Case Studies in Chemical and Environmental Engineering 7 (2023) 100353

Contents lists available at ScienceDirect

Case Studies in Chemical and Environmental Engineering

journal homepage: www.sciencedirect.com/journal/case-studies-in-chemical-and-environmental-engineering

Case Report

# A machine learning approach to understand how accessibility influences alluvial gold mining expansion in the Peruvian Amazon

Gustavo Larrea-Gallegos *, Ramzy Kahhat, Ian Vázquez-Rowe, Eduardo Parodi

*Peruvian Life Cycle Assessment & Industrial Ecology Network (PELCAN), Department of Engineering, Pontificia Universidad Católica del Perú, Avenida Universitaria 1801, San Miguel, 15088, Lima, Peru*

ARTICLE INFO

*Keywords:*
Climate change
Deforestation
Google earth engine
Madre de Dios
Peru
Random forest

ABSTRACT

Alluvial small-scale gold mining (ASGM) mining in the Amazon is expanding fiercely, generating severe environmental degradation, which includes the fast disappearance of primary forests in a highly biodiverse area of the world. Different factors motivate the growth of mining in the areas and understanding this expansion is important to safeguard protected areas or implement strategies to mitigate the related social and environmental impacts. Thus, the goal of this study is to apply machine learning techniques to explore gold mining expansion in Madre de Dios, in the Peruvian Amazon, and to identify possible future hotspots of these activities. Using an unsupervised learning algorithm and a random forest classification model, past expansion trends were analyzed and an explicit geo-spatial model was built. Results demonstrate that proximity to infrastructure is not always indicative of high mining probability. In fact, when analyzing the spatial distribution of model accuracy, it is observed that model performance decreases in clusters where accessibility and mining activity showed opposite trends. In contrast, the models yield accuracies greater than 0.9 when accessibility-related variables stand out as the most important. The model, which is flexible and reproducible, demonstrates to be useful to enhance decision making when implementing geo-spatial policies to address the problem of ASGM expansion in the Amazon.

## 1. Introduction

The thirst for gold has triggered through history multiple extraction systems with heterogeneous outcomes when observing their environmental, social and economic impacts [1,2]. Currently these activities are still common practice in different parts of South America, such as Colombia and Brazil [3]. In the particular case of the Peruvian Amazon, gold mining, especially in the region of Madre de Dios (MDD), is responsible for the use and release to the environment of large amounts of mercury [4] and the deforestation of large areas of primary rainforest [2,5]. Since 1984, more than 100,000 ha. have been deforested in the region, with subsequent impacts on climate change [6,7].

Once promoted by the government, alluvial, informal and illegal, mining activities in MDD have expanded considerably in the past decade, mainly due to the rapid increase of the price of gold [8]. Furthermore, the construction of the Interoceanic Highway (IH) that facilitates accessibility to extraction areas to miners, equipment and machinery, has consolidated mining operations in many zones [9]. Additional triggers of alluvial small-scale gold mining (ASGM) are the lack of labor opportunities in the Peruvian Amazon and the desire of local communities to attain economic prosperity.

However, there are some deterrents related to ASGM activities in MDD, such as interdictions performed by the government intended to eliminate or at least discourage mining operations in restricted areas (i.e., national reserves or parks). Probably the most successful interdiction was “Operación Mercurio”, an ongoing operation initiated in 2019 in La Pampa, a well-known illegal mining area in the buffer zone (BZ) of the Tambopata National Reserve, which managed to reduce deforestation by 90% in the area [10]. Nevertheless, the success of interdictions strategies is difficult, if not impossible, to achieve. Orographic barriers or the lack of rule of law in many areas, among other factors, are some issues that explain the relative failure of these strategies in recent years.

Solutions concerning potential future expansion patterns of mining activities would be helpful in terms of policy support in order to better organize expansion activities, avoid further expansion in protected areas (PAs), or mitigate related environmental impacts. The determination of the main drivers of informal and illegal mining expansion in Peru, particularly in MDD, is crucial to establish effective policies to control the expansion process itself, but also to determine damage mitigation actions in terms of human health [11,12], mainly referred to local

* Corresponding author.
*E-mail address:* glarrea@pucp.pe (G. Larrea-Gallegos).

https://doi.org/10.1016/j.cscee.2023.100353
Received 23 February 2023; Received in revised form 12 April 2023; Accepted 14 April 2023
Available online 15 April 2023

communities, and ecosystem quality [7]. However, understanding the triggers and expected behavior related to the expansion of this complex system is a difficult endeavor. Hence, we argue that exploring the potential expansion of this activity through sophisticated computational tools can allow a more efficient management of the problem.

### 1.1. Understanding ASGM

Some methodological approaches have been applied in the literature with the goal of understanding the patterns of ASGM expansion in the Amazon. For instance, Swenson and colleagues [13] identified deforested areas due to ASGM in MDD for the period 2003–2009 using remote sensing methods and statistical approaches to study the relationship between deforestation, import of mercury and the price of gold. Similarly, Asner and Tupayachi [14] and Asner et al. [15] have identified an increase in the loss of protected forest due to the acceleration of mining activity in the last decade, suggesting accessibility to the area as an important driver of this phenomenon. Another example focusing on the Brazilian Amazon can be found in Lobo and colleagues [16]. The authors identified deforestation due to alluvial gold mining in the lower section of the Tapajos River Basin, analyzing the relationship between the historical deforestation and variables such as proximity to roads and soil types. Finally, Caballero-Espejo et al. [6] used remote sensing methods to identify ASGM areas during the period 1985–2017. The authors analyzed the historic expansion of mining activities throughout MDD considering the evolution of the price of gold, as well as the type of machinery and the geographic location of deforestation. Moreover, in a more recent publication, Alvarez-Berrios et al. [17] studied the effect of formalization of ASGM by modelling changes in deforestation due to formalization between 2001 and 2014 using a parametric regression model.

All the studies mentioned above relied on retrospective approaches and statistical or parametric models to understand alluvial gold mining. However, most of these contributions lack a prospective orientation and lack a spatial explicit interpretation of the findings. Moreover, as far as we were able to ascertain, studies focusing on the expansion of mining activities using machine learning approaches in the Amazon are currently not available in scientific literature.

### 1.2. Machine learning methods

Artificial intelligence applications, such as machine learning (ML), have become useful tools to enhance analytical capabilities in society [18]. In fact, ML has steadily become an important method to be applied in industrial ecology, as it provides an efficient pathway towards data innovation, enhancing the capacity to map energy and material flows and stocks [19], while improving decision support systems [20]. In this context, ML techniques have been applied to predict: i) environmental impacts [21]; ii) changes in crop production [22]; iii) household-induced environmental impacts [23]; or, iv) variations in greenhouse gas (GHG) emissions due to vehicle fleet electrification [24]. In contrast to traditional statistical approaches, ML methods propose the construction of functions that map input variables to an output variable by using a function constructed exclusively from data [25]. This modelling approach has been used extensively in spatially explicit models, especially in the study of deforestation and land-use changes [26–28], but, to the best of our knowledge, there are still no implementations of these methods focusing on alluvial mining and accessibility variables.

In the current study, ML methods have been applied as a novel perspective to analyze the trends of ASGM. More specifically, the main goal is to analyze past mined areas to identify insights in terms of their dependency to spatial variables and to predict potential future hotspots of mining activity expansion in MDD. Firstly, we want to know if mined areas and the selected proximity variables follow any observable distribution or pattern along the whole area of study. Secondly, we aim to evaluate the predictive capacity of the selected variables and their utility to locate future hotspots of mining activity. To this end, we propose the use of an unsupervised learning algorithm (i.e., X-means), and a classification model (i.e., random forest–RF) for the descriptive analysis and hotspots prediction, respectively.

The novelty of the approach proposed is based on two main pillars: i) to analyze in a descriptive and spatial manner the past expansion of mining activities; and, ii) to explore the effectiveness of using a ML model to identify future potential hotspots of mining activity. In both cases, we propose the use of explicit geo-spatial information where physical and anthropogenic publicly available variables are considered. All the analysis and calculations were performed on the cloud-based computing framework, Google Earth Engine – GEE [29]. We expect this approach to be of utility for academics working on prospective ML methods linked to deforestation, especially in other tropical areas of the world. Nevertheless, the case study selected can be considered a useful tool to develop policy support actions in the Peruvian Amazon.

## 2. Materials and methods

### 2.1. Study area

The analysis performed considered the alluvial mining area of the region of MDD, which is mainly located throughout the riverbeds of the tributary rivers of the Madre de Dios river. The region of interest was delimited by an approximate rectangular window with an area of 6200 $km^2$, including important mining settlements, areas with scattered activities, and areas of potential emergence of these. The region of interest described was used as the geographical reference throughout the methodological framework, including sampling, data analysis and model construction. A map with the location of the area of interest can be found in Fig. S1 of Supplementary Information. From a temporal perspective, the study focused on the period 2012–2017 based on two main characteristics of the general trend of mining activities in MDD. On the one hand, the operational phase of the IH began in 2011, triggering accessibility in the region and representing a relevant benchmark of the current infrastructure layout in MDD. On the other hand, the annual mining rates started to show similar trends during this period, showing a resembling distribution of mined area among relevant mining clusters (e.g., Huepetuhe, Delta or La Pampa …).

