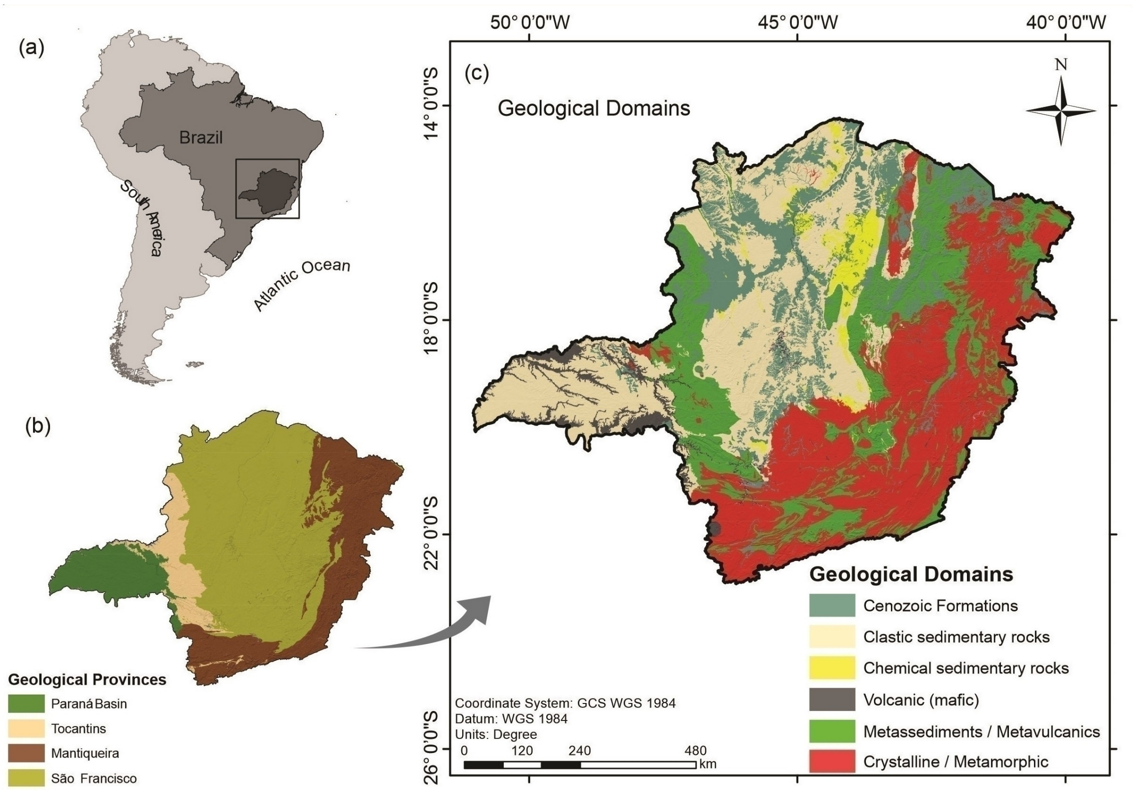


Figure 1 - (a) Location of Minas Gerais state, (b) geological provinces of Minas Gerais, (c) Geological domains of the state of Minas Gerais

# METHODOLOGICAL PROCEDURES

For the analysis of environmental fragility, we set up a methodological framework summarized in Figure 2, with procedures executed in R software (Rcore, 2023). Previously, we applied a model of Potential Environmental Fragility (PEF), which constitutes the insertion of physical-environmental variables (Ross, 1994; Crepani et al., 2001; Spörl et al., 2011). The variables inserted were climate, geology, relief, and soil, with weights assigned to the class of each variable. In this weighting step, we selected environmental fragility values (weights) available in previous studies and assigned the classes

(Ross, 1994; Crepani et al., 2001; Franco et al., 2011; Spörl et al., 2011; Cruz et al., 2017; Campos et al., 2019; Amorim et al., 2021). We applied a multi-criteria analysis (Analytical Hierarchy Process - AHP) to define the relative importance of each variable in relation to environmental fragility, as proposed by Amorim et al. (2021). Finally, we apply an overlap of the variables using the equation PEF=C+G +R+S/4; where PEF: Potential Environmental Fragility, C: climate, G: Geology, R: Relief (slope), S: Soil. At the end, a result is obtained resulting from the arithmetic mean of the fragility values recorded in the classes. The environmental fragility orders obtained are classified as 1: very low, 2: low, 3: medium, 4: strong, 5: very strong.

From the PEF map (previous step), we extracted the values of environmental fragility from this map to a 4,800-point grid randomly distributed with a minimum distance of 3 km. These points constitute the main variable to determine levels of environmental fragility using the Machine Learning algorithm technique (ML); and the result of this procedure is to generate a new map, here called Potential Environmental Fragility by Machine Learning (PEFML). However, to aid prediction by ML, we set up a database of covariates in raster format, structured in a resolution of 1x1 km: fifty-five spatial data from the WorldClim (Hijmans et al., 2005); 42 geomorphometric covariates from Shuttle Radar Topographic Mission (SRTM) (USGS, 2023), extracted using the SAGA software (Olaya e Conrad, 2009; Sena et al., 2020); seven data of gamma-spectrometry and one of gravimetry (Heineck et al., 2003), one NDVI calculated from satellite images of the MODIS (Moderate-Resolution Imaging Spectroradiometer) sensor of August of the year 2019 (USGS, 2023), which is the driest month and shows the most significant differentiation between phytophysinomies. Variables soil, geology, climate, were not included to avoid biased predictions, as these are part of generating PEF.

