Recent Applications of Machine Learning in Labor Economics

Applied Reading Group

Lukas Delgado-Prieto
PhD Candidate, UC3M

October 3, 2023

Today's Presentation

- Background and Framework for estimating Treatment Effects with ML
- Implementation of Causal Forests
- Applications
  - Labor Supply Shocks
  - Job Displacement
- Remarks

Introduction

- A growing number of studies in statistics/economics are using machine learning methods to estimate causal effects
  - I'm not referring to ML methods that *improve* the estimation of causal effects (like double-debiased ML or regularization techniques)
- This can be useful for two main reasons:

1. The estimation of **Heterogeneous Treatment Effects** in a standardized way
   - Can avoid doing multiple regressions for each subgroup and then finding a compelling story of what's driving the main effects
2. The **Estimation of Treatment Effects** in out-of-sample data for policy analysis
   - Generates a trained model that can predict the impact of a certain policy for subgroups that we do not know the outcome of a policy yet!

Introduction

- A growing number of studies in statistics/economics are using machine learning methods to estimate causal effects
  - I'm not referring to ML methods that *improve* the estimation of causal effects (like double-debiased ML or regularization techniques)
- This can be useful for two main reasons:

1. The estimation of **Heterogeneous Treatment Effects** in a standardized way
   - Can avoid doing multiple regressions for each subgroup and then finding a compelling story of what's driving the main effects
2. The **Estimation of Treatment Effects** in out-of-sample data for policy analysis
   - Generates a trained model that can predict the impact of a certain policy for subgroups that we do not know the outcome of a policy yet!

Introduction

- A growing number of studies in statistics/economics are using machine learning methods to estimate causal effects
  - I'm not referring to ML methods that *improve* the estimation of causal effects (like double-debiased ML or regularization techniques)
- This can be useful for two main reasons:

1. The estimation of **Heterogeneous Treatment Effects** in a standardized way
   - Can avoid doing multiple regressions for each subgroup and then finding a compelling story of what's driving the main effects
2. The **Estimation of Treatment Effects** in out-of-sample data for policy analysis
   - Generates a trained model that can predict the impact of a certain policy for subgroups that we do not know the outcome of a policy yet!

Introduction

- A growing number of studies in statistics/economics are using machine learning methods to estimate causal effects
  - I'm not referring to ML methods that *improve* the estimation of causal effects (like double-debiased ML or regularization techniques)
- This can be useful for two main reasons:

1. The estimation of **Heterogeneous Treatment Effects** in a standardized way
   - Can avoid doing multiple regressions for each subgroup and then finding a compelling story of what's driving the main effects
2. The **Estimation of Treatment Effects** in out-of-sample data for policy analysis
   - Generates a trained model that can predict the impact of a certain policy for subgroups that we do not know the outcome of a policy yet!

Background

The main papers we build on today are:

- Athey, S., Imbens, G. (2016). "Recursive partitioning for heterogeneous causal effects". Proceedings of the National Academy of Sciences
- Athey, S., Tibshirani, J., & Wager, S. (2019). "Generalized Random Forests". The Annals of Statistics
  - By a recursive partitioning method (**causal trees**), this framework quantifies individual TEs and can yield predictions
- There is another complementary paper with a different but more general approach: Chernozhukov et al. (NBER WP, 2018)
  - Instead of partitioning the data, it groups the ATEs into subgroups G (**GATEs**)
- We focus today on the empirics. These papers have many technical details that I will not cover

Background

The main papers we build on today are:

- Athey, S., Imbens, G. (2016). "Recursive partitioning for heterogeneous causal effects". Proceedings of the National Academy of Sciences
- Athey, S., Tibshirani, J., & Wager, S. (2019). "Generalized Random Forests". The Annals of Statistics
  - By a recursive partitioning method (**causal trees**), this framework quantifies individual TEs and can yield predictions
- There is another complementary paper with a different but more general approach: Chernozhukov et al. (NBER WP, 2018)
  - Instead of partitioning the data, it groups the ATEs into subgroups G (**GATEs**)
- We focus today on the empirics. These papers have many technical details that I will not cover

Background

The main papers we build on today are:

- Athey, S., Imbens, G. (2016). "Recursive partitioning for heterogeneous causal effects". Proceedings of the National Academy of Sciences
- Athey, S., Tibshirani, J., & Wager, S. (2019). "Generalized Random Forests". The Annals of Statistics
  - By a recursive partitioning method (**causal trees**), this framework quantifies individual TEs and can yield predictions
- There is another complementary paper with a different but more general approach: Chernozhukov et al. (NBER WP, 2018)
  - Instead of partitioning the data, it groups the ATEs into subgroups G (**GATEs**)
- We focus today on the empirics. These papers have many technical details that I will not cover

Framework

Suppose we have the following model with random assignment of a binary treatment:

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_i * D_i + \epsilon_i \tag{1}$$

We can run OLS to estimate the ATE or CATE:

$$ATE = E[Y_i(1) - Y_i(0)], CATE(x) = E[Y_i(1) - Y_i(0)|X_i = x] \tag{2}$$

But what happens when X_i contains many variables, possibly with interactions?

- You will begin the quest to estimate multiple regressions with interactions or sample restrictions (*subject to arbitrary decisions*)
- Even if we show that the effect is higher in certain subgroups, how are we sure what's the main driver of the effects (*apart from statistical noise*)

Framework

Suppose we have the following model with random assignment of a binary treatment:

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_i * D_i + \epsilon_i \tag{1}$$

We can run OLS to estimate the ATE or CATE:

$$ATE = E[Y_i(1) - Y_i(0)], CATE(x) = E[Y_i(1) - Y_i(0)|X_i = x] \tag{2}$$

But what happens when X_i contains many variables, possibly with interactions?

- You will begin the quest to estimate multiple regressions with interactions or sample restrictions (*subject to arbitrary decisions*)
- Even if we show that the effect is higher in certain subgroups, how are we sure what's the main driver of the effects (*apart from statistical noise*)

Framework

Suppose we have the following model with random assignment of a binary treatment:

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_i * D_i + \epsilon_i \tag{1}$$

We can run OLS to estimate the ATE or CATE:

$$ATE = E[Y_i(1) - Y_i(0)], CATE(x) = E[Y_i(1) - Y_i(0)|X_i = x] \tag{2}$$

But what happens when X_i contains many variables, possibly with interactions?

- You will begin the quest to estimate multiple regressions with interactions or sample restrictions (*subject to arbitrary decisions*)
- Even if we show that the effect is higher in certain subgroups, how are we sure what's the main driver of the effects (*apart from statistical noise*)

Causal Tree

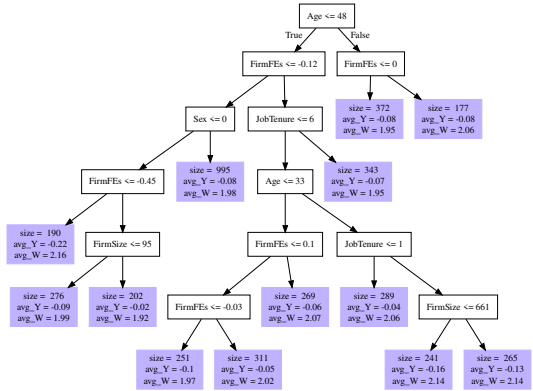
The procedure of Athey & Imbens (2016) and Athey et al. (2019) forms decision trees according to the difference in treatment effects

- Start with specification $Y_i = \tau(x_i)D_i + \epsilon_i$, multiple variables in X_i and sample P

Causal Tree

The procedure of Athey & Imbens (2016) and Athey et al. (2019) forms decision trees according to the difference in treatment effects

- Start with specification $Y_i = \tau(x_i)D_i + \epsilon_i$, multiple variables in X_i and sample P



Algorithm of Causal Trees

1. Use sample P. You can divide it for computational burden and use the remaining out-of-bag (OOB) sample for further prediction
2. Take a random subsample, without replacement, of P and choose a variable randomly from X_i (normally, it chooses all variables, but when there are many, it needs to randomize which one it chooses to start)
3. For every possible value of one variable in X_i, the data is split into two partitions (say P_l and P_r) to estimate treatment effects separately using the main specification. Choose the variable with its cutoff value that maximizes the squared difference in treatment effects:

$$(\tau_l - \tau_r)^2. \tag{3}$$

4. Obs. with a value below (above) or equal to the cutoff value are placed into a new left (right) node of the decision tree
5. Recursively forms the resulting nodes until they reach a min node size, the difference in sample size between the two partitions is large, or when the split would yield a difference relatively small

