Evaluation of forest conservation programs in Amazonia rainforest in Brazil

Masters in Interdisciplinary Data Science Capstone Project Whitepaper

Zhenhua Wang & Vivek Sahukar

Faculty Lead:	Dr. Alex Pfaff
Conservation International:	Dr. Rachel Golden Kroner & Dr. Sebastien Costedoat
Capstone Director:	Dr. Gregory Herschlag
Project Manager:	Dr. Heather Huntington & Dr. Ryan Huang

Date: Apr 15, 2020

Abstract

Rainforest deforestation in Amazonia rainforest in Brazil is alarming and hence, the Brazilian government has implemented forest conservation programs such as Bolsa Verde. We evaluated the effectiveness of Bolsa Verde in Brazilian Amazonia rainforest in reducing forest cover loss in those areas which are also declared as Protected Areas. We used regression methods to determine the effect of various confounding factors on forest cover loss. We used matching techniques to balance the control and treatment groups in terms of values of the covariates. We found that Bolsa Verde is effective in reducing forest cover loss in certain areas only. We also concluded that covariates such as slope, elevation, distance to roads and cities are more helpful than air temperature and precipitation in studying the effects of Bolsa Verde.

Introduction

Rainforest deforestation is the second largest cause of climate change. Brazil has 60% of the Amazonia - which are the largest tropical rainforest in the world. According to the Data released by the Brazil's National Institute for Space Research: the annual forest cover loss in Brazil for 2018 (7900 km^2) was 12 times the size of New York City. This is the highest deforestation rate since 2008 and 30% increase over the deforestation rate in 2017. This forest loss is alarming and detrimental to climate change. Hence, the Brazilian government has implemented various forest conservation programs to reduce forest cover loss.

For this project, we are working with our stakeholder, Conservation International, which is an American non-profit environmental conservation organization. The CI's goal for Amazonia is "***to achieve zero net deforestation rate***" to protect essential resources, mitigate climate change and increase prosperity for people. Therefore, CI is making Global Conservation Atlas, which is a database of area-based conservation systems to identify and map natural capital. Evaluation studies have been done to analyse these forest conservation programs and understand their impact. **Hence, our project goal is to evaluate the forest conservation programs in Amazonia rainforest in Brazil from 2011 - 2018, that would help CI to understand the spatial trends in forest cover loss.**

Protected Areas (PA) and Payment for Ecosystem Services (PES) are popular conservation practices used by the government to reduce deforestation. As per International Union for Conservation of Nature (IUCN) Definition 2008: "A protected area (PA) is a clearly defined geographical space, recognized, dedicated and managed, through legal or other effective means, to achieve the long term conservation of nature with associated ecosystem services and cultural values." (IUCN Definition 2008). Payments for ecosystem services (PES), also known as payments for environmental services (or benefits), are incentives offered to farmers or landowners in exchange for managing their land to provide some sort of ecological service.

Bolsa Verde (BV), a type of PES, is a conditional cash transfer program that was implemented by Brazilian government from 2011 to 2018. The Brazilian government paid extremely poor households in exchange for the protection measures undertaken by those households. The condition for payment in BV was contingent upon maintaining at least 80% of the original forest cover in the area. In existing literature, studies focused on analysing the effect of Bolsa Verde at different scales. For instance, Po Yin et al. (2019) showed that Bolsa verde reduced forest loss by 44-53% using a difference-in-difference approach.

This capstone project will help CI understand factors affecting forest cover loss from 2011 to 2018 in Amazonia forest in Brazil. The end goal of this study is to understand the effectiveness of the Bolsa Verde on reducing forest cover loss in Amazonia rainforest in Brazil.

Literature Review

Spatial evaluation is the technique used for assess forest conservation policy (Blackman 2013). The motivation behind this analysis is to quantify the causal effect of forest conservation policy on forest cover change (FCC).

Different types of analysis such as risk assessment, cost-effectiveness analysis, and cost-benefit analysis are done to evaluate environmental policy programs (Bennear, 2004). However, all this analysis is done before the implementation of the program and relatively little analysis is done after the environmental policy is implemented. The program evaluation is important since it quantifies the changes seen after policy implementation and relevant decisions can be made. But program evaluation is difficult since even if environmental policy is correlated with a specific outcome, it does not imply a causal relationship between the policy and the outcome. Therefore, it's necessary to isolate the causal effects of treatments on outcomes.

The methods to isolate effects of treatment and control variables are random experimental designs, quantifying confounders (variables that are correlated both with treatment and outcomes), regression, matching estimation and propensity score. Regression analysis and matching estimates fail when any one of the confounders is unobservable. Then, differences-in-differences analysis is done where data is required both for pre and post-treatment, which sometimes is not available for environmental policy. Another technique used in such cases is the instrumental variables method. Data availability is another problem for program evaluation since various kinds of independent longitudinal data sets do not exist for environmental policy program evaluation.

Protected areas (PAs) are the major policy tool to protect ecosystem and biodiversity resources (Tesfaw et al 2018). Assigning new protected areas (PAs) does not help if they are in similar areas as the old PAs or located on lands where there is a low threat, which means that they will have a lower impact (Joppa et al 2009). PAs are mostly located where deforestation would not happen even if those areas are not declared PAs. The reason is that these PAs are mostly in locations that have higher elevations, steeper slopes and greater distances to human settlements. Also, legal changes can reduce the extent of and restrictions within the PAs (Kroner et al, 2019). PAs have both internal impacts and external effects called spillovers. The type of government regulation also changes the impact of PAs (Herrera, et al 2019). The causal link between the predictors and the outcomes (forest cover loss) is difficult to establish due to the multiple confounding factors (Lambin et al 2014).

Problem Framing

Bolsa Verde has been implemented in areas that are already PAs. So, the after-effect of the Bolsa Verde on deforestation is also due to the region being declared as PA too. The previous studies on Bolsa Verde did not isolate the effect of PA. Hence, in this project, we have considered BV implemented in PAs only to study the interaction effect between BV and PA and compared with the effects of BV alone.

In addition, we have a hypothesis that there are heterogeneity effects of these interventions. For instance, there is no significant internal impact of PA outside the region's "arc of deforestation" (a curve adjoining the southeastern edge of the Amazon Rainforest where the deforestation is occurring most rapidly), but PAs show relatively high interval impact inside the "arc of deforestation" (Herrera et al. 2019). Therefore, the heterogeneity effects of Bolsa Verde and PAs have also been analysed in this project.

Our project answers the following key questions:

1. **Are PA & PES successful in reducing deforestation in Amazonia Rainforest in Brazil?**

2. **Which variables affect the forest cover loss?**

3. **When is the Bolsa Verde eco-payments program most effective within a protected area?**

Data Collection, Munging, & Description

Protected areas and payment for ecosystem services are extracted from global Conservation Atlas, a database of area-based conservation systems from roughly 350-400 sources around the world. The target variable is the forest loss from 2000 to 2017. Each pixel in forest loss is a region of 30 by 30 square meters, and it represents an indicator of whether this region is deforested. For computing efficiency, we converted this forest loss to 900 by 900 square meter region, where each pixel represents the counts of forested area in this larger region. The covariates in this project (Table 1) include annual air temperature in 2011, annual precipitation in 2011, slope, elevation, distance to road and accessibility to cities. In particular, distance to road and slope were calculated using the Euclidean Distance tool and Slope tool in ArcGIS.

Table 1: Data Sources

	Variables	**Data Source**
Outcome	Forest cover loss	Hansen's Analysis Results of Landsat Images
Treatment	Presence of Bolsa Verde (0 / 1)	Conservation Governance Atlas
Confounders	Elevation (m)	The NASA Shuttle Radar Topographic Mission
	Slope (degree)	Calculated from elevation and is in degree.
	Annual air temperature (℃)	The Climate Data Guide: global precipitation and temperature
	Annual precipitation (mm)	The Climate Data Guide: global precipitation and temperature
	Distance to road (decimal degree)	Calculated from global roads open access data set, Socioeconomic Data and Applications Center
	Estimated travel time to the nearest city of 50,000 or more people (minutes)	Joint Research Center of the European Commission

Study Design

We divided the entire dataset into two groups to study the effect of policy implementation (BV or PES) on forest cover loss:

1. Treatment group: where policy (PES) has been implemented (PES=1)
2. Control group: where policy (PES) has not been implemented (PES=0)

Balance Tables were used to check whether the treatment and control groups have similar distribution of covariates. There is imbalance in terms of both the number and distribution of covariates in the treatment and control group (*as shown in the Balance Tables in the Supplemental Material*). Propensity Score Matching (PSM) with k-nearest (k=1) neighbor has been used for matching to achieve balance in the treatment and control groups (Detailed explanation of PSM is in supplemental material).

The randomized study is not possible since the policy has already been implemented. Therefore, causal inference method - matching with regression is used. Logistic and Poisson regressions were used to model the effect of covariates on the forest cover loss.

