Quantifying Heterogeneous Causal Treatment Effects in World Bank Development Finance Projects

Jianing Zhao[1], Daniel M. Runfola[2], and Peter Kemper[1]

[1] College of William and Mary, Williamsburg, VA 23187-8795, USA
{jzhao,kemper}@cs.wm.edu
[2] AidData, 427 Scotland Street, Williamsburg, VA. 23185 USA
drunfola@aiddata.org

Abstract. The World Bank provides billions of dollars in development finance to countries across the world every year. As many projects are related to the environment, we want to understand the World Bank projects impact to forest cover. However, the global extent of these projects results in substantial heterogeneity in impacts due to geographic, cultural, and other factors. Recent research by Athey and Imbens has illustrated the potential for hybrid machine learning and causal inferential techniques which may be able to capture such heterogeneity. We apply their approach using a geolocated dataset of World Bank projects, and augment this data with satellite-retrieved characteristics of their geographic context (including temperature, precipitation, slope, distance to urban areas, and many others). We use this information in conjunction with causal tree (CT) and causal forest (CF) approaches to contrast 'control' and 'treatment' geographic locations to estimate the impact of World Bank projects on vegetative cover.

1 Introduction

We frequently seek to test the effectiveness of targeted interventions - for example, a new website design or medical treatment. Here, we present a case study of using recent theoretical advances - specifically the use of tree-based analysis [3] - to estimate heterogeneous causal effects of global World Bank projects on forest cover over the last 30 years.

The World Bank is one of the largest contributors to development finance in the world, seeking to promote human well-being through a wide variety of programs and related institutions [1]. However, this goal is frequently at odds with environmental sustainability - building a road can necessitate the removal of trees; building a factory that supplies jobs can lead to the pollution of proximate forests. Multiple environmental safeguards have been put in place to offset these challenges, but relatively little is known about their efficacy across large scales.

We adopt the commonly applied approach of selecting "control" cases (i.e., areas where World Bank projects have very little funding) to contrast to "treated"

cases (i.e., areas where World Bank projects have a large amount of funding). This is analogous to similar approaches in the medical literature, where humans are put into control and treatment groups, and individuals that are similar along all measurable attributes are contrasted to one another after a medicine is administered. This is necessary due to the generalized challenge of all observational studies: it is impossible observe the exact same unit of observation with and without a World Bank project simultaneously - in the same way it would be impossible to examine a patient that was and was not given medication at the same time. Further complicating the challenge presented in this paper is the scope of the World Bank - with tens of thousands of project locations worldwide, there is considerable variation in the aims of different projects, the project's size, location, socio-economic, environmental, and historical settings. This variation makes traditional, aggregate estimates of impact unhelpful, as such aggregates mask variation in where World Bank projects may be helping - or harming - the environment. Following this, we investigate the research question *What is the impact of world bank projects on forest cover?*

To examine this question, we first integrate information on the geographic location of World Bank projects with additional, satellite derived information on the geographic, environmental, and economic characteristics of each project. We apply four different models to this dataset, and contrast our findings to illustrate the various tradeoffs in these approaches. Specifically, we test Transfomed Outcome Trees (TOTs), Causal Trees (CTs), Random Forest TOTs (RFTOTs), and Causal Forests (CFs). We follow the work of Athey and Imbens [3], who demonstrated how regression trees and random forests can be adjusted to estimate heterogenous causal effects. This work is based on the Rubin Causal Model (or potential outcome framework), where causal effects are estimated through comparisons between observed outcomes and the "counterfactual" outcomes one would have observed under the absence of an aid project [9]. While traditional tree-based approaches rely on training with data with known outcomes, Athey and Imbens illustrated that one can estimate the conditional average treatment effect on a subset with regressions trees after an appropriate data transformation process.