Algorithm of Causal Trees

1. Use sample P. You can divide it for computational burden and use the remaining out-of-bag (OOB) sample for further prediction
2. Take a random subsample, without replacement, of P and choose a variable randomly from X_i (normally, it chooses all variables, but when there are many, it needs to randomize which one it chooses to start)
3. For every possible value of one variable in X_i, the data is split into two partitions (say P_l and P_r) to estimate treatment effects separately using the main specification. Choose the variable with its cutoff value that maximizes the squared difference in treatment effects:

$$(\tau_l - \tau_r)^2. \tag{3}$$

4. Obs. with a value below (above) or equal to the cutoff value are placed into a new left (right) node of the decision tree
5. Recursively forms the resulting nodes until they reach a min node size, the difference in sample size between the two partitions is large, or when the split would yield a difference relatively small

Algorithm of Causal Trees

1. Use sample P. You can divide it for computational burden and use the remaining out-of-bag (OOB) sample for further prediction
2. Take a random subsample, without replacement, of P and choose a variable randomly from X_i (normally, it chooses all variables, but when there are many, it needs to randomize which one it chooses to start)
3. For every possible value of one variable in X_i, the data is split into two partitions (say P_l and P_r) to estimate treatment effects separately using the main specification. Choose the variable with its cutoff value that maximizes the squared difference in treatment effects:

$$(\tau_l - \tau_r)^2. \tag{3}$$

4. Obs. with a value below (above) or equal to the cutoff value are placed into a new left (right) node of the decision tree
5. Recursively forms the resulting nodes until they reach a min node size, the difference in sample size between the two partitions is large, or when the split would yield a difference relatively small

Estimation of Treatment Effects

- The estimation of individual treatment effects is more complicated than running a linear regression
  - The algorithm uses two subsamples of the data: one for the splits and one for the estimation, and then it creates similarity weights for weighted estimators (**very intensive computationally!**)

- **Intuition:**

  1. With the OOB data and according to each obs. characteristics, it assigns them into a final node of each tree of the causal forest
  2. For all trees, count the times this obs. falls in the same terminal node as the training sample for the weights (similar to nearest neighbor matching or kernel estimators)
  3. The weighted mean of treatment effects across trees yields the individual treatment effect $\tau(x_i)$

Estimation of Treatment Effects

- The estimation of individual treatment effects is more complicated than running a linear regression
  - The algorithm uses two subsamples of the data: one for the splits and one for the estimation, and then it creates similarity weights for weighted estimators (**very intensive computationally!**)

- **Intuition:**

  1. With the OOB data and according to each obs. characteristics, it assigns them into a final node of each tree of the causal forest
  2. For all trees, count the times this obs. falls in the same terminal node as the training sample for the weights (similar to nearest neighbor matching or kernel estimators)
  3. The weighted mean of treatment effects across trees yields the individual treatment effect $\tau(x_i)$

Implementation

Several pages explain how to implement the algorithm:

- In `R` you can use the package `grf` (see documentation developed by the authors)
- In `Python` you can use the package `EconML` (see documentation)
  - This can also estimate the Chernozhukov et al. (2018) algorithm
- In `Stata` through `R` with the `MLRtime` package (see documentation)

The algorithm takes several tunable parameters as given, but they can be adapted

1. Chosen by default or optimally with validation or cross-validation (*tuning all parameters may lead to issues*)
2. **Number of trees** in the causal forest (need to grow more for CIs)
3. **Minimum node size** in the tree (reduces overfitting)

Implementation

Several pages explain how to implement the algorithm:

- In `R` you can use the package `grf` (see documentation developed by the authors)
- In `Python` you can use the package `EconML` (see documentation)
  - This can also estimate the Chernozhukov et al. (2018) algorithm
- In `Stata` through `R` with the `MLRtime` package (see documentation)

The algorithm takes several tunable parameters as given, but they can be adapted

1. Chosen by default or optimally with validation or cross-validation (*tuning all parameters may lead to issues*)
2. **Number of trees** in the causal forest (need to grow more for CIs)
3. **Minimum node size** in the tree (reduces overfitting)

Keep in mind

- The treatment can be binary or continuous but only works in settings when we have exogenous variation (**the algorithm assumes this**)
  - Still, it works with IV but only with one instrument (possible to predict and directly estimate 2S)

- Standard issues of random forests might apply to causal forests (Athey and Imbens, Annu. Rev. Econ., 2019)

- As the number of values or categories of a predictor increases, it might appear in more splits of the decision trees (Strobl et al., 2007)

