remote sensing

MDPI

Article

Improved Inference and Prediction for Imbalanced Binary Big Data Using Case-Control Sampling: A Case Study on Deforestation in the Amazon Region

Denis Valle [1,*], Jacy Hyde [1], Matthew Marsik [2] and Stephen Perz [3]

[1] School of Forest Resources and Conservation, University of Florida, Gainesville, FL 32611, USA; jlhyde@ufl.edu

[2] UF Health Shands, University of Florida, Gainesville, FL 32611, USA; mmarsik@ufl.edu

[3] Sociology and Criminology & Law, University of Florida, Gainesville, FL 32611, USA; sperz@ufl.edu

[*] Correspondence: drvalle@ufl.edu

Received: 10 March 2020; Accepted: 14 April 2020; Published: 17 April 2020

check for updates

Abstract: It is computationally challenging to fit models to big data. For example, satellite imagery data often contain billions to trillions of pixels and it is not possible to use a pixel-level analysis to identify drivers of land-use change and create predictions using all the data. A common strategy to reduce sample size consists of drawing a random sample but this approach is not ideal when the outcome of interest is rare in the landscape because it leads to very few pixels with this outcome. Here we show that a case-control (CC) sampling approach, in which all (or a large fraction of) pixels with the outcome of interest and a subset of the pixels without this outcome are selected, can yield much better inference and prediction than random sampling (RS) if the estimated parameters and probabilities are adjusted with the equations that we provide. More specifically, we show that a CC approach can yield unbiased inference with much less uncertainty when CC data are analyzed with logistic regression models and its semiparametric variants (e.g., generalized additive models). We also show that a random forest model, when fitted to CC data, can generate much better predictions than when fitted to RS data. We illustrate this improved performance of the CC approach, when used together with the proposed bias-correction adjustments, with extensive simulations and a case study in the Amazon region focused on deforestation.

Keywords: satellite imagery; deforestation; inference; prediction; Amazon; pixel sampling; case-control

1. Introduction

Big data have become ubiquitous in multiple domains. However, the analysis of these data is very challenging because these data are often too large for standard statistical and machine learning software. Sensors are an important source of big data, including for example measurements of carbon dioxide concentration in the ocean and coastal seas (e.g., 14.7 million values in the Surface Ocean CO$_2$ Atlas [SOCAT]) [1], digital camera imagery to monitor vegetation phenology (e.g., almost 750 years of imagery collected typically every 30 minutes by the PhenoCam network) [2], wildlife data from camera traps (e.g., over 4.5 million records in the Wildlife Insights cloud platform) [3], and satellite imagery (e.g., over 46 years of the Landsat mission, with currently ~1200 new images being added per day, each image containing approximately 34 million pixels) [4]. We focus on satellite imagery data (hereafter remote sensing data) in this article as these data have been extensively used in multiple applications, such as the mapping and quantification of large-scale changes in tropical forests, urban areas, ice cover, coral reefs, and surface water [5–9]. Unfortunately, the size of remote sensing data precludes the use of standard statistical models and predictive algorithms without first aggregating or sampling the data.

In this article, we propose a methodology that improves inference and prediction for binary big data in which one category is much rarer than the other (highly imbalanced data). We illustrate this approach by focusing on modeling deforestation risk.

Most studies focused on inferring drivers of deforestation have spatially aggregated the data to reduce its size. For example, researchers often aggregate data within politically defined units (e.g., municipalities or counties) [10–13], aggregate data into super-pixels [14–18], or aggregate data based on the particular driver being investigated (e.g., by creating distance buffers from the road) [19,20]. Unfortunately, this spatial aggregation unavoidably results in some loss of fine-scale spatial information, which might be particularly undesirable if researchers are interested in determining the effect of important distance-based variables (e.g., distance to roads or other infrastructure). Furthermore, when aggregating data based on the particular deforestation driver of interest (e.g., distance to roads), it can be challenging to simultaneously remove the effect of other potential confounders such as proximity to cities, rivers, and markets.

The few studies that have attempted to infer the drivers of deforestation through the modeling of pixel-level data have had to rely on the random sampling of pixels to reduce sample size [21–23]. Similarly, studies that have focused on predicting future deforestation, rather than on inference on drivers of deforestation, have also had to rely on randomly sampled data to convert spatial determinants (e.g., proximity to roads, topography, soil fertility) into deforestation/conversion probabilities [24–26]. The problem of the standard random sampling of pixels in the context of highly imbalanced binary data is that the resulting dataset might contain very few samples with the category of interest, resulting in models with decreased inference and prediction ability. Indeed, several studies have demonstrated the importance of increasing the prevalence of the category of interest (e.g., deforested pixels) through under-sampling of the dominant class or over-sampling the rare class in highly imbalanced data [24,27–32]. However, it has also been widely acknowledged that this differential sampling approach introduces important biases [24], and determining how models can be calibrated to generate accurate probability estimates is often cited as an important problem [33].

In this article, we derive equations that remove the bias from the estimated parameters and predicted probabilities when the prevalence of the category of interest is increased through a case-control (CC) sampling approach, resulting in substantially improved inference and prediction for imbalanced binary data when compared to simple random sampling (RS). We first describe the CC sampling methodology and how parameters and predictions need to be adjusted to improve inference and prediction. Then, we describe how we have tested this methodology using both simulated and real data. In particular, we illustrate this approach with Generalized Additive Models (GAMs) and random forest, a commonly used machine learning algorithm, applied to a case study involving deforestation risk in the Amazon region.

2. Materials and Methods

2.1. Case-Control Sampling

The case-control (CC) approach is widely used in epidemiological studies of rare diseases [34]. In standard cross-sectional or prospective cohort studies, researchers often end up with very few individuals with the rare disease, a major roadblock to determining potential risk factors. The basic idea of the CC approach is to retrospectively sample observations according to their outcome. A CC approach in epidemiological studies typically selects all individuals with the rare disease (i.e., cases) and a sample of individuals without the disease (i.e., controls), thereby circumventing the issue of not having enough individuals with the rare disease. Data that arise from a CC approach will typically yield much better estimates than a random sample as a result of the increased number of cases because the CC approach "concentrates resources where there is the greatest amount of information, namely on the cases" [34].

2.1.1. Inference from Case-Control Data

The idea that a CC sampling approach results in better inference can be formalized following King and Zeng [35]. In a standard logistic regression, the variance of the regression parameters is given by

$$Cov\left(\hat{\beta}\right) = \left(X^T \hat{W} X\right)^{-1} \tag{1}$$

where $\hat{\beta}$ is the vector with the estimated regression coefficients, X is the design matrix containing covariate information, and $\hat{W}$ is a diagonal "weight" matrix with elements given by $\hat{w}_{ii} = \hat{p}_i(1-\hat{p}_i)$. In this expression, $\hat{p}_i$ is the predicted probability for the binary event (e.g., disease or deforestation) for observation i. All else being equal, the variance for $\hat{\beta}$ is minimized for $\hat{p}_i$ values close to 0.5. For a sample mostly comprised of controls, $\hat{p}_i$ will generally be close to 0, yielding high variances. On the other hand, as we increase the proportion of cases, $\hat{p}_i$ will be closer to 0.5 for more observations, decreasing the variance.

Intuitively it might seem that selecting samples based on the response variable is not a valid approach given that the response variable is being purposefully chosen by the modeler. Following Agresti [36], we show that using a standard logistic regression to analyze data that originates from a CC design generates highly biased results for the intercept but yields unbiased inference on slope parameters. Let y indicate whether a subject has the outcome of interest (0 = no, 1 = yes) and let x denote a covariate. A standard logistic regression assumes that

$$y \sim Bern(\pi) \tag{2}$$

$$\pi = p\left(y = 1 \middle| x\right) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} \tag{3}$$

where we are interested in estimating the regression parameters β_0 and β_1. However, when we use a logistic regression on CC data, we are making inference on $p\left(y = 1 \middle| z = 1, x\right)$ instead of $p\left(y = 1 \middle| x\right)$, where z indicates whether a subject was sampled (0 = no,1 = yes).

To relate $p\left(y = 1 \middle| z = 1, x\right)$ to our actual quantity of interest $p\left(y = 1 \middle| x\right)$, note that:

$$\begin{aligned} & p\left(y = 1 \middle| z = 1, x\right) \\ & = \frac{p\left(z = 1 \middle| y = 1, x\right) p\left(y = 1 \middle| x\right)}{p\left(z = 1 \middle| y = 1, x\right) p\left(y = 1 \middle| x\right) + p\left(z = 1 \middle| y = 0, x\right) p\left(y = 0 \middle| x\right)} \end{aligned} \tag{4}$$

Further, note that the sampling probability just depends on the response variable y in a CC study and not on covariate x (i.e., $p\left(z = 1 \middle| y, x\right) = p\left(z = 1 \middle| y\right)$). Therefore:

$$\begin{aligned} & p\left(y = 1 \middle| z = 1, x\right) \\ & = \frac{p\left(z = 1 \middle| y = 1\right) p\left(y = 1 \middle| x\right)}{p\left(z = 1 \middle| y = 1\right) p\left(y = 1 \middle| x\right) + p\left(z = 1 \middle| y = 0\right) p\left(y = 0 \middle| x\right)} \\ & = \frac{p_1 p\left(y = 1 \middle| x\right)}{p_1 p\left(y = 1 \middle| x\right) + p_0 p\left(y = 0 \middle| x\right)} \end{aligned} \tag{5}$$

where $p_1 = p\left(z = 1 \middle| y = 1\right)$ denotes the probability of sampling a case whereas $p_0 = p\left(z = 1 \middle| y = 0\right)$ denotes the probability of sampling a control. After substituting $p\left(y = 1 \middle| x\right) = \frac{\exp(\beta_0 + \beta_1 x)}{1+\exp(\beta_0 + \beta_1 x)}$ and $p\left(y = 0 \middle| x\right) = \frac{1}{1+\exp(\beta_0 + \beta_1 x)}$ and doing some algebraic manipulations, we obtain the following expression:

$$p\left(y = 1 \middle| z = 1, x\right) = \frac{\exp\left(\widetilde{\beta_0} + \beta_1 x\right)}{\exp\left(\widetilde{\beta_0} + \beta_1 x\right) + 1} \tag{6}$$

where $\widetilde{\beta_0} = log\left(\frac{p_1}{p_0}\right) + \beta_0$. This expression reveals that the logistic regression applied to data originated from a CC design will generate valid estimates for the slope coefficient β_1 and that the intercept estimate $\widetilde{\beta_0}$ is a biased version of the true intercept β_0.