The modelling was done in two steps:

1. Firstly for PES alone (without considering the effect of PA)
2. Secondly, the data was subset for those values where PA has been implemented. Then, Step 1 was repeated, so effectively the entire study was for PA=1 values only. This was done to study the interaction of PES with PA.

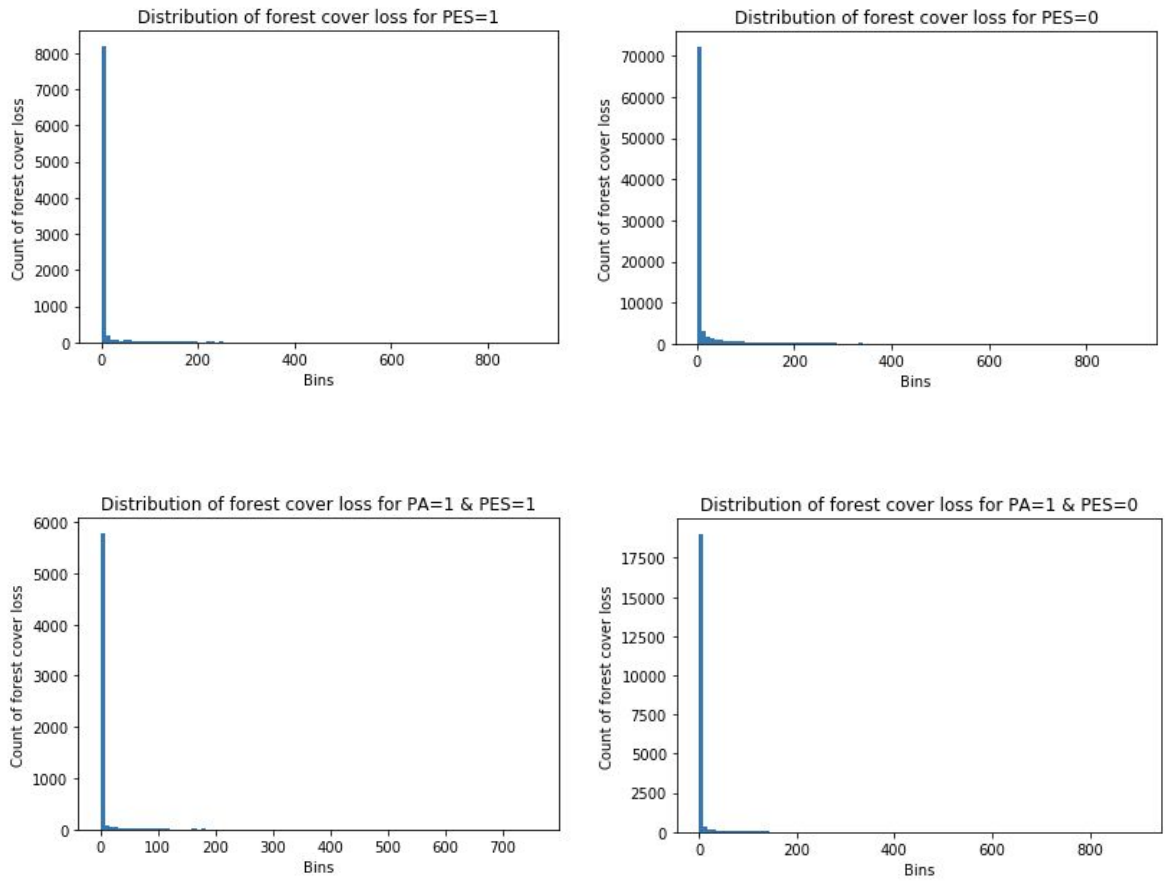

Figure 1: Distribution of forest cover loss for different combinations of PES & PA

We studied the distribution of the outcome variable i.e. forest cover loss for PES only and then both PA and PES (Figure 1). The above figures show that the distribution is not normal. Hence, the assumptions for the linear regression model were not satisfied. Therefore, we used logistic and poisson regression models to study the effect of covariates and treatment variable on the outcome variable.

Methodology

As discussed above, logistic regression is chosen over linear regression. To use logistic regression, we converted forest cover loss into binary variables. The forest cover loss in the original dataset is the count of 30x30 square meter plots that have lost forest cover on 900x900 square meter plots. If any plot has lost forest cover then the forest cover loss is one (1) otherwise it is zero (0). This was necessary to be able to use logistic regression, which would give the probability of log odds of the loss regressed with covariates.

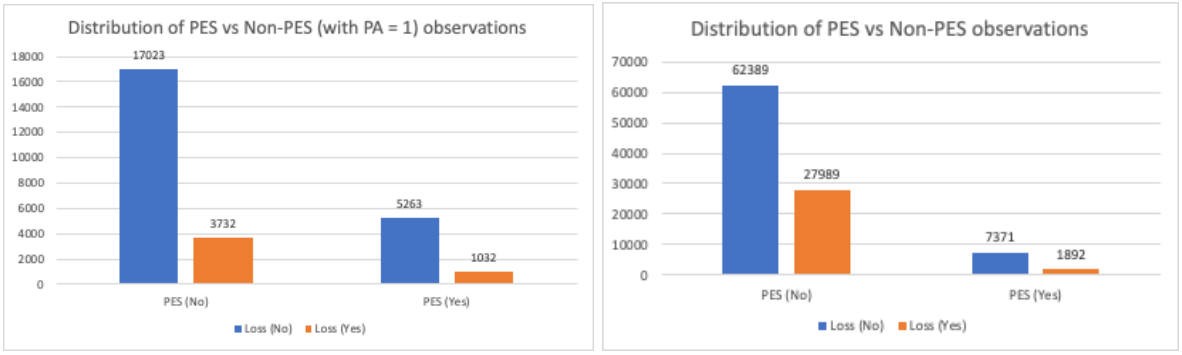

Figure 2: Distribution for forest cover loss when converted to binary variable

We again checked the distribution of the PES variable segmented by binary values of newly converted forest cover loss variables (Figure 2). The distribution is again not balanced. Hence, we first used PSM to achieve balance in the treatment and control groups. Then we did logistic regression analysis first for the entire dataset without considering the effect of PA. Then the second time, we subsetted the dataset where PA has been implemented and applied the same logistic modelling to understand the interaction effect of PES with PA.

After applying the logistic regression and studying the effect of PES on the entire dataset we wanted to study whether we would observe similar trends across different regions. Therefore to analyse the spatial effects of covariates, we divided the data into four different clusters by using k-means clustering. We used a machine learning algorithm instead of a manual method for clustering so as to avoid the bias that would have been introduced by the manual method.

Before we could analyse the data, we had to account for any bias that might be inherent in the location of the Bolsa Verde participants. To do this, we used propensity score matching in each cluster that identified locations that were similar to our treatment based on a number of covariates. We used a poisson regression that combines the results of matching and heterogeneity of settings to estimate the average treatment effect of both Bolsa Verde and Protected Areas.

The region of different physical covariates can be achieved by grouping regions according to their characteristics. We used KMeans clustering, a clustering algorithm that partitions all observations into ‘k’ regions that minimise the within-cluster dissimilarity of their characteristics. The characteristics for clustering used are annual air temperature, annual precipitation, slope, elevation, distance to road and accessibility to cities.

Study has shown that the optimal k could be chosen by analysing the clustering results (Kaczan, 2019). The optimal k should be the one that reduces within-cluster dissimilarity and contains enough treated points. In this project, k > 5 would result in a small cluster that causes propensity score matching not converging. The resulting map of KMeans with k = 4 is shown in Figure 1. All clusters are scattered around our research region and have clear spatial boundaries.