Many approaches to estimating heterogeneous effects have emerged over the last decade. LASSO [14] and support vector machines (SVM) [15] may serve as two popular examples. For this paper, we focus on very recent tree-based techniques that are very promising for causal inference. In [12], Su et al. proposed a statistical test as the criterion for node splitting. In [3], Athey and Imbens derived TOTs and CTs, an idea that is followed up on by Wagner and Athey [16] with CF (causal forest, random forests of CTs), and similarly Denil et al. in [6] who use different data for the structure of the tree and the estimated value within each node. Random forests naturally gave rise to the question of confidence intervals for the estimates they deliver. Following this, Meinshausen introduced quantile regression forests in [10] to estimate a distribution of results, and Wagner et al [17] provided guidance for confidence intervals with random

forests. Several authors, including Biau [4], recognize a gap between theoretical underpinnings and the practical applications of random forests.

The contribution of this paper is twofold: we evaluate and compare a number of proposed methods on simulated data where the ground truth is known and apply the most promising for the analysis of a real world data set. Practical experience results on tree-based causal inference methods are rare. To the best of our knowledge, this is the first investigation on the analysis of a spatial data set of world wide range with a large scale set of projects and dimensions. When it comes to applications for causal inference techniques, A/B testing for websites (such as eBay) is a more common [13]. A/B testing is conducted by diverting some percentage of traffic for a website A to a modified variant B of said website for evaluation purposes. This leads to a large amount of data with clearly defined treated and untreated groups where cases vary mainly by user activity. While the difference between A and B is precisely defined and typically small, the huge number of cases helps to recognize treatment effects. This is very different to the World Bank data which is both much more limited in size, and also spread all over the world (resulting in large diversity across projects). The rest of the paper is structured as follows. In Section 2, we present the basic methodology for the calculation of CT and CF. Section 3 introduces the data set, its characteristics, preprocessing steps and the calculation of propensity scores necessary for the estimation of each type of tree. In Section 4, we present the outcome of the analysis. We conclude in Section 5.

2 Methodology

Causal inference is to a vast part a missing data problem as we can not observe a unit at the same time receiving and not receiving treatment to compare the outcomes. We introduce some notation and recall common concepts to be able to address this problem in a more formal way.

Causal effects Suppose we have a data set with n independently and identically distributed (iid) units $U_i = (X_i, Y_i)$ with $i = 1, \cdots, n$. Each unit has an observed feature vector $X_i \in \mathbb{R}^d$, a response (i.e., the outcome of interest) $Y_i \in \mathbb{R}$ and binary treatment indicator $W_i \in \{0, 1\}$. For a unit-level causal effect, the Rubin causal model considers the treatment effect on unit i being $\tau(X_i) = Y_i(1) - Y_i(0)$, the difference between treated $Y_i(1)$ and untreated $Y_i(0)$ outcome. One can be interested in an overall average treatment effect across all units U or investigate treatment effects of subsets that are characterized by their features X. The latter describes heterogeneous causal effects and is often of particular interest. In our case, it is interesting to identify characteristics of subsets of projects where the environment is affected strongly (positive or negative) by a World Bank project. The heterogenous causal effect is defined as $\hat{\tau}(x) = \mathbb{E}\big[Y_i(1) - Y_i(0) \mid X_i = x\big]$ following [8].

Causal Tree A regression tree defines a partition of a set of units $U_i = (X_i, Y_i)$ as each leaf node holds a subset of units satisfying conditions on X expressed

along the path from root node to leaf. This helps for the condition in $\hat{\tau}(x) = \mathbb{E}\big[Y_i(1) - Y_i(0) \mid X_i = x\big]$. In observational studies, a unit is either treated or not, so we know either $Y_i(1)$ or $Y_i(0)$, but not both. However, one can still estimate $\tau(x)$ if one assumes unconfoundedness: $W_i \perp\!\!\!\perp (Y_i(1), Y_i(0)) \mid X_i$. Athey and Imbens [3] showed that one can estimate the causal effect as:

$$\hat{\tau}(X_i) = \sum_{i \in T} Y_i \cdot \frac{W_i/\hat{e}(X_i)}{\sum_{j \in T} W_j/\hat{e}(X_j)} - \sum_{i \in C} Y_i \cdot \frac{(1 - W_i)/(1 - \hat{e}(X_i))}{\sum_{j \in C}(1 - W_j)/(1 - \hat{e}(X_j))} \quad (1)$$

where $e(X_i)$ is the propensity score of project i which is calculated by logistic regression, T represents treatment units, and C control units. Hence one can adapt the calculation of a regression tree to support calculation of $\hat{\tau}(X_i)$ by (1) by adjusting the splitting rule in the tree generation process.

In a classic regression tree, mean square error (MSE) is often used as the criterion for node splitting, and the average value within the node is used as the estimator. Following Athey and Imbens [3], we use (1) as the estimator and the following equation as the new MSE for any given node J in the causal tree.

$$MSE = \sum_{i \in J}(Y_i(1) - Y_i(0) - \hat{\tau}(X_i))^2 = \sum_{i \in J}\tau(X_i)^2 - \sum_{i \in J}\hat{\tau}(X_i)^2 \quad (2)$$

The right equation follows if one assumes that $\sum_{i \in J}\tau(X_i) = \sum_{i \in J}\hat{\tau}(X_i)$. The key observation is that $\sum_{i \in J}\tau(X_i)^2$ is constant and does not impact ΔMSE. For a split, data in node P is split into a left L and right R node, $\Delta MSE = MSE_P - MSE_L - MSE_R = \sum_{i \in P}\hat{\tau}(X_i)^2 - \sum_{i \in L}\hat{\tau}(X_i)^2 - \sum_{i \in R}\hat{\tau}(X_i)^2$. The ground truth $\tau(X_i)$ cancels out in ΔMSE and we can grow the tree without knowledge of $\tau(X_i)$. However, there is one more constraint we need to add to the splitting rule aside from MSE. To use (1) for the calculation of $\hat{\tau}(X_i)$, neither set T nor C can be empty. Due to characteristics of the data in our applied study, we found that cases where only C or T units existed in children naturally emerged, so we added a corresponding additional stopping criterion to the splitting rule to prevent splits that would lead to situations where T or C had less than a fixed minimum cardinality.

Causal Forest While a single causal tree allows us to estimate the causal effect, it leads to the problem of overfitting and subsequent challenges for pruning the tree. A common solution is to use an ensemble method such as bootstrap aggregating or bagging, namely a variant of Breiman's random forest [5]. If one applies the random forest approach to causal trees, the result is called a causal forest. Computation of a causal forest scales well as it can naturally be run in parallel. The same adjustments for generating a single CT apply to the generation of a random forest of CTs. We implemented a causal forest algorithm with the help of the scikit learn package. We can estimate the causal effect $\tau_{CF}(X_i)$ from a causal forest (a set CF of causal trees) for a unit i as the average across the estimates obtained from its trees: $\hat{\tau}_{CF}(X_i) = \frac{1}{|CF|}\sum_{t \in CF}\hat{\tau}_t(X_i)$.

3 Data

Data pre-processing This analysis relies on three key types of data: satellite data to measure vegetation, data on the geospatial locations of World Bank projects, and covariate datasets[3]. Our primary variable of interest is the fluctuation of vegetation proximate to World Bank projects, which is derived from long-term satellite data [11]. There are many different approaches to using satellite data to approximate vegetation on a global scale, and satellites have been taking imagery that can be used for this purpose for over three decades. Of these approaches, the most frequently used is the Normalized Difference Vegetation Index (NDVI), which has the advantage of the longest continuous time record. NDVI measures the relative absorption and reflectance of red and near-infrared light from plants to quantify vegetation on a scale of -1 to 1, with vegetated areas falling between 0.2 and 1 [7]. While the NDVI does have a number of challenges - including a propensity to saturate over densely vegetated regions, the potential for atmospheric noise (including clouds) to incorrectly offset values, and reflectances from bright soils providing misleading estimates - the popularity of this measurement has led to a number of improvements over time to offset many of these errors. This is especially true of measurements from longer-term satellite records, such as those used in this analysis, produced from the MODIS and AVHRR satellite platforms [11].