### 2.2. Data acquisition and processing

The analysis performed relied on a Geographical Information System (GIS). In this sense, model construction and analysis used a database structured, mainly, from public georeferenced data. The selected variables aimed to be predictors of mining pressure in the area then were associated with accessibility and other geophysical properties of the environment. The importance of accessibility as a driver of deforestation and mining activity has been suggested by several studies associated with mining activities [14-15,30-31], and with other sources of deforestation in similar Amazonian regions [28]. Physiography data were obtained from Theobald and colleagues by combining landforms and lithology [32]. Digital elevation data of 1 arc-second resolution (about 30 m) were obtained from the Shuttle Radar Topography Mission (SRTM) led by NASA in 2000, and globally released in 2015. Clay and sand content were directly obtained as rasters globally gridded at a 250 m resolution based on geostatistics and performed by ML techniques [33,34].

These four raster images are already stored in GEE. Road maps were obtained from the Ministry of Transport as shape files that included the distribution of roads throughout the Peruvian territory, as well as attributes describing their characteristics (e.g., primary, secondary, tertiary, constructed or projected) [35]. PAs, BZs and villages were obtained in shape format from the Ministry of Environment and they indicate the geographic boundaries of these designated areas [36].

Similarly, polygons and vectors that indicated rivers and ponds were obtained from HDX [37]. Finally, mining data were obtained from Caballero-Espejo and colleagues [6] as a shape file. This file considers deforestation due to ASGM during the period 1984–2017, as well as attributes with mining activity characteristics, such as type of mining activity that generated the deforestation (e.g., heavy machinery or suction pumping). These data derived from the use of use of Landsat imagery and the integration of classification algorithms, visual inspection, and manual processing. Any reference to mining activity hereinafter will allude to deforestation due to mining activities. A summary of variables utilized can be found in Table S1 of Supplementary Information.

Once the data were collected, these were entirely handled in GEE. Raster files were directly uploaded and shape files were processed using GIS to extract only the attributes of interest. Thereafter, shape files were converted to rasters and resampled using a resolution of 100 m per pixel, leading to a stack of raster maps with the same resolution. The raster stack contains implicit data which may not provide sufficient information. However, for model construction purposes, the algorithm selected requires input data to be represented in a more explicit manner [38]. For this, different *features* were formulated in order to be included in the database and to improve the final performance of the predictive model.

The *feature* engineering process consisted of a variety of GIS-based calculations that implied, in some cases, a complete modification of the initial data. This followed a rationale like the one found in the literature, where important variables linked to deforestation and other land-use changes are commonly described as a type of proximity or intensity. To ease comprehension, we found it convenient to analyze features following a dual classification based on their temporal characteristics: static and dynamic.

#### 2.2.1. Static features

Static features are those with values that are not time-dependent, or at least those that are not supposed to change in a short period of time (e.g., 5 years). These are, mostly, geographic and physical variables which describe the natural environment and are assumed to remain invariant during modelling and prediction. *Altitude*, *physiography, clay content,* and *sand content* did not require any additional processing and they were imported directly from GEE. *Physiography* presents a map of 15 classes. *Altitude* contains the pixel elevation. *Clay content* and *Sand content* present the fraction of clay and sand content up to 2 m depth. A *slope* feature was calculated based on *altitude* and describes the slope in degrees. *Latitude* and *Longitude features* were also considered to indicate the position of a pixel. Finally, two proximity features, *distance to rivers* and *distance to ponds,* were obtained by calculating the Euclidean distance from each pixel to the nearest river or pond, respectively.

#### 2.2.2. Dynamic features

These features describe changing phenomena linked to anthropogenic activities or elements, such as infrastructure or human settlements. In this sense, dynamic features might have different values depending on the year of observation and constitute the features that can be modified to predict a given scenario. To this effect, a first set of proximity features were calculated from the road maps, PAs, BZs and villages, which represented the Euclidean distance from any pixel to these infrastructure or designated areas. In this case, the available data did not distinguish the year of construction of the roads nor the year of establishment of the villages; therefore, the values of each pixel remained constant for the period 2011–2015. However, because of their nature and for the sake of consistency, these were still classified as dynamic features.

A second set of features were calculated from the mining activity raster. The *mined* feature, which was treated as a dichotomous variable, indicated if there was an occurrence in the corresponding year. This feature was used as the variable to predict, but it also served as an input in the generation of other features. For instance, for every previous year in the window of time, we created rasters that contained the quantity of mined pixels surrounding every pixel in a radius of 400 m. The rasters of mined neighbors and mining activities for previous years were included as the *i_neighbors* feature and the *i_mined* feature, respectively. For both features, the prefix *i* made reference to the previous *i*th year counting from the corresponding *mined* feature year.

Additionally, a mining rate map was generated to represent the tendency of mining activities surrounding a certain pixel in a short window of time (i.e., 4 years). The main assumption for this is that mining is more likely to occur in areas where this activity is already happening. We selected a linear regression (i.e., temporal reduction) of a convolution of mined pixels (i.e., spatial reduction) as a strategy to reduce the multidimensional nature of this tendency. For this, a linear regression was fitted for each pixel using the year and the sum of surrounding mined pixels as variables. This allowed us to extract the parameters of the regression (i.e., slope of the curve and intercept, *beta* and *alpha* hereinafter) and to use them as a *beta* feature and an *alpha* feature (see Fig. S3 in Supplementary Information). For instance, pixels with a positive *beta* value indicate that in that specific window of time, the mining activity in a radius of 400 m has increased. Conversely, a negative value may suggest that the activity has been decreasing. A value close to zero can be associated with no mining activity or recurrent mining activity at a similar rate.

### 2.3. Algorithms selection

The selection of the algorithms responded to initial insights obtained from a descriptive data analysis in which we aimed to explore correlations and insights among variables. In this stage, paired correlation matrices were built for all variables. The use of a clustering algorithm was justified by the presence of multimodal distributions and heterogeneous correlations with the geographical positions that suggested the presence of distinguishable clusters (see Section 2.4). X-means is a clustering algorithm that groups data in a number of clusters, X, that is optimally determined by running multiple instances of the K-means algorithm [39]. In K-means, every pixel is assigned to a cluster where the Euclidean distance between the cluster centroid and the pixel, in a multidimensional space, is the minimum [40]. We selected this clustering algorithm (i.e., X-means based on K-means) because of two main reasons. First, it is an unsupervised classification technique that has been widely used in studies dealing with geo-reference data, especially when dealing with low dimension datasets of large sizes [41,42]. In fact, as discussed in Wu et al. [41], K-means has shown to be a valid alternative when it respects to answering questions related to spatial-clustering and attributes clustering. Secondly, K-means and X-means have a fast and robust code implementation in GEE which integrates conveniently with GEE parallelized architecture. This allowed us to sample, modify and visualize the results of the clustering exercise repeatedly and on-the-fly regardless of the size of the area of interest (i.e., 26 ha).

RF is an ML algorithm where the combination of independent decision trees with low bias and high variance generates an ensemble "forest" with low bias and low variance [43]. This nonparametric algorithm stands out for its robustness in terms of overfitting and multicollinearity, its low sensitivity to outliers, and its capacity to deal with categorical and continuous variables [44,45]. In fact, recent studies focused on spatio-temporal phenomena still rely on RF as a component of the computation pipeline in the study of spatial phenomena in tropical areas (e.g., deforestation) [46,47] or when trying to identify other types of spatial patterns [48,49]. Moreover, when oriented towards spatial prediction, this algorithm can match the performance of feed-forward deep learning models that predict deforestation using accessibility variables [28]. This occurs due to the low dimension of the spatial dataset and the high prediction capacity of accessibility variables. Moreover, RF does not require data normalization and is faster during training and prediction. Secondly, since the objective of the study is to identify hotspots and insights, we prioritize the practicality in the implementation, since

RF has also been implemented in GEE and is suited for performing predictions and visualizing in real time (see SI).

### 2.4. Sampling, analysis and data clustering

The sampling exercise consisted in the construction of a structured database obtained by sampling pixels from the rasters previously generated. This structured database was meant to serve as the source for the data analysis, and as training and validation data for model construction. We built paired correlation matrices in which we observed the presence of noticeable multimodal distributions among the distance variables when paired with location-related variables (i.e., latitude and longitude), meaning the correlation based on proximity was not constant in the whole area of study. To deal with this and to include the heterogeneity of the region of interest into the sampling exercise, we clustered the mined pixels from period 2011–2015 using the X-means algorithm. For this clustering task, we only considered the *static* and all the proximity features. The resulting clusters represent groups of homogenous mined pixels not only in terms of geographical location (i.e., *altitude*, *latitude* and *longitude*), but also related to distance to infrastructure, soil, geomorphology, and hydrological characteristics. The resulting raster image containing pixels' cluster labels was included as a feature and was used for both data analysis and model construction. Moreover, since this database should represent the area of study and the mining phenomenon and may condition the performance of the model, we later decided to extract multiple samples as explained below.

