Quantifying Heterogeneous Causal Treatment Effects in World Bank Aid 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 hundreds of millions of dollars in development finance to countries across the world every year. In order to ensure these funds are being spent as effectively as possible, there is a natural drive to promote better understandings of what projects work and which don't. However, the global extent of these projects results in a great deal of 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), transformed outcome tree (TOT), and random forest with TOT trees approaches to (a) segment the data into relevant 'control' and 'treatment' groups, and (b) examine the impact of World Bank projects on vegetative cover. We contrast our findings with project evaluations from the World Bank, and outcomes of more traditional, empirical econometric models.

1 Introduction

For any serious human activity, there is the natural question: what difference does it make? Identifying and quantifying causal effects from data is one of the most interesting research problem across many disciplines. For example, this arises in measuring the effectiveness of a drug in medical studies, in measuring the impact of changes in an e-commerce website design on customers, in evaluating the effectiveness of public policies. In our case, the world bank is interested in measuring the impact of aid projects it funded and supported all over the world over 30 years.

The world bank's overarching goal is in economic development; it formulates this as to "end extreme poverty by decreasing the percentage of people living on less than $1.90 a day to no more than 3%" and to "promote shared prosperity by fostering the income growth of the bottom 40% for every country" [**?**].

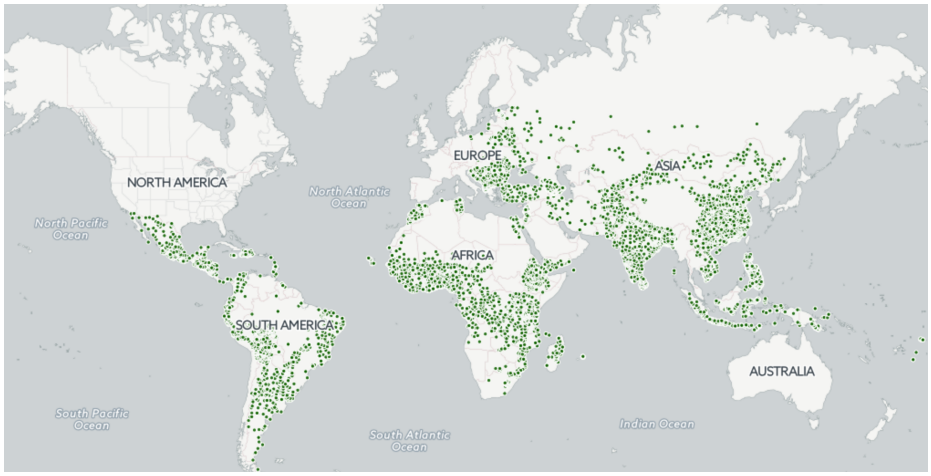


Fig. 1: world bank projects

At an abstract level, the world bank provides funding for aid projects and can rely on a metric of choice to measure differences before a project begins and after a project ends. At the level of an individual project, we face the general crux of all observational studies, i.e., we can not observe the exact same geographic, environmental, social, economic, and historical setting with or without the project. So for measuring a difference, one has to rely on making meaningful comparisons between locations that are sufficiently similar. In addition to that, the world bank's operation is large scale with a large number of projects and worldwide, which creates a huge variability in the specific kind of project, the project's size, location, socio-economic, environmental, and historical setting. Figure 1 shows the locations of a set of world bank projects with 1168 projects in 16415 locations that were performed between 2001 and 2012 on a global map.

The research questions, we investigate in this paper are:

- Can we estimate the impact of a project?
- Can we identify subsets of entities in our data set that are meaningful to compare?
- Can we identify attributes that are indicative of projects that show a positive (or negative) impact?

To do so, we enhance a given data set of world bank projects with additional information about the geographic, environmental, and economic characteristics over a number of years and rely on state-of-the-art techniques to estimate heterogenous causal effects. In our case, it is not interesting to estimate the overall average effect of all aid projects, but to identify subsets of projects by attributes and estimate average effects for individual subsets. To estimate heterogenous causal effects, there are several candidates, such as classification and regression trees [4], random forests [3], LASSO [17], and support vector machines (SVM)

[18]. In this paper, we follow the work of Athey and Imbens [1] who demonstrated how regression trees and in conclusion also random forests can be adjusted to estimate heterogenous causal effects. It is based on the Rubin Causal Model or potential outcome framework where causal effects are comparisons between observed outcomes and counterfactual outcomes one would have observed under the absence of an aid project [9]. Regression trees and random forests in traditional machine learning rely on training with data with known outcomes. Athey and Imbens showed that one can estimate the conditional average treatment effect on a subset with regressions trees after an appropriate data transformation using propensity scores. This leads to the notion of transformed-outcome trees and causal trees that we use for our analysis.