The second primary dataset used in this analysis measures where - geographically - World Bank projects were located. This dataset was produced by [2], relying on a double-blind coding system where two experts independently assign latitude and longitude coordinates, precision codes, and standardized place names to each geographic feature. Disagreements are then arbitrated by a third party.

In addition to the project name, the World Bank provided information on the amount of funding for each project and the year it was implemented, alongside a number of other ancillary variables. The database also provides information on the number of locations associated with each project - i.e., a single project may build multiple schools. These range from $n = 1$ to $n = 649$ project locations for a single project.

Data characteristics The temporal coverage of the covariates is variable across sources. For NDVI, precipitation, and temperature we have highly granular, yearly information on characteristics at each World Bank project location. From this information, we generate additional information regarding the trend (positive or negative) before and after project implementation, as well as simple averages in the pre- and post periods. Many variables only have a single measurement - population density, accessibility to urban areas, slope, and elevation are all measured circa 2000, while distances to roads and rivers are measured circa 2010. Figure 1 a) shows average annual NDVI values of all project locations for each year since 1982. The mean values are non-negative for all projects over

---

[3] For detailed information, check `https://github.com/zjnsteven/appendix`

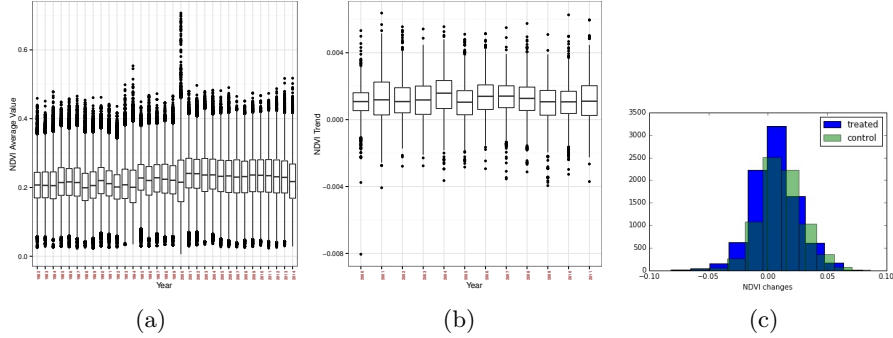(a)   (b)   (c)

Fig. 1: Properties of NDVI values at World Bank project locations

all years , and typical values are around 0.2 which is a lower bound for areas with vegetation. Figure 1 b) shows the distributions of slope values for a time series of NDVI values that starts in 1982 and ends with the year before each project starts. Approximately 75% of all projects have an upward trend in NDVI values across this time period. Figure 1 c) shows that treated and control projects have very similar empirical distributions for changes in NDVI values when the pre- and post- averages are contrasted.

Data interpretation for the context of measuring heterogenous treatment effects
One key attribute of causal attribution is a dataset which distinguishes between treated and untreated cases. In the case of a clinical trial, human beings who receive treatment might be contrasted to a control group of other humans of similar characteristics who do not receive a treatment. Because World Bank projects either exist or not, here we attempt to replicate the treated and untreated conditions by contrasting World Bank projects that were funded at very low levels ("control") to those that were funded at high levels ("treated"). This is reflective of a hypothesis that the observed treatment effect should positively correlate with the amount of funding, i.e., huge amounts of funding are expected to have a bigger effect than small amounts of funding. Following this, we assign $W_i = 1$ if a project's funding is in the upper third of all funded projects. While an imperfect representation of an area at which no World Bank project exists, by leveraging locations where a World Bank project exists but at a very low intensity we mitigate potential confounding sources of bias associated with locations the World Bank chooses to site projects at. Further, we bias our results in the more conservative negative direction - i.e., we will tend to under-estimate the impact of World Bank projects relative to null cases. Future research will consider the difference between this and a true null case in which locations with no aid at them are contrasted.