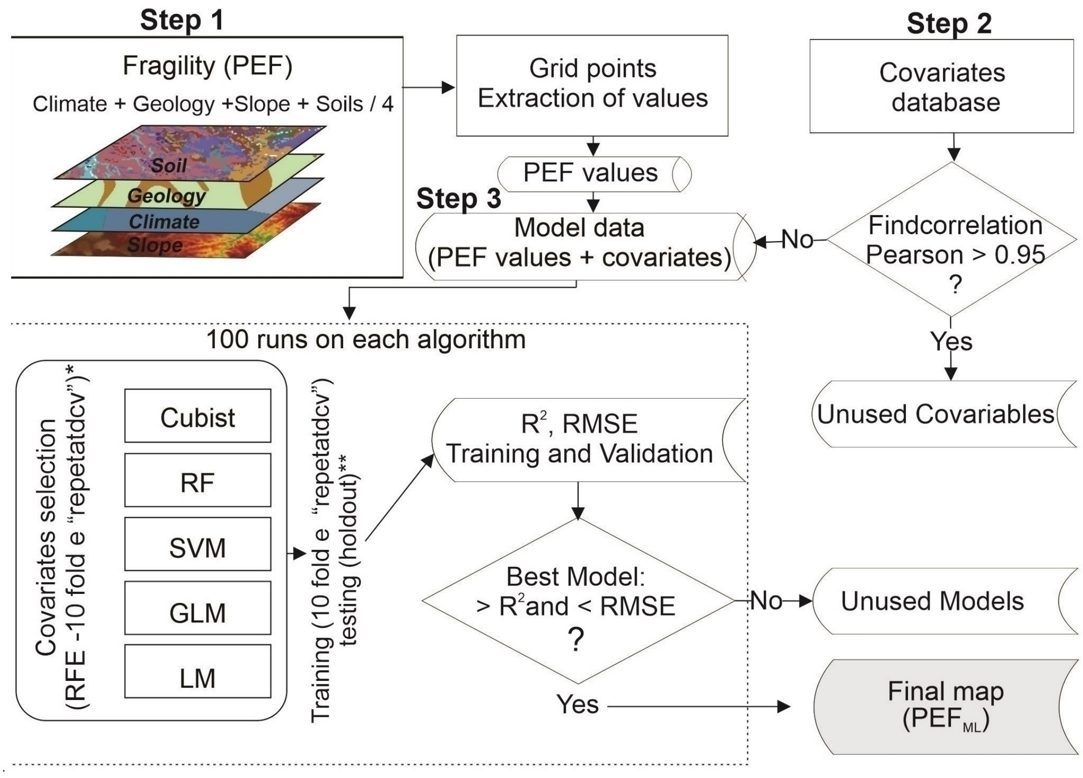


Figure 2 - Flowchart with the sequence: input variable determination (Potential Environmental Fragility - PEF); creation of covariates database, selection of important covariates with RFE function; test of algorithms for PEF prediction with 100 runs. In the end, the most accurate algorithm was selected to model and map PEFML (Potential Environmental Fragility by Machine Learning PEFML). RF: Random Forest; SVM: Support Vector Machine; GLM: Generalized linear models; LM: Linear model. R2: R- Squared; RMSE: Root Mean Square Error

The previous step generates a data structure with variable values (PEF) and covariate values, but the excessive number of covariates also favors creating overestimated models (Gomes et al., 2019; Padarian et al., 2020). Therefore, we apply a cut-of-correlation to eliminate highly correlated covariates, as they have a similar contribution to explaining the distribution of an analyzed variable. The function used was find correlation, parameterized by the correlation cutoff criterion >0.95 Pearson (Kuhn e Johnson, 2013). Moreover, we applied an important covariate selection tool using a tool widely used in prediction models, called Recursive Feature Elimination (RFE), which also avoids overestimated predictions. RFE is a backwards feature selection method whose central concept is based on eliminating unimportant covariates (Kuhn e Johnson, 2013). After removing the least important covariate, the RFE readjusts the model with a smaller set of covariates, restarting the process to eliminate the least important covariate; and this process is repeated several times until stabilization in which the accuracy level determined by R-squared ($R^2$) does not decrease. This process is performed based only on the training dataset that was 75% of the samples, while the other 25% are used in the testing process (holdout-test). In machine learning, the test phase is a process to evaluate the performance of a model trained with a test set (i.e., 25% of the samples), which was not seen by the model in the training phase.

From the data adjusted to the covariates selected by the RFE, these were used to predict the environmental fragility for the study area by testing different algorithms. Moreover, for each algorithm the process was repeated in 100 runs using the proper subset of covariates indicated by RFE. In the prediction step, five machine learning algorithms were used: Cubist (Quinlan, 1992), Generalized linear models-GLM (Hastie e Tibshirani, 1987), Linear Model Regression – LM (Faraway, 2016), Random Forest – RF (Breiman, 2001), and Support Vector Machine – SVM (Cortes e Vapnik, 1995). These models have already been tested in several studies (Brungard et al., 2015; Faraway, 2016; Morellos et al., 2016; Gomes et al., 2019; Souza et al., 2022; Silva et al., 2023). Furthermore, using different algorithms is essential to evaluate the limitations related to the prediction of the target variable by algorithms that have different statistical routines (Kuhn e Johnson, 2013; Padarian et al., 2020). We considered, in the evaluation of the most accurate algorithm, the metrics of the testing phase (25% of the samples) and evaluating the overfitting effect comparing with training data (75% of the samples), using as metrics: R-squared - $R^2$ and root mean squared error – RMSE.

To observe the variation between the PEFML and the original method (PEF), a new points grid was created with 4,800 points at random, with subsequent extraction of the PEFML and PEF values. Also, in this statistical analysis, spatial data of risks and vulnerability available for Minas Gerais were inserted (Maps: Degree conservation, Erosion risk, Natural vulnerability, Conservation priority, Soil vulnerability, Erodibility). These maps are part of the compiled work of the agroecological zoning of Minas Gerais made available in matrix data (raster), with intervals from 1 to 5, which is a range compatible with the PEFML map (Scolforo et al., 2008). On these data, we applied Pearson correlation analysis and K-means cluster analysis to observe the relationships between these maps and similarity clusters (Sena et al., 2020). Furthermore, to have a spatial overview of the correlations, we selected the maps of environmental risk and vulnerability from the base of Scolforo et al. (2008) and applied a subtraction calculation using PEFML as a reference, according to the equation: Relation maps=PEFML-EZMi. Where PEFML: Potential Environmental Fragility Map by Machine Learning, EZM: Environmental zoning map, i: specific map of the EZM.

# RESULTS

## ENVIRONMENTAL COVARIATES AND MODEL SELECTION

The PEFML predictive process involved five algorithms and covariates database with climatic, topographic, geochemistry and vegetation data, with correlation levels below 0.95 in these covariates. In applying the RFE function to select important covariates, we observed that approximately twelve covariates is the maximum number to generate higher $R^2$ and lower RMSE in the algorithms used (Figure 3 a,b). The Random Forests algorithm presented the best performance in the selection of the RFE, selecting only 10 covariates ranked by level of importance (Figure 4). According to the ranking (% overall), the bioclimatic data from WorldClim were predominant and more significant. Among the topographic covariates, only MRRTF (MRRTF: multiresolution ridge-top flatness index), which indicates flat positions in high altitude areas, and MRVBF (MRVBF: multiresolution valley bottom flatness index), which shows flat surfaces at the bottom of the valley, were included in the modeling. In addition to these, a covariate of element K gamma-spectrometry occupied the third level of importance. This covariate, in addition to indicating the gamma ray spectrometry for potassium (40K), can also be correlated to a certain volume of the element on the surface. Therefore, Random Forest was selected for prediction because from 100 runs it generated >$R_2$ (training 0.61, test 0.59)

