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 funds aid projects in third world countries all over the world each year. There is a natural interest in better understanding what makes a project succeed and what not. In technical terms, we are interested in estimating heterogeneity in causal effects and conduct inference about the magnitude of the differences in project effects across subsets of world bank projects. To do so, we analyze a data set with data on world bank projects from 1982 to 2014 that contains project characteristics, geographical, and environmental data such as temperature and precipitation. The key challenge for this analysis is that the observational data does not allow us to directly measure the difference between the result of performing a project and of not performing a project as reality only gives us the choice to do one of the two. Following recent research results by Athey and Imbens, we employ a combination of machine learning techniques such as random forests with techniques from causal inference to measure the average treatment effect, i.e. the average effect of a project, for subsets of geographic locations. We validate our findings with project evaluations from the world bank and outcomes of competing econometric models.

1 introduction

Some notes

- general problem in development: project has a time, space, and economic dimension,
- how to measure success
- how to measure what's going on
- how to measure impact (and when), how to infer causality
- problem present in particular in aid projects for third world
- describe data
- formulate research question that is addressed

Draw causal effects from data is one of the most interesting research problem across many disciplines. For example, people want to know the effects of a drug

in medical studies, companies would like to know the effect of their advertisement on customers, government seeks to evaluate the effect of public policies, for our case, world bank wants to know the effect of the aid projects they investigate around the world over 30 years.
Instead of investigate the causal effects for the whole population, in this paper, we are interested in estimating heterogeneous causal effects for subpopulations by features or covariates. We can estimate heterogeneity by covariates on causal effects and then conducting inference for a distinct unit.
To avoid getting extreme treatment effects which lead to a spurious heterogeneous result, in disciplines such as clinical trial, they use pre-planed subgroup to analyze, for economic, they have pre-analysis plans for randomized experiments. With a data driven approach, the advantage is to discover some other causal effects instead of only the pre-planed subgroups.
To estimate heterogenous causal effects, there are several candidates, for example, classification and regression tree [4], random forest [3], LASSO [15], SVM [16] and so on. In this paper, we use the regression tree, the other methods such as random forest is also good candidate, but we focus on the regression tree in this paper.
In tradition, we can use decision trees to do prediction using the trained data or labeled data. We can build the regression tree to predicting the causal effects with the features as nodes in the tree. However, for the causal inference, the challenge is we do not have such data,in rubin causal model [10], we can only have the treated data or untreated data, but not both at the same time, hence we do not know the ground truth for prediction. We can't follow the traditional supervised machine learning method that we contract the tree with the trained the data and then use the the test data to do prediction based on the constructed tree. Follow the work of Athey and Imbens [1], we use causal tree to do heterogeneous causal effects estimation. However, in practice,for example in our case the world bank data set, within a node, there maybe only treated or untreated data, we will discuss in the paper how to explain such data and other issues

2 methodology

2.1 conditional average treatment effects

Suppose we have a data set with n 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 unit level causal effect, we can use Rubin causal model to estimate the average causal effect as shown in function 1,

$$\tau_i = Y_i(1) - Y_i(0) \tag{1}$$

In this paper, we are interested in heterogenous causal effect as 2, this estimator is proposed by [7],

$$\tau(x) = \mathbb{E}\big[Y_i(1) - Y_i(0) \mid X_i = x\big] \tag{2}$$

The challenge is we know either $Y_i(0)$ or $Y_i(1)$, but not both at the same time. We need to make a unconfoundness assumptions to estimate $\tau(x)$.

$$W_i \perp (Y_i(1), Y_i(0)) \mid X_i \tag{3}$$

Under the unconfoundness assumption, we can get the causal effect as

$$\tau(x) = \mathbb{E}\big[Y^* \mid X_i = x\big] \tag{4}$$

where Y^* is function 5, $e(x)$ is function 6, to estimate the propensity score, there are several ways for calculation such as [12], [9], in this paper, we use logic regression to calculate the pscore.