The rest of the paper is structured as follows. In Section 2, we recall the basic methodology for the calculation of transformed-outcome trees, causal trees, and random forests. Section 3 introduces the data set, its characteristics, preprocessing steps and the calculation of propensity scores. In Section 4, we present the outcome of the analysis. We conclude in Section 5.

2 Methodology

2.1 Conditional average treatment effects

Suppose we have a data set with n independently and identically distributed (iid) units with $i = 1, \cdots, n$, for each unit, it has a feature vector $X_i \in [0,1]^d$, a response $Y_i \in \mathbb{R}$ and treatment indicator $W_i \in \{0,1\}$.
For a unit-level causal effect, we can use the Rubin causal model and consider the treatment effect on unit i being $\tau_i = Y_i(1) - Y_i(0)$. In this paper, we are interested in calculating the heterogenous causal effect, which we define as $\tau(x) = \mathbb{E}\big[Y_i(1) - Y_i(0) \mid X_i = x\big]$ following [6]. Of course, in an observational study, a unit is either untreated or not, so we know either $Y_i(0)$ or $Y_i(1)$ but not both. However, one can still estimate $\tau(x)$ if one assumes unconfoundedness:

$$W_i \perp (Y_i(1), Y_i(0)) \mid X_i \tag{1}$$

Under the unconfoundedness assumption, Athey and Imbens [**?**] show that one can estimate the causal effect as

$$\tau(x) = \mathbb{E}\big[Y^* \mid X_i = x\big] \tag{2}$$

where the transformed outcome Y^* is defined as

$$Y_i^* = Y_i \cdot \frac{W_i - e(X_i)}{e(X_i) \cdot (1 - e(X_i))} \tag{3}$$

and the propensity score function $e(x)$ is defined as $e(x) = \mathbb{E}\big[W_i \mid X_i = x\big]$. Several approaches to estimate the propensity score are known [13], [8] including logic regression which we decided to use in this paper.

2.2 Transformed Outcome Tree

The transformed outcomes tree is a regression tree with Y^* instead of Y. As mentioned above, the transformed outcome is calculated with (3), then we can use traditional regression tree method to estimate the causal effect as $\tau(x) = \mathbb{E}\big[Y^* \mid X_i = x\big]$ based on the average transformed outcome of all nodes in the leaf of the tree where i resides in.

2.3 Causal Tree Model

We use regression tree to estimate the heterogeneous causal effects, the first step is to construct the tree. To construct the regression tree, we recursively partition the node until the size of the node is less than a threshold we set or the gain of split is negative.
In classic regression tree, mean square error (MSE) is often used to as the criterion for node splitting, the average value within the node is used as the estimator. Following Asthey and Imbens [1], we use (4) as the estimator and we calculate the error of the node by summing $Y_i^* - \hat{\tau}(X_i)$.

$$\hat{\tau}^{CT}(X_i) = \sum_{i:X_i \in \mathbb{X}_l} Y_i^{obs} \cdot \frac{W_i/\hat{e}(X_i)}{\sum_{i:X_i \in \mathbb{X}_l} W_i/\hat{e}(X_i)} - \sum_{i:X_i \in \mathbb{X}_l} Y_i^{obs} \cdot \frac{(1-W_i)/(1-\hat{e}(X_i))}{\sum_{i:X_i \in \mathbb{X}_l} (1-W_i)/(1-\hat{e}(X_i))} \tag{4}$$

2.4 Pruning the tree

To avoid overfitting of the tree, we need to prune the tree. We use the minimal cost complexity pruning and we define it as 5. α is the complexity parameter, with it we can construct the regression with the right size.

$$R_\alpha(T) = R(T) + \alpha \left|\tilde{T}\right| \tag{5}$$

where R(T) is the resubstitution error estimate of tree T, $\left|\tilde{T}\right|$ is defined as the complexity of the tree, which is the number of leaves in the tree,
To estimate the error of a node, we use function 6,

$$R(t) = \sum_{i=1}^{N} (Y_i - \hat{\tau}(X_i)) \tag{6}$$

where N is the total units in the nodes.