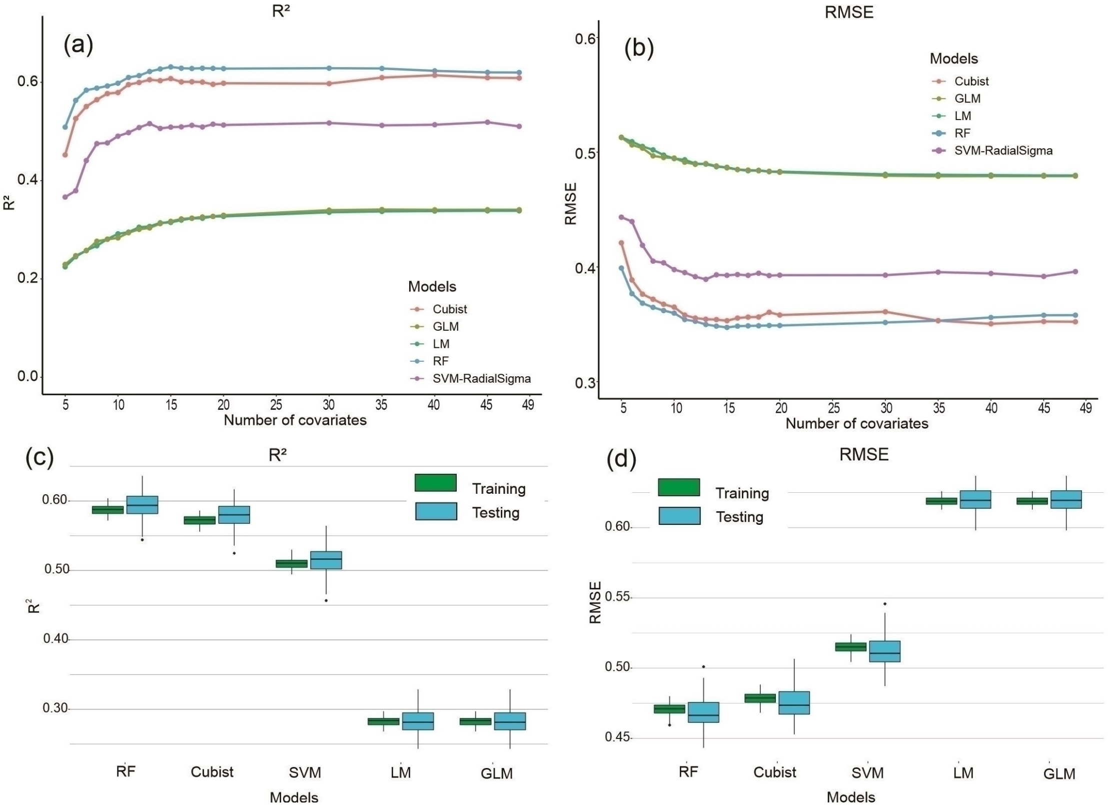


Figure 3 - Performance of the models: (a, b) evolution of the metrics R-squared (R2), root mean squared error (RMSE) in the selection of covariate using Recursive Feature Elimination. (c, d) Metrics of accuracy R2 and RMSE error, respectively, in 100 runs using covariates selected by RFE. RF: Random Forest; SVM: Support Vector Machine; GLM: Generalized linear models; LM: Linear model. R2: R-Squared; RMSE: Root Mean Square Error

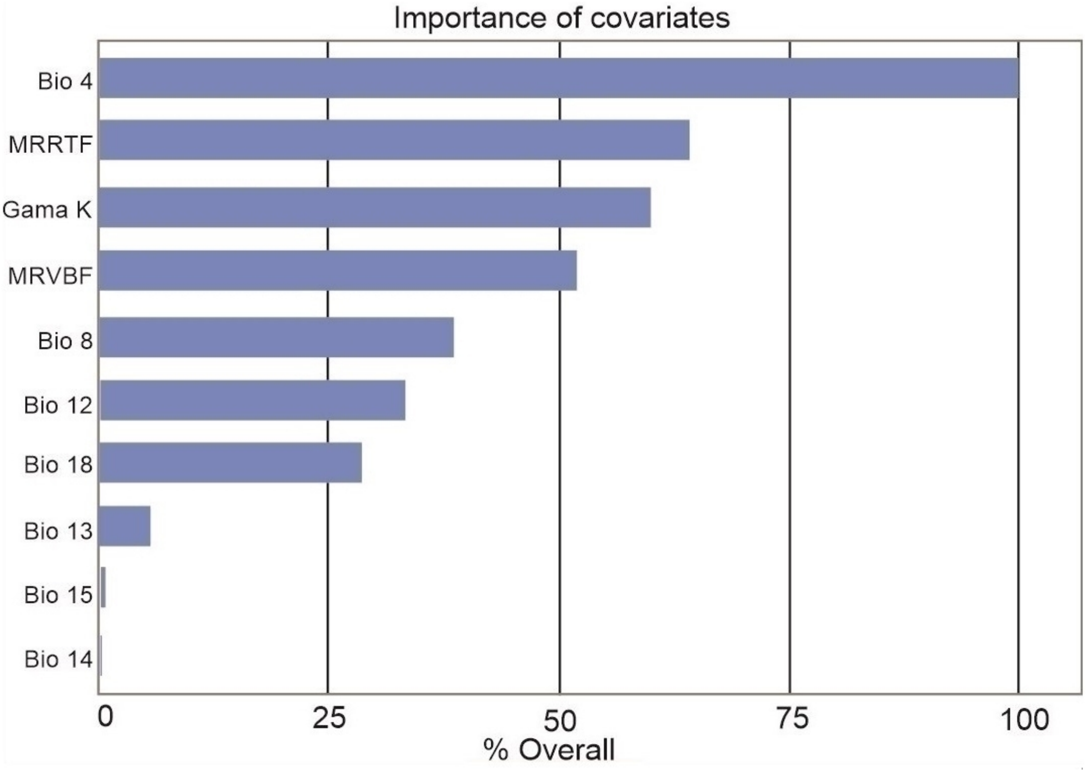


Figure 4 - The relative importance of covariates, given by the overall percent. Bio 4: temperature seasonality (standard deviation  100), MRRTF: multiresolution ridge-top flatness index, Gama K: gamma spectrometry K-element, MRVBF: multiresolution valley bottom flatness index, Bio 4: temperature seasonality, Bio 8: mean temperature of wettest quarter, Bio 12: annual precipitation, BIO18: precipitation of warmest quarter, Bio 13: precipitation of wettest month, BIO15: precipitation seasonality (coefficient of variation), BIO14: precipitation of driest month

## PREDICTION OF ENVIRONMENTAL FRAGILITY

The Potential Environmental Fragility map using ML (PEFML) shows similar spatial patterns with the environmental fragility map derived from overlapping variables (PEF) - (Figure 5). This similarity attests to the robustness of the ML predictions in learning how a variable's distribution occurs, and this learning was done using new covariates (Figure 4), therefore, without adding the variables that gave rise to the map of environmental fragility by the Ross method (i.e., variables shown in Figure 5a).

However, the prediction using the RF algorithm showed the disadvantage by normalizing values that have low territorial expressiveness, thus fragility classes below 1.7 and above 4.3 were eliminated, thus including class 1 (very low) and 5 (very strong). Therefore, the alternative was to apply a reclassification of levels, according to (1.7 to 2.7: low, 2.8 to 3.4: average, 3.5 to 4: strong). This separation criterion prioritizes the class of average fragility, which is predominant in the PEF map (Figure 5 b).

Despite normalizing values, the ML prediction generated a map with more detailed features, as the prediction captures the features of the inserted covariates and did not create large homogeneous zones, coherently indicating areas of environmental fragility (Figure 5 b,c). Therefore, the spatial distribution of PEFML showed spatial patterns that denote the influence of input variables recommended in studies of environmental fragility (Ross, 1994; Spröl e Ross, 2006; Cruz et al., 2017).

For example, the prediction captured the influence of geomorphology in increasing the level of environmental fragility in areas of greater slope in the context of the central region, where steep slope landforms occur (e.g., Espinhaço Mountain Range), and in escarpment areas, such as the plateaus of the northwest region. The geology and soil factors also contribute to areas of medium and strong fragility, especially in areas with greater pedological and geological variation, such as the central and northern regions of the state (Iglesias e Uhlein, 2009; Silva et al., 2018). Furthermore, the northern region has the added factor of low precipitation weather conditions to contribute to higher levels of fragility; being this part of the state included in the limit of the Brazilian semiarid.