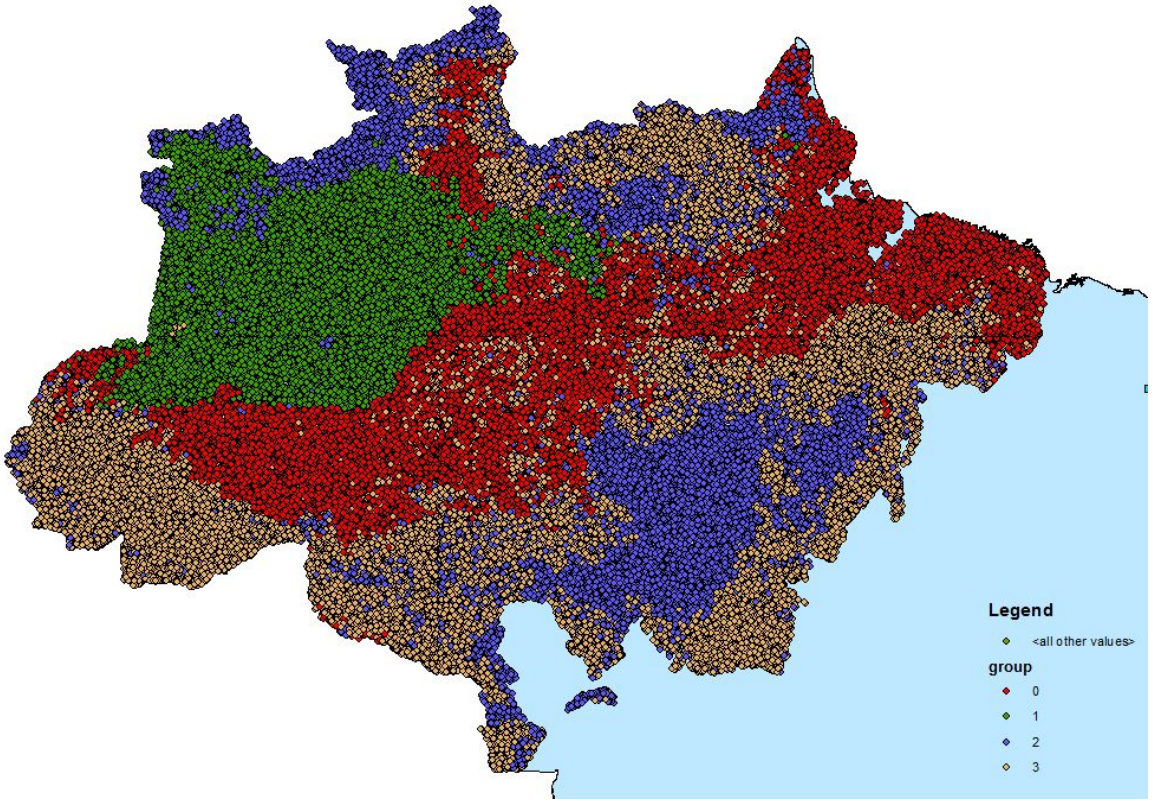

Figure 3: Clustering Result Map

Results & Discussion

The results from the logistic regression are as follows:

1. Matching improved the balance in covariates across the treatment & control group.
2. All confounding variables are important in explaining the forest cover loss (p-value less than 0.05 significance level)
3. Both PA and PES are effective in decreasing the forest cover loss.
4. Forest loss decreases as elevation, distance to roads and cities increase.

We consider two types of heterogeneity:

1. Probability of enrollment in Bolsa Verde, and
2. Effect of Bolsa Verde on forest cover loss

Since these two heterogeneities vary for each cluster, we examined the physical characteristics of each cluster by studying the average covariates for each cluster (Table 2).

Table 2: Average covariates in each cluster

	slope (degree)	*elevation (m)*	*Distance to road (decimal degree)*	*Access to cities (minutes)*	*precipitation* (mm)	*Air temperature* (℃)
Cluster 1	0.058	69.44	0.464	1412	20.78	27.32
Cluster 2	0.023	62.09	2.027	2725	23.85	26.82
Cluster 3	1.050	279.2	1.204	2824	19.92	26.04
Cluster 4	0.243	188.9	0.575	1938	18.15	26.58

Cluster 1 has the highest annual air temperature and is closest to roads and cites. Cluster 2 has the largest distance to road, accessibility to cites and annual precipitation. Cluster 3 has the largest slope and elevation, and smallest air temperature. Cluster 4 has the lowest precipitation and is close to cities and roads.

Heterogeneity of enrollments of treated units

We believe that the hidden variable for clustering may capture social and economic characteristics related with the criteria of Bolsa Verde. At first, we ignored the heterogeneous effects and ran propensity score matching on the entire region. The generated propensity score, given all physical covariates, indicates that even the most treated observations have low probability of enrollment in the PES (Figure 4). The reason for the low probability of enrollment is that the physical covariates might not be highly correlated with poverty, the criteria for enrolling in Bolsa Verde.

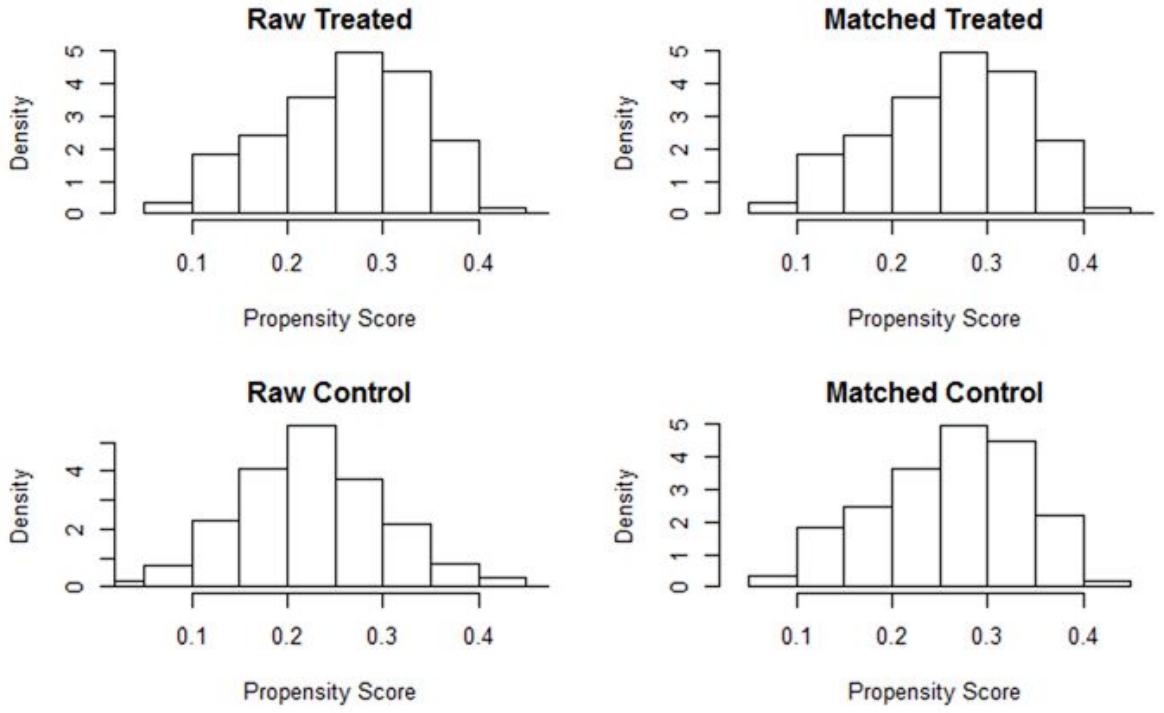

Figure 4: propensity score matching in the entire region

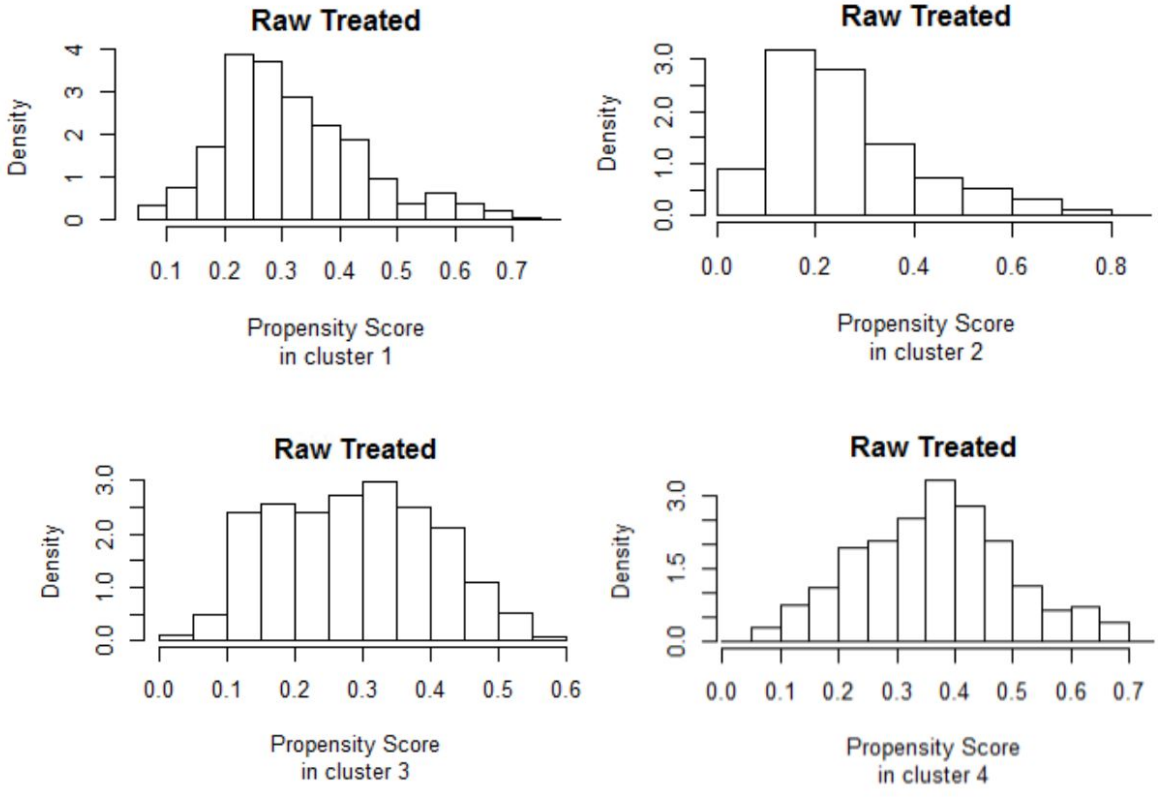

Figure 5: propensity score matching in the each cluster

To explore the heterogeneity of enrollment, we first partitioned the research region into four sub-regions using KMeans clustering. Then, we implemented propensity score matching within each cluster.