Differently from epidemiological studies, in remote sensing studies, we have data from the entire population (i.e., pixels). As a result, we can easily remove the bias from the estimated intercept. More specifically, we can calculate $log\left(\frac{p_1}{p_0}\right)$, where

$$p_0 = \frac{number\ of\ selected\ observations\ where\ y = 0}{overall\ number\ of\ observations\ where\ y = 0} \tag{7}$$

$$p_1 = \frac{number\ of\ selected\ observations\ where\ y = 1}{overall\ number\ of\ observations\ where\ y = 1} \tag{8}$$

In the CC design, p_1 is often equal to 1 as all cases are typically selected whereas p_0 will vary depending on the desired sample size. Using these calculations, we can obtain an unbiased estimate of the true intercept β_0 with the expression $\beta_0 = \widetilde{\beta_0} - log\left(\frac{p_1}{p_0}\right)$. Note that these derivations and results are valid for logistic regression models regardless of the number of covariates. This is an important observation because it reveals that these equations can be readily applied to logistic regression models that allow for nonlinear relationships through the addition of multiple basis functions (e.g., splines or wavelets) such as generalized additive models (GAMs) [37]. Finally, note that the validity of these results rests on the implicit assumption that logistic regression models and their associated assumptions (e.g., logistic link and conditional independence) are reasonable for the data being analyzed.

2.1.2. Prediction Based on Case-Control Data

An apparent limitation of the results described above is that they only pertain to logistic regression type models, precluding the correction of the biases introduced by the CC approach when using other types of models (e.g., probit regression models, random forest, and neural networks). However, our results are more general and are valid for other types of models as long as these models are able to provide consistent estimates of $p(y = 1|z = 1, x)$. Assuming that data are sampled using a CC approach where p_0 and p_1 are known, it is possible to calculate $p(y = 1|x)$. As derived in Equation 5, recall that

$$p(y = 1|z = 1, x) = \frac{p_1 p(y = 1|x)}{p_1 p(y = 1|x) + p_0 p(y = 0|x)} \tag{9}$$

where $p_1 = p(z = 1|y = 1)$ and $p_0 = p(z = 1|y = 0)$. This implies that

$$\frac{p(y = 1|z = 1, x)}{p(y = 0|z = 1, x)} = \frac{p_1 p(y = 1|x)}{p_0 p(y = 0|x)} \tag{10}$$

Recall that $p(y = 0|x) = 1 - p(y = 1|x)$. As a result, we have

$$p(y = 1|x) = \frac{p(y = 1|z = 1, x)p_0}{p(y = 0|z = 1, x)p_1 + p(y = 1|z = 1, x)p_0} \tag{11}$$

Equation (11) reveals how probability estimates based on the CC data $p(y = 1|z = 1, x)$ can be converted to the unbiased probability estimate $p(y = 1|x)$ (i.e., estimates that one would obtain from a random sample of the data). Assuming $p_1 = 1$, Figure 1 shows the relationship between these probabilities given by Equation (11) for different values of p_0. As expected, this figure reveals that bias increases as the fraction of observations sampled from the majority class declines (i.e., p_0 decreases from 1 to 0).

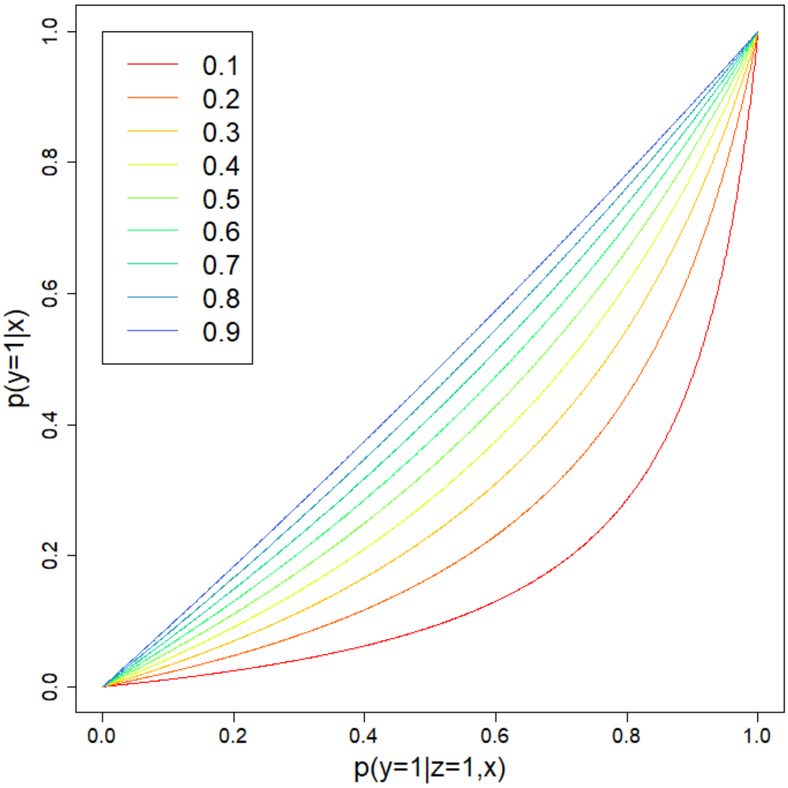


Figure 1. Relationship between the desired probability $p(y = 1 \middle| x)$ and the probability based on a CC sampling approach $p(y = 1 \middle| z = 1, x)$. Assuming $p_1 = 1$, results are shown separately for different rates of sampling the control cases (i.e., p_0, see figure legend). The mathematical expression for this relationship is given by Equation (11).

2.2. Simulations

We created three sets of simulated data to illustrate the benefits of a CC sampling approach relative to an RS approach. In the first set, we were interested in comparing parameter estimates from logistic regression models trained on data sampled according to an RS versus CC sampling approach. To this end, we generate data by assuming that

$$y_i \sim Bernoulli\left(\frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}\right) \tag{12}$$

where we assumed that $\beta_0 \in \{-6, -5, \ldots, -1\}$ and $\beta_1 = -2$. Different values of β_0 were used for each simulated dataset to vary the proportion of cases. The covariate x_i was generated from a uniform distribution between −1 and 1.

In the second set of simulated data, we were interested in evaluating how well the CC approach would work, when compared to the RS approach, when used in conjunction with generalized additive models (GAMs) to capture nonlinear patterns between the response variable and our variable of interest x_1 while simultaneously controlling for the nuisance variable x_2. More specifically, we assumed that

$$y_i \sim Bernoulli\left(\frac{\exp(\beta_0 + \beta_1 \sin(x_{1i}) + \beta_2 x_{2i})}{1 + \exp(\beta_0 + \beta_1 \sin(x_{1i}) + \beta_2 x_{2i})}\right) \tag{13}$$

where x_{1i} is generated from a uniform distribution between 0 and 2π and x_{2i} is generated from a uniform distribution between −1 and 1. Here we assumed that $\beta_0 \in \{-5, -4, -3, -2\}$, $\beta_1 = 1$, and $\beta_2 = -0.1$. Again, each value of β_0 corresponds to a different proportion of cases in the simulated datasets. GAMs were fit using the mgcv package in R [37].

In the third set of simulated data, we generated data based on a complex relationship between $p(y_i|x_{1i}, x_{2i})$ and the covariates x_{1i} and x_{2i}. The estimation of this probability surface is particularly challenging because the dataset also contained information on eight additional covariates $x_{3i}, \ldots, x_{10i}$, making this a more suitable task for very flexible models like random forest. To simulate these data, we assumed a Gaussian kernel for the success probability π_i:

$$y_i \sim Bernoulli(\pi_i) \tag{14}$$

$$\pi_i = \exp\left(-\frac{1}{2\sigma^2(1-\rho^2)}\left[x_{i1}^2 - 2\rho x_{i1}x_{i2} + x_{i2}^2\right]\right) \tag{15}$$

where ρ (set to 0.8) and $\sigma^2 \in \{0.1, 0.6, \ldots, 2.6\}$ are the correlation and variance parameters, respectively. For this set of simulations, we varied the proportion of cases by varying the value of σ^2. All covariates were generated from uniform distributions between -3 and 3. The goal of this set of simulated data was to determine if the random forest model would better estimate the Gaussian probability surface if data were sampled using a CC approach when compared to an RS approach. It is important to note that throughout this article we used the random forest regression (fit with the R package randomForest [38]) rather than the random forest classification algorithm. The reason for this is that the conditional probability estimates provided by the random forest classification algorithm might not be consistent whereas those generated by the random forest regression applied to binary data coded as 0 or 1 are known to be consistent [39].