Low environmental fragility predominates in most of the state, mainly in areas located to the east. This configuration also attests that the PEFML map is spatially consistent, since, in areas located to the east, most of the input variables in the PEF model also have low levels of environmental fragility, except for relief, because where there is relief of the Mares of Morros domain, higher slopes prevail (Figure 5a). The areas of low environmental fragility also have an extension in the northern region, where only the geology variable has higher levels of fragility, therefore, other variables are responsible for this attenuation. This attenuation effect by a set of variables is repeated in the western region (Triângulo Mineiro region), because in this region the geology expresses high environmental fragility; however, the low-slope relief and the predominance of deep soils influence the reduction of values.

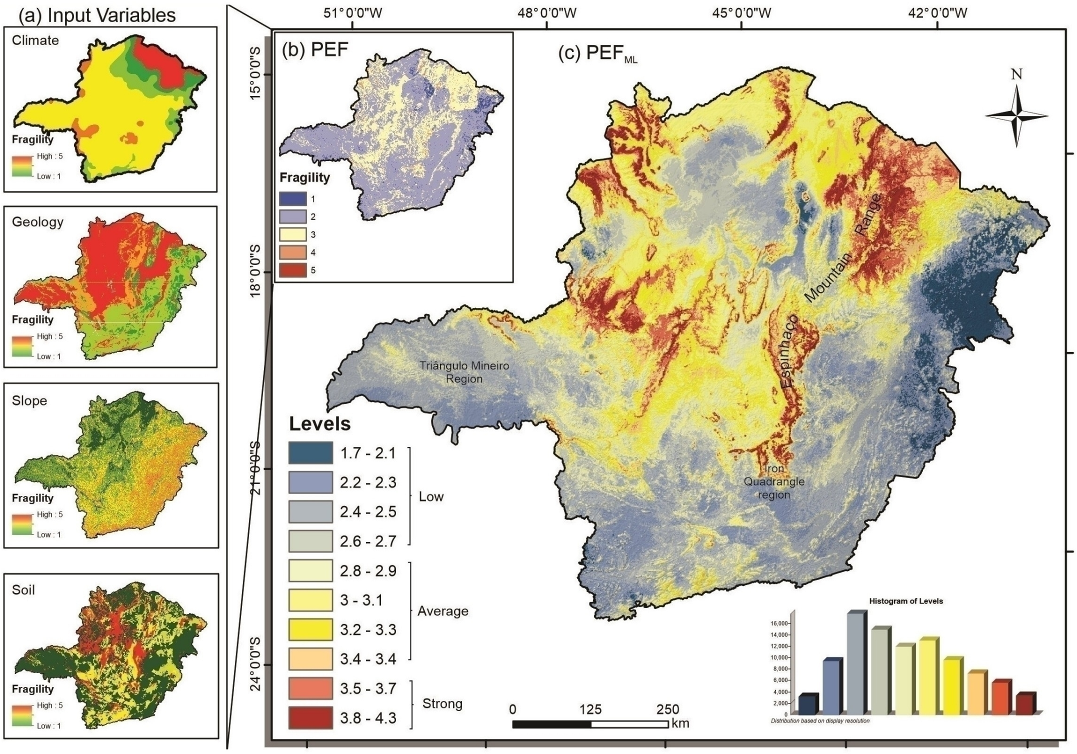
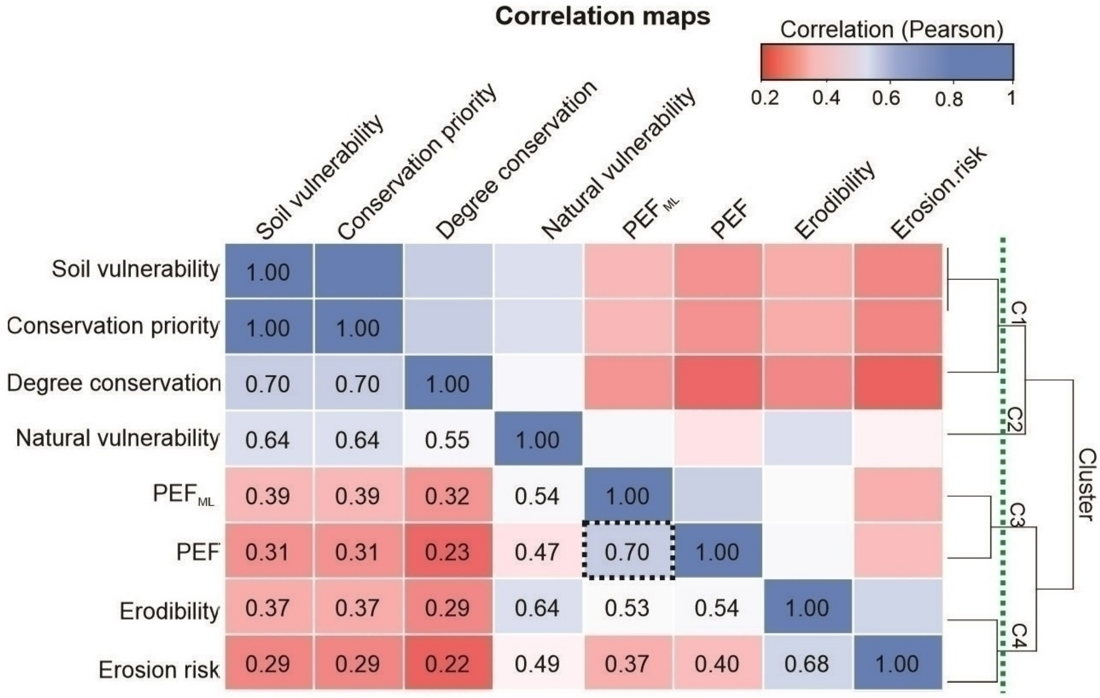


Figure 5 - (a) Input variables with fragility values to generate Potential Environmental Fragility map, (b) Potential Environmental Fragility map by overlapping environmental variables, (c) Potential Environmental Fragility map generated by Machine Learning using the Random Forest model (100 runs) (PEFML), and histogram with distribution of classes

## STATISTICAL COMPARISON OF MAPS

In addition to the similarity of the spatial pattern between the PEF map by the conventional method and the PEFML map derived from machine learning, there was also a strong statistical correlation between these maps (Pearson 0.70); and these two maps are in the same cluster (C3) (Figure 6), confirming the similarity level. Furthermore, we compared the environmental fragility maps of the two methodologies (PEF and PEFML) with other environmental information on environmental risk and vulnerability available for the state of Minas Gerais (Scolforo et al., 2008). The PEFML map always showed a superior correlation with risk and vulnerability maps, indicating that it is potentially more explanatory of other environmental factors: soil vulnerability, conservation priority, degree of conservation, natural vulnerability, erodibility, and erosion risk. On the other hand, the PEF map showed weak values in all correlations with these maps (Figure 6).

