International Journal of Geo-Information

MDPI

Article

Urban–Rural Gradients Predict Educational Gaps: Evidence from a Machine Learning Approach Involving Academic Performance and Impervious Surfaces in Ecuador

Fabián Santos-García [1,*], Karina Delgado Valdivieso [2], Andreas Rienow [3] and Joaquín Gairín [4]

1 Research Center for the Territory and Sustainable Habitat (CITEHS), Technological University Indoamerica, Machala y Sabanilla, Quito 170301, Ecuador
2 Center for Research in Human Sciences and Education (CICHE), Technological University Indoamerica, Machala y Sabanilla, Quito 170301, Ecuador; karinadelgado@uti.edu.ec
3 Institute of Geography, Ruhr University Bochum, Universitätsstraße 150, 44780 Bochum, Germany; andreas.rienow@rub.de
4 Center for Research and Studies for Organizational Development (CRiEDO), Universitat Autònoma de Barcelona, 08193 Bellaterra, Spain; joaquin.gairin@uab.cat
* Correspondence: ernestosantos@uti.edu.ec

Abstract: Academic performance (AP) is explained by a multitude of factors, principally by those related to socioeconomic, cultural, and educational environments. However, AP is less understood from a spatial perspective. The aim of this study was to investigate a methodology using a machine learning approach to determine which answers from a questionnaire-based survey were relevant for explaining the high AP of secondary school students across urban–rural gradients in Ecuador. We used high school locations to construct individual datasets and stratify them according to the AP scores. Using the Boruta algorithm and backward elimination, we identified the best predictors, classified them using random forest, and mapped the AP classification probabilities. We summarized these results as frequent answers observed for each natural region in Ecuador and used their probability outputs to formulate hypotheses with respect to the urban–rural gradient derived from annual maps of impervious surfaces. Our approach resulted in a cartographic analysis of AP probabilities with overall accuracies around 0.83–0.84% and Kappa values of 0.65–0.67%. High AP was primarily related to answers regarding the academic environment and cognitive skills. These identified answers varied depending on the region, which allowed for different interpretations of the driving factors of AP in Ecuador. A rural-to-urban transition ranging 8–17 years was found to be the timespan correlated with achievement of high AP.

Keywords: academic performance; impervious surfaces; urban-rural; Ecuador

Citation: Santos-García, F.; Valdivieso, K.D.; Rienow, A.; Gairín, J. Urban–Rural Gradients Predict Educational Gaps: Evidence from a Machine Learning Approach Involving Academic Performance and Impervious Surfaces in Ecuador. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 830. https://doi.org/10.3390/ijgi10120830

Academic Editor: Wolfgang Kainz

Received: 8 November 2021
Accepted: 3 December 2021
Published: 10 December 2021

1. Introduction

Ensuring inclusive and equitable education is one of the sustainable development goals (SDGs) for promoting higher standards of living through economic, social, and environmental progress [1]. The achievement of primary or secondary school diplomas provides a benchmark for personal development and promotes high future job performance [2]. In this context, evaluating academic performance (AP) is a method of scoring the skills and knowledge accumulated by students throughout their years of study [3]. As AP contributes to the evaluation of students' performance prior to obtaining jobs or engaging in higher studies, it is associated with unequal territorial development and systematic division. This is because access to quality education is often unequal in developing countries [4–6]. These observations have motivated the spatialization of education theory [7,8], in which spatial exclusion from social and economic systems has been described as a key feature explaining AP scores. This is because capital accumulation at educational centers [9] or in households [10] is more evident when its spatial context is observed.

Although this approach can provide valuable insights for better understanding AP scores, it has been more frequently used on a national scale in order to rank AP [11] or explain educational inequities among regions based on their differences [12]. AP scores have been less frequently analyzed on more detailed scales, such as for an educational center and its vicinity [13–15]. This may be because the spatial analysis of AP scores requires specific tools for operating related spatial data and for showing how an educational system can be configured under unequal conditions.

Although little research has considered the spatial dimensions that explain AP scores, such research is no less important than that on related aspects. For instance, socioeconomic status and shortcomings such as a lack of basic necessities [16], poor diet quality [17], and low parental education [18] are examples of how the economic conditions of individuals can negatively affect AP results. Moreover, AP scores have also been explained by research in other fields, which have found cultural, social, and cognitive aspects to be more important predictors [19]. Factors such as language dominance [20], family conditions [21], and social media use [22] are examples of cultural backgrounds explaining effects on AP scores. Despite the scope and methodological differences of these studies, the AP scores they describe remain a complex indicator that interacts with a multitude of factors with complementary effects, though not strictly in a linear fashion. Moreover, if we consider the non-stationary behavior of space [23], the predictive power of these factors could be insignificant for some sites in a study area but strong for others [24]. These discontinuities make the examination of AP scores from a single discipline insufficient. Moreover, the results cannot necessarily be extrapolated to other study areas unless a similarity exists. Therefore, the task of untangling AP score factors could be better addressed through the adoption of an interdisciplinary approach and the use of state-of-the-art geostatistical techniques to distill the common data sources used in educational research (e.g., surveys, censuses, and educational infrastructure). The design of such a data-driven methodology requires critical thinking about how to utilize the best big data-based datasets [25] before they are exhaustively integrated and analyzed.

Advances in social data collection and processing often result in high-dimensional datasets composed of hundreds of thousands of features (f) in each observation (o), where $f > o$ is a common condition. These high-dimensional datasets contain more noise than signals, slow down training algorithms, and require intensive computations to extract meaningful results [26]. A rule of thumb when using these datasets for modeling is to select or sample variables according to the area of knowledge. Nevertheless, such a procedure is not possible or desired in some contexts (see [27], as cited in [28]), causing techniques such as dimensionality reduction to be required. In this respect, principal component analysis and factor analysis are likely the most popular unsupervised learning techniques, despite their inability to adequately handle complex nonlinear data [29] and the fact that they cannot be easily described [30]. Although other alternatives exist (e.g., self-organizing maps, autoencoders, and k-means clustering), feature selection analysis is the supervised learning technique that is the most interesting for the purposes of this research. In this type of analysis, an objective function (e.g., a supervised classification) is defined, and the model accuracy is used to guide the selection of the most compact and informative set of features [31]. Feature selection analysis is recognized as an essential task in data mining [32], and the latest research has focused on using high-dimensional databases, with management, processing, and model interpretability acting as constraints. The first two constraints can be treated as scaling hardware and software resources, whereas an interesting strategy for the third constraint is the wrapper method. This method combines a learning algorithm and a search strategy to guide the selection of the optimal features that maximize model accuracy [33]. An advantage of this method is that popular and powerful learning algorithms such as random forest (RF) and its variable importance measurement [34] can be optimized to identify more confidence-relevant features in a predictive model. Moreover, it can provide extensions for solving the drawbacks of certain algorithms, such as the lack of statistical significance for RF variable importance. In this respect, the Boruta approach [35]

is an interesting feature selection analysis algorithm, as it works around RF to improve it, adding extra randomness to extend its internal variable importance estimates and to strengthen them statistically. This is crucial when high-dimensional datasets are used because the features identified as important may be the results of random fluctuations [36]. Therefore, the application of the Boruta approach can provide an opportunity to analyze AP score factors, especially if survey data from students are used to identify relevant answers that improve AP in a predictive model. Moreover, the processing and collection of such models at the educational-center level allows the model outputs to be examined cartographically. This possibility is interesting for observing the spatial variability of the locally relevant predictors of AP scores, as well as for determining the probabilities of high or low AP. Moreover, these probabilities can be used to hypothesize about AP scores and their relationships with physical capital accumulation [37]. This is made possible by using novel remote sensing products such as annual maps of artificial impervious areas [38]. These data are helpful for differentiating old, urbanized areas from new ones, where the latter are characterized as providing miserable housing and living conditions, especially in Latin America and other developing countries [39].

Considering the aforementioned information, Ecuador is a country of particular interest for this investigation. Considered as one of the Latin American countries that emerged rapidly following the 2008 global economic crisis, its contradictory relationship between natural resource exports and its state-led "knowledge"-based development [40] makes it an interesting case study. Moreover, its national education evaluation instrument, *Ser Bachiller* or Be Bachelor (BCH) [41], is an unprecedented data source that can be qualified as a high-dimensional dataset. It includes a longitudinal survey of microdata with approximately 300 questions, including data on secondary school students' AP scores and their socioeconomic, cultural, academic, and cognitive features since 2014. In addition, it links students' data to their high school locations, making it possible to analyze their relative location with other spatial data sources. In the present study, we first discuss the preprocessing of the BCH instrument and describe the ancillary spatial data. Second, we analyze the concept behind the Boruta approach and its implementation. Third, we explore models results, and build a set of hypotheses to measure correlation between high AP, and urban areas occupation based in the annual artificial impervious area dynamics. To guide this research, we provide answers to our main research questions, which are as follows:

- Which BCH survey answers best predict AP scores, and where are their highest probabilities?
- Are AP scores significantly higher in old urban areas than in new urban areas?

2. Materials and Methods

2.1. The BCH Instrument and Its Preprocessing

The data analyzed in this study were obtained from the BCH instrument and were downloaded freely from the National Institute for Educational Evaluation (INEVAL) [41], which consists of three databases:

- The micro (MCRO) database, which collects students' BCH test scores for four categories: mathematics, language and literature, natural sciences, and social studies, as well as their overall score.
- The associated factors (AFAC) database, which constitutes a survey of 311 questions conducted with students, parents, teachers, and directors of high schools. It describes the socioeconomic and cultural conditions of students, as well as their academic environment and vocational and cognitive attitudes.
- The high school (HSCH) dataset, which comprises a set of 3284 spatial points with associated data about the administrative provinces and natural regions.

As the BCH instrument has been issued every year since 2014, four academic periods were available at the time of this study. However, we selected the period 2016–2017 as the subject of this study, as the INEVAL technical team informed us of data quality and

completeness issues in the other BCH instrument periods. Using these databases, we applied the following preprocessing steps:

1. Transliterate Latin words and remove special characters in the database column names (e.g., "Régimen" became "regimen").
2. Rename the columns of unique identifiers in all databases.
3. Remove irrelevant features in the databases (e.g., high school names, duplicated questions, and categorical scores).
4. Binarize data features by categorizing all answers to the survey questions into 0 (absence) and 1 (presence).
5. Filter the student population to the last year of secondary studies to avoid students older than the 15 to 18-year-old age range.

After following these preprocessing steps and combining the MCRO and AFAC databases, we had 248,252 records (or students) and 1171 data features (or answers to BCH questions). These records were distributed across 3284 spatial points (or high schools) after being combined with the HSCH spatial database. As a result, each high school corresponded to an average of 75.59 $\pm$ 102.51 students.

2.2. Academic Performance and Stratification

After preprocessing the databases, the dependent variable considered in the model could be identified; an overall statistical summary of the dependent variable is presented in Table 1. This information was found in the MCRO database and includes the mean scores for the four cognitive tests: mathematics, language and literature, natural sciences, and social studies. For the remainder of this study, we refer to this score as a measure of AP.

Table 1. The dependent variable and its overall statistical summary.

Name	Description	Min.	Mean	SD	Max.
AP score	Overall test score (0–10) obtained by students for Mathematics, Language and Literature, Natural Sciences, and Social Studies tests	4	7.53	0.81	10

To facilitate modeling, we stratified the AP scores into two ranges, considering that values above seven relate to student enrollment in public higher education in Ecuador. To combine the AP score data with the HSCH spatial database, we averaged students' AP scores according to high school and classified them into "High AP" for schools with average scores greater than or equal to seven and "Low AP" for high schools with average scores below seven. To obtain summaries in the spatial context, as well as to discuss and compare our results with those of other investigations targeting Ecuador at the national level, we labeled high schools and their AP scores according to the three regions occurring in our study area (see Section 2.4) to obtain smaller subsets, as shown in Table 2.

Table 2. Statistical summaries of the average academic performance (AP) scores of students according to their high school, the AP classification, and the regions occurring in our study area.

Region	High AP ($\geq$7)		Low AP (<7)	
	Averag AP Scores (Mean $\pm$ SD [1])	High Schools (Count)	Averag AP Scores (Mean $\pm$ SD)	High Schools (Count)
Amazon	7.47 $\pm$ 0.39	142	6.64 $\pm$ 0.28	178
Andes	7.75 $\pm$ 0.47	1417	6.85 $\pm$ 0.18	224
Coast	7.66 $\pm$ 0.44	858	6.80 $\pm$ 0.19	465

[1] SD: standard deviation.

2.3. Feature Space and Indexing

Since we obtained 1171 data features after the binarization of the BCH questions and answers, we designed a hierarchical conceptual structure based on the literature and expert knowledge to index and facilitate management of the data features. We differentiated

three groups (See Table 3) and 21 "theme groups" of related questions to facilitate the processing and interpretation of the results. A similar number of theme groups was divided among the three groups to obtain an almost equal number of questions and answers for each one. This meant that each group included a pool of data features comprising 103 ± 9 (SD) questions with a total of 391 ± 28 possible answers. A summary of these groups, themes, and number of questions and answers is presented in Table 3.

Table 3. Independent variable groups.