We generated 100,000 observations for each parameter setting (i.e., for each β_0 or σ^2 value). The RS approach consisted of randomly sampling 10,000 observations, regardless of their y_i values. On the other hand, in the CC approach, we included all observations for which $y_i = 1$. If there were more than 5000 such observations, we randomly selected only 5000 observations for which $y_i = 1$. We randomly selected observations for which $y_i = 0$, such that the overall number of observations was equal to 10,000, ensuring that both RS and CC datasets were of the same size. We then fitted the models on data from each approach (CC and RS) and stored these results. This was performed 100 times for each set of simulation parameters to capture the variability in the estimated parameters and relationships.

2.3. Case Study on Deforestation in the Amazon Region

The Amazon region has been widely acknowledged for its remarkable biological and cultural diversity [40] and for its role in regulating global biogeochemical and atmospheric cycles [41–43]. At the same time, the region has been increasingly under pressure by large-scale infrastructure development projects (e.g., paving of existing roads and construction of new roads, hydroelectric dams, and energy transmission lines) [44–46]. Over the past few decades, tree clearing in the region has already removed 15% of the forest [47]. Importantly, there has been a recent surge in forest fires and deforestation in the Brazilian Amazon region. The number of active fires in August 2019 was nearly three times higher than in August 2018, and over 9500 km^2 of forest were lost between August 2018 and July 2019, resulting in the highest annual loss since 2008 [48]. In this context, it is critical to develop methods to better understand the spatial scale of the deforestation impact associated with infrastructure projects and improve predictions of future deforestation hotspots.

We focus on modeling deforestation risk in a tri-national frontier in the Amazon. More specifically, we study how the recent paving of the Inter-Oceanic Highway influences deforestation in the region where Peru, Brazil, and Bolivia meet, known as the MAP frontier (Madre de Dios, Acre, and Pando). In 2005, Madre de Dios, Acre, and Pando had approximately 94%, 82%, and 97% of forest cover, respectively [20]. We selected two originally unpaved road segments that were almost completely forested in 1986 (i.e., 0.3%–0.6% of the area deforested) and that were relatively far from cities and towns (>10 km). One road segment is located in Madre de Dios, in southeastern Peru, and was paved between 2006 and 2010 (the road segment within the red polygon at left, Figure 2). The other road segment is located in Pando, in northern Bolivia (the road segment within the red polygon at right, Figure 2), a region regarded as the last frontier of intact forest in the Bolivian lowlands [49]. This road segment remained unpaved throughout the entire study period (i.e., 1986–2010).

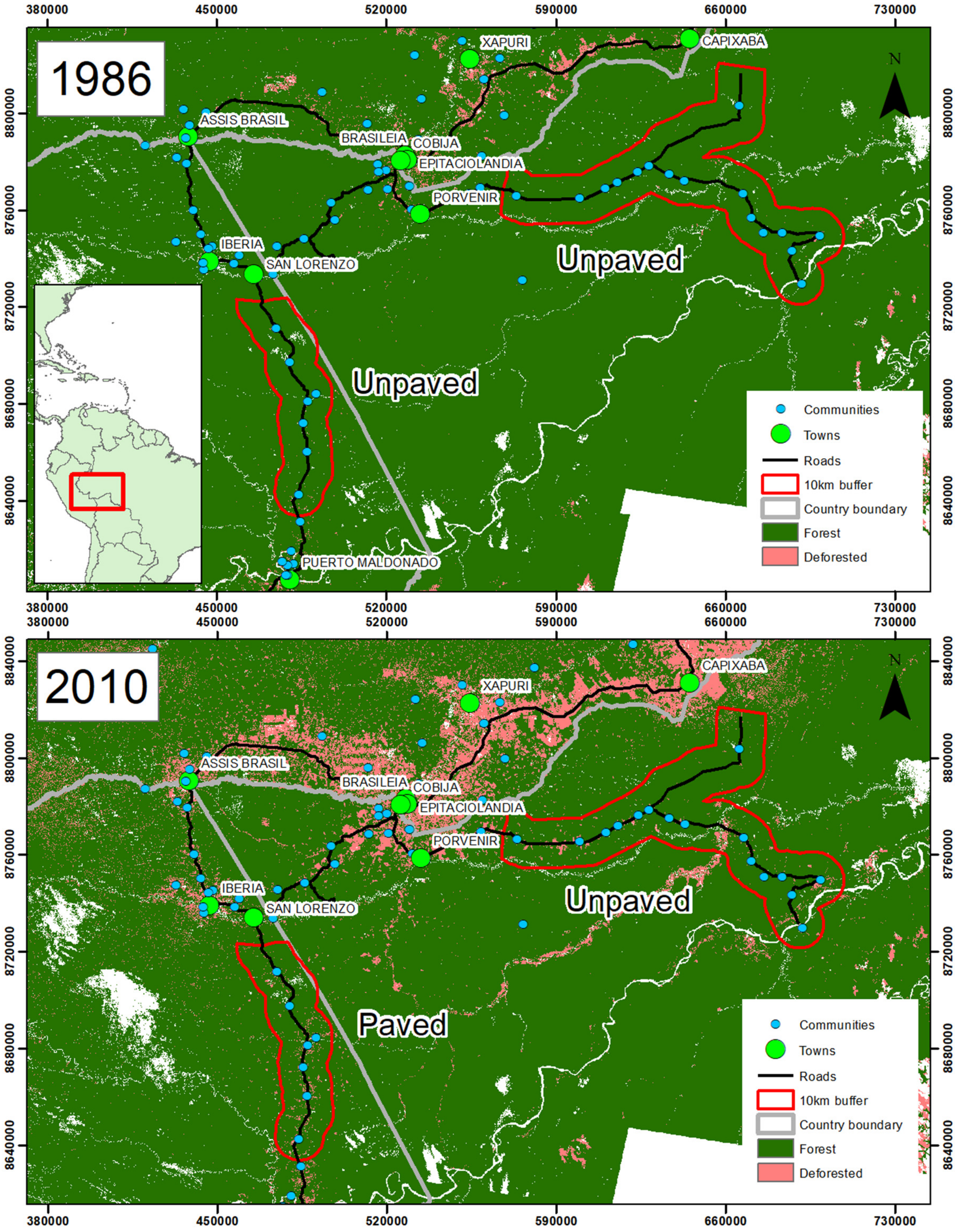


Figure 2. Map of the selected road segments, superimposed on deforestation maps at the beginning and end of the study period (1986 and 2010, respectively). Inset displays the approximate location of the Brazil, Peru, and Bolivia tri-national region. Tick marks show WGS84 UTM Zone 19S projected coordinates. White areas in the background correspond to clouds, water, or shadows that were removed.

The forest/nonforest classification used Landsat images (4 and 5 TM and 7 ETM+) of the area from 1986, 1991, 1996, 2000, 2005, and 2010. Atmospheric and seasonal difference corrections were applied; all images were georeferenced to less than 15 m error and mosaicked; and clouds, shadows, and water were removed with a Principal Component Analysis (PCA) image differencing and thresholding method. Mosaics were classified using a rule-based classifier (Compumine and ERDAS Knowledge Engineer), using an 85/15 split sample to train and test the decision trees. Covariates for the classification included bands 4, 5, and 7; tasseled cap brightness, greenness and wetness indices; a mid-infrared index and a 3 × 3 moving window calculation of image variance to measure texture, helpful to classify forest and non-forest areas. Classification accuracy for the 2005 mosaic was 87.85% based on field data and 97.96% based on ASTER imagery. Additional details regarding this classification can be found in [20]. Information regarding the location of communities, towns, and roads, including the status and timing of road paving, were collected via fieldwork that included interviews with local leaders and long-time residents [50]. Because roads typically increase deforestation rate within 5.5 km from the road [19], we relied on data within 10 km of these road segments to ensure that all the deforestation effects of these road segments were captured (red polygons in Figure 2). Within this buffer area, encompassing a region of approximately 6500 km^2, the proportion of deforested pixels in the data used for modeling varied from 0.3% to 4.4% (Table 1).

Table 1. Percentage of deforested pixels for each year and road segment.

Year	Road Paved after 2005	Unpaved Road
1991	1.3	0.5
1996	2.0	0.3
2000	2.9	1.2
2005	2.5	1.3
2010	4.4	3.8

2.3.1. Inference on the Effect of Road Proximity and Road Paving on Deforestation Risk

Despite the relatively small study area, there were still too many pixels (7.2 million) for our statistical analysis. Therefore, we relied on a CC sampling approach to subsample these pixels by choosing a total of 500,000 pixels in each year and each road segment. For this subsampling, we focused only on pixels that were susceptible to deforestation (i.e., pixels that were forested in the previous time step). Because the goal in this section is to infer the effect of proximity to road and the effect of road paving on deforestation risk, we relied on GAMs. More specifically, we assumed that the deforestation status of pixel i at time t y_{it} (0 = forested, 1 = deforested) arises from a Bernoulli distribution with deforestation probability given by π_{it}:

$$y_{it} \sim Bernoulli(\pi_{it}) \tag{16}$$

$$\pi_{it} = \frac{\exp\left(\widetilde{\beta}_{0t} + \beta_{1t}p_{i(t-1)} + f_t^{Road}\left(D_i^{Road}\right) + f_t^{Commun}\left(D_i^{Commun}\right)\right)}{1 + \exp\left(\widetilde{\beta}_{0t} + \beta_{1t}p_{i(t-1)} + f_t^{Road}\left(D_i^{Road}\right) + f_t^{Commun}\left(D_i^{Commun}\right)\right)} \tag{17}$$

where D_i^{Road} and D_i^{Commun} are the distances from pixel i to the nearest road and community, respectively, and f_t^{Road} and f_t^{Commun} are the corresponding smooth functions based on thin-plate splines. In this equation, the spatially contagious effect of deforestation [51] is accounted for by the term $p_{i(t-1)}$, which is the proportion of the 8 nearest neighbor pixels that were not forested in the previous year. The subscript t in the smooth functions and parameters emphasize that separate GAM models were fit for each year.