$$Y_i^* = Y_i^{obs} \cdot \frac{W_i - e(X_i)}{e(X_i) \cdot (1 - e(X_i))} \tag{5}$$

$$e(x) = \mathbb{E}\big[W_i \mid X_i = x\big] \tag{6}$$

2.2 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 7 as the estimator and we calculate the error of the node by summing $Y_i^* - \hat{\tau}(X_i)$.

$$\hat{\tau}^{CT}(X_i) = \sum_{i:X_i \in \mathbb{X}_l} Y_i^{obs} \cdot \frac{W_i/\hat{e}(X_i)}{\sum_{i:X_i \in \mathbb{X}_l} W_i/\hat{e}(X_i)} - \sum_{i:X_i \in \mathbb{X}_l} Y_i^{obs} \cdot \frac{(1-W_i)/(1-\hat{e}(X_i))}{\sum_{i:X_i \in \mathbb{X}_l} (1-W_i)/(1-\hat{e}(X_i))} \tag{7}$$

2.3 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 8. α is the complexity parameter, with it we can construct the regression with the right size.

$$R_\alpha(T) = R(T) + \alpha \left|\tilde{T}\right| \tag{8}$$

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 9,

$$R(t) = \sum_{i=1}^{N} (Y_i - \hat{\tau}(X_i)) \tag{9}$$

where N is the total units in the nodes.

To get a sequence of α, we minimize function 10,

$$g(t,T) = \frac{R(t) - R(T_t)}{\left|\tilde{T}_t\right| - 1} \tag{10}$$

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 11,

$$Err(\beta; Y_i^v, X_i^v) = \sum_1^N (Y_i^v - \hat{\tau}(X_i^{(v)})) \tag{11}$$

where $N = n/V$, n is the size of the whole data set. Then by using function 12, we can get the estimated error for each β value.

$$R(T(\beta)) = \frac{1}{V}\sum_1^V Err(\beta; Y_i^v, X_i^v) \tag{12}$$

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

Variables	Description	Source
ForestCover	NASA Long Term Data Record measurements of vegetative cover	http://ltdr.nascom.nasa.gov/cgi-bin/ltdr/ltdrPage.cgi
WorldBankProjectLocations	Double-blind geocoded information on the geographic location of each World Bank project	http://aiddata.org/level1/geocoded/worldbank
DistancetoRivers	The calculated average distance to all rivers	http://hydrosheds.cr.usgs.gov/index.php
DistancetoCommercialRivers	Calculated average distance to all commercial rivers	http://hydrosheds.cr.usgs.gov/index.php
DistancetoRoads	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/
AccesibilitytoUrbanAreas	European Commission Joint Research Centre estimation of urban travel times.	http://forobs.jrc.ec.europa.eu/products/gam/download.p
PopulationDensity	Center for International Earth Science estimation of population density, derived from Nighttime Lights	http://sedac.ciesin.columbia.edu/data/collection/gpw-v3
AirTemperature	University of Delware Long term, global temperature data interpolated from weather station measurements.	http://climate.geog.udel.edu/ climate/
Precipitation	University of Delware Long term, global precipitation data interpolated from weather station measurements.	http://climate.geog.udel.edu/ climate/

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 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

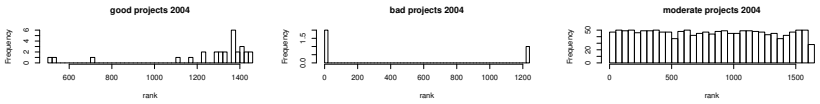
has led to a number of improvements over time to offset many of these errors. This is especially true of measurements from longer-term satellite records, such as those used in this analysis, produced from the MODIS and AVHRR satellite platforms (NASA 2015).

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.