The training and validation sets were built combining different sampling regions and criteria. Regarding the sampling regions, two different regions were considered: the complete area of study (i.e., FULL hereafter) and a smaller rectangular region that encapsulated historical mining areas (i.e., Huepetuhe, La Pampa and Delta), labeled as Historic Region (i.e., HR hereafter). The logic of this approach was to observe whether models trained with data sampled only from this HR could lead to a realistic or reasonable prediction of mining hotspots outside the training region.

Regarding the sampling criteria, our main objective was to reduce the model bias towards predicting a mined or non-mined class only. For this, the training dataset was sampled in a stratified manner. Consequently, we considered two sampling criteria. In the first case, we sampled an evenly distributed number of mined and non-mined pixels from the *mined feature* (i.e., 0 and 1) [50]. In the second case, we sampled the same number of mined pixels from each cluster. This led to an evenly distributed sample of mined and non-mined pixels, where the former was also evenly distributed among their clusters' labels. For the sake of simplicity, hereafter both criteria will be labeled as "stratified-only" and "stratified-from-clusters", respectively.

In all cases, only pixels 600 m apart were considered in order to reduce spatial autocorrelation among samples, similarly to the methodological choice proposed by Mayfield and colleagues [26]. All models were constructed using the combination of these sampling criteria, meaning that the rasters were sampled multiple times. For every sampling combination, data were divided into training and validation sets which included 70% and 30% of the sampled data, respectively. For the specific case of the HR, an additional validation set was sampled from an area that encompassed the mined pixels located outside the training zone.

### 2.5. Modelling stage

Two modelling perspectives were set. The first one, a *static-oriented* perspective, used only *static* and all the distance to infrastructure *features*. For this, we merged mined pixels from the period 2011–2015 into a single image from where data were sampled following the before mentioned sampling strategies. Our target was to observe if the model could detect hotspots of high probability of mining activity under the assumption that, for some areas, these static constraints may influence the decision to initiate (or continue) mining activities. For this perspective, a total of 8000 pixels were randomly sampled, resulting in 6400 and 1600 training and validation pixels, respectively.

The second perspective was *temporal-oriented* and considered all the *static* and *dynamic* variables. The training database was constructed by sampling a subset of mined and non-mined pixels corresponding to years 2011, 2013, and 2015 following the sampling strategies previously stated. The *features* associated to a mined pixel in a given year contained information that corresponded to the past. More specifically, we considered a time span of 2 years in order to prevent the model from depending only on the mined-derived features (i.e., *mining rate*, *mined neighbors*). In this sense, the outcome of each prediction task could be interpreted as the probability of mining activity on a given pixel for a year *t*, based on data from years *t-2, t-4,* and *t-6*. In contrast to the first perspective, the model was evaluated based on its capacity to predict pixels that were going to be mined in the following two years. For this perspective, we trained and evaluated the models to identify mining hotspots for year 2017. For each sampling strategy, a total of 24,000 pixels were sampled (i.e., 8000 per year), resulting in 16,800 and 7200 sampling and training points, respectively.

The combination of sampling regions and strategies led to a total of 4 sampling procedures: FULL + stratified-from-clusters (A); FULL + stratified-only (B); HR + stratified-from-clusters (C); and, HR + stratified (D). These sampling strategies were used to train both *static-* and *temporal-oriented* models varying the hyperparameters such as the number of trees (i.e., from 10 to 200), the maximum amount of leaf nodes in each tree (i.e., from 10 to no-limit), and the percentage of data out-of-bag (i.e., from 0 to 100%). The best set of hyperparameters for every model was selected using accuracy (ACC) and the area under the curve ROC (AUC) as metrics. Finally, the performance of the trained models was analyzed and compared using ACC, AUC, and f1-score (F1), and recall (REC) as metrics.

The RF models were constructed utilizing the SMILE's RF algorithm which is implemented as a built-in function in the GEE web API (i.e., ee.Classifier.smileRandomForest). A graphical representation of the computational pipeline of the modelling stage can be found in Fig. S2 of Supplementary Information.

## 3. Results

### 3.1. Descriptive analysis based on clustering

When the distribution of mined pixels based on the distance to infrastructure *features* is observed, many of them show a multimodal curve (see Cluster 0 in Fig. 1). This occurs mostly because mined pixels are not homogeneously distributed around infrastructure. Instead, they follow a heterogeneous geographic distribution that results in visually distinguishable clusters (see Fig. 1). This heterogeneity hinders the capacity of the model to interpret the geographical distributions and the relationship among mined pixels and the *features* under evaluation, especially because there is not an easily distinguishable pattern that can be valid for the whole area of study. In this sense, we analyzed the training data considering also different clusters that were obtained from the unsupervised classification stage explained above.

The implementation of the *X-means* algorithm resulted in 13 clusters that were used to label each mined pixel with a corresponding *cluster* label. We analyzed the relationship between mining occurrences and the different distance variables (i.e., primary road, secondary road, tertiary road, protected area, ponds, rivers, and villages), *mining type* (i.e., heavy machinery and suction pumping) and the *year of mining* (i.e., from 2011 to 2015) for every resulting cluster. Consequently, only the distance variables were considered during the clustering task, while *mining type* and *year of mining* were included after *cluster labels* were assigned.

Regarding the proximity to roads features (i.e., primary (a), secondary (b) and tertiary (c)) in the different mining sites, two main patterns can be distinguished. On the one hand, in clusters 1, 3–10 and

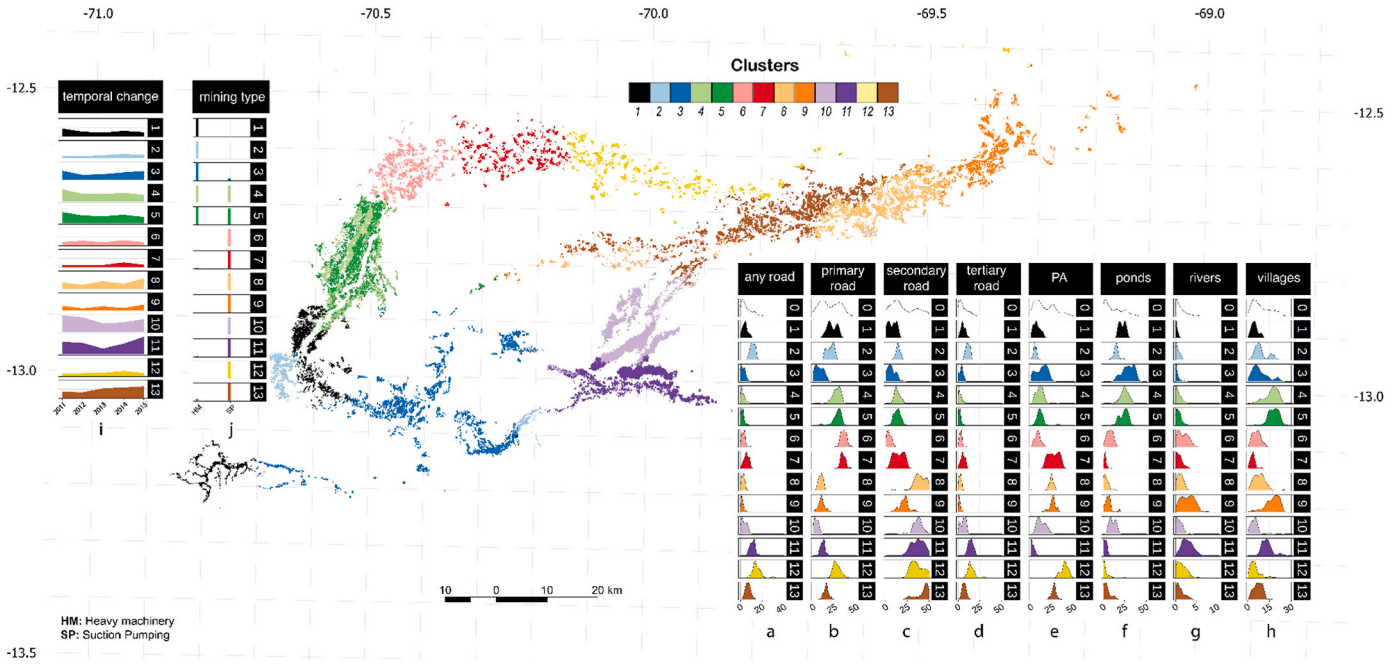


**Fig. 1.** Visualization of 13 mining clusters obtained by using X-means algorithm in Madre de Dios (Peru). Mining activity distributions according to distance to infrastructure features (from (a) to (h)), activity evolution (j), and mining type (i) for each cluster are shown.

13, mining activity occurs approximately 5–10 km away from the access to any type of road. Other important arteries that connect mining activities with the supply of resources and transportation of people are rivers. Most of the clusters had a mean distance to rivers lower than to any type of roads; however, this fact may not only be connected to accessibility but may be also a consequence of the intrinsic characteristics of some alluvial mining methods that find most of their extraction potential near riverbeds and other water bodies (e.g., oxbow lakes).