As a single project typically takes place at several project locations, we consider each project location as an individual unit - i.e., a school may be effective

in one location, but not another, even if they were implemented by the same funding mechanism. Further, to capture potential geographic heterogeneity this might introduce, for each unit's feature vector (i.e., selected covariates), we include the longitude and latitude of the project location. The total length of the feature vector is $d = 40$. All covariates are numerical and their values are not normalized. For our outcome measure (i.e., the variable we seek to estimate the impact on), we contrast the pre-treatment and post-treatment average NDVI values at a project's location. Let $ndvi_i(92, 03)$ denote the average of NDVI values observed for project location i over the years from 1992 to 2003 (the year before the project is implemented, which varies across projects; 2003 is used here for illustration). Let $ndvi_i(05, 12)$ describe the corresponding value for the eight years after the project starts. The response $Y_i = ndvi_i(05, 12) - ndvi_i(92, 03)$ is thus the difference of the two averages. Figure 1 c) shows histograms of Y_i values for treated and control projects. In order to calculate Y^* for Y, we calculate the propensity score $e(x)$, which describes the expected likelihood of treatment W_i for a given unit of observation. As described above, while there are many methods for estimating $e(x)$, here we use logistic regression to provide a better comparison with the econometric approaches commonly employed in the international aid community.

4   Experiments and Results

We follow a two stage procedure to examine the effectiveness of both the CT and CF algorithms, specifically considering our unique context of the effectiveness of World Bank projects. First, we test and evaluate which approach is most suited to our application using simulated synthetic data where we know the ground truth and where we can vary the size of sample data. Second, we apply these algorithms to examine the efficacy of World Bank projects based on satellite imagery. We implemented the CT and CF algorithms as well as Athey and Imbens transformed outcome tree (TOT) approach [3] and a random forest variant of TOT (RFTOT) using scikit-learn. The latter serve as a baseline for the performance of CT and CF algorithms.

Experimental results for simulated data First, we iteratively simulate synthetic datasets with known parameters to evaluate how the estimation of propensity score, dataset size, and degree of similarity between the control and treatment groups impact the accuracy of the result. To do this, we follow a bi-partite data generation process, in which two equations are used (one for treated cases and another for control cases).

We use each of the following two equations to produce one half of all data points. Y_i^1 gives the result for treated cases; Y_i^0 is for the control group. Here, from x_1 to x_8, $x_j \sim \mathcal{N}(0, 1)$ as well as $\varepsilon \sim \mathcal{N}(0, 1)$.

$$Y_i(1) = W_i^1 + \sum_{j=1}^{k} x_j * W_i^1 + \sum_{j=1}^{8} x_j + \varepsilon, \quad Y_i(0) = W_i^0 + \sum_{j=1}^{k} x_j * W_i^0 + \sum_{j=1}^{8} x_j + \varepsilon \quad (3)$$

As used in Table 1, k is defined as the number of covariates which contribute to heterogeneity in the causal effect. The true value of the causal effect is then $\tau(X_i) = Y_i(1) - Y_i(0) = 1 + \sum_{i=1}^{k} x_i$, with $W^1 = 1$ and $W^0 = 0$.

The first scenario we examine considers synthetic datasets with a randomized treatment assignment (each unit has the same probability to be treated, $e(x) = 0.5$). Figure 2 shows corresponding results for n=2000, and includes both single tree and random forest implementations of Transformed Outcome Trees (TOT; [3]) for comparison. The resultant distributions all encompass the true mean results, but with considerable difference in overall metrics of error. The Causal Forest approach is the most accurate across all simulations as well as the tightest overall distribution; this is in contrast to the TOT forest implementation. For single trees, the CT performs much better than the TOT and even outperforms the RFTOT.