In cluster 3 and 4, people are more likely to enroll in Bolsa Verde (Figure 5). These two clusters share a common feature, they all have higher elevation and slope. One possible interpretation is that destroying forest for personal benefit is less profitable in these regions (for example, these regions are not well suited for agricultural activities), therefore, people are less likely to clear forest. By enrolling in Bolsa Verde, they still get paid by the government without actually protecting the forest. Therefore, they are eager to enroll in Bolsa Verde.

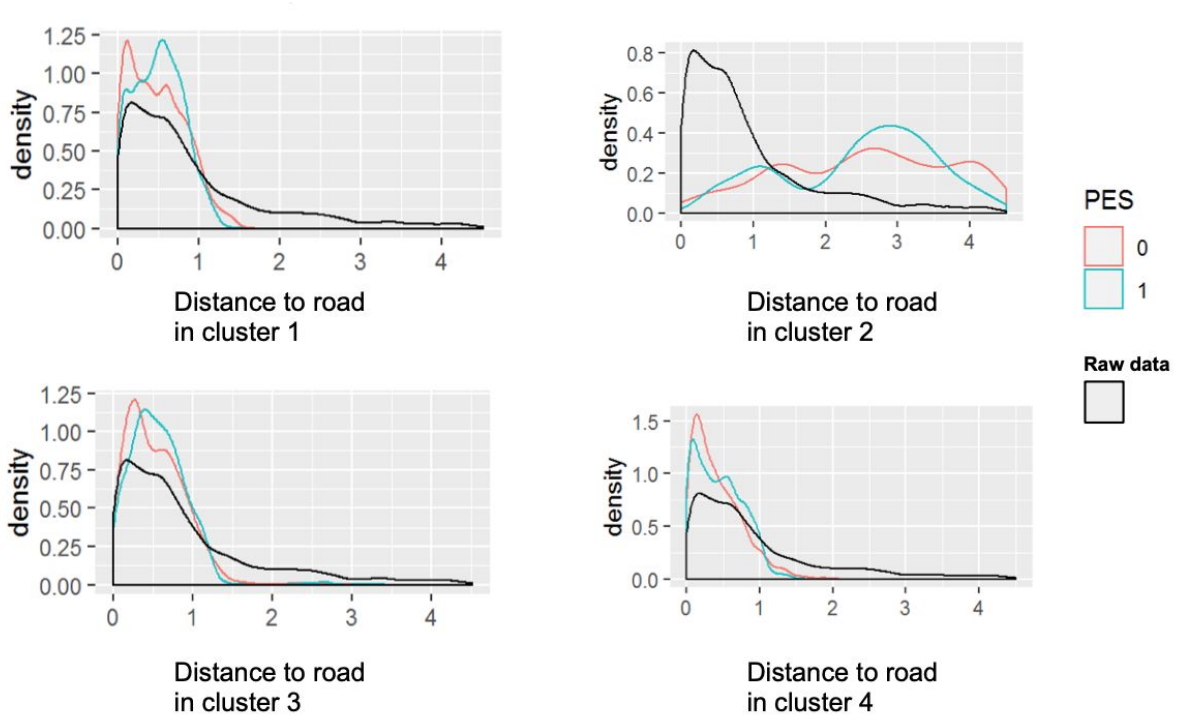

Figure 6: distribution of distance to road in the each cluster

Figure 6 shows the distributions of distance to road in each cluster, that confirms that it is necessary to run propensity score matching within each cluster. For instance, in cluster 2, the raw distribution is similar to other clusters as shown by the black lines. However, the treated units are different from other clusters as shown by blue lines. By matching only within each cluster (say cluster #2), we ensure that treated and matched control units will have balanced distribution.

Interaction effects of Bolsa Verde and protected areas

In this section, we considered the PAs where Bolsa Verde is also implemented to understand when Bolsa Verde is effective in reducing deforestation. To construct a balanced dataset, we considered those areas as treated units where both Bolsa Verde and PAs are implemented and those areas as untreated units where only PA is implemented. Control units are then selected by propensity score matching. Regression results are shown in Table 3.

Table 3: The effectiveness of PES and PA compared with that of PA alone

	Percentage forest loss reduced by PA and PES compared with PA alone
Cluster 1	25.0
Cluster 2	-39.1
Cluster 3	-10.6
Cluster 4	-9.1

We could discuss this result in 2x2 matrix (Table 4).

Table 4: Effects of both Bolsa Verde and PA in different clusters

Confounding Variables		*Slope & Elevation*	
		Low	*High*
Distance to roads & Access to cities	*Low*	**Cluster 1**	Cluster 4
	High	Cluster 2	Cluster 3

Results show that Bolsa Verde tends to be effective when it is implemented in regions that are closer to cities and have lower elevation. This implies that removing forest for personal benefit is profitable in cluster 1 but by compensating people from cluster 1 enrolled in Bolsa Verde, these people become less likely to destroy forest. In other words, profits to be obtained from clearing the forest are lower than our PES payments. On the other hand, Bolsa Verde is not effective in cluster 2, 3, and 4. The reason is that the clusters 2, 3, and 4 might face lower deforestation threats naturally without any need of human intervention.

Studies have shown that people don't find it beneficial to reduce forest cover in areas (i.e. clusters 2,3, and 4) ill suited for deforestation activities (Pfaff, 2008). Hence implementing pes would not make any more positive effect on people's behavior. Therefore, people anyways are eager to enroll in PES in order to get paid since payment can not be obtained from deforestation activities in clusters 2, 3, and 4. Therefore, Bolsa Verde tends to be not effective in clusters 2, 3, and 4.

Conclusion

Our final conclusions are the answers to the key questions mentioned in the problem framing:

1. **Are PA & PES successful in reducing deforestation in Amazonia Rainforest in Brazil?**
   Yes, the overall impact of PA and PES is effective in curbing forest cover loss. However, we cannot state with certainty that Bolsa Verde is effective at reducing deforestation. Hence, we need to analyse the effect of these programs in specific areas defined by clustering techniques.

2. **Which confounding factors affect the success of Bolsa verde?**
   When considering the entire dataset, forest cover loss decreases with increase in slope and decrease in elevation and distance to roads and cities. Also, precipitation and air temperature are not effective in predicting forest cover loss. However these trends differ by clusters. Hence we have analysed the effect of the confounding variables by each cluster. The values of air temperature and precipitation are in the similar range for all the four clusters (Table 1). Hence air temperature and precipitation are not crucial for evaluating the effect of Bolsa Verde and PA. The confounding variables slope, elevation, distance to roads and access to cities are important for evaluation of Bolsa Verde and PA.

3. **When is the Bolsa Verde eco-payments program most effective within a protected area?**
   a. Carrying out deforestation activities in areas having higher slope and elevation is tough even when those areas are closer to cities and accessible by roads.
   b. Areas that are far away from roads and cities are less influenced by human settlement activities, and hence there is less deforestation.
   c. Therefore, implementing BV in clusters 2 & 4 is not effective and very less effective in cluster 3; BV is most effective in cluster 1, which is closer to roads and cities and has lower slope and elevation, because those areas are more prone to deforestation by human activities.

References

1. Bennear, L. S. (2004). Evaluating Environmental Policies Lori Snyder Bennear and Cary Coglianese November 2004 RWP04-049.
2. Blackman, A. (2013). Evaluating forest conservation policies in developing countries using remote sensing data: An introduction and practical guide. *Forest Policy and Economics*, *34*, 1-16.
3. Herrera, D., Pfaff, A., & Robalino, J. (2019). Impacts of protected areas vary with the level of government: Comparing avoided deforestation across agencies in the Brazilian Amazon. *Proceedings of the National Academy of Sciences*, *116*(30), 14916-14925.
4. Joppa, L. N., & Pfaff, A. (2009). High and far: biases in the location of protected areas. *PloS one*, *4*(12), e8273.
5. Kaczan, D. J. (2020). Can roads contribute to forest transitions?. *World Development*, *129*, 104898.
6. Kroner, R. E. G., Qin, S., Cook, C. N., Krithivasan, R., Pack, S. M., Bonilla, O. D., ... & He, Y. (2019). The uncertain future of protected lands and waters. *Science*, *364*(6443), 881-886.
7. Lambin, E. F., Meyfroidt, P., Rueda, X., Blackman, A., Börner, J., Cerutti, P. O., ... & Walker, N. F. (2014). Effectiveness and synergies of policy instruments for land use governance in tropical regions. *Global Environmental Change*, *28*, 129-140.
8. Pfaff, A., Robalino, J. A., & Sanchez-Azofeifa, G. A. (2008). Payments for environmental services: empirical analysis for Costa Rica. Terry Sanford Institute of Public Policy, Duke University, Durham, NC, USA, 40.
9. Tesfaw, A. T., Pfaff, A., Kroner, R. E. G., Qin, S., Medeiros, R., & Mascia, M. B. (2018). Land-use and land-cover change shape the sustainability and impacts of protected areas. *Proceedings of the National Academy of Sciences*, *115*(9), 2084-2089.
10. Wong, P. Y., Harding, T., Kuralbayeva, K., Anderson, L. O., & Pessoa, A. M. (2018). Pay for Performance and Deforestation: Evidence From Brazil.