The spatial comparison between the PEFMLmap with the risk and vulnerability maps showed a correspondence betwween the maps in indicaring areas with a similar degree of fragility; even considering that the PEFML map has a smaller range of classes (1.7 to 4.3) .Therefore, in the largest area of the state, the results show that the variation of environmental fragility values in the PEFML with other maps is in a low range (1 and -1 for each fragility class). This low variation attests that the PEFML map can capture information on environmental risk/vulnerability of isolated factors, suggesting being a more complete model.



| | Soil vulnerability | Conservation priority | Degree conservation | Natural vulnerability | $PEF_{ML}$ | PEF | Erodibility | Erosion.risk |
|---|---|---|---|---|---|---|---|---|
| Soil vulnerability | 1.00 | | | | | | | |
| Conservation priority | 1.00 | 1.00 | | | | | | |
| Degree conservation | 0.70 | 0.70 | 1.00 | | | | | |
| Natural vulnerability | 0.64 | 0.64 | 0.55 | 1.00 | | | | |
| $PEF_{ML}$ | 0.39 | 0.39 | 0.32 | 0.54 | 1.00 | | | |
| PEF | 0.31 | 0.31 | 0.23 | 0.47 | 0.70 | 1.00 | | |
| Erodibility | 0.37 | 0.37 | 0.29 | 0.64 | 0.53 | 0.54 | 1.00 | |
| Erosion risk | 0.29 | 0.29 | 0.22 | 0.49 | 0.37 | 0.40 | 0.68 | 1.00 |

Figure 6 - Correlation between fragility values using the map algebra method (PEF) and machine learning algorithm (PEFML). Also, correlation with risk and vulnerability maps using data from the digital zoning base of Minas Gerais (Scolforo et al., 2008). (All values with a significance level of alpha = 0.05; Green line for separation of clusters (C1 to C4); black square correlation between PFE and PEFML

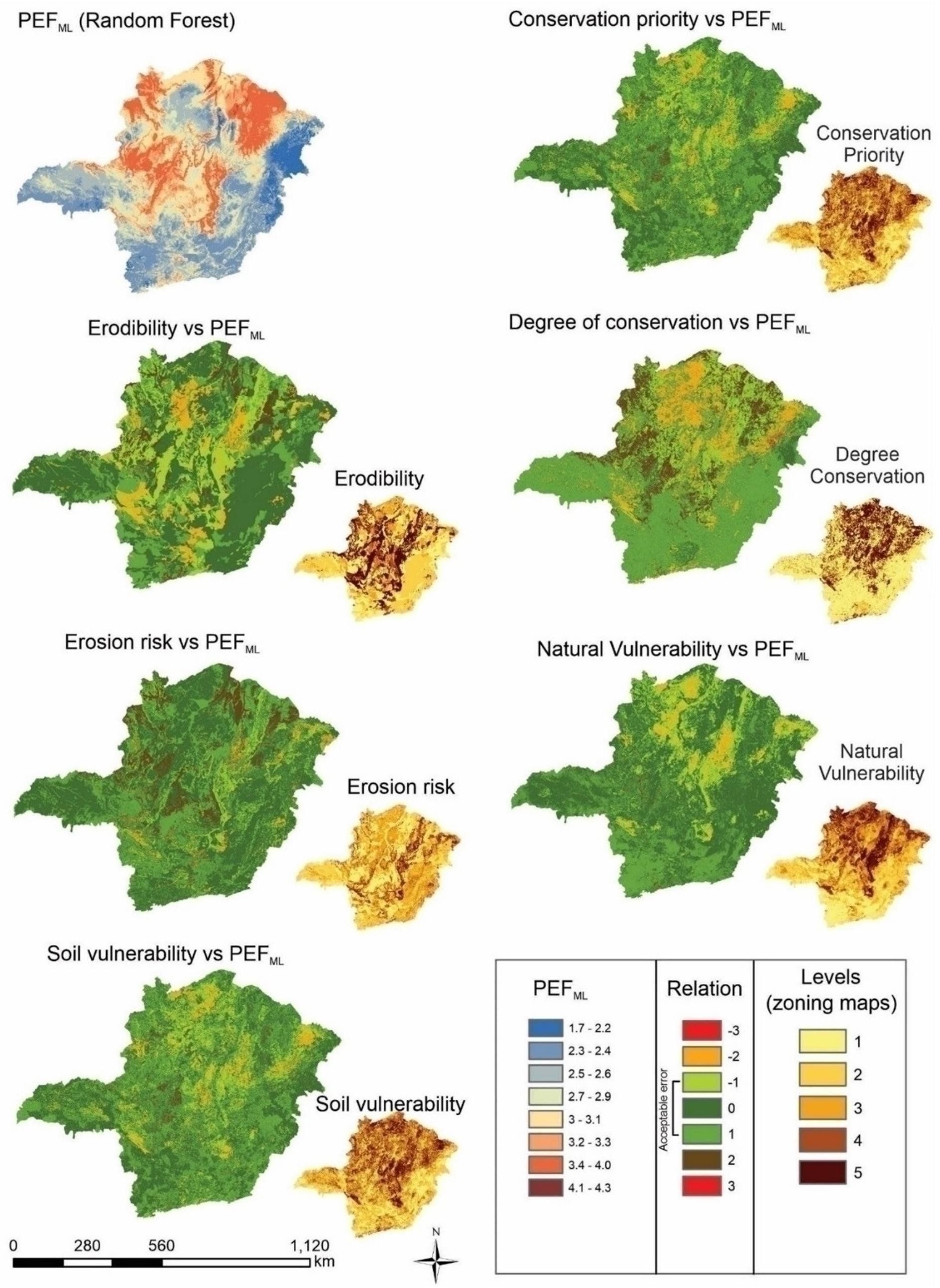


Figure 7 - Maps of the relationship between the environmental fragility map (PEFML) (upper left corner), with each vulnerability and environmental risk map produced from the economic-ecological zoning of the state of Minas Gerais. Small map in each bottom corner is the risk and vulnerability map.

# DISCUSSION

## *MODEL PERFORMANCE*

The spatial prediction of fragility using several ML associated with covariates dataset is a robust spatial analysis procedure and shows different statistical performances according to selected algorithm. The most efficient algorithm for choosing covariates for PEFML prediction was RF, with the RFE tool adjust to this algorithm, and his algorithm handles high-dimensional data well and allows for non-linear relationships between predictors (Breiman, 2001; Gomes et al., 2019). We emphasize the advantage of using the RFE function in the training phase, eliminating covariates that did not improve the prediction performance, and this removal creates a simpler and non-overestimated prediction, obeying the principle of parsimony in modeling (Brungard et al., 2015). The $R^2$ and RMSE metrics were satisfactory (Figure 3 c,d), especially when considering the values of the training and testing phases, which were similar, indicating that in the algorithm execution phase, the model trained and tested satisfactorily (i.e., without overfitting effect), and RF has an advantage in mitigating this adverse effect (Breiman, 2001; Were et al., 2015). Environmental prediction data, especially for large areas, rarely generates $R^2$ greater than 0.70. The limitation of high metrics stems from factors of low resolution of covariates, lack of covariates strictly linked to the analyzed variable, and data that do not have simple linear relationships (Malone et al., 2009; Gomes et al., 2019; Padarian et al., 2020; Souza et al., 2022).