2.3.2. Predicting Deforestation Risk

In this section, we rely on machine learning methods to predict deforestation. More specifically, based on the deforestation data from the MAP frontier region, we used two approaches to assess if the random forest trained with CC data yields better predictions when compared to the same model trained with RS data. The first procedure evaluated predictive ability by relying on a standard 10-fold cross-validation approach. In this approach, we randomly divided the data into 10 non-overlapping parts, where one part was reserved to evaluate out-of-sample predictive skill and 9 parts were available as training data. Based on the observations that were available for training, we sampled 100,000 pixels for each year, either using the RS or the CC approach, and used these data to train our random forest model. Then, the trained models were used to predict the deforestation status of the withheld data.

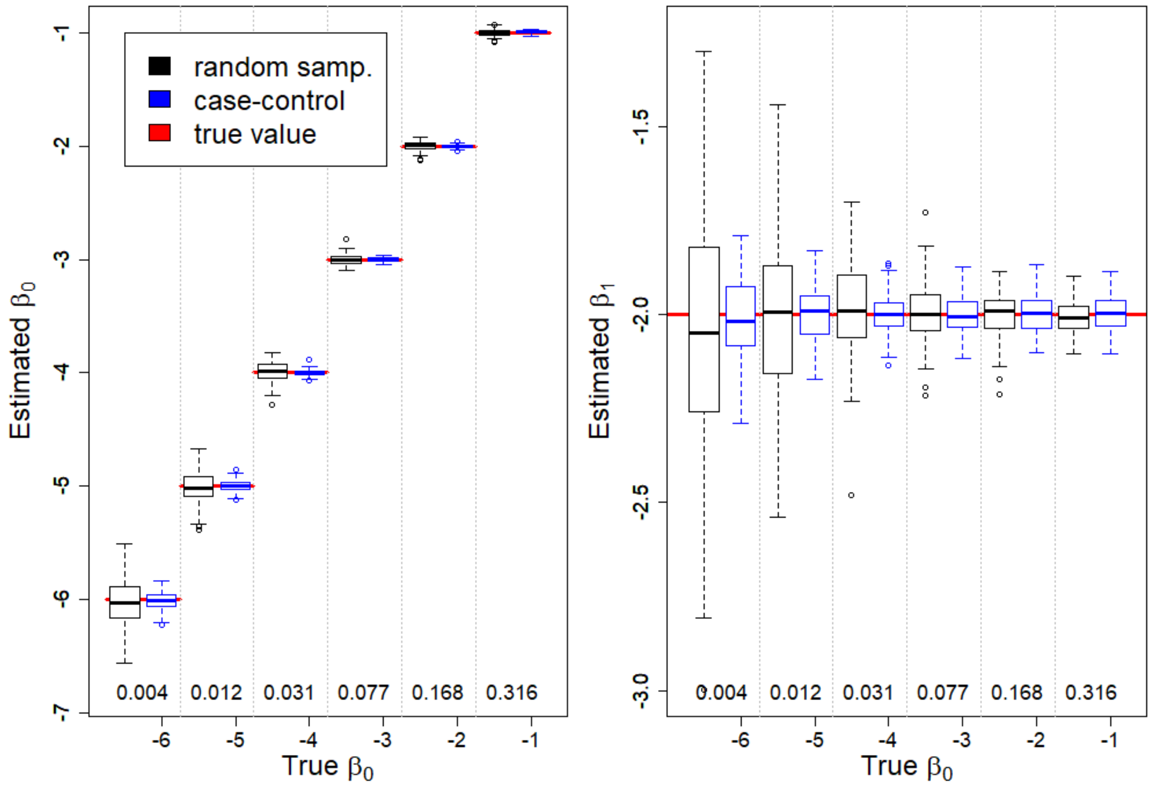
The second procedure evaluated predictive ability by training the model on present deforestation data and using this model to predict future deforestation. More specifically, we first trained our models using predictor variables from 1986 and deforestation status information from 1991. Then, we used this model together with predictor variables from 1991 to predict deforestation status in 1996. Similarly, we also trained our models on predictor variables from 2000 and deforestation status information from 2005. Then, we used this model together with predictor variables from 2005 to predict deforestation status in 2010. To fit these models, we sampled 100,000 pixels from the training data using either the RS or the CC approach. Finally, predictions were made on a random sample of 500,000 pixels. This procedure was repeated 10 times for predicting deforestation in 1996 and 10 times for predicting deforestation in 2010.

For both prediction exercises, predictor variables consisted of distances to the nearest road segment, community and town, the proportion of neighboring deforested pixels at the previous time step, and the latitude and longitude coordinates of the pixel. Only pixels susceptible to deforestation (i.e., that were forested in the previous time step) were used to train and validate the model. Similar to the simulated data example, we rely on the random forest regression algorithm, instead of the usual classification random forest, because it has been shown that this regression approach generates consistent probability estimates [39]. Finally, to compare model results, we calculate the log-likelihood of the withheld data under the assumption that the binary deforestation status observations come from a Bernoulli distribution with the estimated success probability. Models that result in higher log-likelihood have higher out-of-sample predictive skill.

3. Results

3.1. Simulations

Based on the first set of simulated data, which was focused on estimating parameters from the logistic regression, we found that fitting the model to data sampled using a CC approach yields unbiased estimates for β_1. Furthermore, after correcting for the bias in the estimated intercept, we found that it is possible to generate reliable inference for β_0 as well (Figure 3). Importantly, we found that the CC approach generally results in less variable β_0 and β_1 estimates when compared to the RS approach, with increasingly better results when compared to RS as the proportion of cases decreases (i.e., for smaller true β_0).



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

Figure 3. Data sampled using a case-control (CC) approach yielded unbiased estimates of logistic regression parameters (intercept β_0 and slope β_1) and will generally yield more precise estimates when compared to data sampled using a random sampling (RS) approach. Panels (**a**) and (**b**) display the estimated intercepts and slope parameters, respectively. True parameter values are represented by thick horizontal red lines and estimated parameters based on simulated datasets originating from the RS and CC control approaches are shown with black and blue boxplots, respectively. The average proportion of cases in the original dataset before sampling took place is given by the numbers at the bottom. Results are based on 100 simulated datasets for each value of true β_0.

We used GAMs to estimate nonlinear associations in the second set of simulated data. We found that using data sampled through a CC approach can yield unbiased inference regarding the nonlinear relationship between the binary response variable y and the predictor x_1, while statistically adjusting for the effect of the nuisance covariate x_2. Importantly, we also find that data sampled through a CC approach yield much more precise inference about this nonlinear relationship than a similarly sized dataset sampled using the RS approach, particularly as the proportion of cases decreases (i.e., for smaller and smaller β_0; Figure 4).

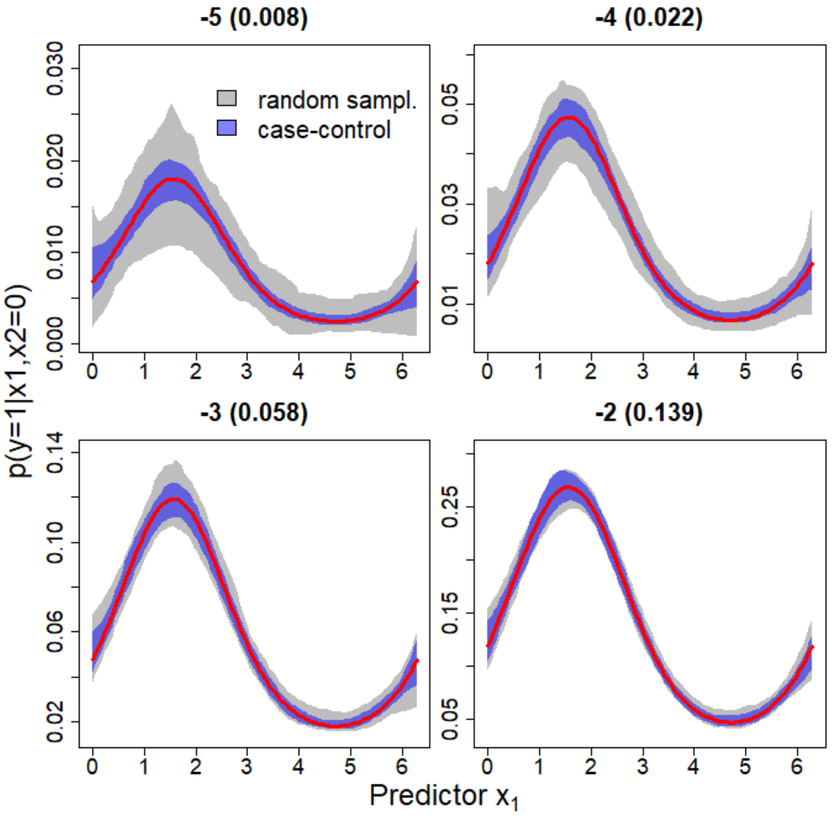


Figure 4. A case-control (CC) approach yielded unbiased and more precise inference on the nonlinear relationship between the probability of the response binary response variable y and the predictor x_1 using GAMs. Each panel displays modeling results for data simulated with a particular value of β_0. The number on top of each panel refers to the value for β_0 while the proportion of cases prior to sampling is given between parentheses. For each β_0 value, grey and blue polygons represent 95% point-wise envelopes based on 100 simulated datasets generated through a random sampling (RS) approach or a CC approach, respectively. The true relationship between $p(y = 1 | x_1, x_2 = 0)$ and x_1 is depicted by the thick red line.