Supplemental Material

Matching and Balance checking

Propensity score matching is a type of matching method for causal inference. Logistic Regression is fitted to all the covariates with treatment variable: PES (0/1) as the outcome variable. The result of the logistic regression is the propensity score which is the probability of being assigned to a treatment or control group given all the values of the covariates. Then, the distribution of propensity scores across the treatment and control groups is seen through the graph called as Region of Common Support (Figure 1, 2). More the overlap of propensity scores in the treatment and control groups, better the balance. After checking for balance, each observation in the treatment group is matched to the observation in the control group that has the closest propensity score. The metric for closeness is decided by k-nearest (k=1) neighbor algorithm. Matching is done without replacement i.e. once an observation has been matched from control group to the treatment group, it's not used again for matching.

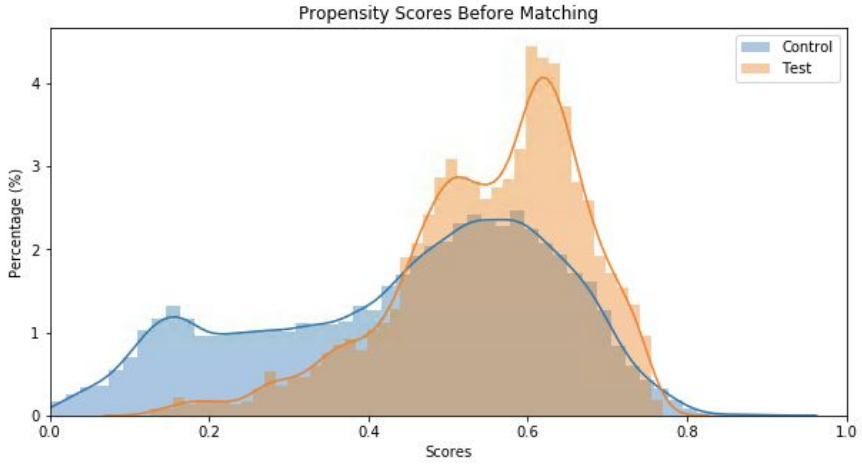

Figure 1: Region of Common Support - Propensity Score Matching for PES (0 / 1)

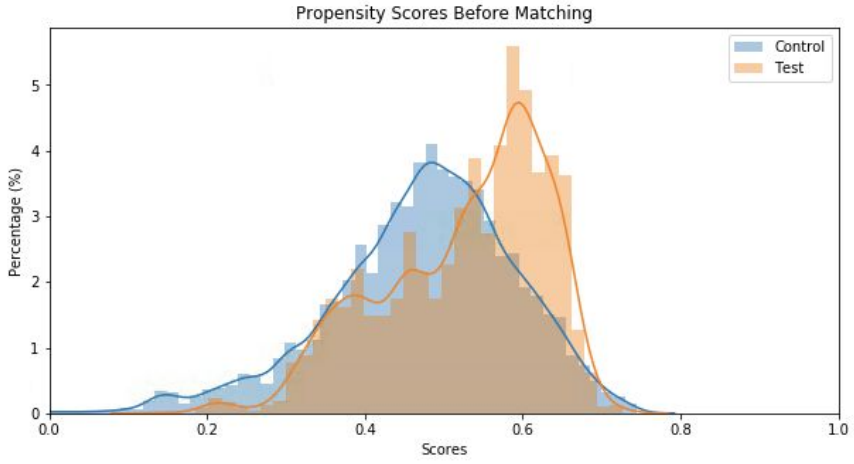

Figure 2 : Region of Common Support - Propensity Score Matching for PA=1 and PES (0 / 1)

Regression without matching for covariates in control or treatment group does not give for correct results. Propensity score combines all the effect of covariates in a single score. Points in the treatment (having policy) and balance (no policy) are matched with the one having the closest propensity score (k=1 in k-nearest neighbors). Matching reduces dependency on the model and balances the effect of covariates. Therefore, we have used propensity score matching for the entire dataset and then in each cluster before running the regression. Balance is achieved both in terms of values of covariates and number of observations in the treatment and control group. Balance tables (Table 1, 2) have been used to study the effect of matching on achieving balance in treatment and control groups.

Table 1: Balance Table - Distribution of mean of covariates for PES (0/1)

Variable	**PES**	**Non-PES** ***(Before matching)***	**P-value for difference** ***(Before matching)***	**Non-PES** ***(After matching)***	**P-value for difference** ***(After matching)***
Slope	0.25	0.32	0.00	0.24	0.11
Elevation	130.02	169.08	0.00	123.29	0.00
Distance to road	0.62	0.95	0.00	0.64	0.00
Access to cities	1743.60	1734.75	0.59	1775.13	0.13
Precipitation 2011	20.00	19.53	0.00	19.95	0.35
Air temperature 2011	26.87	26.38	0.00	26.85	0.09

Table 2: Balance Table - Distribution of mean of covariates for PA=1 & PES (0/1)

Variable	**PES**	**Non-PES** ***(Before matching)***	**P-value for difference** ***(Before matching)***	**Non-PES** ***(After matching)***	**P-value for difference** ***(After matching)***
Slope	0.24	0.34	0.00	0.24	0.41
Elevation	135.20	151.24	0.00	133.84	0.41
Distance to road	0.73	0.98	0.00	0.73	0.90
Access to cities	1947.46	2134.80	0.00	1898.79	0.04
Precipitation 2011	19.55	20.61	0.00	19.58	0.65
Air temperature 2011	26.78	26.74	0.00	26.78	0.96

Logistic Regression results (Table 3 and 4) show that all variables are significant in explaining the outcome variable, forest cover loss.

Table 3: Logistic Regression Results: PES (0 / 1) after matching

Logit Regression Results

Dep. Variable:	loss	**No. Observations:**	18526
Model:	Logit	**Df Residuals:**	18516
Method:	MLE	**Df Model:**	9
Date:	Thu, 02 Apr 2020	**Pseudo R-squ.:**	0.1388
Time:	01:51:45	**Log-Likelihood:**	-9026.9
converged:	True	**LL-Null:**	-10482.
Covariance Type:	nonrobust	**LLR p-value:**	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	10.6719	0.752	14.192	0.000	9.198	12.146
C(within_pes)[T.1]	-0.2866	0.052	-5.469	0.000	-0.389	-0.184
C(within_pa)[T.1]	-0.6289	0.063	-9.955	0.000	-0.753	-0.505
C(within_pes)[T.1]:C(within_pa)[T.1]	0.0901	0.086	1.045	0.296	-0.079	0.259
slope	0.2496	0.044	5.699	0.000	0.164	0.335
elevation	-0.0036	0.000	-12.566	0.000	-0.004	-0.003
DistToRoad	-0.1228	0.041	-2.976	0.003	-0.204	-0.042
acc_50k	-0.0006	2.21e-05	-27.010	0.000	-0.001	-0.001
precip2011	-0.0578	0.006	-9.335	0.000	-0.070	-0.046
air2011	-0.3328	0.028	-11.997	0.000	-0.387	-0.278

Table 4: Logistic Regression Results: PA=1 and PES (0 / 1) after matching

Logit Regression Results

Dep. Variable:	loss	**No. Observations:**	27050
Model:	Logit	**Df Residuals:**	27042
Method:	MLE	**Df Model:**	7
Date:	Thu, 02 Apr 2020	**Pseudo R-squ.:**	0.04430
Time:	01:59:28	**Log-Likelihood:**	-12033.
converged:	True	**LL-Null:**	-12591.
Covariance Type:	nonrobust	**LLR p-value:**	1.329e-236

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	8.9801	0.589	15.235	0.000	7.825	10.135
C(within_pes)[T.1]	-0.2056	0.040	-5.165	0.000	-0.284	-0.128
slope	0.1831	0.030	6.161	0.000	0.125	0.241
elevation	-0.0022	0.000	-10.514	0.000	-0.003	-0.002
DistToRoad	-0.1334	0.023	-5.740	0.000	-0.179	-0.088
acc_50k	-0.0003	1.52e-05	-19.714	0.000	-0.000	-0.000
precip2011	-0.0235	0.005	-4.697	0.000	-0.033	-0.014
air2011	-0.3394	0.022	-15.674	0.000	-0.382	-0.297

The distribution of propensity scores across control and test groups is balanced in all the four clusters (Figures 3, 4, 5, and 6). Indeed, the distributions in each group vary and the treated units in each group has higher probability of enrollment than the control units. Also, the raw control distributions are different from matched control distributions. This indicates that matching attempts to achieve balance between the observed variables.