On the other hand, mining activities identified in Cluster 11, in contrast, have occurred in areas with low accessibility and relatively distant from water bodies (see e, f and g in Fig. 3). Moreover, when analyzing proximity to villages, clusters 4–5, 9 and 11 are those located farthest away from human settlements. Therefore, Cluster 11 is the most distant cluster from all infrastructure elements (i.e., roads and human settlements) evaluated, as shown in Fig. 2.

Regarding PAs, observations suggest that they actually act as repellents of mining activity. Thus, only 5 out of the 13 clusters (i.e., 7–9 and 12–13) present a considerable distance from these areas, while other clusters have a distance, on average, that ranges from 5 to 15 km (see Fig. 3a). These low distances may imply that mining activities are occurring inside BZs, such as the case of the Amarakeri Reserve (see Fig. 3b).

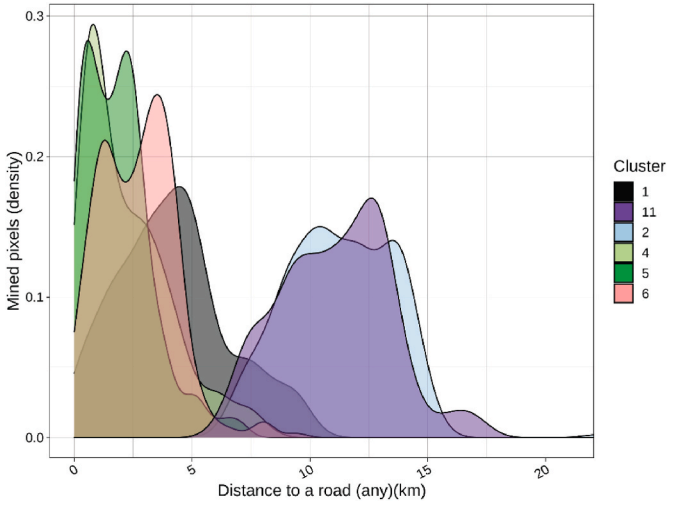


**Fig. 2.** Density curves represent distance of mined pixels to accessibility infrastructure (roads) for clusters 1, 2, 4, 5, 6, and 11.

It is important to note, however, that Cluster 11, which was earlier identified as being the most distant from infrastructure elements and includes most of the area of La Pampa [6], shows mining activity inside BZs and PAs. More specifically, the activities occur inside the PA of the Tambopata National Park and its BZ (see Fig. 4b). This clustering exercise led us to identify areas where the absence of this repellent effect is suggested (see Fig. 4a).

Regarding the *mining type* (see distribution of mining type in Fig. 1), clusters 1–3 and 6–13 are areas where mainly heavy machinery and suction pumping techniques are applied for gold extraction, respectively, whereas clusters 4 and 5 include both types of mining without a clear predominance. Finally, *temporal change* presents the evolution of mining activity (i.e., area) for each year. While the overall mined area in the region of interest has increased, the evolution of mining activity for the different clusters is heterogeneous. For instance, clusters 1–2, 4–7 and 12 show a slight reduction in mined pixels from year 2014–2015, while the remaining clusters show an increase in mining activity during the same period. The rapid increase in mining activity observed in Cluster 11 starting in 2013 constitutes the most remarkable augmentation.

### 3.2. Model training and validation

After varying the number of trees and nodes, we identified that for all sampling strategies, model metrics converged to stable values after reaching 100 trees (see Fig. S5 in the SI). In parallel, we also noted that the number of nodes does not influence the model performance after a value of 400. A feature importance ranking revealed that distance to secondary roads, ponds, parks and villages accounted for more than 50% of the importance in the RF model (see Fig. S4 in the SI).

Fig. 5 shows the scores for models tested on validation data resulting from splitting the sampled dataset and from the additional testing

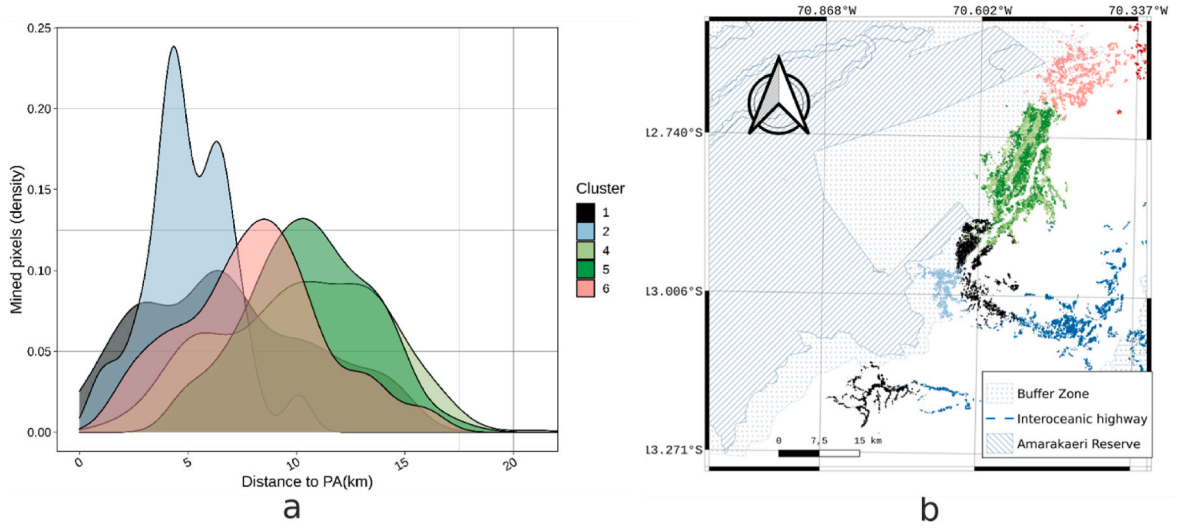


**Fig. 3.** Density curves (a) representing the distribution of mined pixels for specific clusters (1, 2, 4, 5 and 6) in the surroundings of the Amarakaeri Reserve (b).

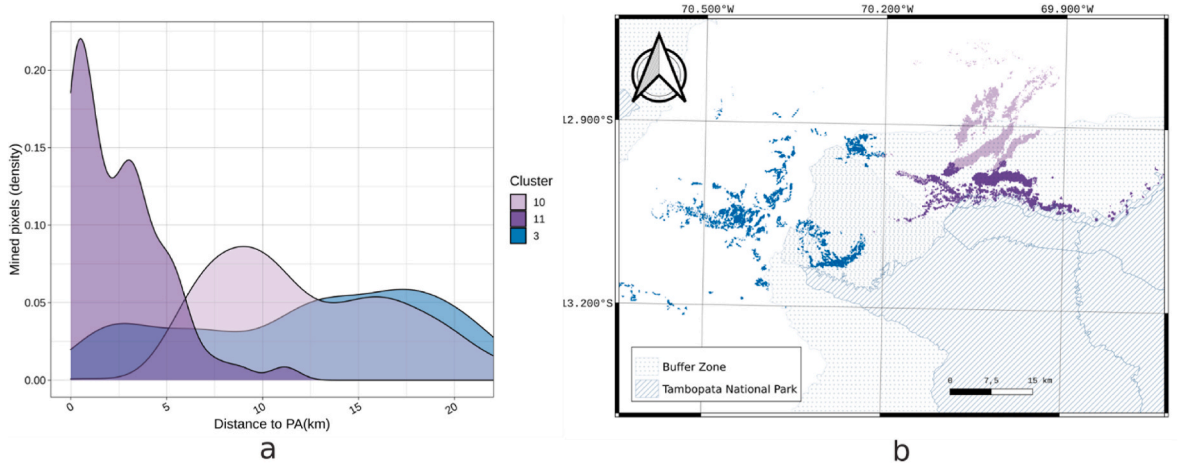


**Fig. 4.** Density curves (a) representing the distribution of mined pixels for specific clusters (3, 10, 11) in the surroundings of the Tambopata National Park (b).

sample (i.e., indicated as *) used in samples C and D. As can be noted, for both *static-oriented* and *temporal-oriented* approaches, taking the FULL area of study as the sampling region always produces models with higher performance for every metric. Similarly, in all cases, predictions performed on test data sampled from inside the HR showed higher scores than predictions obtained using samples from outside the HR. This behavior is expected if we consider that it is easier to predict samples linked to locations closest to the training data.

When focusing on ACC and AUC, it can be noted that the selection of the sampling criteria does not produce significant differences in the performance metrics. However, this trend changes when observing F1 and REC in the *static-oriented* models. In the case of these metrics, the proposed *stratified-from-clusters* criteria did not yield models with better performance if compared with *stratified-only* criteria. In the case of the *temporal-oriented* models, both sampling criteria do affect the metric's values that are more influenced by the sampling region.