Group (Alias)	Theme	Prefix	Description	Data Features (Count)		Reference
				Questions	Answers	
Socioeconomic and cultural (SC) contexts	Cultural activities	CAC	Cultural, sport, and recreational activities undertaken by the student and his/her family.	14	56	[42–44]
	Digital technologies for culture and entertainment	CDC	Technological equipment in the home for cultural and entertainment activities.	15	71	
	Identity and language	CIL	The student's cultural and linguistic identity.	5	24	
	Migration status	CMG	Mobility of the student and their family members.	11	31	
	Educational and work status of parents	SEW	Academic level achieved by the parents and their work activities.	9	87	
	Housing features, goods, and services	SGS	Availability of goods and services in the home, as well as the construction materials used in the dwelling.	20	76	
	Household income	SHI	Social benefits and economic support of the family, as well as the student's work situation.	8	31	
	Household structure	SHS	Family coexistence according to degree of consanguinity.	13	33	
			Total	95	409	
Academic environment (AE)	Digital technologies for education	ADE	Technological equipment in the home for educational activities.	11	42	[45,46]
	Household climate	AHC	The student's attitude towards their home and parental supervision.	10	37	
	Security (classroom and home)	ASC	Insecurity due to robberies or violence at the student's home or high school.	6	24	
	Classroom climate: student vs. high school	ASH	Student's attitude towards their high school's management.	33	99	
	Classroom climate: student vs. students	ASS	Student's attitude towards their peers' coexistence.	22	82	
	Classroom climate: student vs. teachers	AST	Student's attitude toward their teacher's management of the classroom.	32	122	
			Total	114	406	
Cognitive skills (CS)	BCH test preparation	VEP	Type and budget invested in BCH test preparation.	2	10	[47]
	Grade repetition	VGR	Primary or secondary grade repetition.	3	12	
	Hard science skills	VHS	Mathematics, physics, chemistry, and biology skills.	20	44	
	Student's personality and learning attitude	VPL	The student's attitude towards learning and his/her main personality traits.	25	102	
	Availability and promotion of reading	VPR	Student's attitude towards reading and its promotion in the home.	20	72	
	Soft science skills	VSS	Language, literature, history, education for citizenship, and philosophy skills.	20	44	
	Student's vocation	VSV	Student's career preference.	12	75	
			Total	102	359	

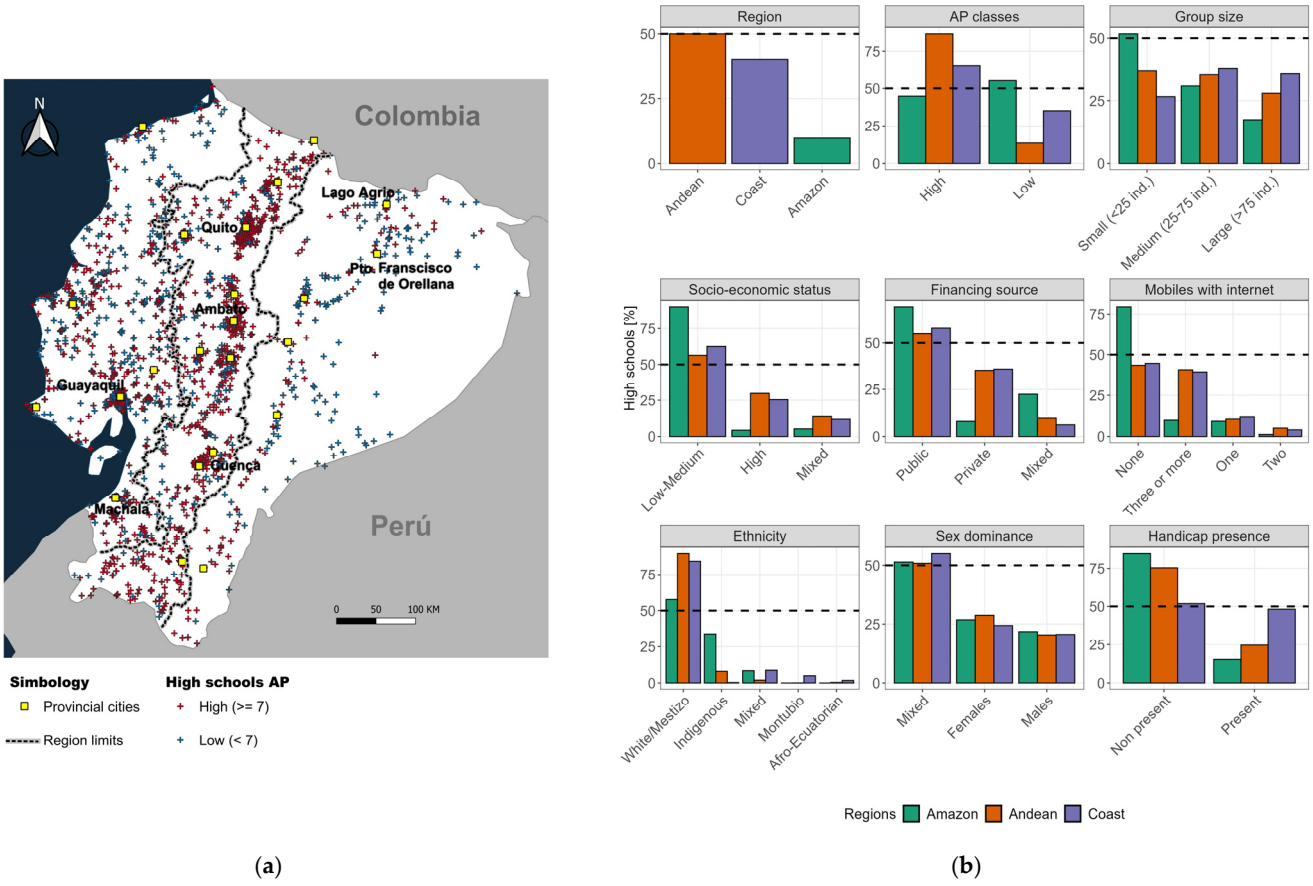
The theme groups were organized in the feature space operated under the Boruta algorithm. Each group represented an area of knowledge with respect to the situations of the students, specifically:

- Socioeconomic and cultural (SC) contexts, which refer to students' household resources, their parents' work and educational status, and their cultural practices and migration status.
- Academic environment (AE), which refers to high school and household climates, feelings of security, and the availability of digital technologies for education.
- Cognitive skills (CS), which refer to students' perceptions of their knowledge and skills, their grade repetition, and BCH test preparation.

The Boruta was indexed according to the prefixes of the report results.

2.4. Study Area and High School Descriptions

Our study was conducted in continental Ecuador, which represents an area of 247,414.5 km^2, or 96.48% of the country's total landmass. It is commonly divided into three natural regions, namely, the coast, the Andes, and the Amazon, which are located in the western, central, and eastern areas of the country, respectively (Figure 1a). According to the National Institute of Statistics and Census [48], the coast is the most densely populated region, with 107.74 inhabitants/km^2. It comprises territories below 1300 m a.s.l. and includes Guayaquil and Machala as the most populous cities. The Andes is the next-most densely populated region, with 101.07 inhabitants/km^2. It comprises territories above 1300 m a.s.l., including the country's highest point (6268 m a.s.l.), and the largest cities in the region are Quito (the capital) and Cuenca. The Amazon is the least densely populated region, with 6.34 inhabitants/km^2. It includes territories below 1300 m a.s.l., and its largest cities are Lago Agrio and Puerto Franscisco de Orellana. Of these areas, the Andean region is the most economically developed, followed by the coastal and Amazonian regions [49]. This was also observed by summarizing the high school data used in this study (Figure 1b), as more than half of all Ecuadorian high schools were located in the Andean and coastal regions. This also corresponds with the high AP and large student group sizes (>75 individuals) observed in these regions. The opposite was observed for the Amazon region, as it had lower AP and smaller student group sizes (<25 individuals). Low and medium socioeconomic statuses were represented in the high schools of all regions, but more so in the Amazon. These socioeconomic trends correspond with the public financing sources of high schools and other indicators of purchasing power (e.g., average number of mobile phones with internet in students' households). Furthermore, white/mestizo ethnicity and mixed-sex dominance were characteristic of most high schools, except in the Amazon, where indigenous people were more predominant. High schools having at least one student with a disability were more common in the coastal region than in the other regions.

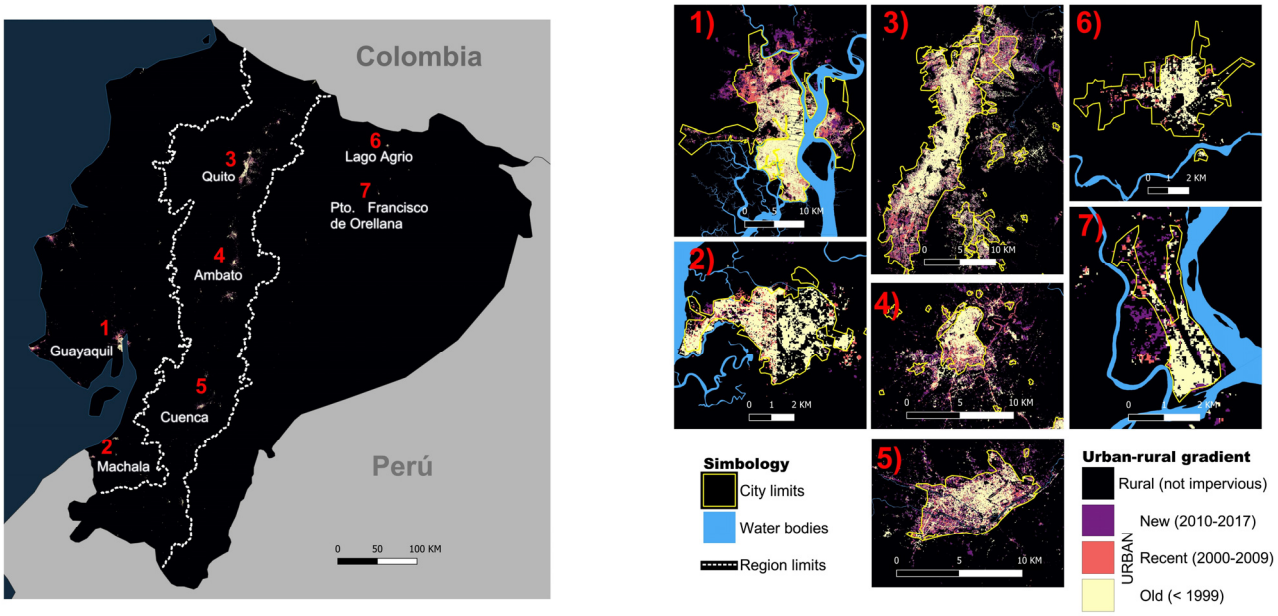


(**a**) (**b**)

Figure 1. (**a**) High schools, regions, and largest cities in the study area. (**b**) Students' answers to selected survey questions, presented as percentages of total high schools by region.

2.5. Urban–Rural Gradient Derivation

To derive a proxy for physical capital accumulation, we acquired a novel remote sensing product called "Annual Maps of Global Artificial Impervious Areas" [38]. Impervious surfaces are surfaces made of any material of a natural or anthropogenic source that prevent infiltration of water into the soil and whose growth is mostly related to human construction (e.g., roofs, paved surfaces, and hardened grounds) [50]. An artificial impervious area dataset at the global scale was derived from the Landsat multi-decadal archive and other ancillary datasets (i.e., nightlight images and Sentinel-1 data); this dataset was mapped annually from 1985 to 2018 with a 30 mts resolution. We extracted our study area from the Google Earth Engine [51] and maintained its spatial resolution. We reclassified our study area into four classes: (1) rural, to refer to nonimpervious areas; (2) old urban, to refer to areas that were impervious before 1999; (3) recent urban, to refer to areas that became impervious between 2000 and 2009; and (4) new urban, to refer to areas that became impervious between 2010 and 2017 (Figure 2). These particular time periods were considered because they correspond with important economic and political events in Ecuador, such as the economic crisis and dollarization [52] as well as Correa reformism and its later debacle [53]. We used this dataset to verify our second research question, as described in Sections 2.9 and 3.5.



(**a**) (**b**)

Figure 2. (**a**) Urban–rural gradient map for continental Ecuador, modified following Gong et al. [38]. Recent and new urban areas are shown in purple, and old areas are shown in yellow. (**b**) Map panels showing the largest cities. The first column shows Guayaquil and Machala in the coastal region (**1**,**2**); the second column shows Quito, Ambato, and Cuenca in the Andes (**3**–**5**); and the third column shows Lago Agrio and Puerto Franscisco de Orellana in the Amazon (**6**,**7**).

2.6. Feature Selection and Prediction Algorithms

We used the Boruta approach to identify the best predictors and RF to predict AP. Therefore, we first introduced preliminary definitions of these algorithms to enable the understanding of our implementation. Interested readers can refer to Biau and Scornet, 2016 [54] and Kursa and Rudnicki, 2010 [35] for detailed descriptions.

2.6.1. Random Forest (RF)

RF is an ensemble algorithm with an unexcelled accuracy for solving classification and regression problems [34]. It is immune to correlated features and can handle large datasets without making assumptions about their structure. During the training phase,

RF applies a bagging approach to construct a set of random samples with replacement from the training dataset, reserving one-third of them for error estimation—these samples are called "out-of-bag" (OBB) samples. The construction of these random samples is controlled by two main parameters: *mtry* (or split rule) and *ntree* (or number of trees). In most cases, these parameters do not require significant tuning, and default values are used (see [55]). With these bagged samples, RF constructs multiple decision trees that are used, for prediction and for calculating the mode of the resulting classes (classification) or their mean prediction (regression). To compute the model performance for classification problems, RF uses the OBB samples to derive a prediction and probability of occurrence. RF also calculates the variable importance (VI), which is a measure of the predictive power of the *j*-th feature in a predictive model. It can be derived using different approaches, but permutation is generally preferred because it is less prone to selection bias [56]. This bias was first noticed when using decision trees; therefore, when using RF, selection bias is carried forward, and it is not recommended when different data types exist [57].