For the third set of simulations, which was focused on predictive skill using random forests, we found that the mean absolute error between the true and estimated probabilities $p(y_i = 1 | x)$ was almost always smaller for results based on the CC versus RS approach for datasets of the same size (Figure 5). Unadjusted probabilities $p(y = 1 | z = 1, x)$ based on the CC data were also substantially worse than the probabilities estimated based on RS data (results not shown), suggesting that our adjustment to convert $p(y = 1 | z = 1, x)$ to $p(y = 1 | x)$ is critical for the generation of well-calibrated probabilities.

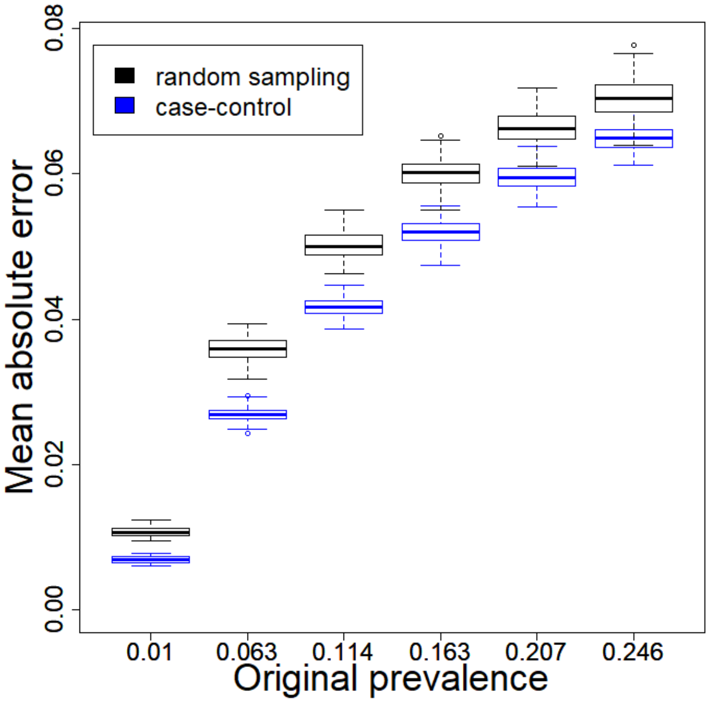


Figure 5. The random forest model trained with case-control (CC) data and using adjusted probabilities yields better probability estimates when compared to the same model trained with randomly sampled (RS) data. Data with an increasing prevalence of cases were generated by systematically varying σ^2 from 0.1 to 2.6. The mean absolute error was calculated by comparing the estimated probabilities to the true probabilities and lower values indicate better performance. The displayed results are based on 100 simulated datasets for each value of σ^2.

3.2. Case study on Deforestation in the Amazon Region

3.2.1. Modeling the Effect of Proximity to Road and the Effect of Road Paving on Deforestation Risk

We found a clear influence of proximity to the road on the probability of deforestation, as expected (Figure 6). Interestingly, we also found substantial temporal variation. For example, for the paved road segment, deforestation probability in 2000 was almost as high as in 2010 despite the road pavement affecting only this latter year (Figure 6a). Differently from what we originally expected, the road segment that was never paved tended to have a much higher deforestation probability in 2005 and 2010 when compared to the paved road (Figure 6f,g), after statistically controlling for potential confounders (e.g., distance to communities). These findings highlight substantial differences between road segments in terms of the effect of proximity to roads on deforestation probability, suggesting that the chosen unpaved road segment might not be a reasonable "control" road for the paved road segment.

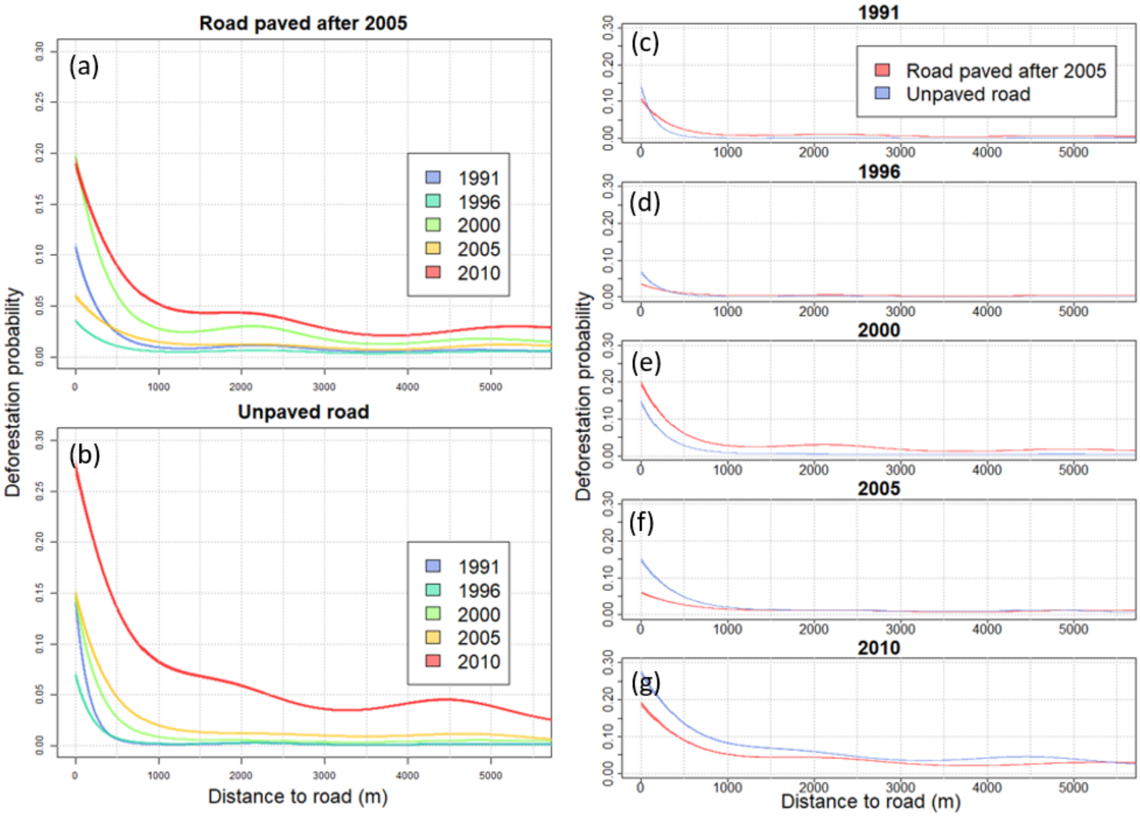
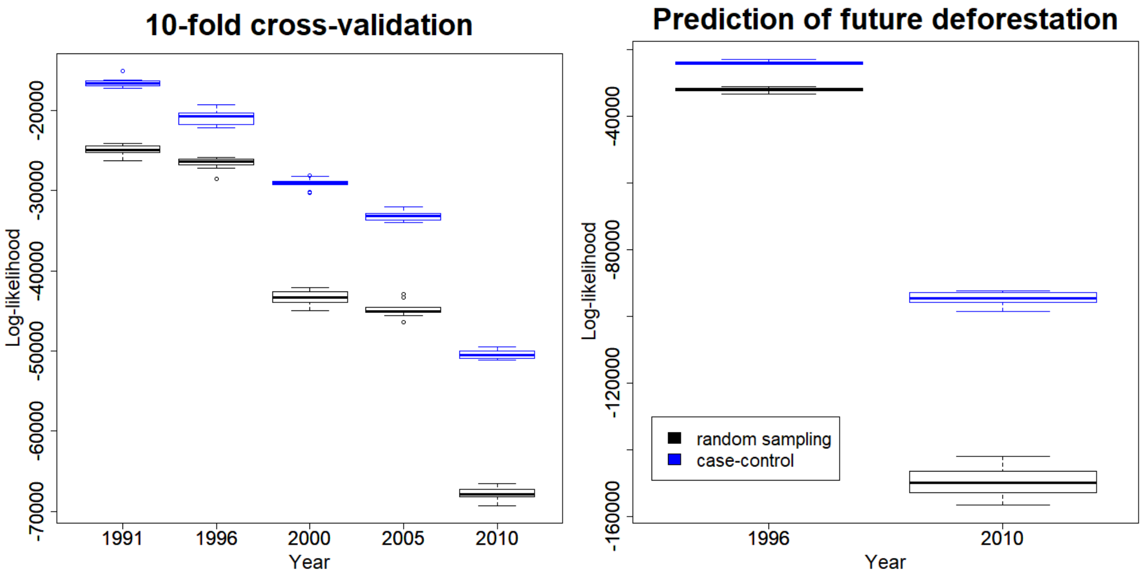


Figure 6. Proximity to roads increase deforestation probability, but this relationship varies substantially from year to year and from one road segment to another in the MAP frontier, southwestern Amazon. Panels (**a**) and (**b**) highlight temporal variability within each road segment while panels (**c**)–(**g**) compare both road segments within individual years. Separate GAM models were fitted to data containing 500,000 pixels for each year and each road segment. To depict the relationship between D_i^{Road} and deforestation risk $\hat{\pi}_{it}$, we created a dataset in which we systematically varied D_i^{Road} while keeping $p_{i(t-1)}$ and D_i^{Commun} fixed to their mean values. We used this dataset, together with the unbiased intercept $\beta_{0t} = \widetilde{\beta}_{0t} - log\left(\frac{p_{1t}}{p_{0t}}\right)$ and the estimated smooth functions, to calculate $\hat{\pi}_{it}$.