4 experiments

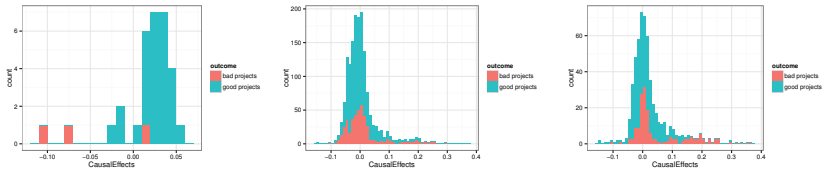
4.1 random forest with TOT trees

- 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.
-



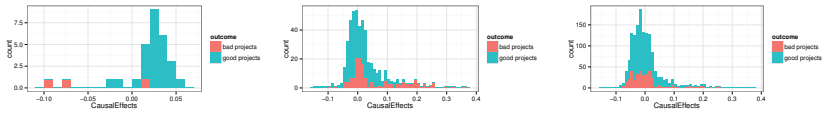
(a) good projects rank (b) bad projects rank (c) moderate projects rank

Fig. 1: original data without considering quantile



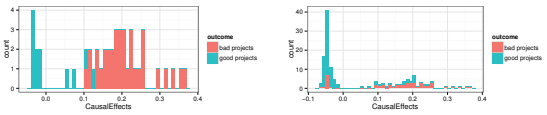
(a) Highly Sat vs Highly Unsat (b) Highly Sat, Sat vs Highly Unsat, Unsat (c) Highly Sat, Sat, Moderately Sat vs Highly Unsat, Unsat, Moderately Unsat

Fig. 2: original data without considering quantile



(a) Highly Sat vs Highly Unsat (b) Highly Sat, Sat vs Highly Unsat, Unsat (c) Highly Sat, Sat, Moderately Sat vs Highly Unsat, Unsat, Moderately Unsat

Fig. 3: data with causal effect quantile from $25\% - 50\%$ without cross 0



(a) Highly Sat, Sat vs Highly Unsat, Unsat (b) Highly Sat, Sat, Moderately Sat vs Highly Unsat, Unsat, Moderately Unsat

Fig. 4: data with causal effect quantile from $0\% - 100\%$ without cross 0

project ID(total 1628)	random forest estimation(low to high)	world bank result	economist result(4271)
P076924	6 ,21,1179	Highly Unsatisfactory	NA()
P082914	484,545	Highly satisfactory	499, 501, 654, 684 ,750, 954
P074872	711,1085 ,1182, 1224 ,1230 ,1288, 1291, 1308, 1313 ,1324, 1333, 1360, 1366, 1367, 1376, 1378 ,1379, 1383, 1396, 1400, 1403, 1417, 1429, 1434, 1443	Highly satisfactory	**??**
P075387	1463	Highly satisfactory	NA

4.2 random forest

4.3 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.

4.4 World Bank aid data

Our method is based on the R package rpart. Rpart support user defined split function, therefor, we can use the split criterion function 7. 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.

name	count	improve
latitude	1339	0.4782316
$pc41\_minbefore\_intercept$	1339	0.4760974
$pc41\_minafter\_intercept$	1339	0.4759779
$at41\_maxbefore\_intercept$	1339	0.4732124
$dbri\_e$	1339	0.4716534

5 related work

Causality [11] 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 [13], 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 [17], 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 [14], 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. [8] As discussed in [2], [18], , [5], there is a gap between theory property and practical use of random forest.

6 Conclusions

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. Geurts, P., Wehenkel, L.: Investigation and reduction of discretization variance in decision tree induction. In: Proc. of the 11th European Conference on Machine Learning (ECML-2000. pp. 162–170. Springer Verlag (2000)
7. 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
8. Hirano, K., Imbens, G.W., Ridder, G.: Efficient estimation of average treatment effects using the estimated propensity score. Econometrica 71(4), 1161–1189 (2003)
9. 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)
10. Imbens, G.W., Rubin, D.B.: Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. Cambridge University Press, New York, NY, USA (2015)
11. Pearl, J.: Causality: Models, Reasoning, and Inference. Cambridge University Press, New York, NY, USA (2000)

12. Rosenbaum, P.R., Rubin, D.B.: The central role of the propensity score in observational studies for causal effects. Biometrika 70, 41–55 (1983)
13. 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
14. Taddy, M., Gardner, M., Chen, L., Draper, D.: A nonparametric bayesian analysis of heterogeneous treatment effects in digital experimentation (2014)
15. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B 58, 267–288 (1994)
16. Vapnik, V.N.: Statistical Learning Theory. Wiley-Interscience (1998)
17. Wager, S., Athey, S.: Estimation and inference of heterogeneous treatment effects using random forests (2015)
18. 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