- The complexity of the tree can induce overfitting (grow more trees), partitions can change in different samples, and it has poor inference with a small number of obs. (Check Escanciano's class)

Keep in mind

- The treatment can be binary or continuous but only works in settings when we have exogenous variation (**the algorithm assumes this**)
  - Still, it works with IV but only with one instrument (possible to predict and directly estimate 2S)

- Standard issues of random forests might apply to causal forests (Athey and Imbens, Annu. Rev. Econ., 2019)

- As the number of values or categories of a predictor increases, it might appear in more splits of the decision trees (Strobl et al., 2007)

- The complexity of the tree can induce overfitting (grow more trees), partitions can change in different samples, and it has poor inference with a small number of obs. (Check Escanciano's class)

Keep in mind

- The treatment can be binary or continuous but only works in settings when we have exogenous variation (**the algorithm assumes this**)
  - Still, it works with IV but only with one instrument (possible to predict and directly estimate 2S)

- Standard issues of random forests might apply to causal forests (Athey and Imbens, Annu. Rev. Econ., 2019)

- As the number of values or categories of a predictor increases, it might appear in more splits of the decision trees (Strobl et al., 2007)

- The complexity of the tree can induce overfitting (grow more trees), partitions can change in different samples, and it has poor inference with a small number of obs. (Check Escanciano's class)

Application 1: Immigration

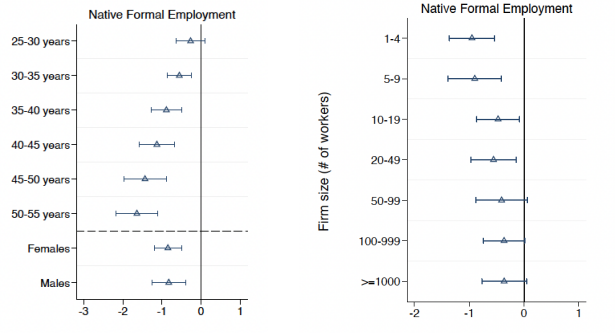
Delgado-Prieto, L. (2023). "Immigration and Worker Responses Across Firms: Evidence from Administrative Records in Colombia."

- I study the impact of a labor supply shock on workers using the Colombian matched employee-employer dataset
  - $N = 6.7M$ for employment outcomes and $N = 4.1M$ for wage outcomes
- The effects are heterogeneous across worker and firm characteristics

Application 1: Immigration

Delgado-Prieto, L. (2023). "Immigration and Worker Responses Across Firms: Evidence from Administrative Records in Colombia."

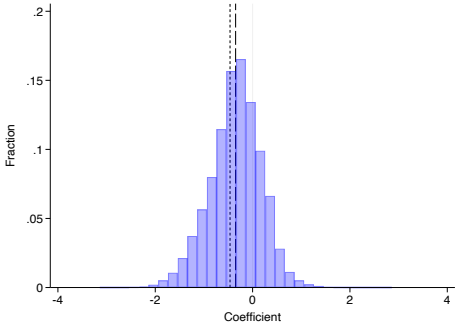
- I study the impact of a labor supply shock on workers using the Colombian matched employee-employer dataset
  - $N = 6.7M$ for employment outcomes and $N = 4.1M$ for wage outcomes
- The effects are heterogeneous across worker and firm characteristics



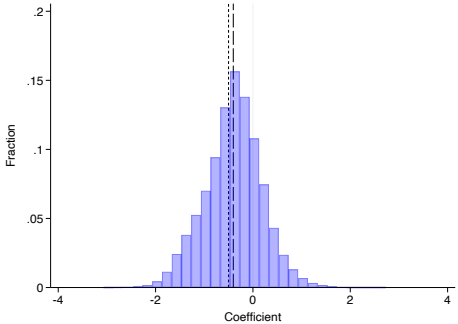
(a) Worker's Age **(b) Firm Size**

Application 1: Immigration

Histogram of treatment effects from the causal forest



(a) Employment



(b) Wages

Note: The short dashed line refers to the coefficient from the benchmark specification, and the long dashed line refers to the average predicted treatment effects that are estimated with the trained causal forest using the OOB sample. The number of trees is 2,000. The minimum node size is 300.