2.6.2. The Boruta Approach

To reinforce the sensitivity of RF to random fluctuations, the Boruta approach was developed [35]. This algorithm constructs artificial features (called "shadow" features) from the original ones, shuffling their values to provide an external reference and to decide the feature relevance. Then, an RF classification is run, and the VI scores are determined using both original and shadow features. The VI scores are compared by observing if the score of the original feature is higher than that obtained for the best shadow. If this is the case, the observed feature achieves a "hit," and such cases are counted until their number becomes significantly higher than what is expected at random (it uses a *p*-value cutoff of 0.01 to decide this). The features are finally classified as "confirmed"; that is, their importance is statistically significant and is not considered a product of a random process. Features that do not achieve a hit are removed from the extended feature system and are qualified as "rejected". Boruta iterates these steps using a top-down search strategy to evaluate all features, and it stops when the designated maximum number of iterations (or *maxRuns* threshold) has been reached. As some features cannot be evaluated in the process, Boruta classifies them as "tentative". Such features require additional iterations or must be solved using other external calculations. A recommended option for this is the *TentativeRoughFix* function in Boruta, which observes that the median importance of a feature is higher than the median importance of the maximal shadow attribute.

2.7. Sampling Strategy Based on Distances to High Schools

To construct spatial samples that could be analyzed by Boruta and predicted with RF, we structured each high school (*i*) as a spatial data point with linked data features from the AFAC and MCRO datasets. These samples were potentially unbalanced and do not necessarily contain enough observations to sum at least 20 students for each AP classification or conserve reliable prediction results [58]. Therefore, a balance-sampling procedure based on distance was used. To implement this procedure, we developed an algorithm in the R language [59] using the GWmodel and raster libraries [60,61] (see Step 3 in Algorithm 1) to perform the following computations:

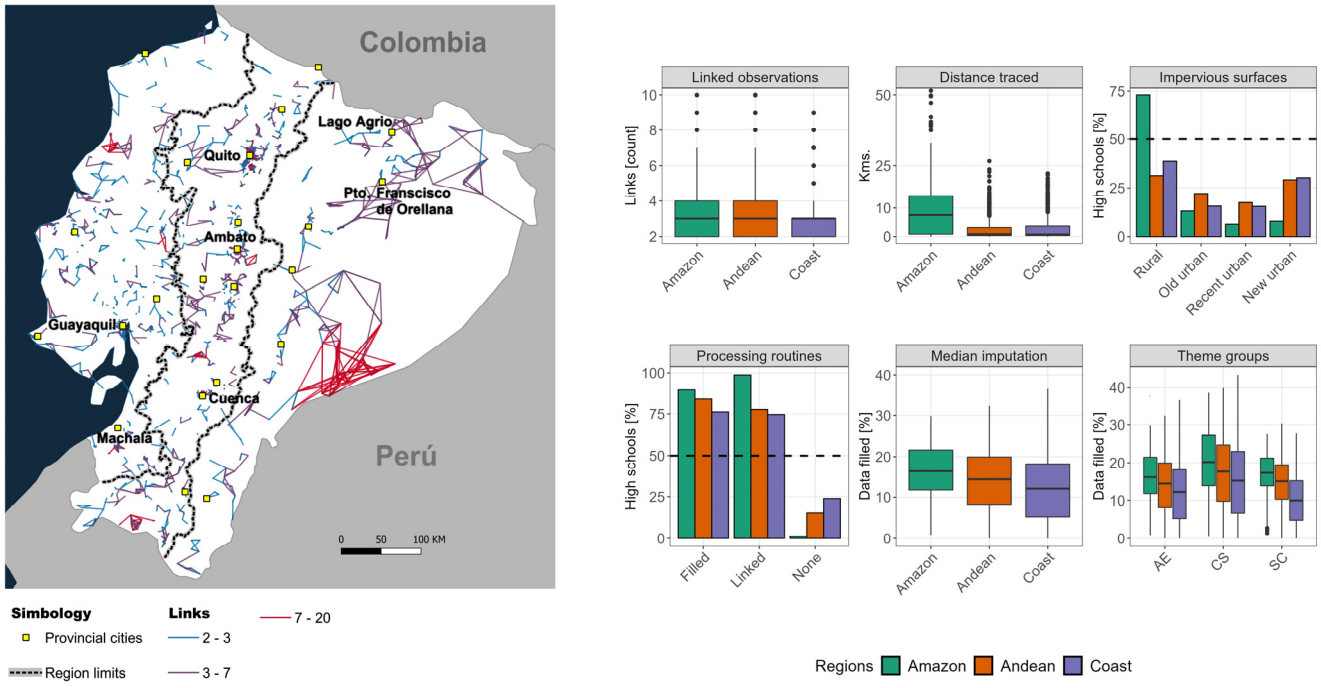
- Tabulate the AP classifications for high school *i*.
- When the AP classification of high school *i* contains >20 students, extract a random sample.
- Otherwise, invoke data to complete the sample. In this case, random samples from the nearest neighboring high schools with respect to *i* were used to complete the sample.

While this procedure uses repeated observations to complete cases, its application enables Boruta and RF to be free of the class imbalance problem [62].

We flagged these high schools and summarized their attributes based on: (1) the number of linked high schools, (2) the distance traced between high school links, (3) null data imputation, (4) urban–rural classification, and (5) frequent routines applied to the high

schools. The link counts are mapped in Figure 3a, and all of the attributes are summarized in Figure 3b. They can be described as follows:

- The Amazon is the region with the most high-schools in rural areas (>60%), and it therefore required larger distances (median: 7.5 km) for sample balancing. This was different for the Andean and coastal regions, where closer distances (median: 0.85 and 0.74 km, respectively) were needed for sample balancing. However, a more frequent association with old and new urban areas was observed for the Andean and coastal regions.
- Almost all high schools in all of the regions (~75–95%) required sample balancing or data filling (we used median imputation for this operation). Only a small percentage of high schools (~20–25%) located in the Andean and coastal regions did not require sample balancing or data filling. These imputed values represented approximately.
- ~8–20% of all the data, and the CS (cognitive skills) group and the Amazon region were the most affected.



(**a**) (**b**)

Figure 3. (**a**) High school linkages based on distance; (**b**) sample feature characterizations by region.

2.8. Implementation of Boruta and RF for the BCH Instrument

Since our distance-based sampling allowed us to systematically balance and operate the high school data, we now describe how we identified and confirmed features with Boruta before they were predicted with RF. First, consider that $i = \{1, \ldots 3312\}$ represents the high school locations associated with the AFAC data according to $j = \{SC, AE, CS\}$ theme groups. An automated approach for modeling i using a j group can be represented by the formula:

$$AP_i \sim AFAC_{ij} \quad (1)$$

where AP_i represents the AP classification (i.e., high and low) of students from high school i, and $AFAC_{ij}$ is an index to subset j features from the AFAC data. Suppose that $j =' SC'$. Then, j corresponds to the balanced sample from high school i with 40 observations (20 students each for the high and low AP classifications) and 409 answers to socioeconomic and cultural (SC)-related questions. Since our objective was to predict the AP classifications for multiple high schools using Boruta to discriminate significant answers

that optimize the RF prediction, we had to expand our processing capabilities. Therefore, we enabled parallel processing to speed up this routine, automating two operations for each high school case i. This included:

- The execution of Boruta and *tentativeRoughFix* functions using $AP_i \sim AFAC_j$ to identify the confirmed features (CF).
- Training a classification RF model with $AP_i \sim AFAC_{ij=CF}$ to derive VI, AP prediction probabilities, model-based accuracy metrics (e.g., Kappa, Breier, and R^2), and cross-validation error rate with the selected features, See [63].

To fulfill these operations, we used a loop to control their recursive execution during calculations for each high school. This allowed multiple attempts with the Boruta and RF iterations, and for all data features in $AFAC_j$ to be explored. Backward elimination of the best predictors was re-run for the modeling routines until nine iterations were completed. This number of iterations was chosen in order to ensure a reasonable computation time. On an eight-core workstation with 32 GB of RAM, the computations for each theme took ~9.5 h. This processing resulted in multiple modeling results, and it allowed the relevant features to be correctly identify with Boruta, as the processing did not find them in the first run (i.e., Kappa was not highest in the first iteration). To describe this implementation more precisely, the pseudocode is shown below.

Algorithm 1: Boruta implementation for the BCH instrument

Input: $HSCH_{i=\{1,...3312\}}$; $MCRO_{i=\{AP\}}$; $AFAC_{ij=\{SC,AE,CS\}}$;
Output: *BorutaBase* database and spatial derivatives: modeling accuracy (ACC), data summaries (SUM), AP probabilities ($PRED$), confirmed features (CF), and sample linkages (LKN).
Step 1: Start parallel computing. For high school i in the $HSCH$ spatial database, do:
Step 2: Filter students' data from $MCRO$ and $AFAC$ according to high school i and feature j. Define this subset as HS_{ij};
Step 3: Obtain a random sample of AP classifications from HS_{ij}. If required, complete unbalanced cases with data from the nearest high schools. Flag and count the number of created linkages. Fill cases with no data with the median value and calculate its proportion. Overwrite HS_{ij} and continue.
Step 4: Run Boruta considering the next steps:
4.1 Set $Boruta_{ijk}$ as an empty list and set *true* to the *Compute* parameter. Set first iteration as $k = 1$.
4.2 While *Compute*:
4.2.1 Run Boruta (default parameters: $maxRuns = 100$), collect confirmed features (CF) for HS_{ij}. If tentative features remain, apply the *tentativeRoughFix* function to solve them.
4.2.2 If there are no results, repeat step 4.2.1. Otherwise, continue;
4.2.3 Filter CF from HS_{ij} and predict the classification model $AP \sim HS_{ij=CF}$ with RF (default parameters: $mtry = \sqrt{\| CF \|}$, $ntree = 500$).
4.2.4 Extract CF importance scores and class probabilities, together with R^2, Kappa, and Breier scores from the model. Conduct 5-fold cross-validation with CF and obtain the error rate. Count $k + 1$ iterations. Store them in $Boruta_{ijk}$.
4.2.5 If $k < 10$, remove $confirmed$ features and overwrite HS_{ij}; otherwise set $Compute$ as $false$.
Step 5: Stop parallel loop processing. Collect and merge $Boruta_{ijk}$ results as *BorutaBase*. Extract data from the Boruta k iteration, which achieved the highest Kappa value, and construct spatial derivatives: ACC, SUM, $PRED$, CF, and LKN. End algorithm.

The spatial outputs of this algorithm are feature summaries (SUM), sample linkages (LKN), modeling accuracies (ACC), confirmed features (CF), and AP prediction probabilities ($PRED$). The first two are summarized as plots in Sections 2.4 and 2.7, while the rest are described in the next sections. To develop this algorithm, we used the R libraries

described in Section 2.7 as well as additional ones, which included ranger [55], random Forest [64], Boruta [35], caret [65], and parallel [66]. An interactive application was developed to show the outputs of these models and extend their results to other spatial units (see Appendix A). We also used QGIS [67], an open-source geographic information system, for cartographic analysis.

2.9. AP Probability Hot Spots and Testing Hypotheses with the Urban–Rural Gradient Map

As the *PRED* output described the probabilities of high schools achieving high or low AP scores, we mapped the output as hot spots with a kernel density estimate. This algorithm has been used to identify traffic accident hotspots and other point patterns [68,69]. It consists of a density function that operates spatial data points to create a smoothed weighted surface. This surface is calculated using a kernel function, which requires a bandwidth parameter to control the amount of smoothing, and an optional set of weights to define the importance of each spatial data point. Therefore, using the R library SpatialKDE [70], we first created a grid with a pixel size of 500 mts to be used for the calculations. This pixel size was determined to be adequate because the processing time was long, and the resolution was sufficient for highlighting hot spots without losing detail or distorting the results. This was verified by looking at the median distances achieved by high schools in the spatial samples (greater than 0.75 km; see Section 2.7). We derived the weights W_{high} for high AP probabilities (P_{high}), differentiating them from low AP probabilities (P_{low}) and removing negative values as follows:

$$W_{high} = \left\{ \left[\left(P_{high} - P_{low} \right) \leq 0 \right] \Rightarrow 0 \right\} \wedge \left\{ \left[\left(P_{high} - P_{low} \right) > 0 \right] \Rightarrow \left(P_{high} - P_{low} \right) \right\} \quad (2)$$

A similar process was adopted for W_{low}:

$$W_{low} = \left\{ \left[\left(P_{low} - P_{high} \right) \leq 0 \right] \Rightarrow 0 \right\} \wedge \left\{ \left[\left(P_{low} - P_{high} \right) > 0 \right] \Rightarrow \left(P_{low} - P_{high} \right) \right\} \quad (3)$$

These two sets of weights emphasized only the relevant high schools for each AP classification in the hot-spot calculations. Following the recommendations of Siloko et al. [71], we chose a tri-weight kernel type because the resulting density estimates were better and the hot spots were sharper than those obtained with other kernel types (e.g., uniform, quartic, Epanechnikov). We tried different bandwidths considering one to five surrounding pixels, and we decided that three surrounding pixels (or 1500 mts) were appropriate for smoothing while also considering the average sampling distances between high schools (see Section 2.7) and providing sufficient detail to identify clusters of high schools with similar AP class probabilities (see Figure A1). As a result, we obtained two hot spot maps, one for P_{high} and another for P_{low}. These maps were only for the theme group that achieved the best prediction accuracy (see Section 3.1). To test our second research question, we used the high school locations to extract values from P_{high} hot spots and $c = \{rural, old, recent, new\}$ classes from the urban–rural gradient map. We applied a Wilcoxon–Mann–Whitney test to determine if the old urban class achieved a significantly greater P_{high} hot spot value than the other c classes; therefore, we evaluated the following hypotheses:

Hypothese 1. *The P_{high} hot spot value for the c class is equal to or greater than that of the old urban class.*

Hypothese 2. *The P_{high} hot spot value for the c class is lower than that for the old urban class.*

In addition, a Wilcoxon effect size test was performed to estimate the effect size and classify the results according to the magnitudes found in Funder et al. [72]: <0.05 (tiny); 0.05–0.1 (very small); 0.1–0.2 (small); 0.2–0.3 (medium); 0.3–0.4 (large); and >0.4 (very large). In Table 4, we show the number of high schools according to the urban–rural gradient map classes. These counts indicate the sample sizes used for conducting the Wilcoxon–Mann–Whitney test.