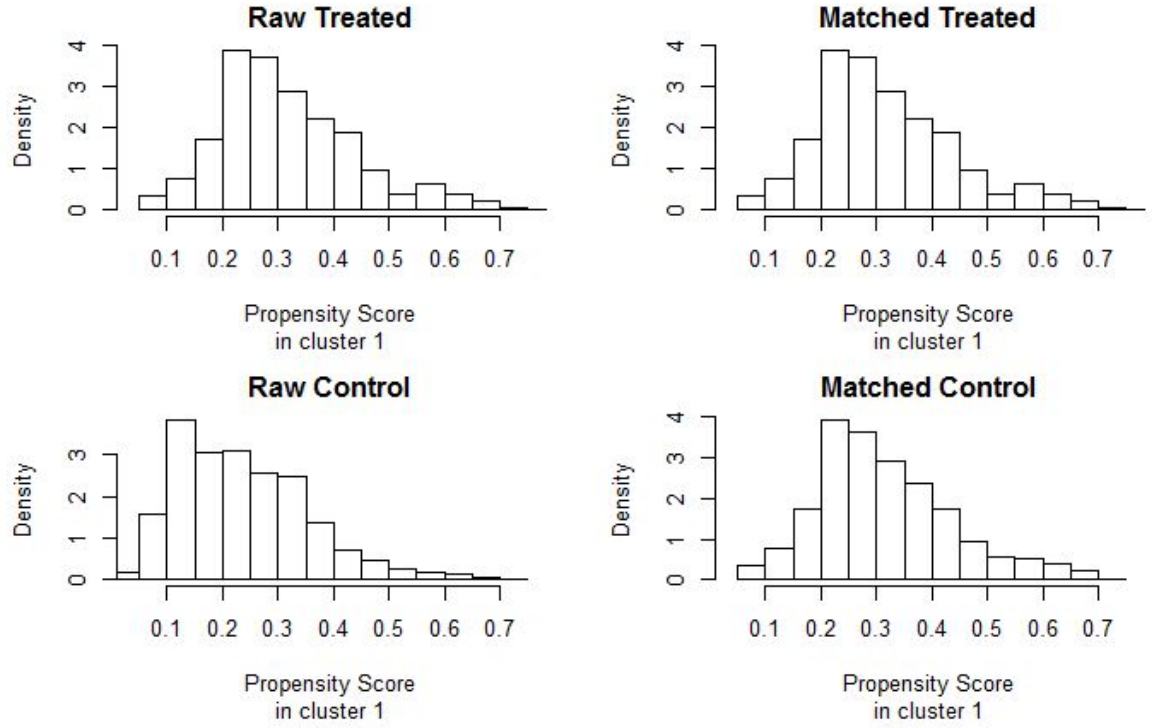

Figure 3: propensity score matching in cluster 1

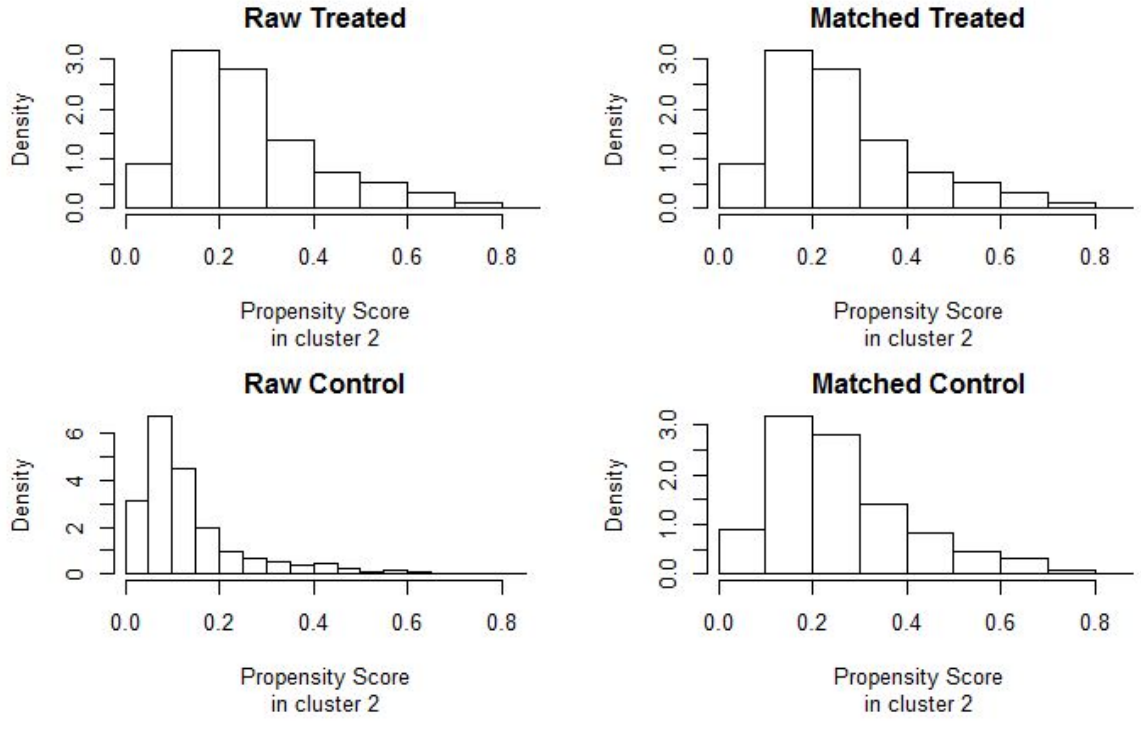

Figure 4: propensity score matching in cluster 2

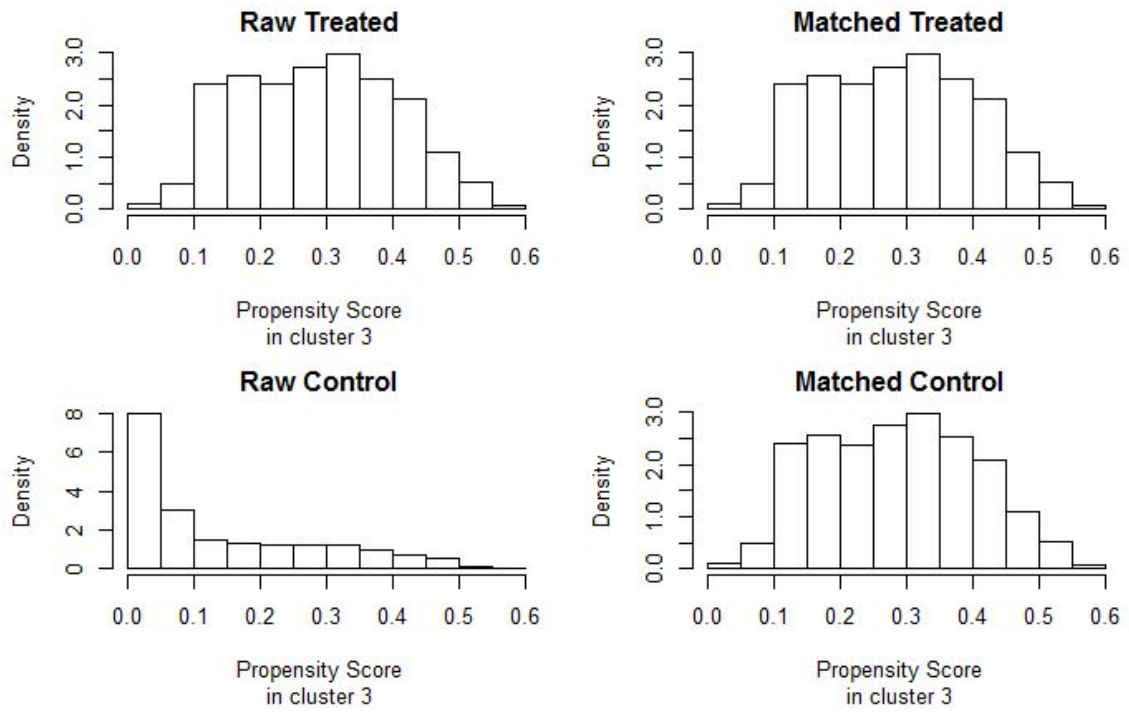

Figure 5: propensity score matching in cluster 3

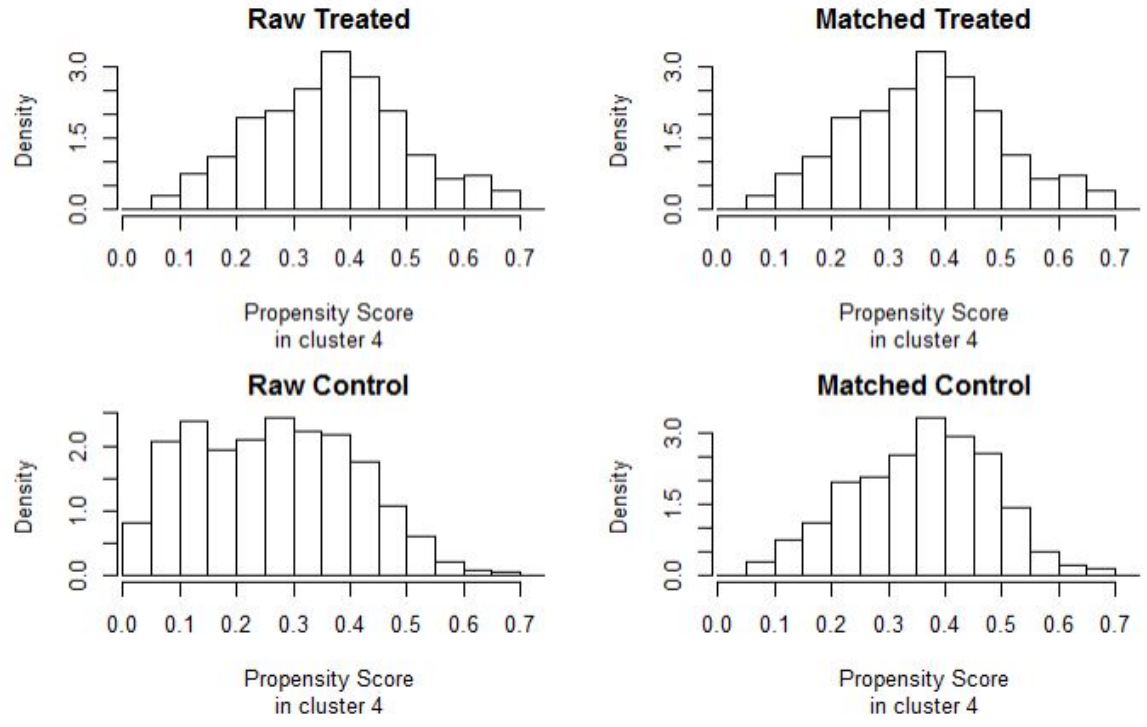

Figure 6: propensity score matching in cluster 4

Table 5,6,7,8 are the balance tables for the result of propensity score matching. The first rows are row data, where the second rows are the matched data. These tables also indicate that the propensity score matching attempt to achieve balance between the observed variables.

Table 5: Balance Table for Cluster 1

Summary of balance for all data:

	Means Treated	Means Control
distance	0.3161	0.2383
slope	0.0530	0.0593
elevation	63.9119	71.3718
DistToRoad	0.5130	0.4468
acc_50k	1386.0872	1420.5406
precip2011	19.7983	21.1162
air2011	27.1590	27.3695

Summary of balance for matched data:

	Means Treated	Means Control
distance	0.3161	0.3150
slope	0.0530	0.0537
elevation	63.9119	65.0951
DistToRoad	0.5130	0.4950
acc_50k	1386.0872	1330.5183
precip2011	19.7983	19.7179
air2011	27.1590	27.1442

Table 6: Balance Table for Cluster 2

Summary of balance for all data:

	Means Treated	Means Control
distance	0.2651	0.1413
slope	0.0342	0.0212
elevation	65.6347	61.4080
DistToRoad	2.4569	1.9444
acc_50k	2673.7960	2734.9959
precip2011	22.7549	24.0635