### 3.3. Prediction maps and empirical validation

Once the visual analysis of maps resulting from predicting mining activity was concluded, the best-performing models that were visually meaningful for the discussion were selected. Fig. 3 shows the prediction of mining activity using the models selected, whereas the complete set of predicted maps can be observed in the SI. In the first place, Fig. 6a represents the results from the *static-oriented* (referenced with the prefix S- hereinafter) model using the FULL area of study as the training zone (i.e., model S-A). Secondly, Fig. 6b and c shows the results of the *static-oriented* model using the HR as training area with different sampling criteria (i.e., S-D and S–C). Finally, Fig. 6d shows the prediction of the *temporal-oriented* (referenced with the prefix T-hereafter) model for year 2017 using HR and stratified-from-clusters criteria (i.e., T-C).

Three prominent zones of analysis were selected (see Z1, Z2, and Z3 in Fig. 6). For instance, Z1 and Z3 are located outside the HR where predictions vary significantly depending on the model. For these zones, model S-A (Fig. 6a) shows higher probability ratios if compared to the results from model S-D (Fig. 6b) and S–C (Fig. 6c). Moreover, in Z1 all models predict similar mining footprints, while in Z3 only model S-A shows a defined footprint. With regards to Z2, located within the HR, the difference between mining probabilities among the three models is less notorious; however, a higher probability and footprint can be found in the first sampling strategy. Nevertheless, none of the models shown was capable of identifying a notorious area of mining in Z2 that corresponded to year 2017 (Fig. 4b), a period that was not considered in the training sample.

When observing the temporal model (i.e., T-C in Fig. 6d), it can be noted that the pixels predicted describe a more demarcated footprint,

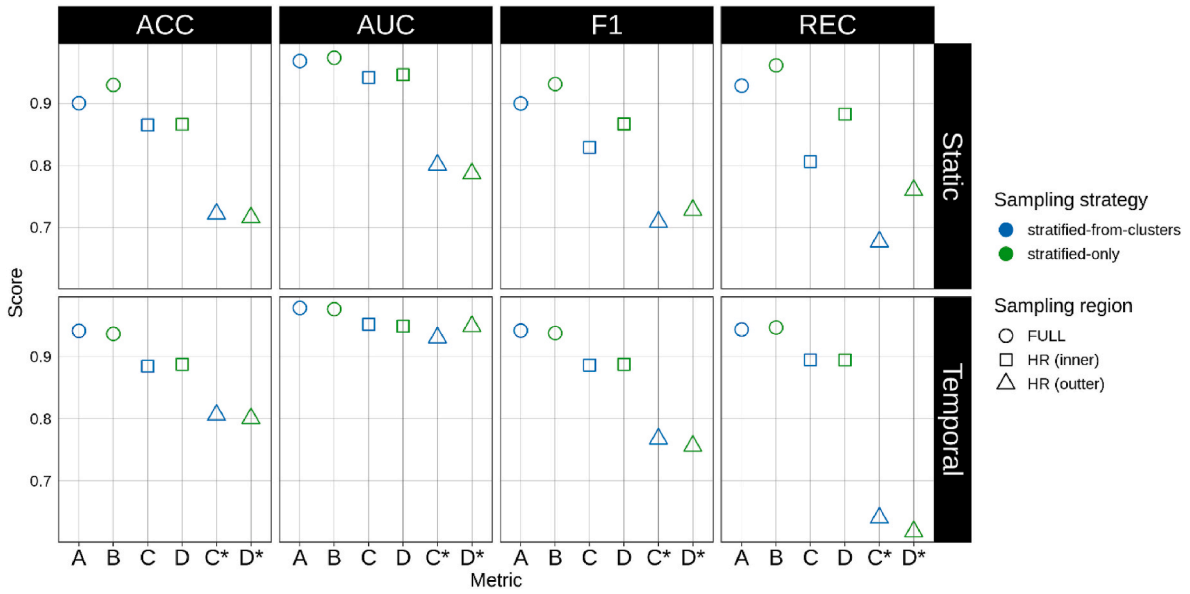


**Fig. 5.** Validation scores for testing datasets constructed following different sampling regions and strategies. * Indicates special testing dataset sampled from a region far away from HR. ACC, AUC, F1 and REC stand for accuracy, area under the curve, F1-score, and recall, respectively.

where the boundaries between deforested and non-deforested areas are clearly distinguishable. This is mainly due to the inclusion of *alpha* and *beta* features into the modeling. Consequently, we can assume that these mining rate proxies have a high influence on the appearance of the prediction map. Moreover, if attention is given to the probability values, distinguishable sectors of intense red (i.e., >0.7) and light orange (i.e., >0.5 and <0.7) match areas of vigorous (i.e., La Pampa) and active but diminishing (i.e., Huepetue) mining, respectively. In contrast to the static models, the model in Fig. 6d cannot easily predict values over 0.5 if there was no previous mining activity nearby. Nevertheless, we identified areas where new mined pixels were predicted despite the absence of mining in previous years.

When observing zones Z1, Z4 and Z5 in Fig. 7 we can distinguish that model S–C has the capacity to identify areas with high probability of mining even when these were completely ignored during the training step. For the case of Z1, the model could grossly predict mining that already occurred years ago, while for Z4 and Z5 the model could predict the footprint of the riverbeds where the mining activity had started in the last year of analysis (i.e., 2017). Interestingly, monitoring and governmental reports indicate that at the time the current study was being written (i.e., late 2021), the Pariamanu riverbed (i.e., Z5) experimented outbreaks of mining activity as previously reported [51].

### 3.4. Prediction error and clustering

Given the inconsistencies in the model's accuracy throughout the area of interest, we conducted an analysis to determine in which clusters the model performed worst. For this, we sampled 4000 random pixels and predicted their probability of mining activity. We later compared the predicted classification (i.e., using a threshold of 0.5) with the real data to separate them in two groups: "correct" and "incorrect". Fig. 8 shows the ratio of correct and incorrect predictions for each cluster. As it can be observed, clusters 3, 11 and 12 are the ones with a higher proportion of incorrect predictions when comparing with the rest of clusters. With respect to clusters 11 and 12, these correspond to the clusters previously identified as distant from infrastructure (i.e., roads and human settlements). It appears that in these areas the variables selected are not as good predictors as they are when evaluating other clusters.

## 4. Discussion

The main areas with high mining probability were identified in the town of Puerto Pariamanu (12°26′0″S; 69°15′0″W) and the surroundings of Puerto Maldonado, the region's capital. Interestingly, for the former there is evidence of recent mining activity in the period 2019–2020, as reported by Mongabay [52]. Our results match with this trend and are plausible if we consider that many mining activities tend to occur close to waterways and human settlements. In fact, despite the variations in results due to model selection, we identified that predicted mining in this area is recurrent in different models (i.e., T-C and S–C). Moreover, we consider that sectors between the city of Puerto Maldonado and Tambopata National Park require particular attention because there is still no established land use in the area [53], and mining activity is bound to continue degrading neighboring areas of one of the most visited national parks in the country [12,54].

Regarding sampling strategies, we argue that these have an important influence on the geographic representation of the predictions, even if they yield models with similar performance metrics. For example, when comparing model S-A with real mining, the predicted mining activity footprint has bigger dimensions than the real mining areas. Moreover, areas that were not previously mined have probability values close to 0 (i.e., color blue in Fig. 3). We hypothesize that this may be due to the selection of the full region of interest as a source of training data, implying, as explained by Ploton and colleagues [55], that the model may be replicating the area of study, expecting that non-mined areas have a low probability of being mined. Thus, we consider that this may hinder the model's capacity to predict future mining activities inside or outside the area of study. In contrast, when observing models S-D and S–C, we noted that these yield a more plausible footprint of the real data and are capable of predicting mining in zones outside or far away from the training area (see Fig. 6).