Table 4. High schools by region and urban–rural gradient map classes.

Region	Urban–Rural Gradient Map Classes (Count)			
	Rural	New Urban	Recent Urban	Old Urban
Amazon	281	-	2	36
Andes	864	123	128	526
Coast	707	86	171	359

3. Results

3.1. Model Predictive Performance

The execution of the Boruta algorithm to identify relevant predictors and RF to predict AP were evaluated before proceeding with interpretations. To accomplish this, we reviewed and plotted accuracy metrics (Figure 4). Here, the SC theme group obtained the highest overall accuracies for each region (0.84 ± 0.079, or 84% correct predictions using test samples), whereas the AE and CS theme groups exhibited slightly lower accuracies, as their results were around 0.83 ± 0.070 and 0.83 ± 0.072, respectively. The overall accuracies were therefore acceptable. The Kappa index (whose value is interpreted as follows: <0 no agreement, 0–0.2 slight, 0.21–0.4 fair, 0.41–0.6 moderate, 0.61–0.8 substantial, and >0.81 almost perfect agreement) indicated a substantial agreement with a similar ranking, with the SC group having a value of 0.67 ± 0.158 and the CS and AE groups achieving similar values (i.e., 0.65 ± 0.145 and 0.65 ± 0.140, respectively). This means that some models failed to predict the AP classification, and we count that 46.4%, 47.2%, and 49.4% of the high school cases they did not match the expected results for the SC, CS, and AE groups, respectively. Spatially, these models were less frequent in the Andean region, followed by the coastal and Amazonian regions. In all three theme groups, the forecasted probabilities indicated fair-to-adequate Brier scores (i.e., 0.14 ± 0.04, where values close to zero means a better prediction). However, some models scored >0.2, indicating that there were random guesses (12.3% of high school cases) that achieved overall accuracies of 0.42 ± 0.106. Furthermore, we observed that some high school cases required more than one Boruta run. Specifically, this was the case when finding the highest Kappa value during backward elimination for 30.97%, 28.3%, and 27.19% of all models for the AE, CS, and SC theme groups, respectively. The average number of iterations observed for this was 1.7 ± 1.6. In some special cases, Boruta could not confirm relevant features, and these represented 0.84%, 0.72%, and 0.66% of all high school cases in the AE, CS, and SC theme groups, respectively. Finally, all modeling with Boruta and RF required in average computation time of 81.84 ± 12.707 s, ranging from 34.5–118.7 s in AE, 29.4–121.6 s in CS, and 3.4–349.1 s in SC.

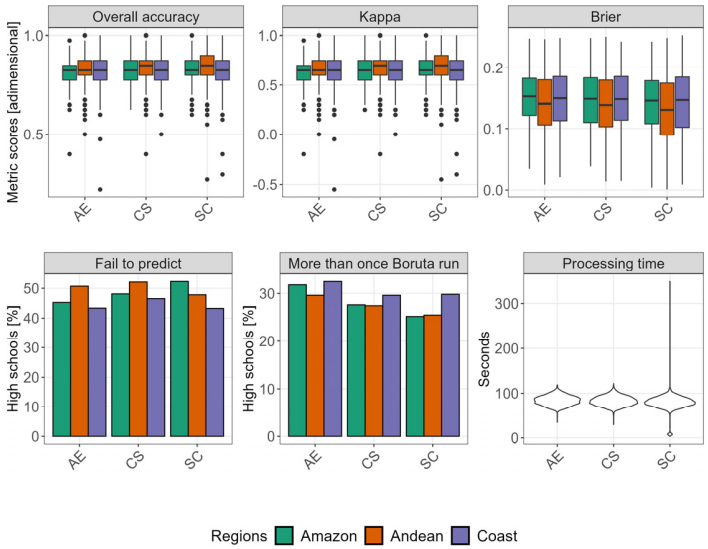
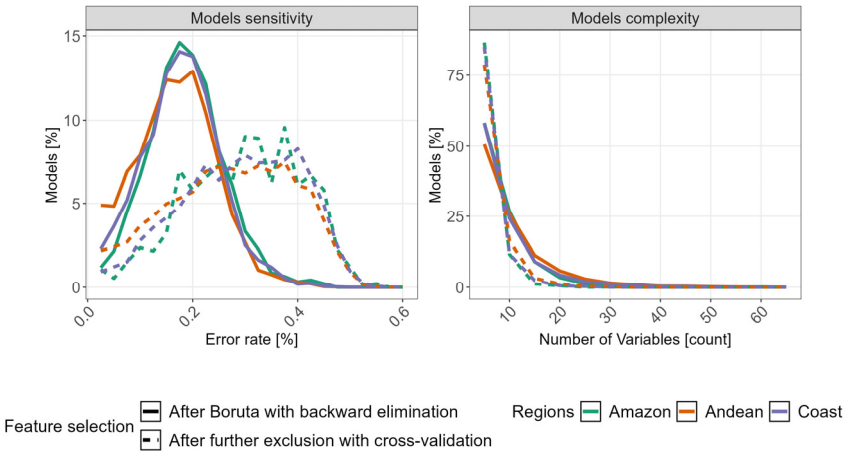


Figure 4. Accuracy metrics, model performance, and processing time summarized by region and theme group.

3.2. Model Sensitivity and Complexity Assessment

To evaluate the sensitivity and complexity of the models to random draws, we extracted the results of the five-fold cross-validation test performed using step 4.2.4 in Algorithm 1 (see Section 2.8). As this test showed the predictive performance of the models with a sequentially reduced number of predictors, we tabulated the error rate and the number of features for all produced models and groups (i.e., AE, CS, and SC) as percentages, considering: (1) models with all selected features from the Boruta algorithm; and (2) models with reduced features after further exclusion with cross-validation, but only for the iterations that obtained the minimum error rates. By comparing these results (Figure 5a), we observed that the error rate in the case of Boruta was lower (median = 0.17 $\pm$ 0.07) than that obtained via cross-validation (median = 0.30 $\pm$ 0.11). This means that the features selected by Boruta were generally relevant, and their elimination degraded the predictive performance of the model. We also plotted the number of features (Figure 5b), and we observed that the Boruta models were highly heterogeneous, ranging from two to sixty features and averaging around 7 $\pm$ 6 features. This was different for the truncated models, as they ranged from one to thirty features and averaged around 3 $\pm$ 3 features, but they did not have improved error rates, as reported above. Nevertheless, it is important to note that the upper quartile of the Boruta models was highly complex (>9 features) and more prone to errors (>0.22 error rate).



(**a**) (**b**)

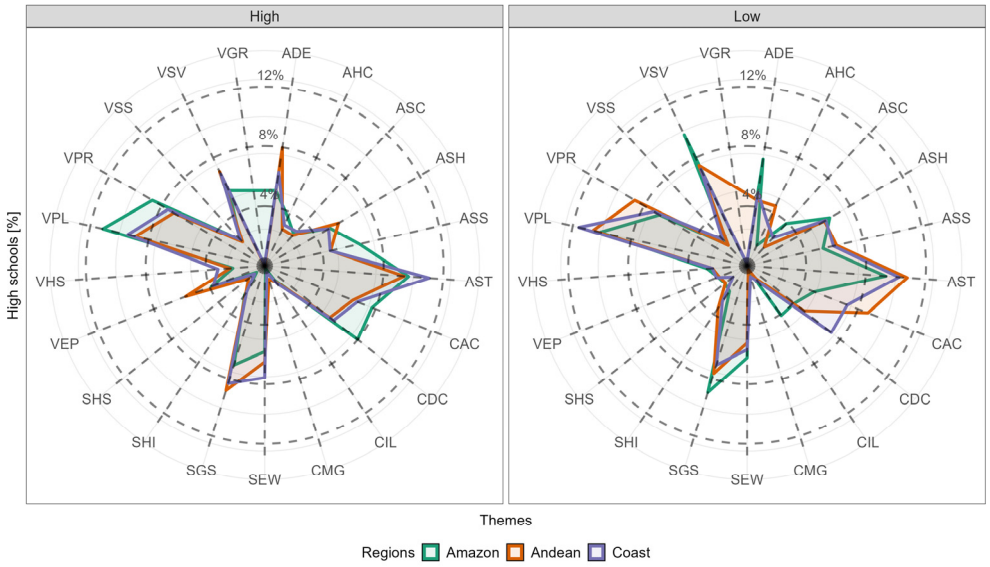
Figure 5. Frequency lines for (**a**) model sensitivity and (**b**) model complexity aggregated by region after analysis using the Boruta algorithm with backward elimination and further exclusion with cross-validation.

3.3. Frequent Best Predictor Themes

As the algorithm output *CF* (see Section 2.8) provided multiple answers, we extracted the one with the highest VI value. This result included the answers in the BCH instrument that best contributed to the AP prediction for each high school. As these results corresponded to a particular theme, we first counted the most frequent answers to highlight those that were relevant for each AP class. This procedure was required because the feature space was extensive, and this operation provided an overview. Then, we created radar plots for each AP class, as shown in Figure 6.

Themes such as VPL (students' personality and learning attitude) and AST (classroom climate: student vs. teachers) were the most frequent (i.e., $\geq$8% of high schools) across regions and AP classifications. Other relevant but less frequent themes (i.e., 6–8% of high schools) across the regions were CAC (cultural activities) for the high AP classification and ASH (classroom climate: student vs. high school) for the low AP classification. For the rest of the themes, differentiated results were observed for the AP classifications and regions. In this respect, the most frequent theme for the high AP classification was VPR (availability

and promotion of reading) in the Amazon, while SGS (housing features, goods, and services) was the most frequent theme for both the Andean and coastal regions. In the case of low AP, VSV (students´ vocation) was the most frequent theme in the Amazon, while VPR was that for the Andes and CDC (digital technologies for culture and entertainment) was that for the coast.



(**a**) (**b**)

Figure 6. Frequent themes derived from the best predictive answer observed in the models for (**a**) high academic performance (AP) and (**b**) low AP. Radar plots are organized by region.

3.4. Frequent Best Predictor Answers to Survey Questions

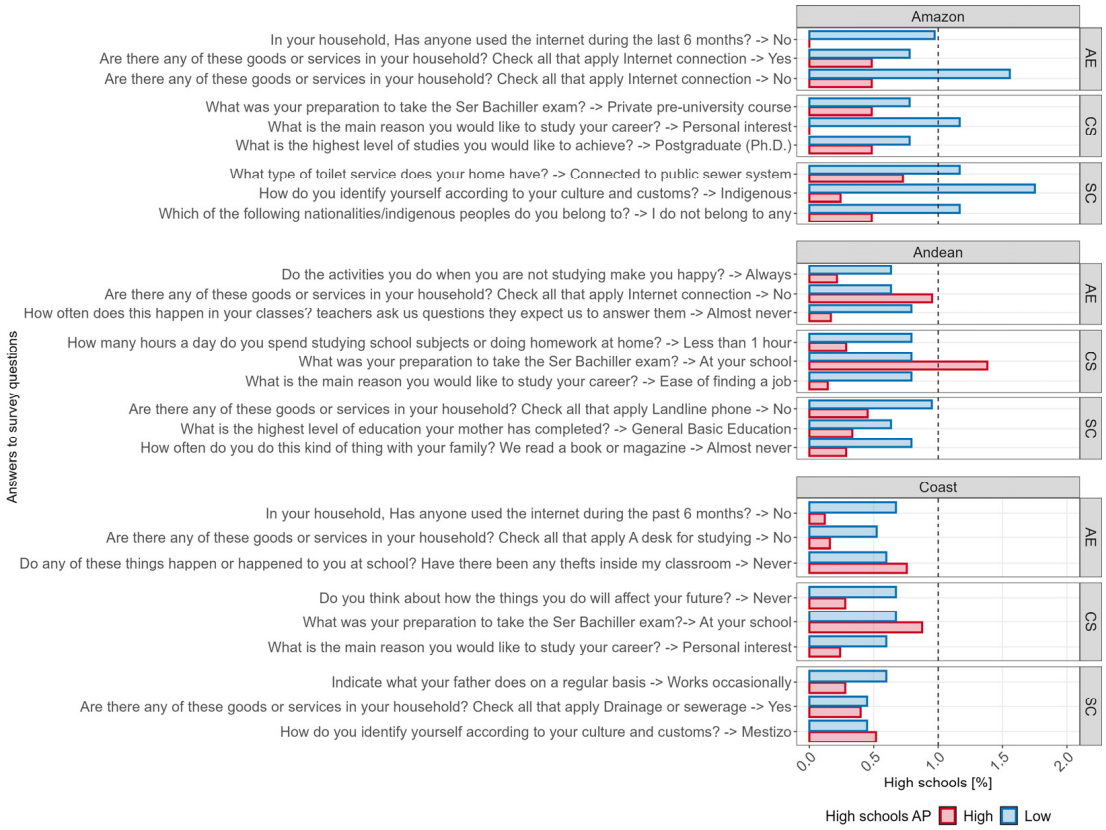
After identifying the relevant theme groups, we proceeded to analyze which answers were important for predicting the AP classification. We counted the most frequent answers for each region and sorted them into the best three for each considered group (see Table 3) to observe them in detail and compare their results. We proportionally normalized their values for later discussion on the model interpretations. We separated the plots by region for high AP (Figure 7a) and low AP (Figure 7b) to improve the readability of each question and its answer. The numeric values are included in Tables A1 and A2. In general, the results indicated proportions of approximately 1% and 2% of high school cases for each AP classification. While these figures do not represent the populations (see Table 2), it should be noted that the survey question answers were frequently chosen from a feature space of 391 $\pm$ 28 options and corresponded with clues for understanding the model feature selection. First, the results for the Amazon region and the academic environment (AE) group indicated that high AP was more frequent in a positive and heartwarming classroom climate and with peer coexistence. In contrast, low AP was related to the perception of a more hostile classroom environment and a lack of internet connectivity. With respect to the cognitive skills (CS) group, the following critical question was asked: How much does a preuniversity course cost? The results show that the cost remained under 100 USD more frequently for high AP schools than for low AP schools. However, low AP was related to students listing "personal interest" as their motivation for studying their intended vocation. Regarding the questions related to socioeconomic conditions (SC), higher AP was associated with electronic devices and internet access, but curiously also to households without garbage collection systems. Furthermore, low AP showed a more frequent association with indigenous identification, which is consistent with our description of high schools in Section 2.4. For the Andean region, the AE group included a slightly higher proportion of internet connections for students in high AP schools, but students in all high schools frequently did not have an internet connection or even a desk