To get a sequence of α, we minimize function 7,

$$g(t,T) = \frac{R(t) - R(T_t)}{\left|\tilde{T}_t\right| - 1} \tag{7}$$

where T_t is a subtree of T rooted at node t.
We use the weakest link cutting to determine α and use it as the complexity parameter when we build the tree for the whole data set.
We use the train data set to construct the tree and then apply weakest link cutting to the tree with α starting with 0. Until there is one node in the tree, we get a series of α, $\alpha_0 < \alpha_1 < \cdots < \alpha_k$. Then we set $\beta_0 = 0$, $\beta_1 = \sqrt{\alpha_1\alpha_2}$, $\cdots$, $\beta_{k-1} = \sqrt{\alpha_{k-1}\alpha_{know}}$

We use V-fold cross validation to estimate the errors for different β and use the β with minimum error. We divide the data into k sets randomly with the same size, we use $1, 2, 3, \cdots, v-1, k$ to represent the v-th data part and (k) as the left part of the data correspond to the v-th part. For each β_k, we useV-fold cross validation to get the estimated error. The error is calculated by function 8,

$$Err(\beta; Y_i^v, X_i^v) = \sum_1^N (Y_i^v - \hat{\tau}(X_i^{(v)})) \tag{8}$$

where $N = n/V$, n is the size of the whole data set. We can then calculate error $R(T(\beta))$for each β value as the average error:

$$R(T(\beta)) = \frac{1}{V}\sum_1^V Err(\beta; Y_i^v, X_i^v) \tag{9}$$

2.5 Random forest with TOT trees

3 Data

3.1 Data sources and collection

We leverage the following data sources in this analysis:

3.2 Data pre-processing

This analysis uses three key types of data: satellite data to measure vegetation, data on the geospatial locations of World Bank projects, and covariate datasets (the sources of which are detailed above). Our primary variable of interest is the fluctuation of vegetation proximate to World Bank projects, which is derived from long-term satellite data (NASA 2015). 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). The NDVI is a metric that has been used since the early 1970s, and is one of the simplest and most frequently used approaches to approximating vegetative biomass. 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

Variables	Description	Source
Forest Cover	NASA Long Term Data Record measurements of vegetative cover	http://ltdr.nascom.nasa.gov/cgi-bin/ltdr/ltdrPage.cgi
World Bank Project Locations	Double-blind geocoded information on the geographic location of each World Bank project	http://aiddata.org/level1/geocoded/worldbank
Distance to Rivers	The calculated average distance to all rivers	http://hydrosheds.cr.usgs.gov/index.php
Distance to Commercial Rivers	Calculated average distance to all commercial rivers	http://hydrosheds.cr.usgs.gov/index.php
Distance to Roads	Distance to nearest road	http://sedac.ciesin.columbia.edu/data/set/groads-global-roads-open-access-v1
Elevation	Elevation data measured from the Shuttle Radar Topography Mission	http://www2.jpl.nasa.gov/srtm/
Slope	Slope data calculated based on the Shuttle Radar Topography Mission	http://www2.jpl.nasa.gov/srtm/
Accessibility to Urban Areas	European Commission Joint Research Centre estimation of urban travel times.	http://forobs.jrc.ec.europa.eu/products/gam/download.php
Population Density	Center for International Earth Science estimation of population density, derived from Nighttime Lights	http://sedac.ciesin.columbia.edu/data/collection/gpw-v3
Air Temperature	University of Delaware Long term, global temperature data interpolated from weather station measurements.	http://climate.geog.udel.edu/~climate/
Precipitation	University of Delaware Long term, global precipitation data interpolated from weather station measurements.	http://climate.geog.udel.edu/~climate/

Table 1: Covariates of the data sets for World Bank projects

1. The reflectance by chlorophyll is correlated with plant health, and multiple studies have illustrated that it is generally also correlated with plant biomass. In other words, healthy vegetation and high plant biomass tend to result in high NDVI values (Dunbar 2009). Using NDVI as an outcome measure has a number of other benefits, including the long and consistent time periods for which it has been calculated. 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 (NASA 2015).

The second primary dataset used in this analysis measures where - geographically - World Bank projects were located. This dataset was produced by AidData (2016), relying on a double-blind coding system where two experts employ a defined hierarchy of geographic terms and independently assign uniform latitude and longitude coordinates, precision codes, and standardized place names to each geographic feature. If the two code rounds disagree, the project is moved into an arbitration round where a geocoding project manager reconciles the codes to assign a master set of geocodes for all of the locations described in the available project documentation. This approach also captures geographic information at several levels?coordinate, city, and administrative divisions?for each location, thereby allowing the data to be visualized and analyzed in different ways depending upon the geographic unit of interest. Once geographic features are assigned coordinates, coders specify a precision code that varies from 1 (exact point) to 9 (national-level project or program). AidData performs many procedures to ensure data quality, including de-duplication of projects and locations, correcting logical inconsistencies (e.g. making sure project start and end dates are in proper order), finding and correcting field and data type mismatches, correcting and aligning geocodes and project locations within country and administrative boundaries, validating place names and correcting gazetteer inconsistencies, deflating financial values to constant dollars across projects and years (where appropriate), strict version control of intermediate and draft data products, semantic versioning to delineate major and minor versions of various geocoded datasets, and final review by a multidisciplinary working group.