air2011	26.8477	26.8126

Summary of balance for matched data:

	Means Treated	Means Control
distance	0.2651	0.2645
slope	0.0342	0.0321
elevation	65.6347	63.3920
DistToRoad	2.4569	2.4797
acc_50k	2673.7960	2546.2653
precip2011	22.7549	22.6210
air2011	26.8477	26.8948

Table 7: Balance Table for Cluster 3

Summary of balance for all data:

	Means Treated	Means Control
distance	0.2846	0.1418
slope	1.0037	1.0595
elevation	230.0934	288.9027
DistToRoad	0.5700	1.3296
acc_50k	2448.2046	2899.0813
precip2011	20.7473	19.7522
air2011	26.3966	25.9705

Summary of balance for matched data:

	Means Treated	Means Control
distance	0.2846	0.2844
slope	1.0037	1.1115
elevation	230.0934	234.7314
DistToRoad	0.5700	0.5390
acc_50k	2448.2046	2480.2707
precip2011	20.7473	20.7861
air2011	26.3966	26.3009

Table 8: Balance Table for Cluster 4

Summary of balance for all data:

	Means Treated	Means Control
distance	0.3665	0.2662
slope	0.1860	0.2673
elevation	188.6921	188.9821
DistToRoad	0.4465	0.6290
acc_50k	2056.8876	1888.2470
precip2011	17.7797	18.3050
air2011	26.5375	26.5989

Summary of balance for matched data:

	Means Treated	Means Control
distance	0.3665	0.3599
slope	0.1860	0.1941
elevation	188.6921	186.1331
DistToRoad	0.4465	0.4284
acc_50k	2056.8876	1932.6115
precip2011	17.7797	17.8450
air2011	26.5375	26.5675

The distribution of covariates in each cluster are shown in Figure 7, 8, 9, and 10.