for studying. For low AP schools, it was highlighted that teachers did not encourage much student participation in class; moreover, students seemed more comfortable not studying than studying. With respect to the CS group, a similar combination of survey answers was observed in the Andean region as was observed in the Amazon region, but with a more contrasting effect between high and low AP. Nevertheless, there was a higher proportion of high AP when private preuniversity course enrollment was confirmed, while low AP indicated low interest in homework and career studies beyond the desire to find a job. Regarding the SC group, it was indicated that populations with high AP were associated with higher education of mothers (e.g., PhD), domestic activities, and households with a public sewage system. In contrast, low AP was more related to mothers having only a basic education, fewer telephone connections, and reading habits. Finally, we present frequent responses for the coastal region. Regarding the AE group, students with high AP did not use computers at their high schools. Furthermore, a lack of internet connections in students' households was related to both high and low AP, but there was a greater correlation with the latter. Furthermore, high AP was related to a slightly greater proportion of classrooms without robberies than was low AP.

Other socioeconomic indicators described in the CS group indicated that exam preparation with a preuniversity private course (<100 USD) but also preparation located at the school were more frequently associated with high AP. In the case of schools with low AP, the personal interests of students highlighted that their purpose of study was for a career, although students also reported not caring about their actions and the related consequences. The final SC group indicated how the basic general education and stable paid job status of parents, as well as the existence of a public sewage system at home, were associated with high AP. In contrast, only unstable job situations of parents seemed to be more frequent for schools with low AP.

(**a**)

Figure 7. *Cont.*

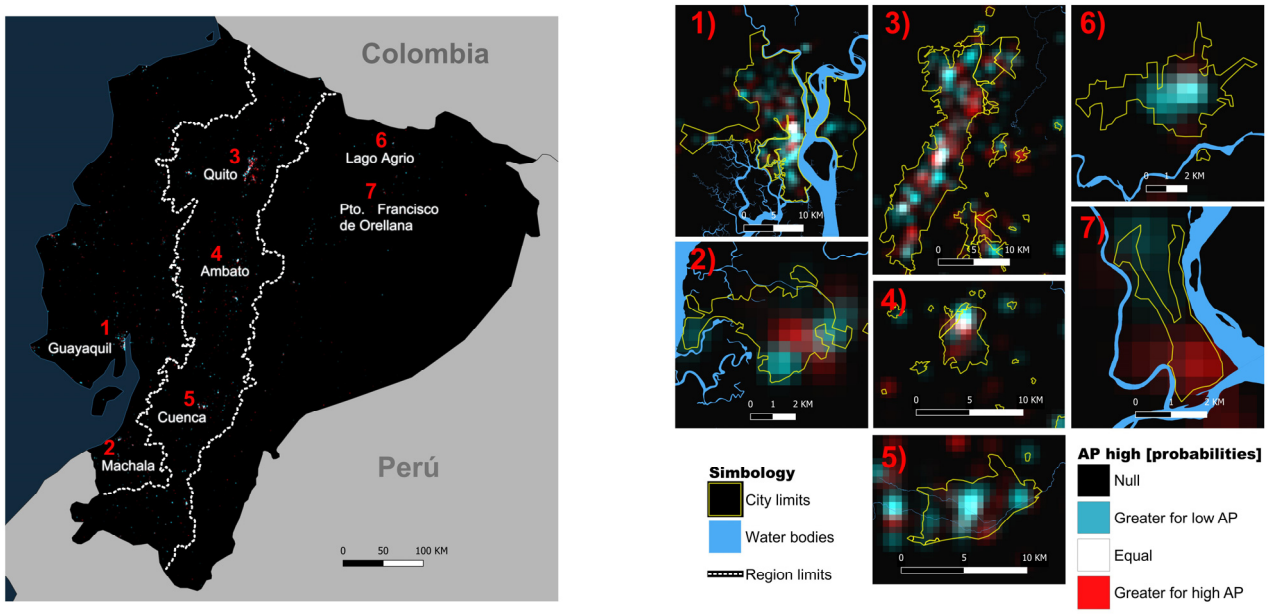


(**b**)

Figure 7. Frequent answers to survey questions derived from feature selection with the Boruta algorithm for (**a**) high and (**b**) low academic performance (AP) classifications. Answers are organized according to region and theme group.

3.5. AP Probability Hot Spots and Significance

We produced two hot spot maps and stacked and visualized them according to a color gradient where: (1) red or blue indicate that only one hot spot AP classification value is higher (red for P_{high} and blue for P_{low}), (2) white represents similar values for both hot spot AP classifications, and (3) black represents the absence of AP classification probability. These results are presented in Figure 8, which describes the results in detail for the largest cities. Satellite cities and the northern section of Quito are distinguished by P_{high} hot spots, whereas they are scattered in the other large cities such as Guayaquil, which has more frequent P_{low} hot spots. In medium and small cities (especially in the Amazon region), these hot spots are not as clear, as few high schools are located within city limits; however, some of them are still appreciable. To clarify the differences between the P_{high} hot spot values and the urban–rural gradient map classes, the Wilcoxon–Mann–Whitney test results can be considered. Figure 9 shows boxplots using P_{high} hot spot values (we normalized them to 0 and 1 to improve readability), which were calculated for each c class (i.e., rural, old, recent, and new urban classes) and region. Additional information has been added to these boxplots to show statistical significance ($sig.$), p-values (p), effect sizes (ez), sample sizes (n), and mean AP score ($\overline{AP}$).



(**a**) (**b**)

Figure 8. (**a**) Academic performance (AP) probability hot spot map for continental Ecuador. High AP probabilities are shown in red and low probabilities in blue. Equal probabilities are shown in white and their absence in black. (**b**) Maps show the largest cities. The first column shows Guayaquil and Machala in the coastal region (**1**,**2**); the second column shows Quito, Ambato, and Cuenca in the Andes (**3**–**5**); and the third column shows Lago Agrio and Puerto Franscisco de Orellana in the Amazon (**6**,**7**).

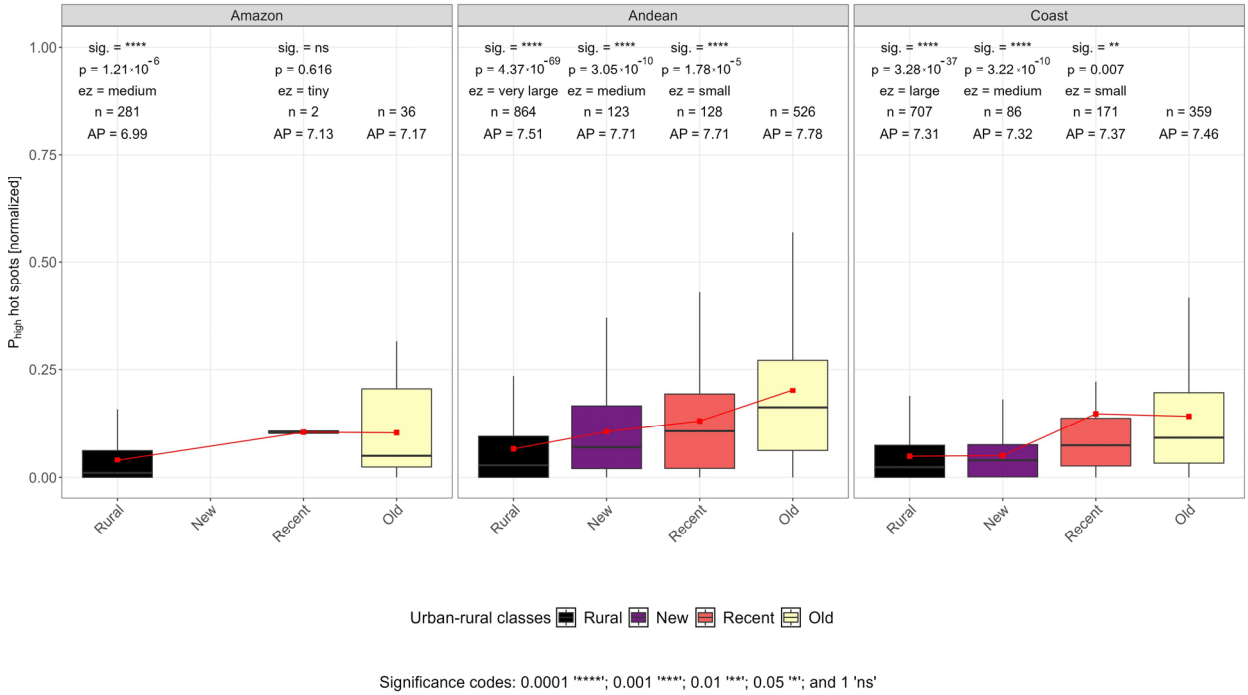


Figure 9. Boxplots of the normalized high academic performance (AP) probability (P_{high}) hot spot values for each urban–rural gradient map class and region. Text in plots refer to the results of the Wilcoxon–Mann–Whitney test, showing significance (*sig.*), *p*-values (*p*), effect sizes (*ez*), sample sizes (*n*), and mean AP scores ($\overline{AP}$). Red lines connect the mean values. Outliers are hidden to facilitate visualization.

Most urban–rural classes exhibited strong significance ($p \leq 0.0001$ or '****'); that is, the old urban class achieved a significantly greater P_{high} than the other *c* classes, and there were

different effect sizes. In this respect, the rural class was very large in the Andean region and large in the coastal region, but it was medium-sized in the Amazon. The latter suggests a less dissimilar P_{high} between the classes, which corresponds well with their differences in $\overline{AP}$ with respect to the other regions. The new urban class was characteristic of the Andean and coastal regions, and it had a medium *ez*. For this class, it is remarkable that no high schools were located in the Amazon region; thus, the test could not be accomplished. The recent urban class exhibited a small *ez* for the Andean and coastal regions, but there was a lower significance ($p = 0.007$) for the latter. This can be explained by the recent and old urban class achieving a similar probability of high AP in the coastal region, which was not as pronounced in the Andean region. Finally, the Amazon region indicated no significance. However, this result was not well supported, as sampling of only two high schools was possible and because fewer high schools seemed to be located in the Amazon region.

4. Discussion

In this study, we designed and implemented a novel approach based on machine learning to evaluate the spatial drivers of AP using a high-dimensional dataset, such as the BCH instrument. Moreover, we were able to calculate high AP probabilities, which allowed us to hypothesize about AP with a spatial proxy of capital accumulation, such as an urban–rural gradient map. Our study helps to better describe the drivers of AP scores, which have commonly been reported in other studies but have not been located or analyzed from a probabilistic approach, as done here. To discuss these findings, we examine the advances and limitations of our methodology, provide answers to our research questions, and explore the implications of this work.