In addition to the project name, the World Bank provided information on donors and amounts of funding for each project. For a subset of project, there is data on the performance evaluation of projects for evaluating effectiveness of staff. We also learnt that a single project rarely resides on a single location. It typically spreads out over a number of n locations with typical values of $n = 2$ but ranging up to $n = 10$.

3.3 Data characteristics

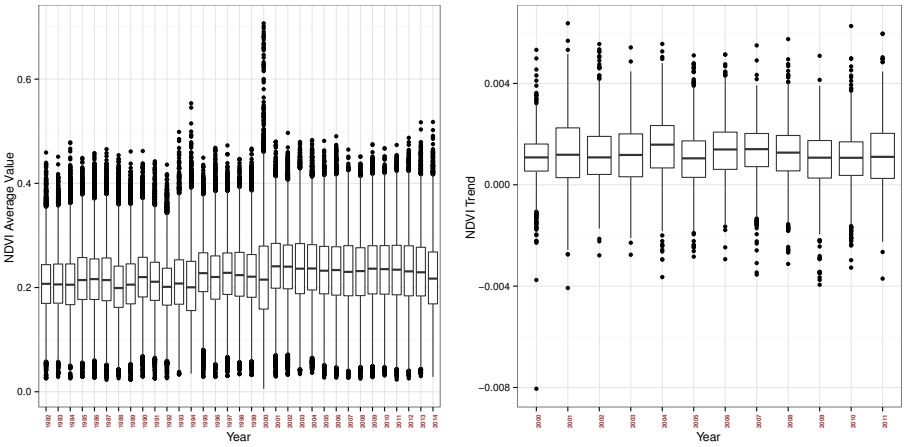
For variables listed in the Table 1 excluding the geographic location, population density and NDVI, we have minimum, maximum, and average values for each year between 1992 and 2012. For NDVI and population density, we only have a time series with average annual values. The project's geographic location is assumed constant. While annual values for covariates such as elevation and slope are typically constant over time, for others we expect to see change over time, in particular those in response to economic development such as distance to roads, accessibility to urban areas and population density. Air temperature and precipitation describe environmental conditions that may change subject to long-term trends due to a change in climate, subject to multi-year seasonal effects due to the El Nino Southern Oscillation, and subject to local or regional environmental changes, e.g. a massive deforestation. For lack of space, we only describe some

characteristics of NDVI data. Figure 2 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 all years and typical values are around 0.2 which is a lower bound for areas with vegetation. For each project location, the average annual NDVI values give a time series for which we can do a linear regression fitting to estimate the slope of an underlying trend. Figure 2 b) shows the distributions of slope values for a time series of NDVI values that starts in 1982 and ends with the year before the project starts. We see that regardless of the year the projects start, more than 75% of all projects face a starting situation where NDVI values are slightly on the rise.

- for covariates, we should provide some outline about type (categorical, ordinal,numerical), ranges of values, and if some normalization is applied. If data gives a time series, its granularity (time and space), concerns about precision (no space for too much details about the data, we can refer them to aiddata website for details)
- set of world bank projects covers a broad range of topics and individual projects do not necessarily directly target deforestation. We want to recognize projects that show an impact on forest cover, be it positive or negative. (too many projects to check, can only guess from the title)
- discuss projects and project locations: a project can take place at more than one geographic location.(The data contains projects that have between 1 and 10 locations.) One project may contains sub projects in different locations from one to hundreds. This raises the question on how to aggregate average treatment effects across a set of project locations into a single value for the overall project.