Consequently, all the sampling strategies selected yield an overall mining footprint delimitation that encompasses real mining activity (see Fig. 1). However, *static-oriented* models S-D and S–C show a more refined and feasible representation of the real data when compared with S-A. Moreover, when observing the *temporal-oriented* model (i.e., T-C), it can be noted that the predicted footprint has a more precise delimitation, but it is less sensitive to identify new potentially mined areas. In contrast, *static-oriented* models S-D and S–C produce less precise footprints but are capable of predicting mined areas outside and far away

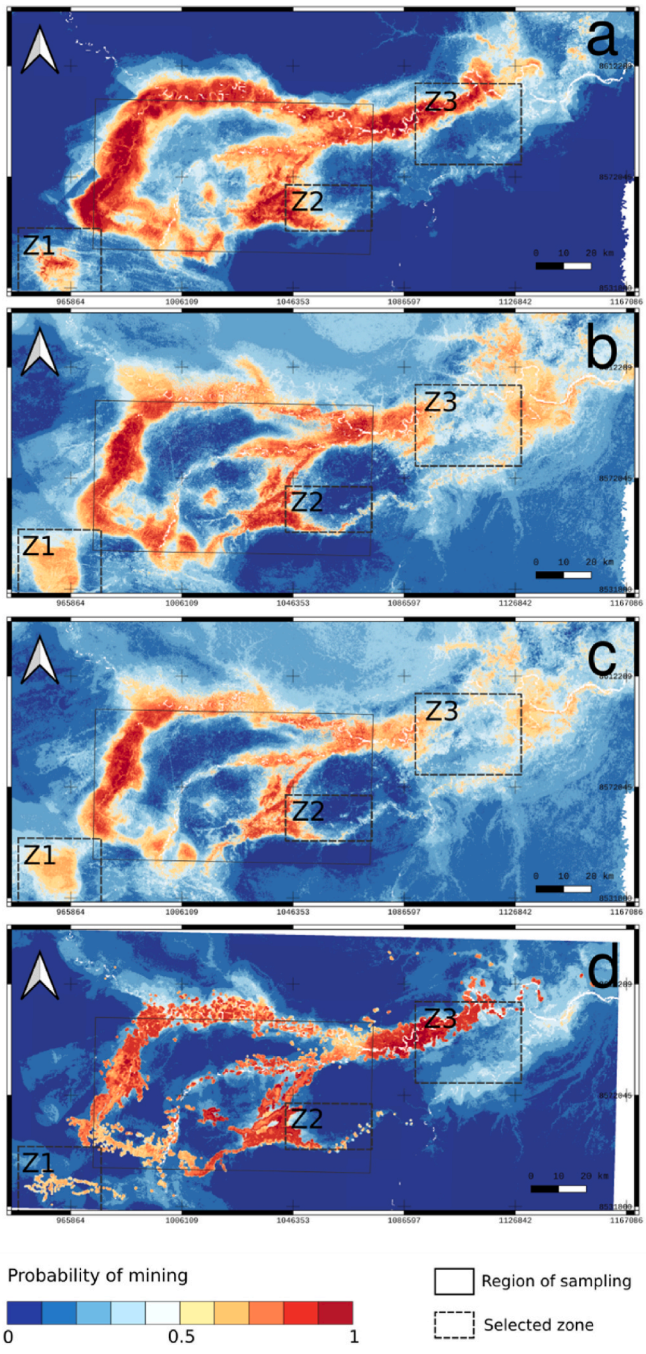


**Fig. 6.** Maps of mining probability for static and temporal models using different sampling strategies. **a**: *static* FULL + stratified-from-clusters (S–A). **b**: *static* HR + stratified-only (S–D). **c**: *static* HR + stratified-from-clusters (S–C). **d**: *temporal* HR + stratified-from-clusters (T–C). Z1, Z2 and Z3 indicate prominent zones selected for inspection.

from the training zone, or that were not mined before. In this sense, we argue that it is more practical and useful to rely on *static-oriented* models to conduct our analysis and discussion since they deliver information more aligned with the exploratory nature of our study. Having said this, we consider that a *temporal-oriented* approach would be more suitable if the goal is to foresee with detail how the mining activity expands.

Our results demonstrate that proximity to infrastructure is not always indicative of high mining probability. In fact, when analyzing the spatial distribution of model accuracy (i.e., correct/incorrect predictions per cluster), we observe that the model's performance decreases in those clusters where accessibility and mining activity showed opposite trends (i.e., 11 and 12). For instance, model predictions in areas nearby and downstream the Malinowski river do not show important hotspots due to their contiguity to the Tambopata National Park, far from human settlements and infrastructure. However, in real life (i.e., 2017 - present), areas like La Pampa and Malinowski have experienced aggressive mining activity. This fact demonstrates the existence of opposing trends in terms of dependency to accessibility, making mining a potential outlier as compared to other land uses (e.g., urban sprawl, logging, cattle ranching, aquaculture …) in the Amazon rainforest, which tend to be highly correlated to road proximity and accessibility [56,57]. Curie and colleagues [58] have related poverty of Amazonian communities with remoteness, identifying that forest products become a trigger to mitigate poverty in many remote areas. In this sense, it is plausible to assume that mining may also be a mechanism to boost income in remote communities in the region of interest. Consequently, we recommend future research to focus on the behavior of mining activities in inaccessible and remote areas and their relationship to socio-economic indicators in the local population.

Considering that static and dynamic features were selected for the analysis, on the one hand, it has been shown that geographical conditions can be associated with the occurrence of mining for particular regions. On the other hand, for some regions, the sudden appearance of mining does not match previous patterns and might suggest the presence of external factors that were not considered in this study. Nevertheless, the results lead us to state that ML techniques allow anticipating new mining areas, and can serve as a benchmark to estimate GHG emissions and other environmental impacts (e.g., toxicity) from the deforestation process [59]. A perfect match between past data and predictive values does not appear to be a plausible outcome. However, the current model provides insights of potential mined areas, based on our knowledge of the past. It is important to mention that we were not able to obtain a numeric metric that represented the consistency of the prediction of expansion mining areas. In this sense, we trained models to identify hotspots and we analyzed where the mining predictions were recurrent.

Regarding the spatial distribution of the predicted pixels in the *static-oriented* models, neighboring mined pixels tended to have similar probability and it was not common to observe drastic changes among their values. This trend, however, was not always reflected in the real data where non-mined pixels could be found inside regions with high mining activity. Our interpretation is that it is logical that neighboring pixels should have the same probability of mining since they have similar geographical *features*. In this sense, we hypothesize that the presence of these non-mined pixels could be explained by *features* not identified in this study (e.g., miner behavior and/or perception on mining potential in a given piece of land) or may correspond to some randomness intrinsic to this anthropogenic activity.

Beyond the geographical distribution of the predictions, including more explanatory *features* could aid in the interpretation of the prediction and in understanding the causality of the phenomenon. For instance, ASGM activities are highly dependent on governance conditions [5]. In fact, command and control actions, such as interdictions implemented by the Peruvian Government in the past, have been used as a recurrent strategy to enforce laws. This policy has shown scarce benefits when solving the issues related to mining sprawling and has accelerated the propagation of this activity in unexpected areas, some within the area of study, although others may occur in other piedmont alluvial areas in other Peruvian regions [60]. This behavior has been observed in clusters 2 and 11 (see Fig. 1), sectors with history of enforcement activity which resulted in an expansion phenomenon. Other drivers that lead ASGM expansion are the learning process of informal miners, restrictions linked to technological factors of the supply chain (e.g., mercury imports or innovation through new technologies), socioeconomic parameters in the neighboring highlands, or even aspects related to safety and risk. However, these aspects are difficult to model as variables that can be added to the predictive process of ML applied herein, although they may induce variations or changes in

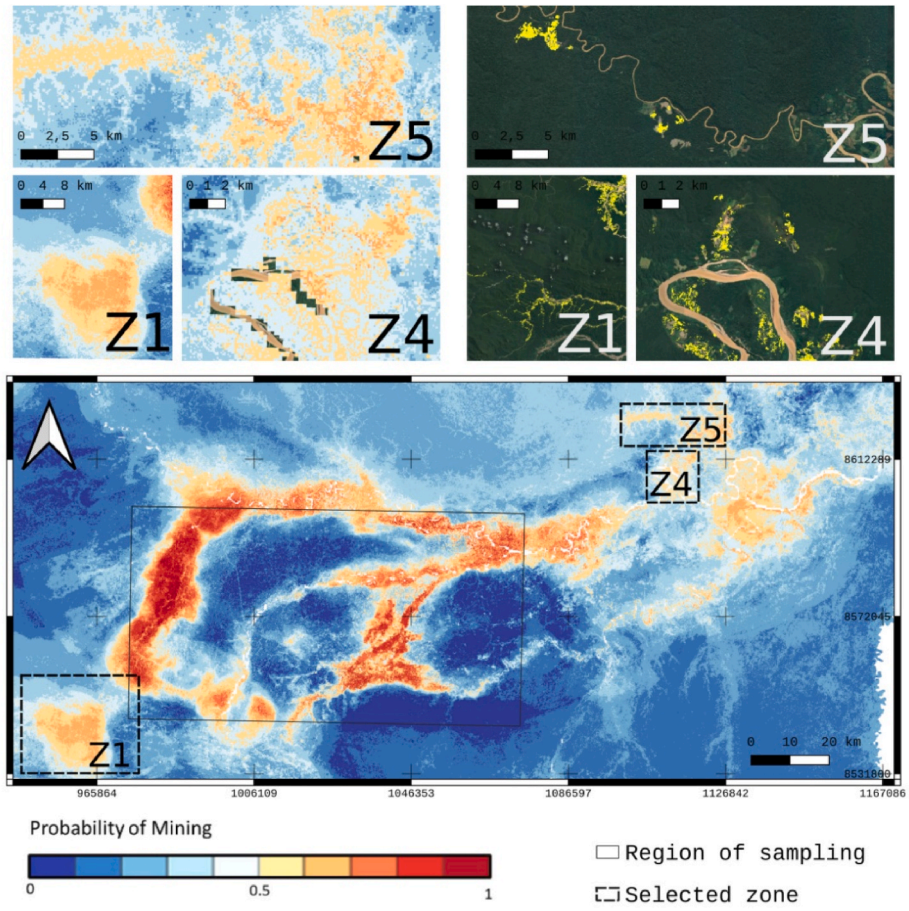


**Fig. 7.** Comparison of predictions of model *static* HR + stratified-from-clusters (S–C) and real mining activity for regions not considered during model training (i.e., zones Z1, Z4 and Z5).