Application 1: Immigration

Subgroups most affected regarding employment

Table: Most affected native workers, 2018-2015

	Q1	Q2	Q3	Q4	Q5
Male (%)	0.7	0.6	0.5	0.5	0.5
Age of worker	42.8	40.3	38.5	35.1	31.1
Job tenure (1-9 years)	2.3	3.6	4.4	4.1	2.8
Monthly wages (USD)	324.8	462.6	521.8	478.4	336.2
Median firm size	79	105	276	510	1109
Quantiles of firm FEs (1-7)	3.8	5.3	6.0	6.3	6.5

Note: These tables report the descriptive statistics for quintiles of treatment effects according to the predictions of the trained causal forest using the OOB sample. The wages are transformed from Colombian pesos to USD using 2020 exchange rates from the World Bank.

Application 1: Immigration

Subgroups most affected regarding wages

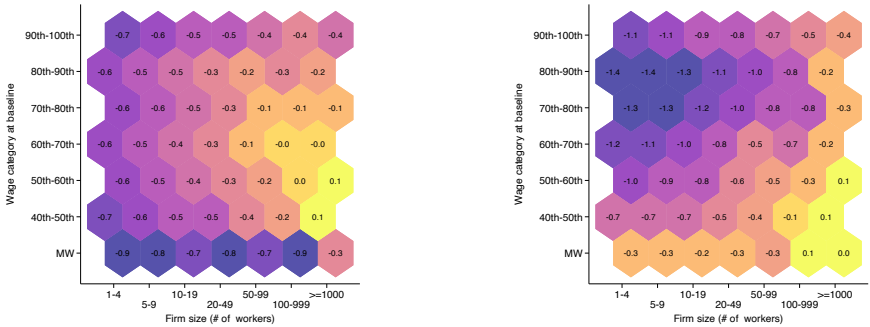
Table: Most affected native workers, 2018-2015

	Q1	Q2	Q3	Q4	Q5
Male (%)	0.6	0.6	0.6	0.6	0.5
Age of worker	36.6	38.5	38.8	38.1	37.5
Job tenure (1-9 years)	3.2	3.9	4.0	3.8	3.5
Monthly wages (USD)	559.5	466.2	419.3	379.0	393.7
Median firm size	86	189	242	309	892
Quantiles of firm FEs (1-7)	5.7	5.8	5.6	5.5	5.5

Note: These tables report the descriptive statistics for quintiles of treatment effects according to the predictions of the trained causal forest using the OOB sample. The wages are transformed from Colombian pesos to USD using 2020 exchange rates from World Bank.

Application 1: Immigration

Heat plot of treatment effects

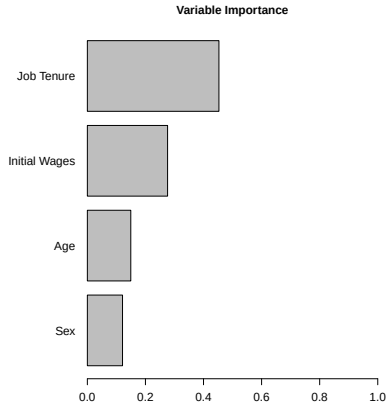


(a) Employment **(b) Wages**

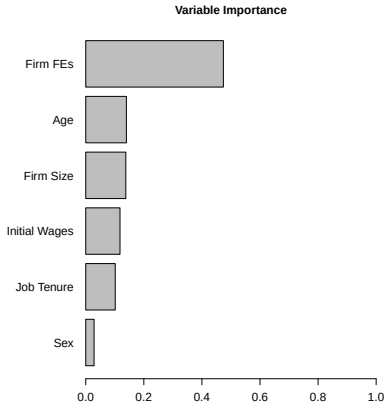
Note: Each hexagon is the mean treatment effect for that subgroup according to the trained causal forest. The sample is restricted to natives between 25 and 55 years old. I use clusters of FUAs for the causal forest estimation. The causal forest uses 50% of the main sample due to computational burden.