3.4 Data interpretation for the context of measuring heterogenous treatment effects

CT and TOT methods expect a separation of the data set into treated and untreated cases. A treatment in our setting is of course the fact that a project takes place. As our data set only contains project data, we make the assumption 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. In this way, we assign $W_i = 1$ if a project's funding exceeds a fixed threshold of $ xxx. The data set contains 1168 projects in 16415 locations that were performed between 2001 and 2012. Many covariates are described with time series data which leads to the question on how to compare projects that started in different years. One possibility is to retain absolute values for points in time, i.e. precipitation in 2002, which has the benefit to keep numerical comparisons of actual precipitation values across projects meaningful (if that year was a draught season in a large region) but creates the difficulty that projects have a variable set of features as we can not include covariates for the time after the project begins. Another possibility is to use time stamps relative to



(a) (b)

Fig. 2: NDVI features of all the projects

the project start, i.e. have values for $1, 2, 3, ..., k$ years before the project starts. The benefit is that all units can share the same set of covariates (after some truncation), but for example the comparison of values for precipitation 3 years before a project stars across some projects may ignore hidden global constraints. This effect gets more pronounced if one compares the outcome, the NDVI values for years following the project, in the computation of the response Y. A further option is to aggregate time series data into fewer features, e.g. to represent it by a linear model and use intercept and slope. This incorporates further assumptions and shares the same issues mentioned before for relative time stamps. In order to avoid this issue, we decided to focus on a subset of projects that all started in the same year. We selected the year 2004. As a single project typically takes place at several project locations, we decided to consider each project location as an individual unit. This means, the overall data set of 114 projects that started in 2004 at a total of 1628 project locations gives $n = 1628$ units. For the feature vector, we include longitude and latitude of the geolocation and time series data for other variables till the beginning of a project. As CT and TOT methods tolerate large numbers of covariates but can not perform functions such as average over a set of features, we decided to include annual values for time series data plus estimates of average and slope (as for example shown for NDVI in Fig. 2). The idea is to recognize if the analysis method has a preference for a particular year or for average or slope. The total length of the feature vector is $d = XXX$. All covariates are numerical and their values are taken as is and not normalized. Let $ndvi_i(92, 03)$ denote the average of NDVI values observed for project location i over the years from 1992 to 2003 before the project starts. 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 the difference of the two averages. In order to calculate Y^* for Y, we need to obtain the propensity score $e(x)$, which describes the expected value for treatment W_i for a given feature vector x. Note that the propensity score for each unit is calculated in a preprocessing step that is independent from the applied CT or TOT method.

- describe how we calculate propensity scores: We use linear model to calculate the pscore.

4 Experiments

In 2004, there are totally 114 projects contain 1628 subproject in different locations.

4.1 Software packages

Our method is based on the R package rpart [16]. Rpart support user defined split function, therefore, we can use the split criterion function 4. To improve the efficiency of the r program, we use rcpp and call C++ functions inside the split and evaluation function for each node in the tree. To further improve the c++ functions, we use openmp inside the C++ functions. To avoid the extreme cases, such as only treated or untreated data in the internal nodes, we would not split under in such condition. We use the randomForest [10] R package to build the forest.

4.2 TOT Results

- observation: calculation of $/alpha$ for pruning is computationally expensive but can be parallelized. This can be done for all methods. Discuss speed up for one of the methods.
- Resulting average treatment effect obtained from leaf node. In order to have a reasonable estimate, there is a trade off in being specific (have a refined tree with little variation but fewer units per leaf node) or being more general and have more diverse units in a leaf node but a have more values to support an estimate for an average value
- TOT allows us to compute average treatment effect for each unit i
- TOT allows us to rank covariates according to the order of variables in the tree
- TOT allows us to find similar projects to compare based on sets of units in each leaf
- Can not measure how representative the tree, unclear how to measure confidence in judgement etc
- discuss the actual outcome for a single example project that serves as an illustrating example all along the other methods (may be a particular good project)

4.3 CT Results

- Resulting average treatment effect obtained from leaf node. In order to have a reasonable estimate, one needs at least one treated and one untreated unit per leaf. In order to ensure this, one can configure the splitting rule accordingly, however this may come for the price that units remain in the same set that are not very similar.
- CT allows us to compute average treatment effect for each unit i if at least one treated, untreated unit is present
- CT allows us to rank covariates according to the order of variables in the tree
- CT allows us to find similar projects to compare based on sets of units in each leaf
- Can not measure how representative the tree, unclear how to measure confidence in judgement etc
- discuss the actual outcome for a single example project that serves as an illustrating example all along the other methods (may be a particular good project)

4.4 Random Forest Results

- Resulting average treatment effect obtained from averaging the outcomes of all TOT trees in the forest.
- RF allows us to compute average treatment effect for each unit i
- RF allows us to rank covariates according to the order of variables in the tree and quantify this in terms of quantiles (percentage of trees that use covariate among its top k covariates or alike)
- RF allows us to find similar projects to compare based on sets of units in each leaf (
- Can not measure how representative the tree, unclear how to measure confidence in judgement etc
- discuss the actual outcome for a single example project that serves as an illustrating example all along the other methods (may be a particular good project)