Considering that the RF carried out the training using covariates indicated by the RFE, some of these covariates are potentially explanatory for levels of environmental fragility in Minas Gerais (Figure 4). In general, the selected covariates have some relationship with the input variables to determine PEF (i.e., climate, geology, relief, soil). Therefore, WorldClim data was predominant and higher in importance, especially bioclimatic, which are more significant in explaining climate trends (Hijmans et al., 2005; Gomes et al., 2019). Topographic information from the SRTM also composes the covariates selected for prediction; MRRTF and MRVBF that emphasize flat areas in altitude or valleys are part of this selection. These topographical covariates correlate with hydrological processes of erosion and deposition (Gallant e Dowling, 2003), factors associated with areas of environmental fragility. The third most significant covariate was the potassium element gamma-spectrometry data (K gamma-spectrometry). Potassium is a component of minerals (feldspar, biotite, and muscovite), present at prominent levels in some rocks and soils with low pedogenetic development and/or eutrophic soil (Guevara et al., 2018). Therefore, the selection of K gamma-spectrometry was efficient because it combines two pieces of information (soil and geology), which are variables for PEF modeling.

## ENVIRONMENTAL FRAGILITY OF THE STATE OF MINAS GERAIS

Regarding the spatial distribution of PEFML values, there is a conjunction of factors that act to attenuate or increase levels of environmental fragility. Several regions with higher environmental fragility are associated with areas of greater slope, where the morphogenesis process predominates, and this aspect is well-marked in the context of Serra do Espinhaço or plateau escarpment zones, a typical geomorphological feature in Minas Gerais (Callisto et al., 2016; Costa, 2021). In addition, other factors compete to amplify the level of fragility, for example, in Serra do Espinhaço, there are environments sustained by the interdependence of dynamic processes between types of vegetation, climate, roughness, slope, and hydrological flow. An example is the ferruginous rocky "campos rupestres" grasslands “canga ecosystems”, with acidic soils of low fertility and high levels of metallic cations, sustaining flora with a high degree of endemism (Urriago-Ospina et al., 2021).

Geological factors also contribute to increased environmental fragility, especially in areas with great lithological variation, where landscapes are sculpted by processes of differential denudation. For example, the Iron Quadrangle area predominates resistant rocks to the vertical lowering process (downwearing). Still, they are fragile areas concerning lateral retraction processes of escarpments (backwearing), by erosion of more fragile lithotypes, which form the base of the escarpments (Salgado et al., 2006). Similar evolutionary processes, by differential erosion involving other lithological types, are recorded in other areas of the state, creating sloping reliefs (Simões et al., 2020; Souza et al., 2020; Costa, 2021). Another geological context that often has higher levels of fragility are areas of carbonate rocks, particularly in a Karst geomorphology zone, which tend to present several vulnerability problems due to the presence of fractures and cavities produced by karst processes (Pessoa et al., 2020). The attenuation of environmental fragility in areas of carbonate rocks occurs when there is an association of low-slope relief and the presence of deep soils (Oxisols) and/or relatively eutrophic soils (Oxisols - Kandic) – (Ker, 1997), creating more stable environments to morphogenetic processes and predominating pedogenesis. Moreover, the geological and topographic conditions that indicate a high degree of environmental fragility may have an additional contribution from the climatic factor, which seems to be the case for the state's northern region, especially as it is in areas subject to desertification (Barros et al., 2018). Therefore, in this region, the condition of drier climate implants moderate levels of fragility in an area of flat relief and higher levels when there is association with sloping relief.

The lowest levels of environmental fragility occur in the eastern portion, a fact resulting from the geological influence with the presence of granitic and metamorphic rocks (gneisses) from the Mantiqueira geological province (Figure 1), rock systems with low vulnerability (Cruz et al., 2017). Deep soils located even in sloping relief areas also contribute to attenuating environmental fragility, as they are more stable soils in the landscape and the pedogenesis process is preponderant (Ker, 1997; Nunes et al., 2001). However, the levels of environmental fragility can increase in the condition of the high slope, associated with the presence of shallow soils, such as Inceptisols (Nunes et al., 2001). Low and average fragility values also extend to the western part of the state, involving a large part of the Paraná Basin. In this region, flat morphology predominates, with deep soils of low fertility (Ker, 1997). Although soils are dystrophic, they have good physical attributes, and pedogenesis processes supplant morphogenesis (Motta et al., 2002). Therefore, the fragility values obey the configuration of the terrain; only in higher slope areas do fragility values increase (Martins e Rodrigues, 2012).

## CORRELATION MAPS

The fragility model by overlapping maps (variables) (Ross, 1994; Spröl e Ross, 2006), widely applied in environmental studies, remains an efficient method to identify areas of environmental fragility (Spröl e Ross, 2006; Franco et al., 2011; Spörl et al., 2011; Martins e Rodrigues, 2012; Campos et al., 2019; Anjinho et al., 2021). However, the environmental fragility modeling by machine learning (PEFML), applied in this study, also proved to be a reliable method, including results highly correlated with the original method (Figure 6). In addition, the PEFML map was more correlated with other data on fragility, vulnerability, and environmental risk for the state of Minas Gerais (Scolforo et al., 2008). This higher correlation suggests that the ML prediction is as a more explanatory model of other environmental factors not present in the PEF modeling, which only uses overlapping variables with assigned weights. Presumably, this higher correlation is a contribution of the dozens of covariates selected in the RFE function step in the training of the model, because in ML predictive methods, the map result is influenced by the information of the various covariates that help in the predictions (Brungard et al., 2015; Gomes et al., 2019; Souza et al., 2022).

The results presented indicate a methodological gain in modeling potential environmental fragility using ML. Basically, ML is a robust analysis method, and complex models tend to produce more accurate predictions than simpler models, because in addition to testing different well-developed algorithms in the field of statistical science, the ML prediction has in its structure the incorporation of covariates that help the prediction (Kuhn e Johnson, 2013; Brungard et al., 2015; Morellos et al., 2016; Souza et al., 2022). Therefore, the insertion of new covariates that could explain the spatial distribution of potentially fragile areas is also an essential factor, especially when considering that environmental vulnerability is associated with a multiplicity of physical and anthropogenic variables (Cruz et al., 2017), being a promising path for studies of environmental fragility.

# CONCLUSIONS

The potential Environmental Fragility Model obtained by the machine learning algorithm (PEFML) is statically like the Potential Environmental Fragility model (PEF) acquired by the overlap of physical environmental variables.

The Random Forest model was the most efficient in predicting PEFML, using a set of significant covariates, with satisfactory performance levels in the validation phase ($R^2$ 0.59 and RMSE 0.47 testing).The PEFML model proved more robust than PEF as it presented a higher level of correlation with other risk factors and environmental vulnerability, being a more explanatory map of other factors that influence environmental fragility.

PEFML indicated fragility levels derived from topographic, geological, climatic, and pedological effects. The areas of high fragility are associated with mountainous geomorphology and presence of escarpments in plateaus.

We highlight that the study area has a regional scale dimension, with covariates projected on a small cartographic scale. Therefore, replicating the model (PEFML) at the level of small watersheds with field validation is a way to ratify the modeling efficiency.