Application 1: Immigration

Causal Forest of Formal Employment



(a) Without Firm Variables

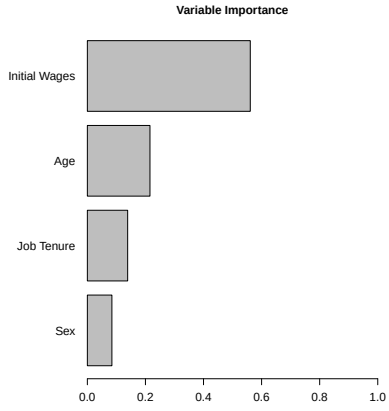


(b) With Firm Variables

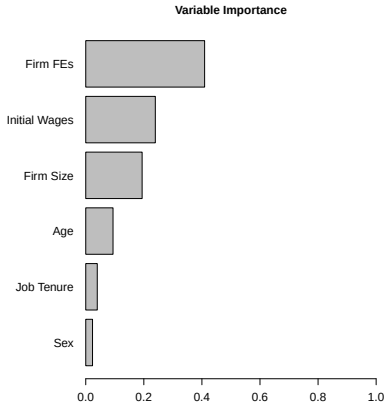
Note: Variable importance is a weighted sum of how many times the feature f appears in the split of each leaf of every tree in the forest. Number of trees=2,000. The importance measure sum up to 1. Minimum node size=300.

Application 1: Immigration

Causal Forest of Formal Wages



(a) Without Firm Variables



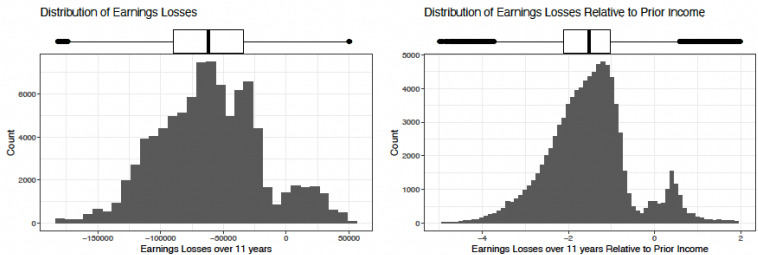
(b) With Firm Variables

Note: Variable importance is a weighted sum of how many times the feature f appears in the split of each leaf of every tree in the forest. Number of trees=2,000. The importance measure sum up to 1. Minimum node size=300.

Application 2: Job Displacement

Gulyas, A., & Pytka, K. (2021). "Understanding the sources of earnings losses after job displacement: A machine-learning approach."

- Using Austrian administrative data and mass layoffs, they quantify substantial heterogeneity in the individual cost of job displacement
- Test which channel is the most important in explaining these losses with causal forests



(a) Absolute Losses (in Euros) (b) Relative to prior income

Figure 3: Distribution of cumulative earnings losses over the first 11 years after job displacement. Estimates from a generalized random forest

Application 2: Job Displacement

- The most important being firm wage premia and the avg. firm premia in the region
  - From 15 variables they include in the algorithm
- Mean reversion in firm wage premia and losses in match quality

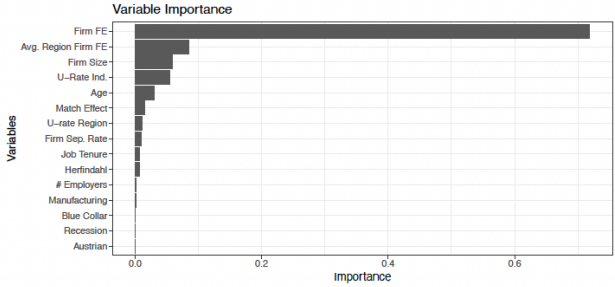
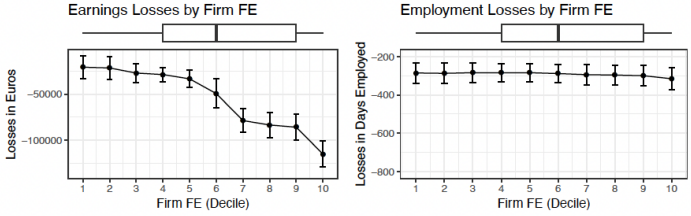


Figure 8: Depth-adjusted variable frequency in splits in the GRF with a decay exponent equal to -2 and the maximum depth level of nodes equal to 4. All values sum to 1.

Application 2: Job Displacement



(a) Earnings Losses

(b) Employment Losses

(c) Wage Losses

(d) Losses in Firm Wage Premia

Figure 9: GRF estimates with 95% CI of losses in earnings, employment, wages, and firm premia by deciles of firm fixed effect. All other variables are set to their median values. The boxplots present the distribution of the partitioning variable in the dataset

Application 3: Job Displacement

Athey, S., Simon, L. K., Skans, O. N., Vikstrom, J., & Yakymovych, Y. (2023). "The Heterogeneous Earnings Impact of Job Loss Across Workers, Establishments, and Markets"

- Same question as before using causal forests but with administrative data in Sweden
- Extensive heterogeneous effects of job displacement