	CF		CT		TOT		RFTOT	
sample size	mean	std	mean	std	mean	std	mean	std
1000	0.60	0.001	1.27	0.02	9.96	0.24	7.74	0.13
5000	0.58	0.001	0.99	0.02	7.95	0.03	5.61	0.05
10000	0.51	0.00001	0.86	0.005	7.45	0.02	5.14	0.02

Table 1: Mean square error (forest has 1000 trees, feature ratio = 0.8)

The second scenario considers synthetic datasets with varying numbers of observations (n=1000, 5000, and 10,000). We calculate the mean square error for CT, CF, TOT and RFTOT. The results in Table 1 show that - as expected - the error gets smaller as the number of observations increases. Of particular importance, we note that in the case of smaller datasets, the CF implementation strongly outperforms the single-tree CT implementation under all the scenarios we test.
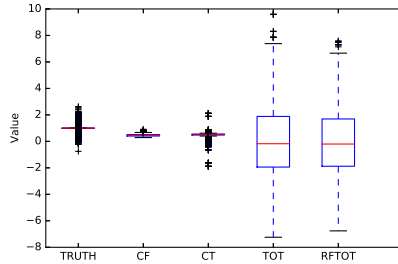

Fig. 2: Estimated treatment effects for randomized assignment, $e(x) = 0.5$
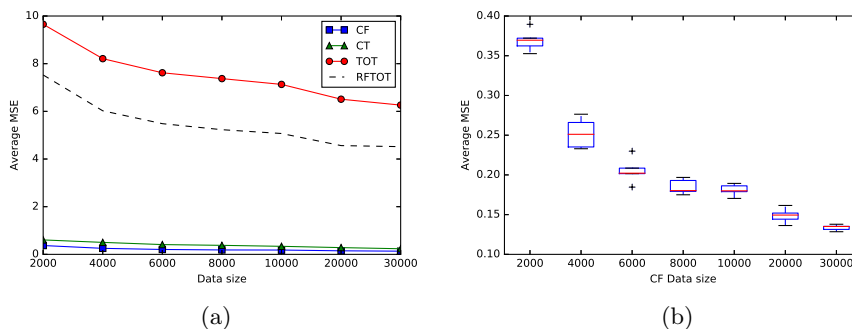
Fig. 3: MSE changes with data size

We also test the convergence of each method as the size of data increases, as shown in Figure 3. Figure 3a shows the MSE of each methods with increasing data size, while Figure 3b shows a zoomed-in version of the MSE of the CF approach (due to the lower magnitude of MSE observed). At least for this specific data generation process, the CF and CT outperform other approaches, which is why we focus on them for the analysis of the World Bank data set where we can not measure accuracy.

Results for World Bank data Following the simulation results, we seek to identify and contrast the benefits and drawbacks associated with applying CT and CF approaches to a real-world scenario. In this case study, we identify the impact of international aid - specifically, World Bank projects - on forest cover. First, we use a single CT to estimate the causal effect $\hat{\tau}(X_i)$ of a single project i with (equation 1) applied to the leaf where the project is located. Second, we implement a Causal Forest.

While our simulations, as well as the existing literature, suggest the Causal Tree has many drawbacks relative to a Causal Forest, it can enable practitioners to make inferences that are precluded by forest-based approaches. Most notably, the structure of single trees can provide insight into the explicit drivers of impacts - in this case, of World Bank projects. As an example, in the Causal Tree implementation here, we find that the year a project started was an important driver of effectiveness - specifically, projects starting before 2005 were more effective than those after 2005. This type of insight is particularly helpful, as it allows for analysis into the causes of impact heterogeneity. However, the lack of information on the robustness of findings in a single tree approach, coupled with the relative inaccuracy of CT as contrasts to CF, indicates that such findings should be approached with caution until better methods for identifying the robustness of CT tree shapes are derived.