4.1. Stratification and Urban–Rural Gradient Map Derivation

In this study, we did not operate a global model, but rather operated hundreds of smaller models to better explain their outputs based on spatial summaries. This required a semiautomated approach that merged and cleaned the databases prior to systematic analysis. As our study area was highly heterogeneous, this procedure allowed us to observe local patterns that a global model might not have revealed, despite adequately predicting the phenomena. The spatial variability of the associations between AP and the selected survey answers could, therefore, be explored, contributing to the determination of the topics that could enhance or degrade AP at the high school level. This provides advantages over other analytical methods as it observes the spatial neighborhood in the construction of predictive models whose outputs can be mapped and integrated with other spatial data sources such as the urban–rural gradient map. This also helped us to better understand how inequities in education are represented beyond the high school level. Furthermore, feature selection analysis focuses on the predictive power of data features rather than prediction itself; therefore, a small set of trained spatial models may produce large quantities of outputs with high-dimensional datasets. This can be overwhelming considering a national-scale spatial dimension. Therefore, summarizing the results according to spatial units is likely the best way to describe them while maintaining awareness of the aggregation level effect [73,74]. This requires a proper description and report of the spatial units used (as provided throughout this paper) to better understand the context of the models before interpreting their results. Moreover, a second criterion, namely the hierarchical conceptual structure, allowed us to further classify the data features into theme groups. This also facilitated processing, the procuring of reasonable sets of predictors for the Boruta algorithm, and interpretation. Nevertheless, our results indicated that the theme groups VPL and AST were the most frequent for low and high AP, which was consistent with the large number of answers collected for these groups across the studied regions. This type of bias responds to the BCH data structure, as some themes were more preferred than others. Therefore, future research should adopt caution when interpreting models in these scenarios, as artifacts such as these can bias the results to specific feature sets [57]. Despite this, other theme groups were identified by region and allowed us to

compare their VI and observe the probabilities for each AP classification. In this regard, the urban–rural gradient map was an interesting input that helped us to test our research question and find new applications for Earth observation satellite products. Multi-decadal monitoring capabilities allowed us to subdivide artificial impervious areas in terms of time ranges, which allowed the production of the urban–rural gradient map. While we did not differentiate vertical cities, future research should consider new scientific advances that promise these results [75]. More detailed scales should take care to avoid omission errors or differential results (e.g., the city of Machala in Figure 2 seems half-divided as a result of differential satellite data densities) with impervious maps, as time series satellite products are more prone to errors in areas with high cloud cover and reduced data collection [76].

4.2. Distance-Based Sampling and Boruta Implementation

We implemented an algorithm that utilizes spatial points as individual high-dimensional datasets to conduct a feature selection analysis. This algorithm includes the Boruta approach and optimizes the RF prediction. We performed this algorithm recursively by applying a backward elimination of the best predictors to further explore the BCH feature space. As we focused on a classification problem to obtain the best result with RF, we required a sampling strategy to balance the results. In this respect, our approach helped us to complete unequal cases by automatically sampling data from neighboring high schools. As there was little possibility of having complete samples, most of the high schools were linked. This exchange of student data among neighboring cases blurred the results, especially in areas where large distances were required (see Figure 3). Nevertheless, this operation allowed us to avoid unbalanced sampling [77] and to prepare data for feature selection and model training. Our results indicated that the selected features of the models using Boruta resulted in RF predictions with overall scores of around 0.83–0.84% and Kappa values of 0.65–0.67%. These values are interesting if we consider that these models received data from a unique source. Moreover, the removal of the best predictors in each iteration allowed us to obtain additional models for analysis. As these results were extensive, we describe here only the best-scoring results. However, this study placed more attention on correctly identifying significant features rather than on predicting AP. In this respect, Boruta enhanced RF, as feature selection was completed in one or two iterations. Nevertheless, recursive execution of *tentativeRoughFix* during our Boruta implementation may have forced the feature selection process, as it simplified the significance tests in complex cases. The multiple Boruta executions demonstrated that less than 0.84% of all high school cases were not solved, while an average of 1.7 ± 1.6 Boruta executions were required to solve those that were. Therefore, for feature spaces with approximately 400 dimensions, one or two Boruta runs accompanied by the *tentativeRoughFix* function seems sufficient for identifying the relevant features in approximately one minute. Nevertheless, conducting further Boruta runs and eliminating the results of the first and second rounds is a recommended practice to ensure that the results are truly the best. This can also be corroborated with cross-validation, as splitting model features can validate whether their presence determines prediction errors in less complex models. As large computations are required for this, users should consider parallelizing the processing routines when working with this approach, as vectorial spatial data can be easily split into parts and operated with multiple cores. Moreover, it is also important that future studies experiment with a dimensional reduction technique before applying feature selection, as this could reduce uncertainties and processing time, although it can also obscure interpretations.

4.3. Hot Spot-Based Hypothesis and Spatial Organization of AP

After calculating P_{high}, postprocessing the data as hot spots allowed us to construct a hypothesis and compare it with the urban–rural gradient map. Our results emphasized the differences using probabilities compared to original AP scores (see Figure A2) and confirmed our expected hypothesis, i.e., that the old urban class has the highest AP probability, followed by recent urban, new urban, and rural areas. These findings help to

identify where AP is magnified and to better understand the context of unequal educational opportunities. In this regard, old urban centers seem to concentrate more on educational services, and students attending these high schools have more chances to access higher education. This was also observed in recently urbanized areas, as a rural-to-urban transition ranging 8–17 years seems to be related to high AP, but the AP remains slightly lower than that in old urban areas. This spatio-temporal pattern is consistent in different regions in Ecuador, but at different levels, which may explain the localization of future professionals and other education externalities that contribute to unequal development [78]. This was the major trend observed in our results, but future research should also pay attention to the new emerging centralities typically found in recent urban areas (e.g., satellite cities around Quito, see Figure 8b), where there exist important accumulation spots of not only high AP, but also of greater economic success. However, these findings are correlations, and future research is encouraged to explore them further. While there is evidence that satellite nightlights and impervious surfaces explain urbanization and economic growth [79], this evidence does not necessarily explain academic performance [80]. Moreover, impervious surfaces (i.e., an input for our urban–rural gradient map) are discussed as scale-dependent drivers [81], and it is possible that AP may not be related to impervious surfaces when a scale other than the national scale is applied. Therefore, researchers should take caution to not describe this relationship as causation, but rather as a suspected correlation that contributes to a better understanding and theorization of AP drivers in the context of the urbanization process. Furthermore, as some high schools' spatial points may not have been correctly located, hot spots may further hide this type of error. Such artifacts may have introduced errors in the urban–rural class labeling of high schools, which in turn could have affected the hypothesis testing. Despite this, the regional approach and the scale of the study did not contradict the expected results after describing the regional characteristics (see Section 2.4). Future research focusing on this technique should consider location uncertainty and the kernel size blurring effect before estimating hot spots. Finally, the formulated hypothesis demonstrated varying sample and effect sizes according to the region, and it could therefore be better interpreted according to these contexts. In this sense, we observed that the Amazon region is an exception to the proposed hypothesis, as its high schools were barely located in new and recent urban areas. Therefore, its comparison with other regions may ignore features not included in the urban–rural gradient, which could better explain the AP scores of this region. Therefore, researchers focusing on the summarization of spatial models should ensure that spatial units and periods enclose the temporal and spatial aspects of the studied phenomena.

4.4. Best Predictive Answers and Implications

Based on the most frequent answers qualified as significant by Boruta, we identified sets of the best predictive answers in the BCH instrument. These results provided a comprehensive set of theme groups and answers by region, which helped us to observe what made the regions unique with respect to their AP scores and classification. While we focused on high AP, our implementation also provided outputs for observing low AP. Their comparison described gains and losses between the classifications. Despite theme bias, we observed various answers that improved the prediction of AP classification. More specifically, answers related to (1) accessibility to digital technologies, (2) positive attitudes between students and teachers, (3) feelings of security at high school, (4) parents' educational level, (5) payment of a preuniversity course, and (6) enjoyment of basic services and an internet connection were more related to high AP than low AP. However, some answers differed between regions, and self-identification responses were one such example. In the Amazon region, a contrasting AP between classes indicated that ethnic minorities were more prone to low AP, while more relaxed and bully-free high school environments were associated with high AP. Furthermore, high AP was associated with employed parents with higher educational levels (e.g., PhD) in the Andean region, and this was true in the coastal region to a lesser extent (i.e., basic education). All of these best survey answers contribute

to the discussion of urban–rural stratification during the historic period analyzed. The period 1999–2010 is interesting for understanding how new urban areas reflect increased AP scores after 10-year of installation. This was evidently achieved by an expansion of services in new urban areas, which also reflects the economic stabilization of the country during this period [82]. Additional interpretations of our results may be possible, but researchers may consider AP scores as a predictor worth observing in cases where high AP follows infrastructure development. However, our findings simply contribute to theories regarding this relationship rather than demonstrate causation. Segregated areas with little or very low probability are locations that require more detailed discussion, as such areas include students with diminished reading habits, lack of basic services, poor internet connections, and degraded high school environments. In this respect, it is concerning that high schools with ethnic minorities seem to be below the global average, as most are classified as having low AP. These results were principally observed in the Amazon region, where cultural and educational systems have been historically excluded, as they offer different knowledge from the dominant Hispanic-mestizo-occidental basis [83]. This evidences the core–periphery structures observed in our cartographic results, which better describe the disadvantages and poverty of ethnic minorities due to lower education and consequent labor discrimination [84,85].

5. Conclusions

We demonstrated that feature selection analysis and supervised classification can be automated using Boruta and RF to analyze complex questionnaire-based surveys at a national scale. Using the locations of high schools, we filtered and trained classification models to observe their probabilities and describe the variable importance. By presenting the model outputs as frequencies and surface representations, we showed regional AP patterns along a rural–urban gradient derived from impervious surfaces. This allowed us to theorize that older urban areas have higher AP than recent and new areas. We concluded that there is a threshold ranging 8–17 years for recently urbanized areas to exhibit a higher AP than that of newly urbanized areas and rural areas. We found core–periphery structures and patterns of regional uniqueness that suggested social class division and deficits in education. Further applications of this approach may provide new possibilities for interpreting machine learning models from a spatial reference for complex social phenomena.

Author Contributions: Conceptualization, Fabián Santos-García; methodology, Fabián Santos-García; software, Fabián Santos-García; validation, Fabián Santos-García, Karina Delgado Valdivieso and Andreas Rienow; formal analysis, Fabián Santos-García; investigation, Fabián Santos-García and Karina Delgado Valdivieso; resources Fabián Santos-García; data curation, Fabián Santos-García; writing—original draft preparation Fabián Santos-García; writing—review and editing Karina Delgado Valdivieso, Andreas Rienow and Joaquín Gairín; visualization, Fabián Santos-García; supervision, Joaquín Gairín; project administration, Fabián Santos-García; funding acquisition, Fabián Santos-García. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Indoamerica University Research Program (grant number: INV-0010-004).

Data Availability Statement: Data used in this research is available in the public repositories cited in the text; while the algorithm and its processing outputs are available at the link described in Appendix A.

Acknowledgments: The authors are thankful to INEVAL and its research team for making the Be Bachelor data available at no cost and for participating in meetings and organized workshops (13 November 2020). Thanks to David Vibas, Eduardo Salgado, Maya Fárez, Pablo Pesantes, Jorge Gómez, Mario Albán, Santiago Bonilla and four anonymous reviewers for their invaluable comments, suggestions and talks, which have improved the quality of the manuscript. The authors are also thankful to Google Earth Engine for its free cloud platform and the R community for all its software possibilities. Special thanks to the Center for International Migration and Develpment program (CIM) for accompanying the training and work activities of F.S.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Details of the interactive application for summarizing the model outputs, as well as the routines and data used in this analysis are available at: https://github.com/FSantosCodes/Urban-rural-gradients-predict-educational-gaps (accessed on 13 October 2021).

Appendix B

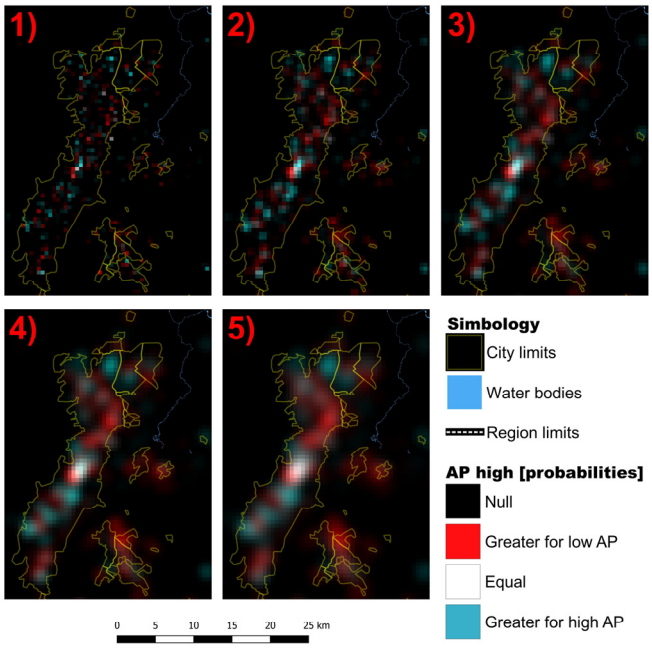


Figure A1. Example of the smoothing effect in the hot spot calculations using different kernel bandwidths: (**1**) one surrounding pixel (500 mts), (**2**) two surrounding pixels (1000 mts), (**3**) three surrounding pixels (1500 mts), (**4**) four surrounding pixels (2000 mts), and (**5**) five surrounding pixels (2500 mts).

Table A1. Best predictive questions and answers for the high AP classification according to groups, themes, and regions. The questions are sorted by their respective counts.