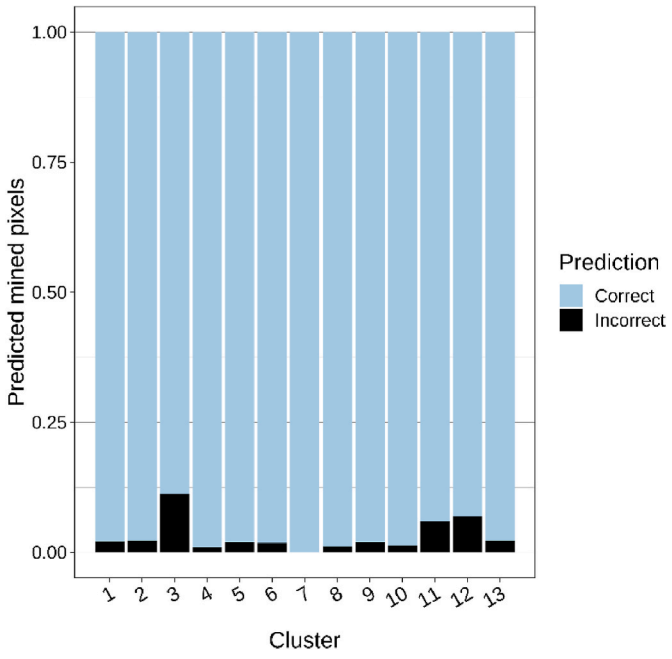


**Fig. 8.** Ratio of correct and incorrect predictions per cluster of a random sample of 4000 pixels.

relationship trends between clusters.

From a predictive perspective, miners can also be independent individuals that can take erratic and unpredictable decisions. Hence, accurately determining the location of future specific mining activities would require understanding miners' motivations and decision criteria. Despite the lack of granularity in terms of social patterns, these are, in many cases, implicitly portrayed in the set of variables used to model mining expansion. However, we do not have the mechanisms to isolate social interactions from other effects. Thus, we do not have the certainty that these social patterns will be replicated in the future and, therefore, these may change in time as part of the nature of a complex system. With the data obtained, we hypothesize that these behaviors may be a driving factor that is observed in Zone 2 (see Fig. 6c), where the model was unable to predict the appearance of new mining activity. Based on our understanding, this particular cluster has experienced intense activity that is not fully reflected if we observe data from previous years.

## 5. Conclusions

Motivated by the aggressive deforestation in the Peruvian Amazon, specifically MDD, due to alluvial mining activities and the lack of computational tools to anticipate new alluvial gold mining areas, this study uses ML techniques to explore gold mining expansion, and to identify possible future hotspots of these activities. ML application to ASGM expansion shows that the most appropriate modelling approach to analyze this activity can only be successfully trained to recognize previous patterns. However, we argue that our model can identify areas of high probability, especially those surrounding the main cities (e.g., Puerto Maldonado) and nearby areas of the Madre de Dios river, aiding

in the anticipation of preventive and mitigation measures. Therefore, we consider that our work contribution is twofold. On the one hand, the model constructed and the results can be used to enhance decision making when implementing geo-spatial policies for planning strategies to address the problem of ASGM expansion in the Amazon. Derived environmental, social and economic policies, such as deforestation, climate change or human health would also benefit directly from these data. On the other hand, the selected computational framework is flexible and reproducible, allowing the methodological pipeline proposed to adapt to alternative case studies in the Amazon (e.g., agricultural or cattle ranching expansion), to replicate informal mining-related deforestation patterns in other areas of the world or to admit the inclusion of additional variables in the present case study, which could become particularly relevant when modelling future scenarios. Indeed, while the method is proposed for a local region in the Peruvian Amazon, we consider that the methodological framework we present is sufficiently flexible to be applied to other areas of the world.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

Data will be made available on request.

## Acknowledgements

The authors wish to thank the Dirección de Fomento de la Investigación (DFI) at the Pontificia Universidad Católica del Perú (PUCP) for funding Project 455. Gustavo Larrea-Gallegos wishes to thank the Consejo Nacional de Ciencia, Tecnología e Innovación Tecnológica (CONCYTEC) for his Master's program scholarship in the period 2017–2019 at PUCP.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cscee.2023.100353.

## References

[1] S.A. Beal, B.P. Jackson, M.A. Kelly, J.S. Stroup, J.D. Landis, Effects of historical and modern mining on mercury deposition in southeastern Peru, Environ. Sci. Technol. 47 (22) (2013) 12715–12720.
[2] S.E. Diringer, A.J. Berky, M. Marani, E.J. Ortiz, O. Karatum, D.L. Plata, H. Hsu-Kim, Deforestation due to artisanal and small-scale gold mining exacerbates soil and mercury mobilization in Madre de Dios, Peru, Environ. Sci. Technol. 54 (1) (2019) 286–296.
[3] I.F. Machado, S. Figueirôa, Mining history of Brazil: a summary, Miner. Econ. (2022), 1-1.
[4] M. Moreno-Brush, J. Rydberg, N. Gamboa, I. Storch, H. Biester, Is mercury from small-scale gold mining prevalent in the southeastern Peruvian Amazon? Environ. Pollut. 218 (2016) 150–159.
[5] M.M. Veiga, O. Fadina, A review of the failed attempts to curb mercury use at artisanal gold mines and a proposed solution, Extr. Ind. Soc. 7 (3) (2020) 1135–1146.
[6] J. Caballero Espejo, M. Messinger, F. Román-Dañobeytia, C. Ascorra, L. Fernandez, M. Silman, Deforestation and forest degradation due to gold mining in the Peruvian Amazon: a 34-year perspective, Rem. Sens. 10 (12) (2018) 1903.
[7] R. Kahhat, E. Parodi, G. Larrea-Gallegos, C. Mesta, I. Vázquez-Rowe, Environmental impacts of the life cycle of Alluvial gold mining in the Peruvian Amazon rainforest, Sci. Total Environ. 662 (2019) 940–951.
[8] J.R. Kuramoto, La Minería Artesanal e Informal en el Perú. Mining, Minerals and Sustainable Development (82), 53. Retrieved from, http://www.detrasdelacortina.com.pe/download/Miner%EDa artesanal e Informal en el Per%FA (1).pdf, 2001.
[9] -B.J. Dammert, Financing infrastructure projects in the southern Amazon of Peru: its relation with environmental and social safeguards. https://www.bu.edu/gdp/files/2018/10/GEGI_GDP-Peru-WP.pdf, 2018.
[10] MAAP, MAAP #130: Illegal Gold Mining DOWN 78% in Peruvian Amazon but Still Threatens Key Areas, 2020, December 2, MAAP, 2020, https://maaproject.org/2020/gold-mining-peru/.
[11] B. Fraser, Peruvian gold rush threatens health and the environment, Environ. Sci. Technol. 43 (2009) 7162–7164.
[12] B. Fraser, Peru's gold rush prompts public-health emergency, Nature 534 (2016) 162.
[13] J.J. Swenson, C.E. Carter, J.C. Domec, C.I. Delgado, Gold mining in the peruvian amazon: global prices, deforestation, and mercury imports, PLoS One 6 (4) (2011).
[14] G.P. Asner, R. Tupayachi, Accelerated losses of protected forests from gold mining in the Peruvian Amazon, Environ. Res. Lett. 12 (9) (2017), 094004.
[15] G.P. Asner, W. Llactayo, R. Tupayachi, E.R. Luna, Elevated rates of gold mining in the Amazon revealed through high-resolution monitoring, Proc. Natl. Acad. Sci. USA 110 (46) (2013) 18454–18459.
[16] F.D.L. Lobo, M. Costa, E.M.L.D.M. Novo, K. Telmer, Distribution of artisanal and small-scale gold mining in the Tapajós River Basin (Brazilian Amazon) over the past 40 years and relationship with water siltation, Rem. Sens. 8 (7) (2016) 579.
[17] N. Alvarez-Berríos, J. L'Roe, L. Naughton-Treves, Does formalizing artisanal gold mining mitigate environmental impacts? Deforestation evidence from the Peruvian Amazon, Environ. Res. Lett. 16 (6) (2021), 064052.
[18] F. Donati, S.M. Dente, C. Li, X. Vilaysouk, A. Froemelt, R. Nishant, S. Hashimoto, The future of artificial intelligence in the context of industrial ecology, J. Ind. Ecol. 26 (4) (2022) 1175–1181.
[19] H. Arbabi, M. Lanau, X. Li, G. Meyers, M. Dai, M. Mayfield, D. Densley Tingley, A scalable data collection, characterization, and accounting framework for urban material stocks, J. Ind. Ecol. 26 (1) (2022) 58–71.
[20] B. Alavi, M. Tavana, H. Mina, A dynamic decision support system for sustainable supplier selection in circular economy, Sustain. Prod. Consum. 27 (2021) 905–920.
[21] A. Nabavi-Pelesaraei, S. Rafiee, S.S. Mohtasebi, H. Hosseinzadeh-Bandbafha, K. W. Chau, Integration of artificial intelligence methods and life cycle assessment to predict energy output and environmental impacts of paddy production, Sci. Total Environ. 631 (2018) 1279–1294.
[22] H. Lee, J. Wang, B. Leblon, Using linear regression, random forests, and support vector machine with unmanned aerial vehicle multispectral images to predict canopy nitrogen weight in corn, Rem. Sens. 12 (13) (2020) 2071.
[23] A. Froemelt, R. Buffat, S. Hellweg, Machine learning based modeling of households: a regionalized bottom-up approach to investigate consumption-induced environmental impacts, J. Ind. Ecol. 24 (3) (2020) 639–652.
[24] H. Cai, M. Xu, Greenhouse gas implications of fleet electrification based on big data-informed individual travel patterns, Environmental science Science & technologyTechnology 47 (16) (2013) 9035–9043.
[25] L. Breiman, Statistical modeling: the two cultures (with comments and a rejoinder by the author), Stat. Sci. 16 (3) (2001) 199–231.
[26] H. Mayfield, C. Smith, M. Gallagher, L. Coad, M. Hockings, Using Machine Learning to Make the Most Out of Free Data: A Deforestation Case Study, 8th International Congress on Environmental Modelling and Software, Toulouse, France, 2016. July 2016.
[27] M.A. Brovelli, Y. Sun, V. Yordanov, Monitoring forest change in the amazon using multi-temporal remote sensing data and machine learning classification on Google Earth Engine, ISPRS Int. J. Geo-Inf. 9 (10) (2020) 580.
[28] G. Larrea-Gallegos, I. Vázquez-Rowe, Exploring machine learning techniques to predict deforestation to enhance the decision-making of road construction projects, J. Ind. Ecol. 26 (1) (2022) 225–239.
[29] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, R. Moore, Google Earth engine: planetary-scale geospatial analysis for everyone, Remote Sens. Environ. 202 (2017) 18–27.
[30] D. Cortés-McPherson, Expansion of small-scale gold mining in Madre de Dios: 'capital interests' and the emergence of a new elite of entrepreneurs in the Peruvian Amazon, Extr. Ind. Soc. 6 (2) (2019) 382–389.
[31] S.G. Perz, E.R. Mendoza, A. dos Santos Pimentel, Seeing the broader picture: stakeholder contributions to understanding infrastructure impacts of the Interoceanic Highway in the southwestern Amazon, World Dev. 159 (2022), 106061.
[32] D.M. Theobald, D. Harrison-Atlas, W.B. Monahan, C.M. Albano, Ecologically-relevant maps of landforms and physiographic diversity for climate adaptation planning, PLoS One 10 (12) (2015), e0143619.
[33] T. Hengl, Clay content in % (kg/kg) at 6 standard depths (0, 10, 30, 60, 100 and 200 cm) at 250 m resolution (Version v02), Zenodo (2018), 10.5281/zenodo.1476854.
[34] T. Hengl, Sand content in % (kg/kg) at 6 standard depths (0, 10, 30, 60, 100 and 200 cm) at 250 m resolution (Version v02), Zenodo (2018), 10.5281/zenodo.1476851.
[35] MTC, Ministerio de Transportes y Comunicaciones, 2016, Información Espacial SINAC, 2016 (D.S. N° 011-2016-MTC), https://portal.mtc.gob.pe/transportes/caminos/normas_carreteras/informacion_espacial.html. (Accessed 24 November 2019).
[36] MINAM, Ministerio del Ambiente, 2019, Sistema Nacional de Información Ambiental, 2019, https://sinia.minam.gob.pe/content/capas-geograficas. (Accessed 24 November 2019).
[37] HDX, Hidrografía del Perú. Información hidrográfica del Perú. The humanitarian data exchange, HDX. https://data.humdata.org/dataset/hidrografia-de-peru, 2015. (Accessed 24 November 2019).
[38] A. Zheng, A. Casari, Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists, O'Reilly Media, Inc.", 2018.
[39] D. Pelleg, A.W. Moore, X-means: extending k-means with efficient estimation of the number of clusters, in: Icml 1, 2000, pp. 727–734.