## REFERENCES

AB'SÁBER, A.N. Províncias Geológicas e Domínios Morfoclimáticos no Brasil. São Paulo: Universidade de São Paulo, Instituto de Geografia, 1970. 26 p. 1970.

AMORIM, A.T., LOPES, E.R.N., SOUSA, J.A.P., SILVA, R.C.F.D., SOUZA, J.C., LOURENÇO, R.W. Geomorphometric Environmental Fragility of a Watershed: A Multicriteria Spatial Approach. Environmental Monitoring and Assessment, v.193, p.850. 2021.

ANJINHO, P.D.S., BARBOSA, M.A.G.A., COSTA, C.W., MAUAD, F.F. Environmental Fragility Analysis in Reservoir Drainage Basin Land Use Planning: A Brazilian Basin Case Study. Land Use Policy, v.100, p.104946. 2021.

BARROS, K.O., RIBEIRO, C.A.A.S., MARCATTI, G.E., LORENZON, A.S., CASTRO, N.L.M., DOMINGUES, G.F., CARVALHO, J.R., SANTOS, A.R. Markov Chains and Cellular Automata to Predict Environments Subject to Desertification. Journal of Environmental Management, v.225, p.160-167. 2018.

BERGEN, K.J., JOHNSON, P.A., DE HOOP, M.V., BEROZA, G.C. Machine Learning for Data-Driven Discovery in Solid Earth Geoscience. Science, v.363, p.eaau0323. 2019.

BREIMAN, L. Random Forests. Machine Learning, v.45, p.5-32. 2001.

BRUNGARD, C.W., BOETTINGER, J.L., DUNIWAY, M.C., WILLS, S.A., EDWARDS, T.C. Machine Learning for Predicting Soil Classes in Three Semi-Arid Landscapes. Geoderma, v.239, p.68-83. 2015.

CALLISTO, M., GONÇALVES, J.F., LIGEIRO, R. Water Resources in the Rupestrian Grasslands of the Espinhaço Mountains. In: G.W. Fernandes (Eds.) Ecology and Conservation of Mountaintop Grasslands in Brazil. Cham: Springer International Publishing, 2016. p. 87-102.

CAMPOS, J.A., AIRES, U.R.V., SILVA, D.D.D., CALIJURI, M.L. Environmental Fragility and Vegetation Cover Dynamics in the Lapa Grande State Park, Mg, Brazil Anais da Academia Brasileira de Ciências, v.91, p.e20170940. 2019.

CORTES, C., VAPNIK, V. Support-Vector Networks. Machine Learning, v.20, p.273-297. 1995.

COSTA, E.M., DOS ANJOS, L.H.C., PINHEIRO, H.S.K., GELSLEICHTER, Y.A., MARCONDES, R.A.T. Spatial Bayesian Belief Networks: A Participatory Approach for Mapping Environmental Vulnerability at the Itatiaia National Park, Brazil. Environmental Earth Sciences, v.79, p.359. 2020.

COSTA, L.R.F. Considerações Sobre as Macrounidades Geomorfológicas do Estado de Minas Gerais–Brasil. William Morris Davis–Revista de Geomorfologia, v.2, p.11-18. 2021.

CREPANI, E., MEDEIROS, J.D., HERNANDEZ FILHO, P., FLORENZANO, T.G., DUARTE, V., BARBOSA, C.C.F. Sensoriamento Remoto e Geoprocessamento Aplicados ao Zoneamento Ecológico-Econômico e ao Ordenamento Territorial. São José dos Campos: Inpe 2001. 124. 2001.

CRUZ, B.B., MANFRÉ, L.A., RICCI, D.S., BRUNORO, D., APPOLINARIO JR., L., QUINTANILHA, J.A. Environmental Fragility Framework for Water Supply Systems: A Case Study in the Paulista Macro Metropolis Area (Se Brazil). Environmental Earth Sciences, v.76, p.441-453. 2017.

DIAS, S.H.B., FILGUEIRAS, R., FERNANDES FILHO, E.I., ARCANJO, G.S., SILVA, G.H.D., MANTOVANI, E.C., CUNHA, F.F.D. Reference Evapotranspiration of Brazil Modeled with Machine Learning Techniques and Remote Sensing. PLOS ONE, v.16, p.e0245834. 2021.

FARAWAY, J. Linear Models with R. Chapman Hall. New York: Chapman and Hall, 2016. 270. 2016.

FRANCO, G.B., MARQUES, E.A.G., GOMES, R.L., CHAGAS, C.S., SOUZA, C.M.P., BETIM, L.S. Fragilidade Ambiental da Bacia Hidrográfica do Rio Almada-Bahia. Revista Geografia, v.28, p.187 - 205. 2011.

GALLANT, J.C., DOWLING, T.I. A Multiresolution Index of Valley Bottom Flatness for Mapping Depositional Areas. Water Resources Research, v.39, 2003.

GOMES, L.C., FARIA, R.M., SOUZA, E., VELOSO, G.V., SCHAEFER, C.E.G.R., FERNANDES FILHO, E.I. Modelling and Mapping Soil Organic Carbon Stocks in Brazil. Geoderma, v.340, p.337-350. 2019.

GUEVARA, Y.Z.C., SOUZA, J.J.L.L.D., VELOSO, G.V., VELOSO, R.W., ROCHA, P.A., ABRAHÃO, W.A.P., FERNANDES FILHO, E.I. Reference Values of Soil Quality for the Rio Doce Basin. Revista Brasileira de Ciência do Solo, v.42, p.1-16. 2018.

HASTIE, T., TIBSHIRANI, R. Generalized Additive Models: Some Applications. Journal of the American Statistical Association, v.82, p.371-386. 1987.

HEINECK, C.A., LEITE, C.A.S., SILVA, M., VIEIRA, V.S. Mapa Geológico do Estado de Minas Gerais, Escala 1: 1.000.000. Belo Horizonte: Convênio COMIG/CPRM, v.1, 2003.

HIJMANS, R.J., CAMERON, S.E., PARRA, J.L., JONES, P.G., JARVIS, A. Very High Resolution Interpolated Climate Surfaces for Global Land Areas. International Journal of Climatology, v.25, p.1965-1978. 2005.

IGLESIAS, M., UHLEIN, A. Estratigrafia do Grupo Bambuí e Coberturas Fanerozóicas no Vale do Rio São Francisco, Norte De Minas Gerais. Revista Brasileira de Geociências, v.39, p.256-266. 2009.

KER, J.C. Latossolos do Brasil: Uma Revisão. Revista Geonomos, v.5, p.17 - 40. 1997.

KUHN, M., JOHNSON, K. Applied Predictive Modeling. New York: Springer, 2013. 600. 2013.

MALONE, B.P., MCBRATNEY, A.B., MINASNY, B., LASLETT, G.M. Mapping Continuous Depth Functions of Soil Carbon Storage and Available Water Capacity. Geoderma, v.154, p.138-152. 2009.

MARTINS, T.I.S., RODRIGUES, S.C. Análise e Mapeamento dos Graus de Fragilidade Ambiental da Bacia do Médio-Baixo Curso do Rio Araguari, Minas Gerais. Caderno de Geografia, v.22, 2012.