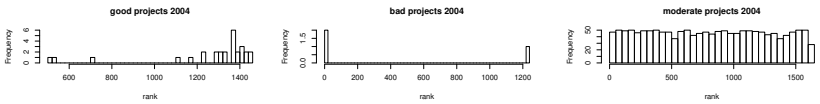
4.5 Comparison

- variable importance, important variables should be in the top levels of the tree, important to the causal effects
- regression variability, interval, in a forest, how stable the effect of a project is, if the variance is small, we can trust the result
- validate the result, one of the challenges is we do not have golden truth for the projects, we have partial result from world bank IEG which is for the assessment of the implement of the overall projects, as each projects usually have more than two sub projects on different locations, the estimation is coarse. Another source of result is from the economist result, based on these two evaluation, we can validate our work to some extent.

WB region	CT	TOT	RF	ECON
AFRICA	-0.0034007384	-0.002259834	-0.002070222	-0.037396489
EAST ASIA AND PACIFIC	-0.0004073951	-0.002602419	-0.002338640	-0.054120740
EUROPE AND CENTRAL ASIA	-0.0077536156	0.026434682	0.027010477	0.030465111
LATIN AMERICA AND CARIBBEAN	0.0089683258	0.014169258	0.010182891	-0.013437384
MIDDLE EAST AND NORTH AFRICA	-0.0025123259	-0.017290754	-0.015723577	-0.008182447
SOUTH ASIA	-0.0096351385	-0.010894742	-0.010095676	-0.081190785

Table 2: causal effect by regions

– About 83% projects in 2004 have more than 1 locations, the IEG outcomes take all of the them as a whole, in our random forest, we have causal results for sub projects, hence, one project may have both good and bad causal effects.
– quantile for each project location, variability of the causal effects, uncertainty of the result, 8% data have all causal effects have same effect, either positive or negative result, 90% have either positive or negative result from quantile 25% to 75%



(a) good projects rank (b) bad projects rank (c) moderate projects rank

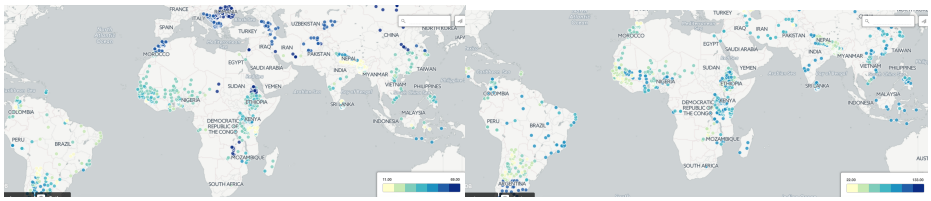
Fig. 3: original data without considering quantile

(a) Econ (b) Random Forest (c) IEG outcome

Fig. 4: Nepal area 2004

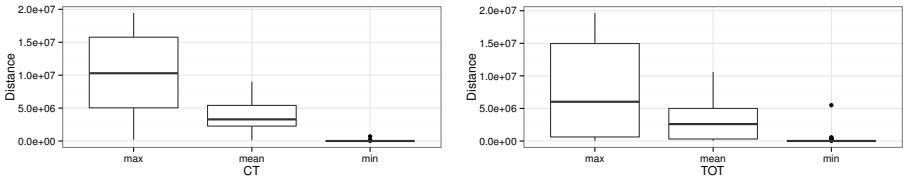
(a) Econ (b) Random Forest (c) IEG outcome

Fig. 5: East Europe



(a) CT (b) TOT

Fig. 6: projects colored by leaf they fall into



(a) CT (b) TOT

Fig. 7: distance within in the same leaf

4.6 new data set

Instead of the old data set with time series covariates, we establish the new data set which is the subset of the whole projects, which share the same project starting year, the starting years is between 2000 to 2012, project that started at 2000 has the largest number. We use projects start at 2000 to build the new data set.
In the new data set, we transform the time series covariate to the trend before the projects started and the trend after the project started along with the covariates with no time series. Then we use cross validation to choose the optimal complexity parameter and then use it to the new data set.