3.2.2. Predicting Deforestation Risk

The 10-fold cross-validation exercise performed for each year revealed that the random forest model, trained on CC data and with adjusted deforestation probabilities, consistently yielded better predictions when compared to the same model trained on RS data (left panel in Figure 7). However, a better test of predictive skill consists of using models trained on past data to make predictions about future deforestation. We found that the random forest model trained on CC data using the adjusted probabilities was also the best performing model in predicting future deforestation in 1996 and 2010 (right panel in Figure 7). In all scenarios, using the adjusted probability for the model fitted to CC data consistently resulted in better performance when compared to using the raw unadjusted probabilities from the same model (results not shown), reiterating the importance of our adjustment to obtain well-calibrated probabilities.

The depiction of the predicted deforestation probabilities reveals a spatially heterogeneous pattern of deforestation probabilities that is not explained by distance to road alone. For example, the random forest model identified substantially increased deforestation risk at the southern edge for the road segment that was paved after 2005, close to the regional capital Puerto Maldonado, a pattern that is particularly evident in the 2010 predictions (top panels in Figure 8). Similarly, for the unpaved road segment (bottom panels in Figure 8), aside from the greater deforestation risk in the immediate vicinity of the road, there was relatively little spatial pattern in 1996 whereas in 2010 there was greater deforestation risk in the southeastern edge of this road segment, close to the town of Puerto Rico and the port town of El Sena.



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

Figure 7. The random forest model fitted using case-control (CC) data and with adjusted probabilities (blue boxes) consistently had better performance when compared to the same model fitted to randomly sampled (RS) data (black boxes). Results from the 10-fold cross-validation exercise and the prediction of future deforestation are shown in panels (**a**) and (**b**), respectively. For each year, 10 different models were fitted to 10 different samples from the training data containing 100,000 pixels. Higher values for the log-likelihood indicate better predictive performance.

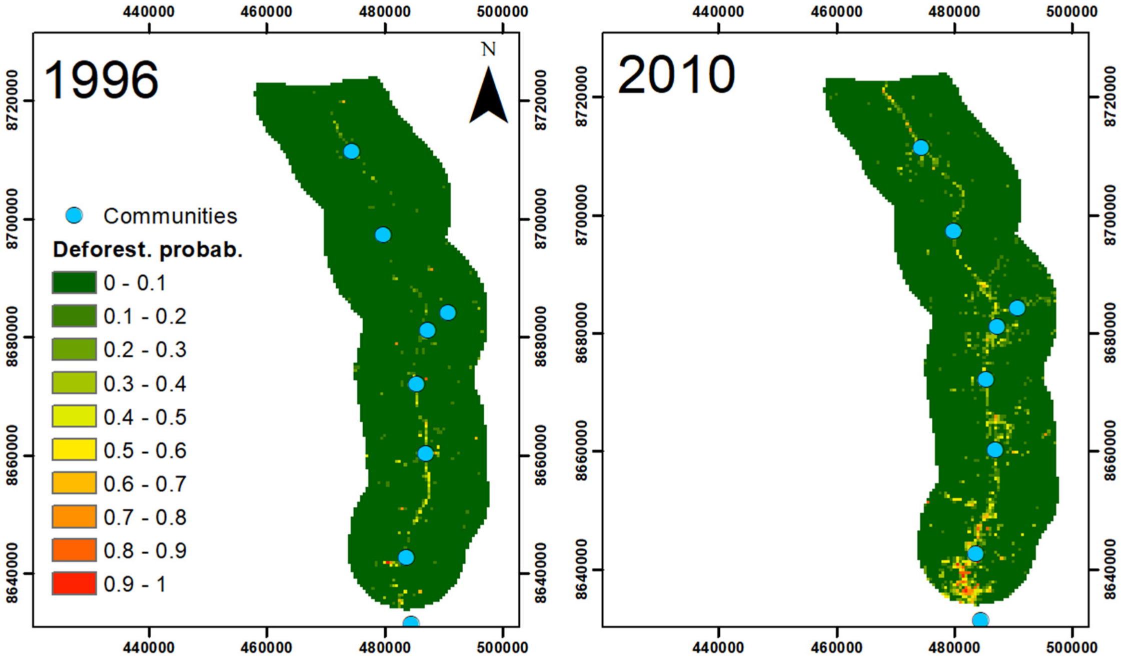


Figure 8. *Cont.*

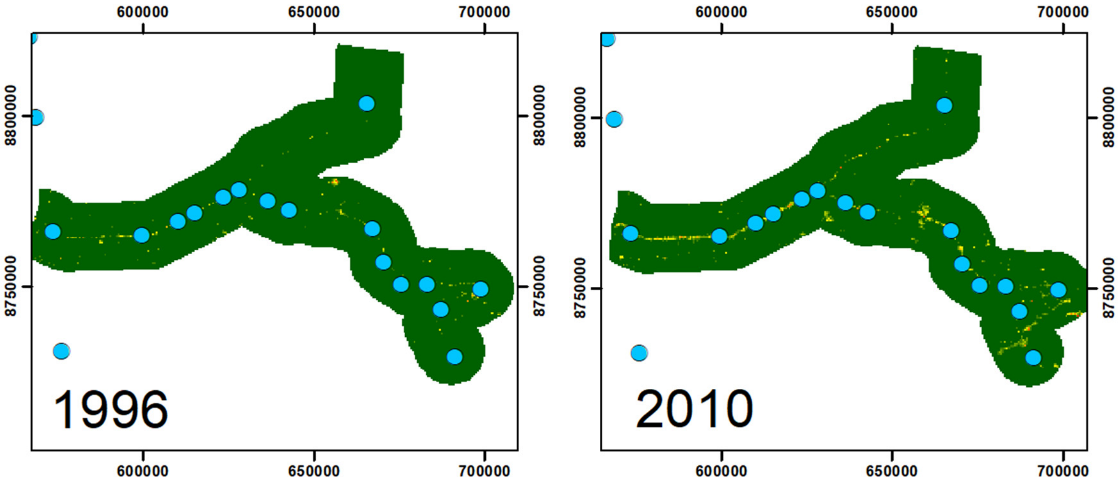


Figure 8. Spatial depiction of the deforestation probabilities predicted by the random forest model fitted using case-control (CC) data. Top panels display deforestation predictions for the road segment that was paved after 2005 while the bottom panels display predictions for the unpaved road segment. Left and right panels display predictions for 1996 and 2010, respectively. Tick marks show the WGS84 UTM Zone 19S projected coordinates. Deforestation probabilities are shown with the background heat map.

4. Discussion

We have shown how inference and prediction based on massive binary remote sensing data can be substantially improved by sampling pixels using a CC approach, particularly when the land-use category of interest (e.g., recently deforested pixels) is rare, if parameters and predictions are adjusted using the bias-correction equations that we derived (Equations (6) and (11)). More specifically, we show both with simulations and our case study how the parameters from a semiparametric model based on logistic regression (i.e., GAMs) can be adjusted to flexibly model the relationship between deforestation risk and covariates in an unbiased fashion. Finally, we show how predictions are improved when the deforestation probabilities estimated by machine learning methods (i.e., random forest) are adjusted to avoid the bias introduced by the CC sampling approach.

Our case study highlighted substantial year-to-year variability regarding the effect of proximity to roads on deforestation risk. As a result, determining the effect of road paving by making before-and-after comparisons is challenging. Importantly, our results reveal substantial spatial differences. The paved road segment in Madre de Dios is to a large extent surrounded by forest concessions for the extraction of castanha nuts (*Bertholletia excelsa*), a valuable non-timber forest product, where deforestation is illegal. Indeed, Perz, *et al.* [52] reveal that communities in this zone had a much lower deforestation rate than elsewhere in the region. While the unpaved road segment in Pando also has a high density of castanha trees, the government of Evo Morales identified this region as appropriate for agricultural colonization. Furthermore, this area is much closer to the Brazilian frontier and likely suffers from a greater influence of the Brazilian ranching culture than the paved road segment. As a result, greater deforestation might have occurred due to greater pasture formation associated with Bolivians adopting Brazilian ranching practices or Brazilians informally buying lands across the border [53]. Despite the relatively small region analyzed, this cross-border process, together with differences in land-tenure (e.g., forest concessions), highlight the complexity associated with understanding how proximity to roads and road paving influence land-use in this highly dynamic tri-national region.

The importance of increasing the prevalence of the rare category in highly imbalanced data has long been acknowledged in the literature focused on predictive modeling [24,27–32]. Indeed, a wide variety of sampling methods have been proposed in the literature to deal with class imbalance, including the CC approach advocated here as well as random oversampling of instances of the minority class (ROS), one-side selection (OSS), cluster-based oversampling (CBOS), Wilson's editing (WE), Synthetic Minority Oversampling Technique (SMOTE), and borderline-SMOTE (BSMOTE) [28,29,54–56]. van Hulse et al. [28] performed a comprehensive comparative study of these sampling approaches, involving 35 different benchmark datasets, seven sampling techniques and 11 commonly used learning algorithms. In their study, they found that "intelligent" sampling techniques (e.g., SMOTE, BSMOTE, WE, OSS, and CBOS) often have inferior performance and that CC was overall the best method, followed by ROS. We experimented with ROS in a preliminary analysis but found that, in the absence of adjustments similar to the ones we have developed for the CC approach, ROS yields substantially biased predictions (data not shown). Appropriately accounting for the biases introduced by these alternative sampling approaches to increase prevalence remains an important challenge [24,33].