[40] A.K. Jain, Data clustering: 50 years beyond K-means, Pattern Recogn. Lett. 31 (8) (2010) 651–666.
[41] X. Wu, C. Cheng, R. Zurita-Milla, C. Song, An overview of clustering methods for geo-referenced time series: from one-way clustering to co-and tri-clustering, Int. J. Geogr. Inf. Sci. 34 (9) (2020) 1822–1848.
[42] H.H. Bock, Clustering methods: a history of k-means algorithms, Sel. Contrib. Data Anal. Classif. (2007) 161–172.
[43] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[44] G. Louppe, Understanding Random Forests: from Theory to Practice, 2014 arXiv preprint arXiv:1407.7502.
[45] Z. Khan, A. Gul, A. Perperoglou, M. Miftahuddin, O. Mahmoud, W. Adler, B. Lausen, Ensemble of optimal trees, random forest and random projection ensemble classification, Adv. Data Anal. Classif. 14 (1) (2020) 97–116.
[46] V. Bax, W. Francesconi, Environmental predictors of forest change: an analysis of natural predisposition to deforestation in the tropical Andes region, Peru, Appl. Geogr. 91 (2018) 99–110.
[47] S. Talukdar, K.U. Eibek, S. Akhter, S.K. Ziaul, A.R.M.T. Islam, J. Mallick, Modeling fragmentation probability of land-use and land-cover using the bagging, random forest and random subspace in the Teesta River Basin, Bangladesh, Ecol. Indicat. 126 (2021), 107612.
[48] A. Affandi, S. Sumpeno, Clustering spatial temporal distribution of fishing vessel based LON VMS data using K-means, 2020, November, in: 2020 3rd International Conference on Information and Communications Technology (ICOIACT), IEEE, 2020, pp. 1–6.
[49] Q. Liu, X. Zheng, H.E. Stanley, F. Xiao, W. Liu, A spatio-temporal co-clustering framework for discovering mobility patterns: a study of manhattan taxi data, IEEE Access 9 (2021) 34338–34351.
[50] P. Olofsson, G.M. Foody, M. Herold, S.V. Stehman, C.E. Woodcock, M.A. Wulder, Good practices for estimating area and assessing accuracy of land change, Remote Sens. Environ. 148 (2014) 42–57.
[51] M. Finer, N. Mamani, Nuevo Foco de Minería Ilegal en la Amazonía Peruana: Río Pariamanu (Madre de Dios), 137, MAAP, 2021.
[52] Mongabay, *Madre de DIOS: Nuevo FOCO de minería Ilegal amenaza a Indígenas Del pariamanu*. Noticias ambientales, November 30th 2020, https://es.mongabay.com/2020/08/madre de-dios-mineria-ilegal-boca-pariamanu-indigenas-peru, 2020.
[53] C.A. Nobre, G. Sampaio, L.S. Borma, J.C. Castilla-Rubio, J.S. Silva, M. Cardoso, Land use and climate change risks in the Amazon and the need of a novel sustainable development paradigm, Proc. Natl. Acad. Sci. USA 113 (39) (2016) 10759–10768.
[54] M.J. Weisse, L.C. Naughton-Treves, Conservation beyond park boundaries: the impact of buffer zones on deforestation and mining concessions in the Peruvian Amazon, Environ. Manag. 58 (2) (2016) 297–311.
[55] P. Ploton, F. Mortier, M. Réjou-Méchain, N. Barbier, N. Picard, V. Rossi, et al., Spatial validation reveals poor predictive performance of large-scale ecological mapping models, Nat. Commun. 11 (1) (2020) 1–11.
[56] G.R. Gallice, G. Larrea-Gallegos, I. Vázquez-Rowe, The threat of road expansion in the Peruvian Amazon, Oryx 53 (2) (2019) 284–292.
[57] T. Vilela, A.M. Harb, A. Bruner, V.L. da Silva Arruda, V. Ribeiro, A.A.C. Alencar, R. Botero, A better Amazon road network for people and the environment, Proc. Natl. Acad. Sci. USA 117 (13) (2020) 7095–7102.
[58] K.B. Curie, K. Mertens, L. Vranken, Tenure regimes and remoteness: when does forest income reduce poverty and inequality? A case study from the Peruvian Amazon, For. Pol. Econ. 128 (2021), 102478.
[59] O. Csillik, G.P. Asner, Aboveground carbon emissions from gold mining in the Peruvian Amazon, Environ. Res. Lett. 15 (1) (2020), 014006.
[60] K.A. Roach, N.F. Jacobsen, C.V. Fiorello, A. Stronza, K.O. Winemiller, Gold mining and mercury bioaccumulation in a floodplain lake and main channel of the Tambopata River, Peru, J. Environ. Protect. 4 (2013), 27180.