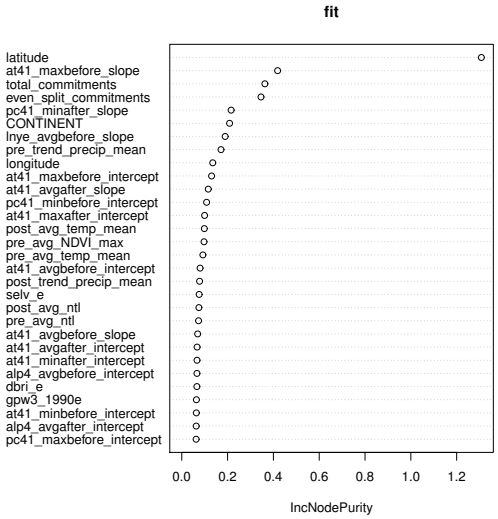


Fig. 8: variable importance 2004 projects

4.7 interpretation of results

- Interpretation of data, what can we get out of the random forest? From the random forest, we can observe the importance of each covariate as shown in 8 of year 2004. We can see that latitude is the most important among all the covariates, the other important factors includes the fund of the projects, the max temperature trend before the projects started.
- Anecdotal evidence, discuss best and worst project, what is it about, what happens here.
- Numerical values for CATE? What is the exact interpretation of the calculated values (difficult thanks to propensity score weighting), but estimate of

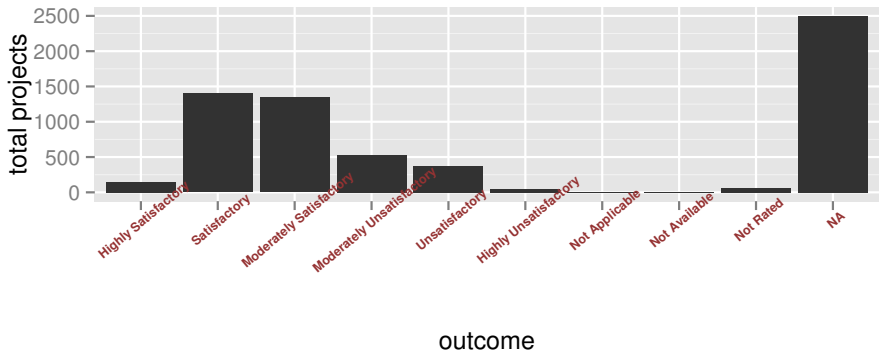


Fig. 9: IEG outcome overview

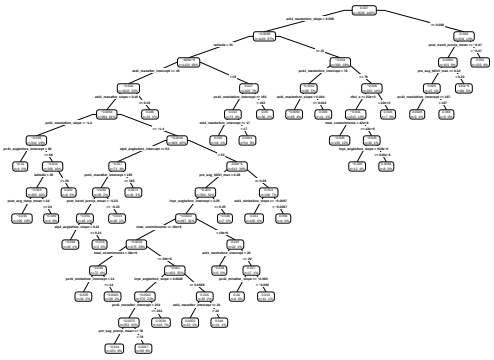


Fig. 10: causal tree of projects in 2004

average should have a direct interpretation right? Interpret value for best and worst project.
- Ranking of projects with respect to CATE?
- Selection of variables / covariates are commonly selected among trees in the random forest? Provide details, interpret results. E.g. geographic location, show maps.
- Comparison with an econometric model that looks at carbon foot print(?) Comparison with a world bank project evaluation from a human resources point of view. In figure 4c, take the Nepal area for example, most of the projects are in the medium, neither good nor bad from the IEG outcome. In the random forest model, in figure 4b, blue points are project in Nepal and the red points are projects in India, from this model, the cause effect in India is better that projects in Nepal, random forest model and economist model both have negative causal effect in Nepal, the difference between these two models is that projects have negative effect by economist model, but good effect by the random forest model. (what's the shortcoming of the econ model?) In Nepal, the projects from the title are eduction and poverty alleviation projects, while in India they are Uttaranchal Decentralized Watershed project(how to explain?)Another example is the east Europe example, in both the random forest and economist model, they estimate Romain and Serbia projects achieve good effect while the effect is bigger in the random forest model than that in the economist model.In the IEG outcome, they evaluate the projects as moderate or unsatisfactory. The projects are Transport Restructuring Project, Mine Closure, Environment and Socio-Economic Regeneration Project, Modernizing Agricultural Knowledge and Information Systems Project (MAKIS). In Bosnia and Herzegovina, random forest evaluate the projects has negative effect while the econ model rate them as positive, the projects title is Urban Infrastructure and Service Delivery Project. (No idea if these projects will cut trees or not or something else related with NDVI). The IEG outcome is Moderately Unsatisfactory for projects in this country.
- What do projects fall into the same leaf in the tree show in the map? In **??**, we can observe that both causal tree and transformed outcome tree would group projects nearby together, and we believe geographic information has big impact to the causal effect, which also in consistent with the important variable in the random forest.