Our results illustrate how the biases introduced by the CC approach can be accounted for, and thus offset, when estimating parameters using logistic regression models. While the reliance on logistic regression for inference might seem to be an important limitation, we have shown that flexible modeling frameworks that build off the basic logistic regression structure, such as geospatial models and GAMs [37,57], can still be successfully used together with our CC methodology. Importantly, we also show that the CC sampling approach, together with the proposed adjustments to the estimated probabilities, can lead to improved predictions from a wide range of models as long as these models are able to provide consistent probability estimates. Examples of machine learning models that have this characteristic include some versions of random forest, k-nearest neighbors, and support vector machines [33,39].

5. Conclusions

To our knowledge, this is the first time that bias-correction adjustments for data sampled through a CC approach have been proposed and used for modeling highly imbalanced land-use data, a common phenomenon in land science. We have shown the utility of the CC approach over the RS approach in generating substantially improved inference and predictions. While we focused on the topic of land cover change, we emphasize that our equations together with the CC sampling approach are applicable to other domains of inquiry that involve big imbalanced datasets. For example, this methodology can help scientists better understand the drivers of, and create predictions for, other relatively rare landscape phenomena, such as remotely sensed fires [58,59] and landslides [60]. In ecology, the proposed methodology might improve the development of species occurrence maps given that these maps are based on spatial predictions from species distribution models [31,61]. More generally, accurate predictions based on large but highly imbalanced data are also needed in fields outside ecology and land sciences, such as predictions of credit card fraud detection [62] and customers leaving a company [54]. In short, the proposed equations together with the CC sampling approach will be of wide use to scientists in multiple fields interested in understanding the drivers of rare outcomes and predicting them in space and time.

Author Contributions: Conceptualization, D.V.; methodology, D.V.; software, D.V.; validation, D.V.; formal analysis, D.V.; investigation, D.V.; resources, D.V.; data curation, M.M.; writing—original draft preparation, D.V.; writing—review and editing, D.V., J.H., M.M., and S.P.; visualization, D.V.; supervision, D.V.; project administration, D.V.; funding acquisition, S.P. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Science Foundation Human and Social Dynamics Program, Grant #0527511, and by the US Agency for International Development, Latin America and Caribbean program in Environment, Cooperative Agreements RLA-A-00-06-00071-00 and 512-A-00-08-00003-00. The conduct of the analysis and interpretation of the findings are the responsibility of the authors.

Acknowledgments: We thank the comments provided by Gilson Shimizu, Rafael Izbicki, Kok Ben Toh, and Josh Cullen.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Bakker, D.C.E.; Pfeil, B.; Landa, C.S.; Metzl, N.; O'Brien, K.M.; Olsen, A.; Smith, K.; Cosca, C.; Harasawa, S.; Jones, S.D.; et al. A multi-decade record of high-quality fCO2 data in version 3 of the Surface Ocean CO2 Atlas (SOCAT). *Earth Syst. Sci. Data* **2016**, *8*, 383–413. [CrossRef]
2. Richardson, A.D.; Hufkens, K.; Milliman, T.; Aubrecht, D.M.; Chen, M.; Gray, J.M.; Johnston, M.R.; Keenan, T.F.; Klosterman, S.T.; Kosmala, M.; et al. Tracking vegetation phenology across diverse North American biomes using PhenoCam imagery. *Sci. Data* **2018**, *5*, 180028. [CrossRef]
3. WCS. A New Cloud Platform Unveils the Most Diverse Camera Trap Database in the World. Available online: https://newsroom.wcs.org/News-Releases/articleType/ArticleView/articleId/13593/A-New-Cloud-Platform-Unveils-the-Most-Diverse-Camera-Trap-Database-in-the-World.aspx (accessed on 6 February 2020).
4. Wulder, M.A.; Loveland, T.R.; Roy, D.P.; Crawford, C.J.; Masek, J.G.; Woodcock, C.E.; Allen, R.G.; Anderson, M.C.; Belward, A.S.; Cohen, W.B.; et al. Current status of Landsat program, science, and applications. *Remote. Sens. Environ.* **2019**, *225*, 127–147. [CrossRef]
5. Zhou, Y.; Smith, S.J.; Zhao, K.; Imhoff, M.; Thomson, A.; Bond-Lamberty, B.; Asrar, G.R.; Zhang, X.; He, C.; Elvidge, C.D. A global map of urban extent from nightlights. *Environ. Res. Lett.* **2015**, *10*, 054011. [CrossRef]
6. Asner, G.P.; Knapp, D.E.; Broadbent, E.N.; Oliveira, P.J.C.; Keller, M.; Silva, J.N. Selective logging in the Brazilian Amazon. *Science* **2005**, *310*, 480–482. [CrossRef] [PubMed]
7. Pekel, J.-F.; Cottam, A.; Gorelick, N.; Belward, A.S. High-resolution mapping of global surface water and its long-term changes. *Nature* **2016**, *540*, 418–422. [CrossRef] [PubMed]
8. Parkinson, C.L. A 40-y record reveals gradual Antarctic sea ice increases followed by decreases at rates far exceeding the rates seen in the Artic. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 14414–14423. [CrossRef] [PubMed]
9. Bunting, P.; Rosenqvist, A.; Lucas, R.M.; Rebelo, L.-M.; Hilarides, L.; Thomas, N.; Hardy, A.; Itoh, T.; Shimada, M.; Finlayson, C.M. The global mangrove watch—A new 2010 global baseline of mangrove extent. *Remote Sens.* **2019**, *10*, 1669. [CrossRef]
10. Southgate, D.; Sierra, R.; Brown, L. The causes of tropical deforestation in Ecuador: A statistical analysis. *World Dev.* **1991**, *19*, 1145–1151. [CrossRef]
11. Pfaff, A.S.P. What drivers deforestation in the Brazilian Amazon? *J. Environ. Econ. Manag.* **1999**, *37*, 26–43. [CrossRef]
12. Jusys, T. Fundamental causes and spatial heterogeneity of deforestation in Legal Amazon. *Appl. Geogr.* **2016**, *75*, 188–199. [CrossRef]
13. Soares-Filho, B.S.; Nepstad, D.; Curran, L.M.; Cerqueira, G.C.; Garcia, R.A.; Ramos, C.A.; Voll, E.; McDonald, A.; Lefebvre, P.; Schlesinger, P. Modelling conservation in the Amazon basin. *Nature* **2006**, *440*, 520–523. [CrossRef] [PubMed]
14. Aguiar, A.P.D.; Camara, G.; Escada, M.I.S. Spatial statistical analysis of land-use determinants in the Brazilian Amazonia: Exploring intra-regional heterogeneity. *Ecol. Model.* **2007**, *209*, 169–188. [CrossRef]
15. Laurance, W.F.; Albernaz, A.K.M.; Schroth, G.; Fearnside, P.M.; Bergen, S.; Venticinque, E.M.; da Costa, C. Predictors of deforestation in the Brazilian Amazon. *J. Biogeogr.* **2002**, *29*, 737–748. [CrossRef]
16. Chomitz, K.M.; Gray, D.A. Roads, land use, and deforestation: A spatial model applied to Belize. *World Bank Econ. Rev.* **1996**, *10*, 487–512. [CrossRef]
17. Ludeke, A.K.; Maggio, R.C.; Reid, L.M. An analysis of anthropogenic deforestation using logistc regression and GIS. *J. Environ. Manag.* **1990**, *31*, 247–259. [CrossRef]
18. Green, J.M.H.; Larrosa, C.; Burgess, N.D.; Balmford, A.; Johnston, A.; Mbilinyi, B.P.; Platts, P.J.; Coad, L. Deforestation in an African biodiversity hotspot: Extent, variation and the effectiveness of protected areas. *Biol. Conserv.* **2013**, *164*, 62–72. [CrossRef]
19. Barber, C.P.; Cochrane, M.A.; Souza, C.M., Jr.; Laurance, W.F. Roads, deforestation, and the mitigating effect of protected areas in the Amazon. *Biol. Conserv.* **2014**, *177*, 203–209. [CrossRef]
20. Southworth, J.; Marsik, M.; Qiu, Y.; Perz, S.; Cumming, G.; Stevens, F.; Rocha, K.; Duchelle, A.; Barnes, G. Roads as drivers of change: Trajectories across the tri-national frontier in MAP, the southwestern Amazon. *Remote Sens.* **2011**, *3*, 1047–1066. [CrossRef]
21. Sales, M.; de Bruin, S.; Herold, M.; Kyriakidis, P.; Souza, C., Jr. A spatiotemporal geostatistical hurdle model approach for short-term deforestation prediction. *Spat. Stat.* **2017**, *21*, 304–318. [CrossRef]