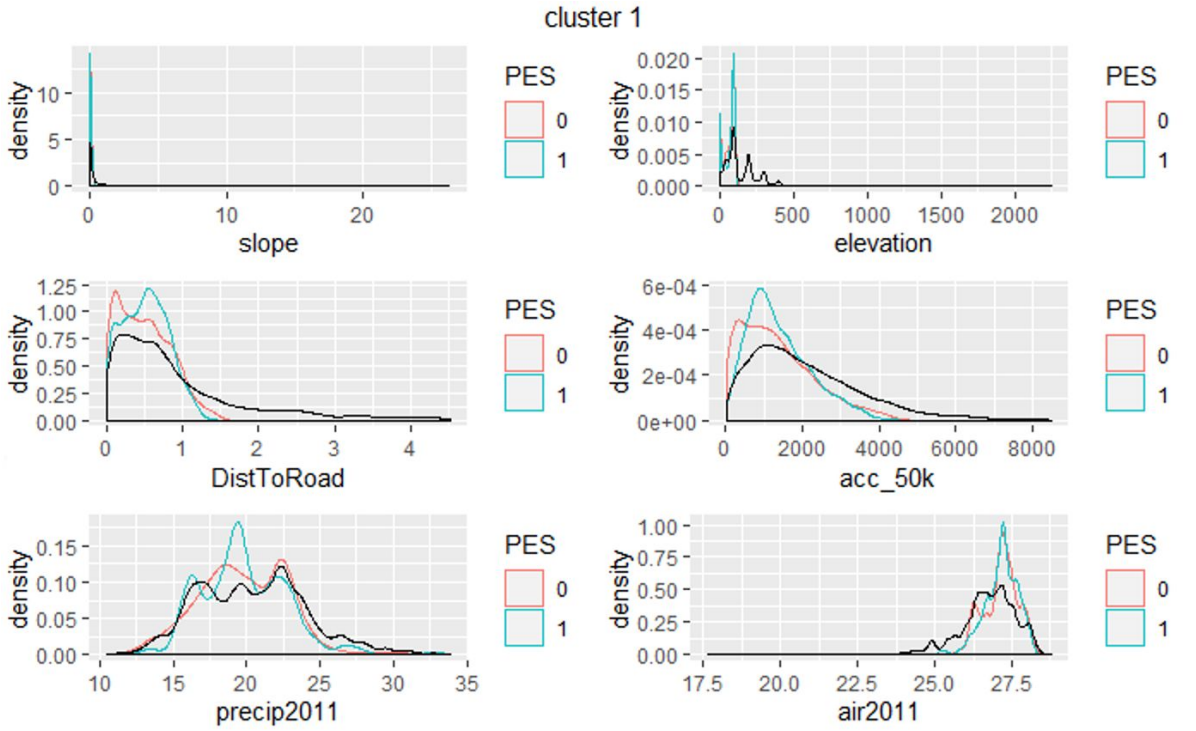

Figure 7: Distribution of covariates for cluster 1

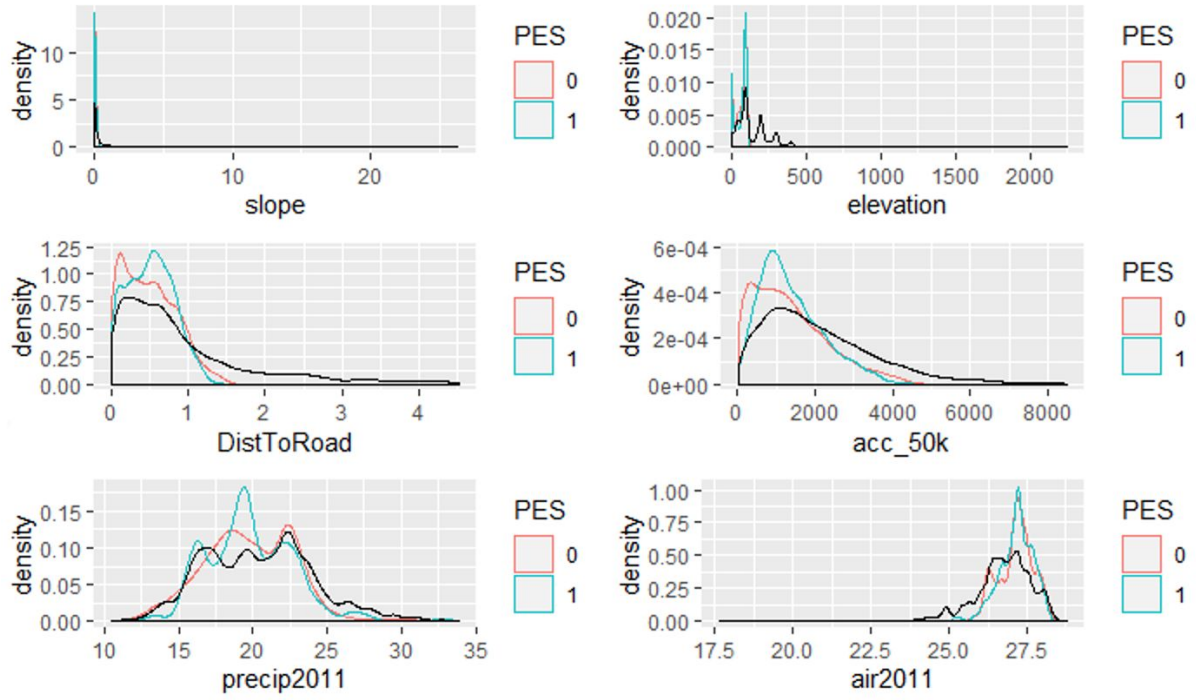

Figure 8: Distribution of covariates for cluster 2

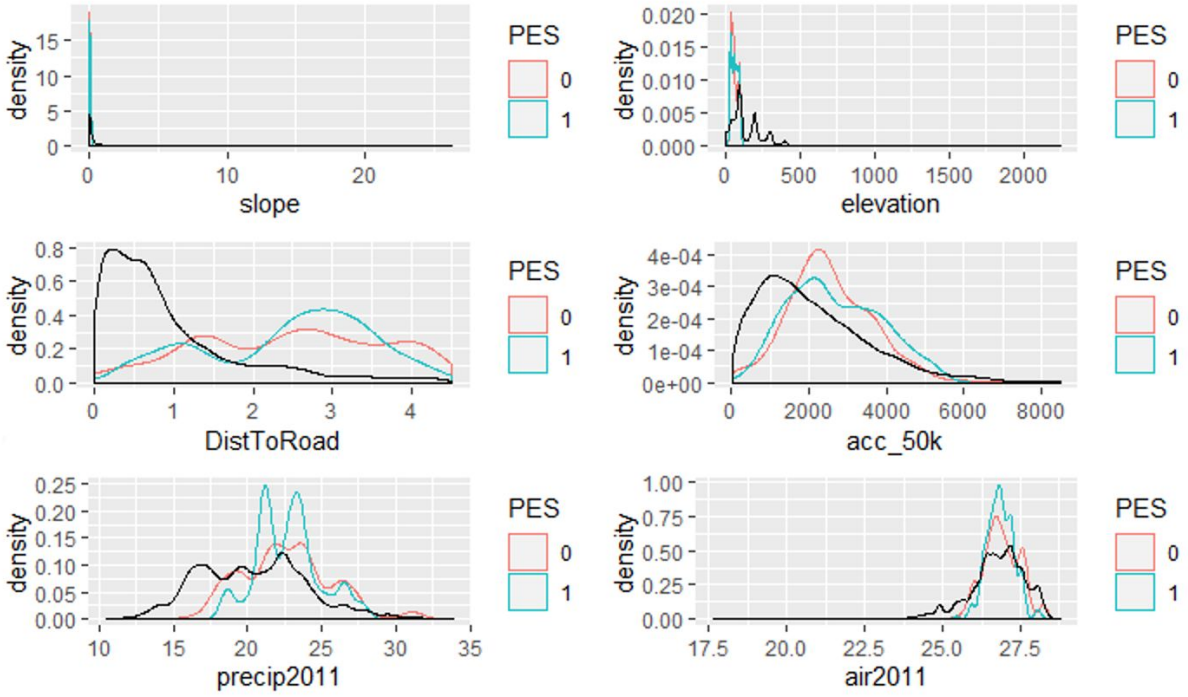

Figure 9: Distribution of covariates for cluster 3

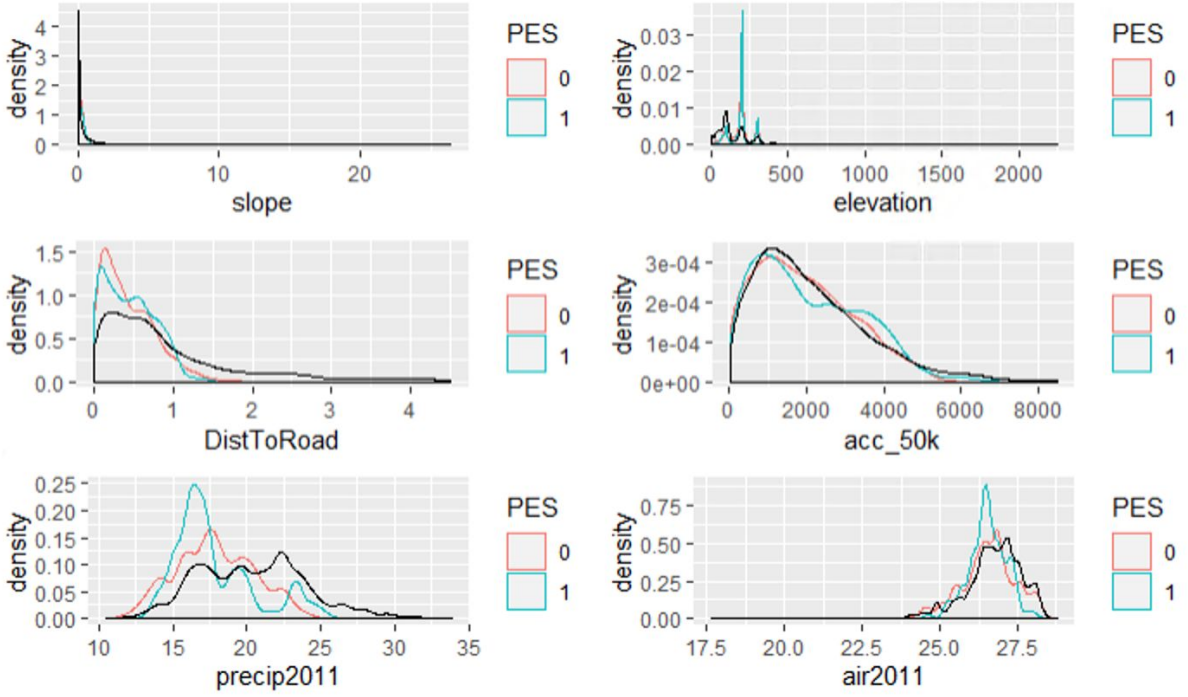

Figure 10: Distribution of covariates for cluster 4

Interactions between treated units clusters

To study the heterogeneity of Bolsa Verde, we run a seperate Poisson regression with interaction terms between the treated variable (presence of Bolsa Verde) and cluster (Table 9).

Table 9: Poisson Regression with interaction terms

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	1.740e+00	8.076e-03	215.401	< 2e-16	***
slope_mean	2.610e-01	8.243e-03	31.661	< 2e-16	***
elevation_mean	-4.898e-03	5.050e-05	-96.985	< 2e-16	***
DistToRoad_mean	-1.319e-01	9.162e-03	-14.401	< 2e-16	***
acc_50k_mean	-7.303e-04	4.583e-06	-159.352	< 2e-16	***
precip2011_mean	-7.251e-02	1.071e-03	-67.680	< 2e-16	***
air2011_mean	-4.660e-01	4.130e-03	-112.818	< 2e-16	***
as.factor(treat)1	-2.883e-01	8.911e-03	-32.354	< 2e-16	***
as.factor(group)1	-4.179e-01	2.724e-02	-15.344	< 2e-16	***
as.factor(group)2	-2.086e-01	1.931e-02	-10.801	< 2e-16	***
as.factor(group)3	-2.917e-02	1.007e-02	-2.896	0.00378	**
as.factor(treat)1:as.factor(group)1	6.183e-01	3.069e-02	20.148	< 2e-16	***
as.factor(treat)1:as.factor(group)2	3.892e-01	1.941e-02	20.055	< 2e-16	***
as.factor(treat)1:as.factor(group)3	3.758e-01	1.316e-02	28.567	< 2e-16	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1