The Causal Forest (CF) implementation represents a set of CTs and thus creates a distribution of values for each World Bank project i. These distributions
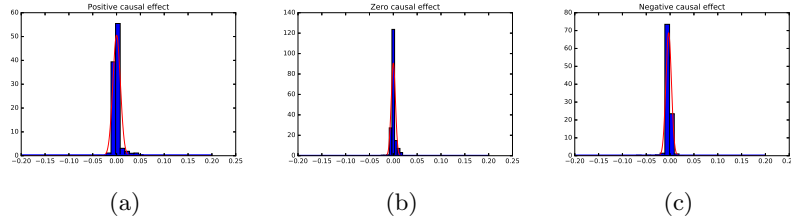
Fig. 4: CF calculated distributions of treatment effect estimates for specific projects: (a) Saint Lucia Hurricane Tomas Emergency Recovery Loan; (b) Sustainable Tourism Development Project; (c) Emergency Infrastructure Reconstruction Project.

are then aggregated to a single value to estimate $\hat{\tau}(X_i)$, or the distributions themselves are analyzed to examine the robustness of a given finding. In Figure 4, we show the detailed distributions for selected example projects. These examples provide an illustration of how applied CF results can provide indications not only of what projects are likely having a negative impact on the environment, but also the robustness of these estimates. By writing a second-stage algorithm which identified projects with distributions following certain characteristics (i.e., a mean centered around 0 with a Gaussian distribution; a negative-centered mean with a left-skewed distribution), it is possible to highlight the subset(s) of projects for which more robust findings exist. Figure 5 (a) shows a histogram of CF calculated $\hat{\tau}(X_i)$ values for all world bank projects in our data set. Most of the projects have a slightly negative to no impact on the forest cover, which is in line with World Bank objectives to offset potential negative environmental outcomes. Figure 5 (b) provides evidence that while the World Bank is generally successful in meeting it's goal of mitigating environmental impacts, the rate at which positive and negative deviation occurs is highly variable by geographic region. We can see that most outliers are in the positive direction, with Asia being a notable exception. The projects in Oceania are in a narrow range, however, projects in other continents have a wide range.

   While both the CT and CF approaches allow for the examination of the relative importance of factors in driving heterogeneity, the interpretation and robustness of these findings is highly variable. In the case of the CT, the position of a variable in the single tree can be interpreted as importance; i.e., splits higher in the tree are more influential on the results, and path-dependencies can be examined. However, the robustness of the shape of the CT approach is unknown, and both our simulations and existing literature suggest CT findings are likely to be less accurate than CF implementations. Conversely, in a CF each covariate can be ranked across all trees in terms of the purity improvements it can provide, giving a relative indication of importance across all trees[3]. While these findings are more robust, they do not enable the interpretation of explicit thresholds (i.e., the year variable may be important, but the explicit year that is
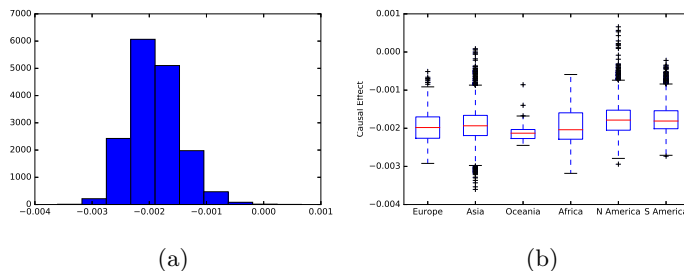
Fig. 5: (a) Causal effect distribution of all World Bank projects combined and (b) separated by continents

split on may change in the RF approach), and path dependencies are not made explicit. In our case study, we find that the first five variables in the CT and CF cases are stable between approaches, but we identify significant variance in deeper areas of the tree. For a practitioner, this allows an understanding of what the major drivers of aid effectiveness are; for example here, the purity metric highlights the dollars committed and environmental conditions as major drivers of forest cover loss, and also highlights a disparity between projects located at different latitudes; all factors which can enable a deeper understanding of what is causing success and failure in World Bank environmental initiatives. This is consistent with past findings which illustrate a stable set of covariates in the top-level of trees across a CF [13]. Further, we note that the 15 most highly ranked covariates in the CF approach are generally uncorrelated, providing an indication that the information they provide is not redundant[3]. However, we leave the interpretation of the shape of the random forest, and the insights that can be gained from it, to future research.