22. Mertens, B.; Poccard-Chapuis, R.; Piketty, M.-G.; Lacques, A.-E.; Venturieri, A. Crossing spatial analyses and livestock economics to understand deforestation processes in the Brazilian Amazon: The case of Sao Felix do Xingu in south Para. *Agric. Econ.* **2002**, *27*, 269–294.
23. Echeverria, C.; Coomes, D.A.; Hall, M.; Newton, A.C. Spatially explicit models to analyze forest loss and fragmentation between 1976 and 2020 in southern Chile. *Ecol. Model.* **2008**, *212*, 439–449. [CrossRef]
24. Cushman, S.A.; Macdonald, E.A.; Landguth, E.L.; Malhi, Y.; Macdonald, D.W. Multiple-scale prediction of forest loss risk across Borneo. *Landsc. Ecol.* **2017**, *32*, 1581–1598. [CrossRef]
25. Voight, C.; Hernandez-Aguilar, K.; Garcia, C.; Gutierrez, S. Predictive modeling of future forest cover change patterns in southern Belize. *Remote Sens.* **2019**, *11*, 823. [CrossRef]
26. Pijanowski, B.C.; Tayyebi, A.; Doucette, J.; Pekin, B.K.; Braun, D.; Plourde, J. A big data urban growth simulation at a national scale: Configuring the GIS and neural network based Land Transformation Model to run in a High Performance Computing (HPC) environment. *Environ. Model. Softw.* **2014**, *51*, 250–268. [CrossRef]
27. Kuhn, M.; Johnson, K. Chapter 16. Remedies for Severe Class Imbalance. In *Applied Predictive Modeling*; Springer: New York, NY, USA, 2016.
28. Van Hulse, J.; Khoshgoftaar, T.M.; Napolitano, A. Experimental perspectives on learning from imbalanced data. In Proceedings of the 24th International Conference on Machine Learning, Corvallis, OR, USA, 20–24 June 2007; pp. 935–942.
29. Lemaitre, G.; Nogueira, F.; Aridas, C.K. Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning. *J. Mach. Learn. Res.* **2017**, *18*, 1–5.
30. Salas-Eljatib, C.; Fuentes-Ramirez, A.; Gregoire, T.G.; Altamirano, A.; Yaitul, V. A study on the effects of unbalanced data when fitting logistic regression models in ecology. *Ecol. Indic.* **2018**, *85*, 502–508. [CrossRef]
31. McPherson, J.M.; Jetz, W.; Rogers, D.J. The effects of species' range sizes on the accuracy of distribution models: Ecological phenomenon or statistical artefact? *J. Appl. Ecol.* **2004**, *41*, 811–823. [CrossRef]
32. Maggini, R.; Lehmann, A.; Zimmermann, N.E.; Guisan, A. Improving generalized regression analysis for the spatial prediction of forest communities. *J. Biogeogr.* **2006**, *33*, 1729–1749. [CrossRef]
33. Kruppa, J.; Liu, Y.; Biau, G.; Kohler, M.; Konig, I.R.; Malley, J.D.; Ziegler, A. Probability estimation with machine learning methods for dichotomous and multicategory outcome: Theory. *Biom. J.* **2014**, *4*, 534–563. [CrossRef]
34. Breslow, N.E. Statistics in epidemiology: The case-control study. *J. Am. Stat. Assoc.* **1996**, *91*, 14–28. [CrossRef] [PubMed]
35. King, G.; Zeng, L. Logistic regression in rare events data. *Political Anal.* **2001**, *9*, 137–163. [CrossRef]
36. Agresti, A. *Categorical Data Analysis*; John Wiley & Sons: Hoboken, NJ, USA, 2003.
37. Wood, S.N. *Generalized Additive Models: An Introduction with R*; CRC Press: Boca Raton, FL, USA, 2017; p. 476.
38. Liaw, A.; Wiener, M. Classification and regression by randomForest. *R News* **2002**, *2*, 18–22.
39. Malley, J.D.; Kruppa, J.; Dasgupta, A.; Malley, K.G.; Ziegler, A. Probability machines: Consistent probability estimation using nonparametric learning machines. *Methods Inf. Med.* **2012**, *51*, 74–81. [CrossRef] [PubMed]
40. Mittermeier, R.A.; Mittermeier, C.G.; Brooks, T.M.; Pilgrim, J.D.; Konstant, W.R.; da Fonseca, G.A.B.; Kormos, C. Widerness and biodiversity conservation. *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 10309–10313. [CrossRef] [PubMed]
41. Davidson, E.A.; Artaxo, P. Globally significant changes in biological processes of the Amazon Basin: Results of the Large-scale Biosphere–Atmosphere Experiment. *Glob. Chang. Biol.* **2004**, *10*, 519–529. [CrossRef]
42. Foley, J.A.; Asner, G.P.; Costa, M.H.; Coe, M.T.; DeFries, R.; Gibbs, H.K.; Howard, E.A.; Olson, S.; Patz, J.; Ramankutty, N.; et al. Amazonia revealed: Forest degradation and loss of ecosystem goods and services in the Amazon Basin. *Front. Ecol. Environ.* **2007**, *5*, 25–32. [CrossRef]
43. Malhi, Y.; Roberts, J.T.; Betts, R.A.; Killeen, T.J.; Li, W.; Nobre, C.A. Climate change, deforestation, and the fate of the Amazon. *Science* **2008**, *319*, 169–172. [CrossRef]
44. Tundisi, J.G.; Goldemberg, J.; Matsumura-Tundisi, T.; Saraiva, A.C.F. How many more dams in the Amazon? *Energy Policy* **2014**, *74*, 703–708. [CrossRef]
45. Hyde, J.; Bohlman, S.; Valle, D. Transmission lines are an under-acknowledged conservation threat to the Brazilian Amazon. *Biol. Conserv.* **2018**, *228*, 343–356. [CrossRef]

46. Spring, J. Bolsonaro-backed Highway Targets Heart of Brazil's Amazon. Available online: https://www.reuters.com/article/us-brazil-environment-highway-insight/bolsonaro-backed-highway-targets-heart-of-brazils-amazon-idUSKBN1WH0Z3 (accessed on 28 February 2019).
47. Amigo, I. The Amazon's fragile future. *Nature* **2020**, *578*, 506–507.
48. Barlow, J.; Berenguer, E.; Carmenta, R.; Franca, F. Clarifying Amazonia's burning crisis. *Glob. Chang. Biol.* **2020**, *26*, 319–321. [CrossRef] [PubMed]
49. Marsik, M.; Stevens, F.R.; Southworth, J. Amazon deforestation: Rates and patterns of land cover change and fragmentation in Pando, northern Bolivia, 1986 to 2005. *Prog. Phys. Geogr.* **2011**, *35*, 353–374. [CrossRef]
50. Perz, S.G.; Cabrera, L.; Carvalho, L.A.; Castillo, J.; Chacacanta, R.; Cossio, R.E.; Solano, Y.F.; Hoelle, J.; Perales, L.M.; Puerta, I.; et al. Regional integration and local change: Road paving, community connectivity, and social-ecological resilience in a tri-national frontier, southwestern Amazonia. *Reg. Environ. Chang.* **2012**, *12*, 35–53. [CrossRef]
51. Rosa, I.M.D.; Purves, D.; Souza, C., Jr.; Ewers, R.M. Predictive modelling of contagious deforestation in the Brazilian Amazon. *PLoS ONE* **2013**, *8*, e77231. [CrossRef] [PubMed]
52. Perz, S.G.; Qiu, Y.; Xia, Y.; Southworth, J.; Sun, J.; Marsik, M.; Rocha, K.; Passos, V.; Rojas, D.; Alarcon, G.; et al. Trans-boundary infrastructure and land cover change: Highway paving and community-level deforestation in a tri-national frontier in the Amazon. *Land Use Policy* **2013**, *34*, 27–41. [CrossRef]
53. Perz, S.; Chavez, A.B.; Cossio, R.; Hoelle, J.; Leite, F.L.; Rocha, K.; Rojas, R.O.; Shenkin, A.; Carvalho, L.A.; Castillo, J.; et al. Trans-boundary infrastructure, access connectivity, and household land use in a tri-national frontier in the Southwestern Amazon. *J. Land Use Sci.* **2015**, *10*, 342–368. [CrossRef]
54. Burez, J.; Van den Poel, D. Handling class imbalance in customer churn prediction. *Expert Syst. Appl.* **2009**, *36*, 4626–4636. [CrossRef]
55. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minotiry Over-sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
56. Weiss, G.M. Mining with rarity: A unifying framework. *ACM SIGKDD Explor. Newsl.* **2004**, *6*, 7–19. [CrossRef]
57. Paciorek, C.J. Computational techniques for spatial logistic regression with large datasets. *Comput. Stat. Data Anal.* **2007**, *51*, 3631–3653. [CrossRef] [PubMed]
58. Adeney, J.M.; Christensen, N.L.; Pimm, S.L. Reserves protect against deforestation fires in the Amazon. *PLoS ONE* **2009**, *4*, e5014. [CrossRef] [PubMed]
59. Zhang, H.; Qi, P.; Guo, G. Improvement of fire danger modelling with geographically weighted logistic model. *Int. J. Wildland Fire* **2014**, *23*, 1130–1146. [CrossRef]
60. Mathew, J.; Jha, V.K.; Rawat, G.S. Application of binary logistic regression analysis and its validation for landslide susceptibility mapping in part of Garhwal Himalaya, India. *Int. J. Remote Sens.* **2007**, *28*, 2257–2275. [CrossRef]
61. Barbet-Massin, M.; Jiguet, F.; Albert, C.H.; Thuiller, W. Selecting pseudo-absences for species distribution models: How, where and how many? *Methods Ecol. Evol.* **2012**, *3*, 327–338. [CrossRef]
62. Chan, P.K.; Stolfo, S.J. Towards scalable learning with non-uniform class and cost distributions: A case study in credit card fraud detection. In Proceedings of the KDD: Fourth International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 27–31 August 1998; pp. 164–168.