Region	Group	Theme	Question	High Schools	
				Count	Percentage
Andean	CS	VEP	How much did the pre-university course cost? -> 100 USD or less	80	1.91
Andean	AE	ADE	Are there any of these goods or services in your household? Check all that apply: Internet connection -> Yes	67	1.6
Andean	CS	VEP	What was your preparation to take the Ser Bachiller exam? -> At your school	58	1.38
Andean	SC	SGS	What type of toilet service does your home have? -> Connected to the public sewer system	50	1.19
Andean	AE	ADE	Are there any of these goods or services in your household? Check all that apply: Internet connection -> No	40	0.95
Andean	CS	VEP	What was your preparation to take the Ser Bachiller exam? -> Private pre-university course	39	0.93
Coast	SC	SEW	What is the highest level of education your mother has completed? -> General Basic Education	35	1.39
Coast	CS	VEP	How much did the pre-university course cost? -> 100 USD or less	34	1.35
Andean	AE	ADE	Are there any of these goods or services in your household? Check all that apply: A desk for studying -> No	31	0.74
Andean	SC	SEW	Point out what your mother regularly does -> She does domestic work in our home	30	0.71
Andean	SC	SEW	Indicate the highest level of education your mother has completed -> Doctorate (PhD)	28	0.67

Table A1. *Cont.*

Region	Group	Theme	Question	High Schools	
				Count	Percentage
Coast	CS	VEP	What was your preparation to take the Ser Bachiller exam? -> At your school	22	0.88
Coast	AE	ADE	In general, how many hours a day do you use the computer at my school? -> Never	21	0.84
Coast	SC	SGS	What type of toilet service does your home have? -> Connected to the public sewer system	21	0.84
Coast	AE	ASC	Have there ever been any thefts inside your classroom -> Never	19	0.76
Coast	SC	SEW	Indicate what your father does on a regular basis -> Has a stable or permanent paid job	19	0.76
Coast	CS	VEP	What was your preparation to take the Ser Bachiller exam? -> Private pre-university course	17	0.68
Coast	AE	ADE	Are there any of these goods or services in your household? Check all that apply: Internet connection -> No	16	0.64
Amazon	CS	VEP	What was your preparation to take the Ser Bachiller exam? -> At your school	6	1.45
Amazon	CS	VPR	Indicate how much you like to read the following types of text: Academic textbooks -> Not very much	4	0.97
Amazon	SC	CDC	In general, how many hours a day do you use electronic devices to check social networks? -> 1 h maximum	4	0.97
Amazon	AE	AST	How often does this happen in your classes? Teachers allow students to explain to their classmates -> Almost always	3	0.73
Amazon	AE	ASS	How often did these things happen in your classes? I felt comfortable doing group work -> Always	3	0.73
Amazon	AE	ASS	Do you make friends easily? -> Almost always	3	0.73
Amazon	CS	VEP	How much did the pre-university course cost? -> 100 USD or less	3	0.73
Amazon	SC	CAC	How often do you do these kinds of things with your family? We watch cultural programs on TV -> Always	3	0.73
Amazon	SC	SGS	Are there any of these goods or services in your household? Check all that apply: Garbage collection -> No	3	0.73

Table A2. Best predictive questions and answers for the low AP classification, according to groups, themes, and regions. The questions are sorted by their respective counts.

Region	Group	Theme	Question	High Schools	
				Count	Percentage
Amazon	SC	CIL	How do you identify yourself according to your culture and customs? -> Indigenous	9	1.75
Coast	AE	ADE	In your household, has anyone used the Internet in the past 6 months? -> No	9	0.67
Coast	CS	VEP	What was your preparation to take the Ser Bachiller exam?-> At your school	9	0.67
Coast	CS	VSV	Do you think about how the things you do will affect your future? -> Never	9	0.67
Amazon	AE	ADE	Are there any of these goods or services in your household? Check all that apply: Internet connection -> No	8	1.56
Coast	AE	ASC	Have there ever been any thefts inside your classroom -> Never	8	0.6
Coast	CS	VSV	What is the main reason you would like to study your career? -> Personal interest	8	0.6
Coast	SC	SEW	Indicate what your father does on a regular basis -> Works occasionally	8	0.6
Coast	AE	ADE	Are there any of these goods or services in your household? Check all that apply: A desk for studying -> No	7	0.52
Amazon	CS	VSV	What is the main reason you would like to study your career? -> Personal interest	6	1.17
Amazon	SC	CIL	Which of the following nationalities/indigenous peoples do you belong to? -> I do not belong to any	6	1.17

Table A2. *Cont.*

Region	Group	Theme	Question	High Schools	
				Count	Percentage
Amazon	SC	SGS	What type of toilet service does your home have? -> Connected to public sewer system	6	1.17
Andean	SC	SGS	Are there any of these goods or services in your household? Check all that apply: Landline phone -> No	6	0.95
Coast	SC	CIL	How do you identify yourself according to your culture and customs? -> Mestizo	6	0.45
Coast	SC	SGS	Are there any of these goods or services in your household? Check all that apply: Drainage or sewerage -> Yes	6	0.45
Amazon	AE	ADE	In your household, has anyone used the Internet in the last 6 months? -> No	5	0.97
Andean	AE	AST	How often does this happen in your classes? Teachers ask us questions and expect us to answer them -> Almost never	5	0.79
Andean	CS	VSV	What is the main reason you would like to study your career? -> Ease of finding a job	5	0.79
Andean	CS	VEP	What was your preparation to take the Ser Bachiller exam? -> At your school	5	0.79
Andean	CS	VPL	How many hours a day do you spend studying school subjects or doing homework at home? -> Less than 1 h	5	0.79
Andean	SC	CAC	How often do you do this kind of thing with your family? We read a book or magazine -> Almost never	5	0.79
Amazon	AE	ADE	Are there any of these goods or services in your household? Check all that apply Internet connection -> Yes	4	0.78
Amazon	CS	VSV	What is the highest level of studies you would like to achieve? -> Postgraduate (PhD)	4	0.78
Amazon	CS	VEP	What was your preparation to take the Ser Bachiller exam? -> Private pre-university course	4	0.78
Andean	AE	ADE	Are there any of these goods or services in your household? Check all that apply Internet connection -> No	4	0.63
Andean	AE	AHC	Do the activities you do when you are not studying make you happy? -> Always	4	0.63
Andean	SC	SEW	What is the highest level of education your mother has completed? -> General Basic Education	4	0.63

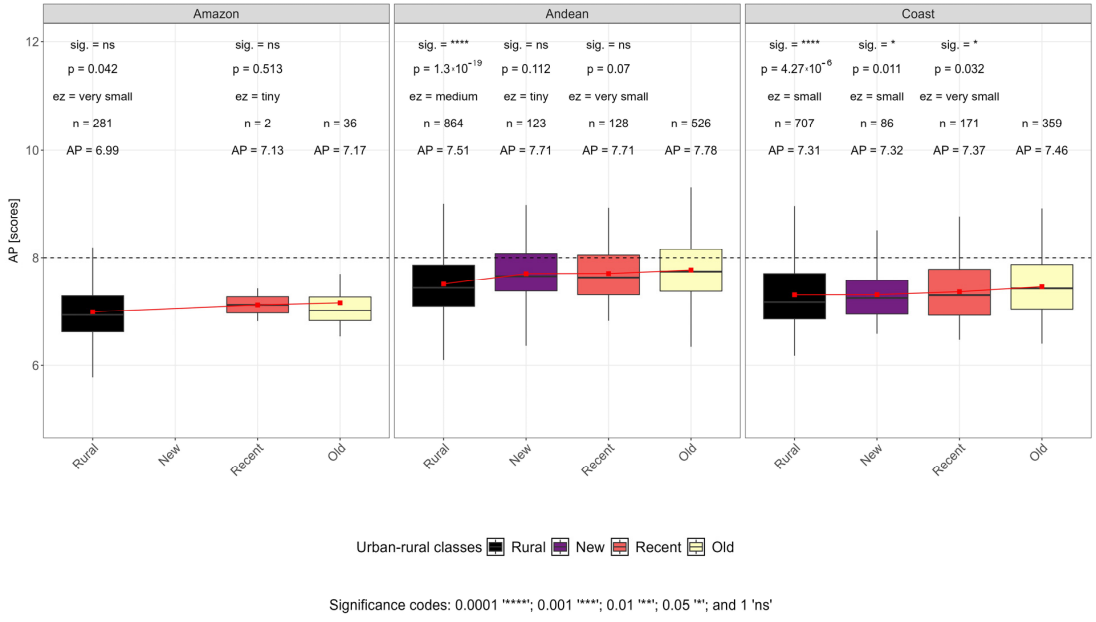


Figure A2. Boxplots using original AP scores for each urban–rural gradient map class and region. The text in the plots refers to the results of the Wilcoxon–Mann–Whitney test, showing significance (***sig.***), *p*-values (*p*), effect sizes (***ez***), sample sizes (*n*), and mean AP scores ($\overline{AP}$). Red lines indicate the mean values. Outliers are hidden to facilitate visualization.

References

1. Pradhan, P.; Costa, L.; Rybski, D.; Lucht, W.; Kropp, J.P. A Systematic Study of Sustainable Development Goal (SDG) Interactions. *Earth's Future* **2017**, *5*, 1169–1179. [CrossRef]
2. Bowles, S.; Gintis, H.; Meyer, P. The Long Shadow of Work: Education, the Family, and the Reproduction of the Social Division of Labor. *Insurg. Sociol.* **1999**, *25*, 286–305.
3. Ward, A.; Stoker, H.W.; Murray-Ward, M. Educational Measurement: Theories and applications. In *Educational Measurement*; University Press of America: Lanham, MD, USA, 1996; p. 3. ISBN 9780761803850.
4. Peet, E.D.; Fink, G.; Fawzi, W. Returns to Education in Developing Countries: Evidence from the Living Standards and Measurement Study Surveys. *Econ. Educ. Rev.* **2015**, *49*, 69–90. [CrossRef]
5. Harvey, D. Uneven geogrpahical developments and the production of space. In *Seventeen Contradictions and the End of Capitalism*; Oxford University Press: Oxford, UK, 2014; pp. 146–163.
6. Mehretu, A.; Pigozzi, B.W.; Sommers, L.M. Concepts in Social and Spatial Marginality. *Geogr. Ann. Ser. B Hum. Geogr.* **2000**, *82*, 89–101. [CrossRef]
7. Ford, D.R. Spatializing Marxist Educational Theory: School, the Built Environment, Fixed Capital and (Relational) Space. *Policy Future Educ.* **2014**, *12*, 784–793. [CrossRef]
8. Ferrare, J.J.; Apple, M.W. Spatializing Critical Education: Progress and Cautions. *Crit. Stud. Educ.* **2010**, *51*, 209–221. [CrossRef]
9. Murillo, F.J.; Román, M. School Infrastructure and Resources Do Matter: Analysis of the Incidence of School Resources on the Performance of Latin American Students. *Sch. Eff. Sch. Improv.* **2011**, *22*, 29–50. [CrossRef]
10. Contreras, D.; Delgadillo, J.; Riveros, G. Is Home Overcrowding a Significant Factor in Children's Academic Performance? Evidence from Latin America. *Int. J. Educ. Dev.* **2019**, *67*, 1–17. [CrossRef]
11. Schleicher, A. *PISA 2018 Insights and Interpretations*; OECD Publishing: Paris, France, 2019.
12. Cox, C. Educational inequality in Latin America: Patterns, policies and issues. In *Growing Gaps. Educational Inequality around the World*; Attewell, P., Newman, K.S., Eds.; Oxford University Press: New York, NY, USA, 2010; pp. 33–58.
13. Misra, K.; Grimes, P.W.; Rogers, K.E. Does Competition Improve Public School Efficiency? A Spatial Analysis. *Econ. Educ. Rev.* **2012**, *31*, 1177–1190. [CrossRef]
14. Farwick, A.; Hanhörster, H.; Lobato, I.R.; Striemer, W. Neighbourhood-Based Social Integration. The Importance of the Local Context for Different Forms of Resource Transfer. *Raumforsch. Raumordn. Spat. Res. Plan.* **2019**, *77*, 417–434. [CrossRef]
15. Ramos Lobato, I. School Segregation in Urban Contexts: Socio-Spatial Dynamics and Educational Inequalities. *Urbaria Summ. Ser.* **2020**, *4*, 1–12.
16. Abou, P.E. Does Domestic Work Affect the Academic Performance of Girls in Primary School in Côte d'Ivoire? Empirical Evidence from Probit Model. *Eur. Sci. J. ESJ* **2016**, *12*, 368. [CrossRef]
17. Florence, M.; Asbridge, M.; Veugelers, P. Diet Quality and Academic Performance. *J. Sch. Health* **2008**, *78*, 239–241. [CrossRef] [PubMed]
18. Farooq, M.S.; Chaudhry, A.H.; Shafiq, M.; Berhanu, G. Factors Affecting Students' Quality of Nacademic Performance: A Case of Secondary School Level. *J. Qual. Technol. Manag.* **2011**, *VII*, 1–14.
19. Akukwe, B.; Schroeders, U. Socio-Economic, Cultural, Social, and Cognitive Aspects of Family Background and the Biology Competency of Ninth-Graders in Germany. *Learn. Individ. Differ.* **2016**, *45*, 185–192. [CrossRef]
20. Buriel, R.; Perez, W.; De Ment, T.L.; Chavez, D.; Moran, V. The Relationship of Language Brokering to Academic Performance, Biculturalism, and Self-Efficacy among Latino Adolescents. *Hisp. J. Behav. Sci.* **1998**, *20*, 283–297. [CrossRef]
21. Kisilevsky, M.; Veleda, C. *Dos Estudios Sobre El Acceso a La Educación Superior En La Argentina*; UNESCO, Instituto Internacional de Planeamiento de la Educación: Paris, France, 2002; pp. 19–27.
22. Kirschner, P.A.; Karpinski, A.C. Facebook®and Academic Performance. *Comput. Hum. Behav.* **2010**, *26*, 1237–1245. [CrossRef]
23. Tobler, W.R. A Computer Movie Simulating Urban Growth in the Detroit Region. *Econ. Geogr.* **1970**, *46*, 234–240. [CrossRef]
24. Brunsdon, C.; Fotheringham, A.; Charlton, M.E. Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity. *Geogr. Anal.* **1996**, *28*, 281–298. [CrossRef]
25. Kitchin, R. Big Data and Human Geography: Opportunities, Challenges and Risks. *Dialogues Hum. Geogr.* **2013**, *3*, 262–267. [CrossRef]
26. Giraud, C. *Introduction to High-Dimensional Statistics*; Taylor & Francis Group, LLC.: Abingdon, UK, 2015; ISBN 978-1-4822-3794-8.
27. Freedman, D.A. A Note on Screening Regression Equations. *Am. Stat.* **1983**, *37*, 152–155. [CrossRef]
28. Lukacs, P.M.; Burnham, K.P.; Anderson, D.R. Model Selection Bias and Freedman's Paradox. *Ann. Inst. Stat. Math.* **2010**, *62*, 117–125. [CrossRef]
29. Van Der Maaten, L.J.P.; Postma, E.O.; Van Den Herik, H.J. Dimensionality Reduction: A Comparative Review. *J. Mach. Learn. Res.* **2009**, *10*, 1–41. [CrossRef]
30. Vilenchik, D.; Yichye, B.; Abutbul, M. To Interpret or Not to Interpret PCA? This Is Our Question. In Proceedings of the International AAAI Conference on Web and Social Media, Münich, Germany, 11–14 June 2019; pp. 655–658.
31. Bolón-Canedo, V.; Sánchez-Maroño, N.; Alonso-Betanzos, A. *Feature Selection for High-Dimensional Data*; O'Sullivan, B., Wooldridge, M., Eds.; Springer Publishing Company: Cham, Switzerland, 2015; ISBN 978-3-319-21857-1.
32. Kumar, V. Feature Selection: A Literature Review. *Smart Comput. Rev.* **2014**, *4*, 211–229. [CrossRef]
33. Kohavi, R.; John, G.H. Wrappers for Feature Subset Selection. *Artif. Intell.* **1997**, *97*, 273–324. [CrossRef]

34. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
35. Kursa, M.B.; Rudnicki, W.R. Feature Selection with the Boruta Package. *J. Stat. Softw.* **2010**, *36*, 1–13. [CrossRef]
36. Kursa, M.B. Robustness of Random Forest-Based Gene Selection Methods. *BMC Bioinform.* **2014**, *15*, 8. [CrossRef]
37. Galor, O.; Moav, O. From Physical to Human Capital Accumulation: Inequality and the Process of Development. *Rev. Econ. Stud.* **2004**, *71*, 1001–1026. [CrossRef]
38. Gong, P.; Li, X.; Wang, J.; Bai, Y.; Chen, B.; Hu, T.; Liu, X.; Xu, B.; Yang, J.; Zhang, W.; et al. Annual Maps of Global Artificial Impervious Area (GAIA) between 1985 and 2018. *Remote Sens. Environ.* **2020**, *236*, 111510. [CrossRef]
39. Angotti, T. Introduction Urban Latin America Violence, Enclaves, and Struggles for Land. *Lat. Am. Perspect.* **2013**, *40*, 5–20. [CrossRef]
40. Purcell, T.F.; Fernandez, N.; Martinez, E. Rents, Knowledge and Neo-Structuralism: Transforming the Productive Matrix in Ecuador. *Third World Q.* **2017**, *38*, 918–938. [CrossRef]
41. INEVAL Descarga de Datos: Exámen Nacional de Evaluación Educativa Ser Bachiller. Available online: http://evaluaciones.evaluacion.gob.ec/BI/bases-de-datos-ser-bachiller/ (accessed on 13 October 2021).
42. Galobardes, B.; Shaw, M.; Lawlor, D.A.; Lynch, J.W.; Smith, G.D. Indicators of Socioeconomic Position (Part 1). *J. Epidemiol. Community Health* **2006**, *60*, 7–12. [CrossRef] [PubMed]
43. Bhugra, D.; Becker, M.A. Migration, Cultural Bereavement and Cultural Identity. *World Psychiatry* **2005**, *4*, 18–24. [PubMed]
44. Jian, W. The Relationship between Culture and Language. *ELT J.* **2000**, *54*, 328–334. [CrossRef]
45. Blackwell, C.K.; Lauricella, A.R.; Wartella, E. Factors Influencing Digital Technology Use in Early Childhood Education. *Comput. Educ.* **2014**, *77*, 82–90. [CrossRef]
46. Jeong, S.; Kwak, D.-H.; Moon, B.; San Miguel, C. Predicting School Bullying Victimization: Focusing on Individual and School Environmental/Security Factors. *J. Criminol.* **2013**, *2013*, 1–13. [CrossRef]
47. Hong, Z.R.; Lin, H.S. An Investigation of Students' Personality Traits and Attitudes toward Science. *Int. J. Sci. Educ.* **2011**, *33*, 1001–1028. [CrossRef]
48. INEC Censo de Población y Vivienda 2010. Available online: http://www.inec.gob.ec/estadisticas/ (accessed on 21 March 2018).
49. Hidalgo, R.J. *Economic Growth and Regional Inequality in Ecuador*; Erasmus University Rotterdam: Burgemeester Oudlaan, The Netherlands, 2018.
50. Slonecker, E.T.; Jennings, D.B.; Garofalo, D. Remote Sensing of Impervious Surfaces: A Review. *Remote Sens. Rev.* **2001**, *20*, 227–255. [CrossRef]
51. Gorelick, N.; Hancher, M.; Dixon, M.; Ilyushchenko, S.; Thau, D.; Moore, R. Google Earth Engine: Planetary-Scale Geospatial Analysis for Everyone. *Remote Sens. Environ.* **2017**, *202*, 18–27. [CrossRef]
52. Beckerman, P.; Solimano, A. *Crisis and Dollarization in Ecuador: Stability, Growth, and Social Equity*; World Bank: Washington, DC, USA, 2002.
53. Clark, P.; García, J. Left Populism, State Building, Class Compromise, and Social Conflict in Ecuador's Citizens' Revolution. *Lat. Am. Perspect.* **2019**, *46*, 230–246. [CrossRef]
54. Biau, G.; Scornet, E. A Random Forest Guided Tour. *Test* **2016**, *25*, 197–227. [CrossRef]
55. Wright, M.N.; Ziegler, A. Ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *arXiv* **2015**, arXiv:1508.04409. [CrossRef]
56. Strobl, C.; Malley, J.; Gerhard, T. An Introduction to Recursive Partitioning: Rationale, Application Psychol Methods. *Psychol. Methods* **2009**, *14*, 323–348. [CrossRef]
57. Strobl, C.; Boulesteix, A.L.; Zeileis, A.; Hothorn, T. Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution. *BMC Bioinform.* **2007**, *8*, 25. [CrossRef] [PubMed]
58. Luan, J.; Zhang, C.; Xu, B.; Xue, Y.; Ren, Y. The Predictive Performances of Random Forest Models with Limited Sample Size and Different Species Traits. *Fish. Res.* **2020**, *227*, 105534. [CrossRef]
59. R Core Team: The R Project for Statistical Computing; Version 4.0.5. 31 March 2021. Available online: https://www.r-project.org/ (accessed on 13 October 2021).
60. Hijmans, R.; van Etten, J.; Cheng, J.; Mattiuzzi, M.; Sumner, M.; Greenberg, J.A.; Lamigueiro, O.P.; Bevan, A.; Racine, E.B.; Shortridge, A.; et al. Raster: Geographic Data Analysis and Modeling; Version 2.6-7.2017. Available online: https://CRAN.R-project.org/package=raster (accessed on 13 October 2021).
61. Binbin, L.; Harris, P.; Charlton, M.; Bruns-don, C.; Nakaya, T.; Gollini, I. GWmodel: Geographically-Weighted Models; Version 2.0-5. 2017. Available online: https://CRAN.R-project.org/package=GWmodel (accessed on 13 October 2021).
62. Japkowicz, N.; Stephen, S. The Class Imbalance Problem: A Systematic Study. *Intell. Data Anal.* **2002**, *6*, 429–449. [CrossRef]
63. Svetnik, V.; Liaw, A.; Tong, C.; Wang, T. Application of Breiman's Random Forest to Modeling Structure-Activity Relationships of Pharmaceutical Molecules. *Lect. Notes Comput. Sci.* **2004**, *3077*, 334–343. [CrossRef]
64. Breiman, L.; Cutler, A.; Liaw, A.; Wiener, M. RandomForest: Breiman and Cutler's Random Forests for Classification and Regression; Version 4.6-14. 2018. Available online: https://CRAN.R-project.org/package=randomForest (accessed on 13 October 2021).
65. Kuhn, M. Caret: Classification and Regression Training; Version 6.0-71. 2016. Available online: https://CRAN.R-project.org/package=caret (accessed on 13 October 2021).

66. R Core Team: Parallel: Support for Parallel Computation in R; Version 4.0.3. 2020. Available online: https://www.r-project.org/ (accessed on 13 October 2021).
67. QGIS Development Team: QGIS Geographic Information System; Version 3.16.11-Hannover. 2019. Available online: https://qgis.org/en/site/ (accessed on 13 October 2021).
68. Xie, Z.; Yan, J. Detecting Traffic Accident Clusters with Network Kernel Density Estimation and Local Spatial Statistics: An Integrated Approach. *J. Transp. Geogr.* **2013**, *31*, 64–71. [CrossRef]
69. Anderson, T.K. Kernel Density Estimation and K-Means Clustering to Profile Road Accident Hotspots. *Accid. Anal. Prev.* **2009**, *41*, 359–364. [CrossRef]
70. Caha, J. SpatialKDE: Kernel Density Estimation for Spatial Data; Version 0.6.2. 2020. Available online: https://CRAN.R-project.org/package=SpatialKDE (accessed on 13 October 2021).
71. Siloko, I.U.; Siloko, E.A.; Ikpotokin, O. A Mini Review of Dimensional Effects on Asymptotic Mean Integrated Squared Error and Efficiencies of Selected Beta Kernels. *Jordan J. Math. Stat.* **2020**, *13*, 327–340.
72. Funder, D.C.; Ozer, D.J. Evaluating Effect Size in Psychological Research: Sense and Nonsense. *Adv. Methods Pract. Psychol. Sci.* **2019**, *2*, 156–168. [CrossRef]
73. Russo, L.; Beauguitte, L. Aggregation Level Matters: Evidence from French Electoral Data. *Qual. Quant.* **2014**, *48*, 923–938. [CrossRef]
74. Amrhein, C.G. Searching for the Elusive Aggregation Effect: Evidence from Statistical Simulations. *Environ. Plan. A Econ. Sp.* **1995**, *27*, 105–119. [CrossRef]
75. Frantz, D.; Schug, F.; Okujeni, A.; Navacchi, C.; Wagner, W.; van der Linden, S.; Hostert, P. National-Scale Mapping of Building Height Using Sentinel-1 and Sentinel-2 Time Series. *Remote Sens. Environ.* **2021**, *252*, 112128. [CrossRef] [PubMed]
76. Wulder, M.A.; White, J.C.; Loveland, T.R.; Woodcock, C.E.; Belward, A.S.; Cohen, W.B.; Fosnight, E.A.; Shaw, J.; Masek, J.G.; Roy, D.P. The Global Landsat Archive: Status, Consolidation, and Direction. *Remote Sens. Environ.* **2016**, *185*, 271–283. [CrossRef]
77. Gruszczyński, M. On Unbalanced Sampling in Bankruptcy Prediction. *Int. J. Financ. Stud.* **2019**, *7*, 28. [CrossRef]
78. Acerenza, S.; Gandelman, N. Household Education Spending in Latin America and the Caribbean: Evidence from Income and Expenditure Surveys. *Educ. Financ. Policy* **2019**, *14*, 61–87. [CrossRef]
79. Henderson, J.V.; Storeygard, A.; Weil, D.N. Measuring Economic Growth from Outer Space. *Am. Econ. Rev.* **2012**, *102*, 994–1028. [CrossRef]
80. Wolf, A. Education and Economic Performance: Simplistic Theories and Their Policy Consequences. *Oxf. Rev. Econ. Policy* **2004**, *20*, 315–333. [CrossRef]
81. Ma, Q.; He, C.; Wu, J. Behind the Rapid Expansion of Urban Impervious Surfaces in China: Major Influencing Factors Revealed by a Hierarchical Multiscale Analysis. *Land Use Policy* **2016**, *59*, 434–445. [CrossRef]
82. Horn, P.; Grugel, J. The SDGs in Middle-Income Countries: Setting or Serving Domestic Development Agendas? Evidence from Ecuador. *World Dev.* **2018**, *109*, 73–84. [CrossRef]
83. Rodríguez Cruz, M. Construir La Interculturalidad. Políticas Educativas, Diversidad Cultural y Desigualdad En El Ecuador. *Íconos Rev. Cienc. Soc.* **2018**, *60*, 217–236. [CrossRef]
84. García-Aracil, A.; Winter, C. Gender and Ethnicity Differentials in School Attainment and Labor Market Earnings in Ecuador. *World Dev.* **2006**, *34*, 289–307. [CrossRef]
85. Álvarez-Gamboa, J.; Cabrera-Barona, P.; Jácome-Estrella, H. Financial Inclusion and Multidimensional Poverty in Ecuador: A Spatial Approach. *World Dev. Perspect.* **2021**, *22*, 100311. [CrossRef]