Figure 2: Displacement effects interacted with standardized continuous variables

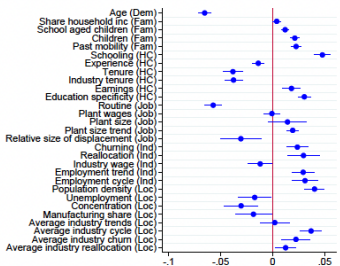
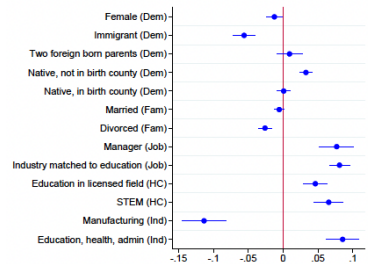


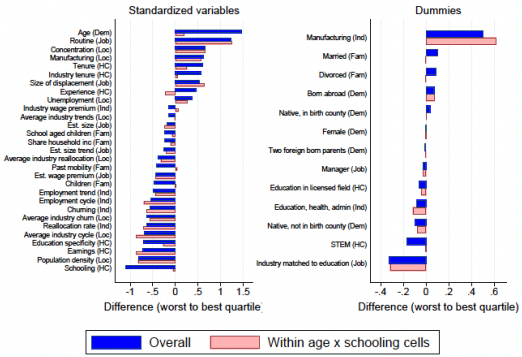
Figure 3: Displacement effects interacted with dummy variables



Application 3: Job Displacement

- They do not corroborate the importance of firm wage premia in the earnings losses
- Instead, it is a combination of individual and job-industry factors
  - Different from before, they use the 1st year of displacement as the outcome, take many more variables in the algorithm, and construct firm wage premia differently

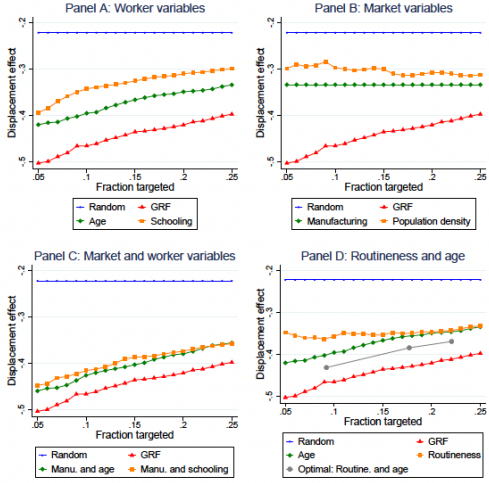
Figure 6: Differences in characteristics across CATE quartiles



Application 3: Job Displacement

Creating Policy Targets

Figure 9: Displacement effects on earnings for workers selected by different targeting policies



Conclusion

- The benefits of these methods are evolving and seem to be getting more attention from economists

- The costs of applying these methods are also decreasing as computers are getting better, and more packages and useful guides are available

- The decision to apply them seems very case-specific
  1. A setting with rich datasets, not only in the number of observations but in the number of variables
  2. A question where HTE can arise prominently and want to uncover relationships
  3. Predict impact of a policy

- It is still being determined whether an entire applied paper can be based on this method. There is still some tension between simplicity from standard regressions to algorithmic modeling (*can combine both*)

Conclusion

- The benefits of these methods are evolving and seem to be getting more attention from economists

- The costs of applying these methods are also decreasing as computers are getting better, and more packages and useful guides are available

- The decision to apply them seems very case-specific
  1. A setting with rich datasets, not only in the number of observations but in the number of variables
  2. A question where HTE can arise prominently and want to uncover relationships
  3. Predict impact of a policy

- It is still being determined whether an entire applied paper can be based on this method. There is still some tension between simplicity from standard regressions to algorithmic modeling (*can combine both*)

Conclusion

- The benefits of these methods are evolving and seem to be getting more attention from economists

- The costs of applying these methods are also decreasing as computers are getting better, and more packages and useful guides are available

- The decision to apply them seems very case-specific
  1. A setting with rich datasets, not only in the number of observations but in the number of variables
  2. A question where HTE can arise prominently and want to uncover relationships
  3. Predict impact of a policy

- It is still being determined whether an entire applied paper can be based on this method. There is still some tension between simplicity from standard regressions to algorithmic modeling (*can combine both*)

Thank you!

Please write me for more information or my own codes: **ludelgad@eco.uc3m.es**