5 related work

Causality [12] plays an important role in many area. In this paper, we focus on the heterogeneous causal effects. Some paper in the literature use tree based machine learning technique to estimate heterogeneous causal effects. In [14], they use statistical test as the criterion for node splitting. In [1], they use causal trees to estimate heterogeneous treatment effect. However, they do not show what if in some nodes, there is only treated units or only untreated units and then how

to estimate the heterogeneous causal effects.
Some paper use forest based machine learning technique to estimate heterogeneous causal effects. In [19], they use casual forest to do heterogeneous causal effects estimation, and they share the same idea in paper [5] that they use difference data for the structure of the tree and the estimation value within each node.
In [15], they change the item image size on ebay and observer the treatment effect as how much money people spent during the experiment. The difference between their work and ours is that they only change one factor, however, for aid data, a project may change several factor which is more complex compared to IT data. [7] Regression random forest can give estimation of the conditional means, in [11], they use quantile regression forest to estimate the distribution of the result instead of mean and they prove the algorithm is consistent. As discussed in [2], [20], , [5], there is a gap between theory property and practical use of random forest.

6 Conclusions

issues: number of control units in leaves
precision of answers
robustness of random forest for TOT
spill over effect in spatial data (big project next to untreated case)
spatial diversity in leaf nodes (not just diversity in values)

References

1. Athey, S., Imbens, G.: Recursive partitioning for heterogeneous causal effects (2015)
2. Biau, G.: Analysis of a random forests model. J. Mach. Learn. Res. 13(1), 1063–1095 (Apr 2012), http://dl.acm.org/citation.cfm?id=2503308.2343682
3. Breiman, L., Friedman, J., Stone, C., Olshen, R.: Classification and Regression Trees. The Wadsworth and Brooks-Cole statistics-probability series, Taylor & Francis (1984), https://books.google.com/books?id=JwQx-WOmSyQC
4. Breiman, L.: Random forests. Mach. Learn. 45(1), 5–32 (Oct 2001), http://dx.doi.org/10.1023/A:1010933404324
5. Denil, M., Matheson, D., de Freitas, N.: Narrowing the gap: Random forests in theory and in practice. In: International Conference on Machine Learning (ICML) (2014)
6. Hirano, K., Imbens, G., Ridder, G.: Efficient estimation of average treatment effects using the estimated propensity score. Econometrica 71(4), 1161–1189 (2003), http://EconPapers.repec.org/RePEc:ecm:emetrp:v:71:y:2003:i:4:p:1161-1189
7. Hirano, K., Imbens, G.W., Ridder, G.: Efficient estimation of average treatment effects using the estimated propensity score. Econometrica 71(4), 1161–1189 (2003)
8. Ho, D., Imai, K., King, G., Stuart, E.: Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. Political Analysis 15, 199–236 (2007)

9. Imbens, G.W., Rubin, D.B.: Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. Cambridge University Press, New York, NY, USA (2015)
10. Liaw, A., Wiener, M.: Classification and regression by randomforest. R News 2(3), 18–22 (2002), http://CRAN.R-project.org/doc/Rnews/
11. Meinshausen, N.: Quantile regression forests. J. Mach. Learn. Res. 7, 983–999 (Dec 2006), http://dl.acm.org/citation.cfm?id=1248547.1248582
12. Pearl, J.: Causality: Models, Reasoning, and Inference. Cambridge University Press, New York, NY, USA (2000)
13. Rosenbaum, P.R., Rubin, D.B.: The central role of the propensity score in observational studies for causal effects. Biometrika 70, 41–55 (1983)
14. 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), http://dblp.uni-trier.de/db/journals/jmlr/jmlr10.html#SuTWNL09
15. Taddy, M., Gardner, M., Chen, L., Draper, D.: A nonparametric bayesian analysis of heterogeneous treatment effects in digital experimentation (2014)
16. Therneau, T., Atkinson, B., Ripley, B.: rpart: Recursive Partitioning and Regression Trees (2015), http://CRAN.R-project.org/package=rpart, r package version 4.1-9
17. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B 58, 267–288 (1994)
18. Vapnik, V.N.: Statistical Learning Theory. Wiley-Interscience (1998)
19. Wager, S., Athey, S.: Estimation and inference of heterogeneous treatment effects using random forests (2015)
20. Wager, S., Hastie, T., Efron, B.: Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. J. Mach. Learn. Res. 15(1), 1625–1651 (Jan 2014), http://dl.acm.org/citation.cfm?id=2627435.2638587