MORELLOS, A., PANTAZI, X.-E., MOSHOU, D., ALEXANDRIDIS, T., WHETTON, R., TZIOTZIOS, G., WIEBENSOHN, J., BILL, R., MOUAZEN, A.M. Machine Learning Based Prediction of Soil Total Nitrogen, Organic Carbon and Moisture Content by Using Vis-Nir Spectroscopy. Biosystems Engineering, v.152, p.104-116. 2016.

MOTTA, P.E., CURI, N., FRANZMEIER, D.P. Relation of Soils and Geomorphic Surfaces in the Brazilian Cerrado. In: P. Oliveira, R.J. Marquis (Eds.) The Cerrados of Brazil: Ecology Natural History of a Neotropical Savanna. New York: Columbia University Press, 2002. p. 13-32.

MURRAY, A.B., LAZARUS, E., ASHTON, A., BAAS, A., COCO, G., COULTHARD, T., FONSTAD, M., HAFF, P., MCNAMARA, D., PAOLA, C., PELLETIER, J., REINHARDT, L. Geomorphology, Complexity, and the Emerging Science of the Earth's Surface. Geomorphology, v.103, p.496-505. 2009.

NUNES, W.A.G.A., KER, J.C., SCHAEFER, C.E.G.R., FERNANDES FILHO, E.I., GOMES, F.H. Relação Solo-Paisagem-Material de Origem e Gênese de alguns Solos no Domínio do "Mar De Morros", Minas Gerais Revista Brasileira de Ciência do Solo, v.25, p.341-354. 2001.

OLAYA, V., CONRAD, O. Chapter 12 Geomorphometry in Saga. In: T. Hengl, H.I. Reuter (Eds.) Developments in Soil Science. Elsevier, 2009. p. 293-308.

PADARIAN, J., MINASNY, B., MCBRATNEY, A.B. Machine Learning and Soil Sciences: A Review Aided by Machine Learning Tools. SOIL, v.6, p.35-52. 2020.

PESSOA, P., ATMAN, D., KIMURA, G. Environmental Problems in the Lagoa Santa Karst. In: A. S. Auler, P. Pessoa (Eds.) Lagoa Santa Karst: Brazil's Iconic Karst Region. Cham: Springer International Publishing, 2020. p. 283-303.

QUINLAN, J.R. Learning with Continuous Classes. (Eds.), Australian joint conference on artificial intelligence, pp. 343-348. 1992.

RCORE, T. R: A Language and Environment for Statistical Computing. Vienna, Austria (2016). http://www.R-project.org/ Access Date: 15 jan 2023. 2023.

ROSS, J.L.S. Análise Empírica da Fragilidade dos Ambientes Naturais Antropizados. Revista do departamento de geografia, v.8, p.63-74. 1994.

SALGADO, A.A.R., BRAUCHER, R., COLIN, F., NALINI, H.A., VARAJÃO, A.F.D.C., VARAJÃO, C.A.C. Denudation Rates of the Quadrilátero Ferrífero (Minas Gerais, Brazil): Preliminary Results from Measurements of Solute Fluxes in Rivers and in Situ-Produced Cosmogenic 10be. Journal of Geochemical Exploration, v.88, p.313-317. 2006.

SCOLFORO, J.R., OLIVEIRA, A., CARVALHO, L.D., MARQUES, J.J.G., LOUZADA, J.N., MELLO, C., PEREIRA, J.R., REZENDE, J.B., VALE, L.C.C. Zoneamento Ecológico-Econômico de Minas Gerais. Lavras: UFLA, 2008. 2008.

SENA, N.C., VELOSO, G.V., FILHO, E.I.F., FRANCELINO, M.R., SCHAEFER, C.E.G.R. Analysis of Terrain Attributes in Different Spatial Resolutions for Digital Soil Mapping Application in Southeastern Brazil. Geoderma Regional, p.e00268. 2020.

SILVA, L.A.P., SOUZA, C.M.P., SILVA, C.R., BOLFE, E.L., ROCHA, A.M. Projection of Climate Change Impacts on Net Primary Productivity of the Legal Amazon – Brazil Caderno de Geografia, v.33, p.1-22. 2023.

SILVA, L.C.L., OLIVEIRA, F.S., RAMOS, V.D.V., SCHAEFER, C.E.G.R. Pedodiversidade No Estado De Minas Gerais-Brasil. Caderno de Geografia, v.28, p.18-38. 2018.

SIMÕES, P.L., VALADÃO, R.C., OLIVEIRA, C.V., OLIVEIRA, F.S. Uso de Atributos Pedológicos na Compreensão da Gênese de Superfícies Geomorfológicas Escalonadas da Borda Oeste do Planalto do Espinhaço Meridional / Minas Gerais – Brasil. Revista Brasileira de Geomorfologia, v.21, 2020.

SOUZA, C.M.P., FIGUEIREDO, N.A., COSTA, L.M., VELOSO, G.V., ALMEIDA, M.I.S., FERREIRA, E.J. Machine Learning Algorithm in the Prediction of Geomorphic Indices for Appraisal the Influence of Landscape Structure on Fluvial Systems, Southeastern - Brazil. Revista Brasileira de Geomorfologia, v.21, p.363-378. 2020.

SOUZA, C.M.P., VELOSO, G.V., MELLO, C.R., RIBEIRO, R.P., SILVA, L.A.P., LEITE, M.E., FERNANDES FILHO, E.I. Spatiotemporal Prediction of Rainfall Erosivity by Machine Learning, Minas Gerais State, Brazil. Geocarto International, v.37, p.1-19. 2022.

SPÖRL, C., CASTRO, E., LUCHIARI, A. Aplicação de Redes Neurais Artificiais na Construção de Modelos de Fragilidade Ambiental. Revista do Departamento de Geografia, v.21, p.113-135. 2011.

SPRÖL, C., ROSS, J.L.S. Análise Comparativa da Fragilidade Ambiental com Aplicação de Três Modelos. GEOUSP: Espaço e Tempo (Online), p.39-49. 2006.

TRICART, J. Ecodynamic. Rio de Janeiro: IBGE, 1977. 91. 1977.

UFV, CETEC, UFLA, FEAM. Mapa de Solos do Estado de Minas Gerais - Escala 1: 500.000. p.49. 2010.

URRIAGO-OSPINA, L.M., JARDIM, C.M., RIVERA-FERNÁNDEZ, G., KOZOVITS, A.R., LEITE, M.G.P., MESSIAS, M.C.T.B. Traditional Ecological Knowledge in a Ferruginous Ecosystem Management: Lessons for Diversifying Land Use. Environment, Development and Sustainability, v.23, p.2092-2121. 2021.

USGS. Earthexplorer Disponível em:< http://earthexplorer.usgs.gov> Access Date: 27 jan 2023. 2023.

WANG, Z., SHI, W., ZHOU, W., LI, X., YUE, T. Comparison of Additive and Isometric Log-Ratio Transformations Combined with Machine Learning and Regression Kriging Models for Mapping Soil Particle Size Fractions. Geoderma, v.365, p.114214. 2020.

WERE, K., BUI, D.T., DICK, Ø.B., SINGH, B.R. A Comparative Assessment of Support Vector Regression, Artificial Neural Networks, and Random Forests for Predicting and Mapping Soil Organic Carbon Stocks across an Afromontane Landscape. Ecological Indicators, v.52, p.394-403. 2015.