5 Discussion and Conclusions

This paper sought to examine the research question *What is the impact of world bank projects on forest cover?* To examine this, we contrasted four different approaches all based on variations of regression trees and random forests of trees: Transfomed Outcome Trees (TOTs), Causal Trees (CTs), Random Forest TOTs (RFTOTs), and Causal Forests (CF). We found that the method selected can have significant influence on the causal effect (or lack thereof) estimated, and provide evidence suggesting CF is more accurate than alternatives in our study context. By applying the CF approach to the case of World Bank projects, we were able to compute estimates for causal effects of individual projects; further, the prominent appearance of some covariates in trees provided us with guidance on which covariates were most important in mediating the impacts of World Bank projects. While - for most projects - the effect on forest cover is close to zero, we identified some notable exceptions, positive as well as negative ones. We

also identified two key questions that have not yet been answered in the academic literature. The first of these is how to select proper limitations on the makeup of terminal nodes - i.e., if splits that result in nodes without both control and treatment cases should be prevented, omitted, or otherwise constrained. Even after propensity score adjustments, terminal nodes with no adequate comparison cases become difficult (if not impossible) to interpret. Second, there is little literature in the machine learning space regarding how to cope with spatial spillover between treated and control cases. The Stable Unit Treatment Value Assumption (SUTVA) is common practice, but in practice the effects of a project can not be expected to be purely local in nature when observations are geographically situated.

References

1. World bank group, `http://www.worldbank.org/en/about/what-we-do`
2. AidData: (2016), `http://aiddata.org/subnational-geospatial-research-datasets`
3. Athey, S., Imbens, G.: Recursive partitioning for heterogeneous causal effects (2015)
4. Biau, G.: Analysis of a random forests model. JMLR 13(1), 1063–1095 (Apr 2012)
5. Breiman, L.: Random forests. Mach. Learn. 45(1), 5–32 (Oct 2001)
6. Denil, M., Matheson, D., de Freitas, N.: Narrowing the gap: Random forests in theory and in practice. In: ICML (2014)
7. Dunbar, B.: Ndvi: Satellites could help keep hungry populations fed as climate changes (2015), `http://www.nasa.gov/topics/earth/features/obscure_data.html`
8. Hirano, K., Imbens, G., Ridder, G.: Efficient estimation of average treatment effects using the estimated propensity score. Econometrica 71(4), 1161–1189 (2003)
9. Imbens, G.W., Rubin, D.B.: Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. Cambridge University Press (2015)
10. Meinshausen, N.: Quantile regression forests. JMLR 7, 983–999 (Dec 2006)
11. NASA: The land long term data record (2015), `http://ltdr.nascom.nasa.gov/cgi-bin/ltdr/ltdrPage.cgi`
12. Su, X., Tsai, C.L., Wang, H., Nickerson, D.M., Li, B.: Subgroup analysis via recursive partitioning. Journal of Machine Learning Research 10, 141–158 (2009)
13. Taddy, M., Gardner, M., Chen, L., Draper, D.: A nonparametric bayesian analysis of heterogeneous treatment effects in digital experimentation (2014)
14. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B 58, 267–288 (1994)
15. Vapnik, V.N.: Statistical Learning Theory. Wiley-Interscience (1998)
16. Wager, S., Athey, S.: Estimation and inference of heterogeneous treatment effects using random forests (2015)
17. Wager, S., Hastie, T., Efron, B.: Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. JMLR 15(1), 1625–1651 (Jan 2014)