Annex I.
Alternatives to Regression: Partial Dependence Plots

Louis Anthony (Tony) Cox, Jr.

Our main findings show that that growth rates for some specific forms of regulation are negatively associated with land productivity growth (i.e., yield growth). However, our analysis makes heavy use of specific econometric models such as equations (1)-(3). It is natural to wonder how robust our conclusions are to possible errors in model specification and, more generally, to model uncertainty. To find out, we applied a useful innovation from machine learning: an ensemble of several hundred non-parametric (classification and regression tree) models fit to the data, with the frequency distribution of predictions of yield growth over the entire population of models providing a quantitative indicator of the extent of model uncertainty. This appendix briefly explains the method and illustrates specific results. They support the conclusion that growth in some forms of growth is negatively associated with yield growth even when non-parametric methods and large model ensembles are used to avoid possible model specification errors and to assess model uncertainty.

Random Forest

We applied one of the best-known and most widely used machine learning techniques for predictive analytics: the Random Forest algorithm as implemented in the *randomForest* R package and made available for general users via the Causal Analytics Toolkit (CAT) platform.

Links to these resources are as follows:

- General concepts explained for Wikipedia readers: https://en.wikipedia.org/wiki/Random_forest
- *randomForest* R package: https://cran.r-project.org/web/packages/randomForest/randomForest.pdf
- Causal Analytics Toolkit (CAT): http://cox-associates.com:8899/

The CAT software integrates randomForest with other packages to generate partial dependence plots (PDPs), uncertainty bands, and individual conditional expectation (ICE) plots. The key ideas of these techniques are as follows.

- A *partial dependence plot* (PDP) shows how the predicted value of a dependent variable (here, *yield_growth* for NAICS 111110—the Soybean Farming industry) varies as an explanatory variable is changed (here, a measure of regulatory growth. We use "*l_rel_restrictions_growth21*"—the one-year lagged value of growth rate in regulatory restrictions associated with command-and-control regulations—for purposes of illustration) while holding all other variables fixed at their current values in the data set.[1] (This corresponds to a "natural direct effect" of the predictor on the dependent variable, in the terminology of causal analysis.) The PDP in Figure 1 averages predicted values of *yield growth* from 500 trees in a random forest ensemble of predictive models as the value of the predictor *l_rel_restrictions_growth21* is varied.

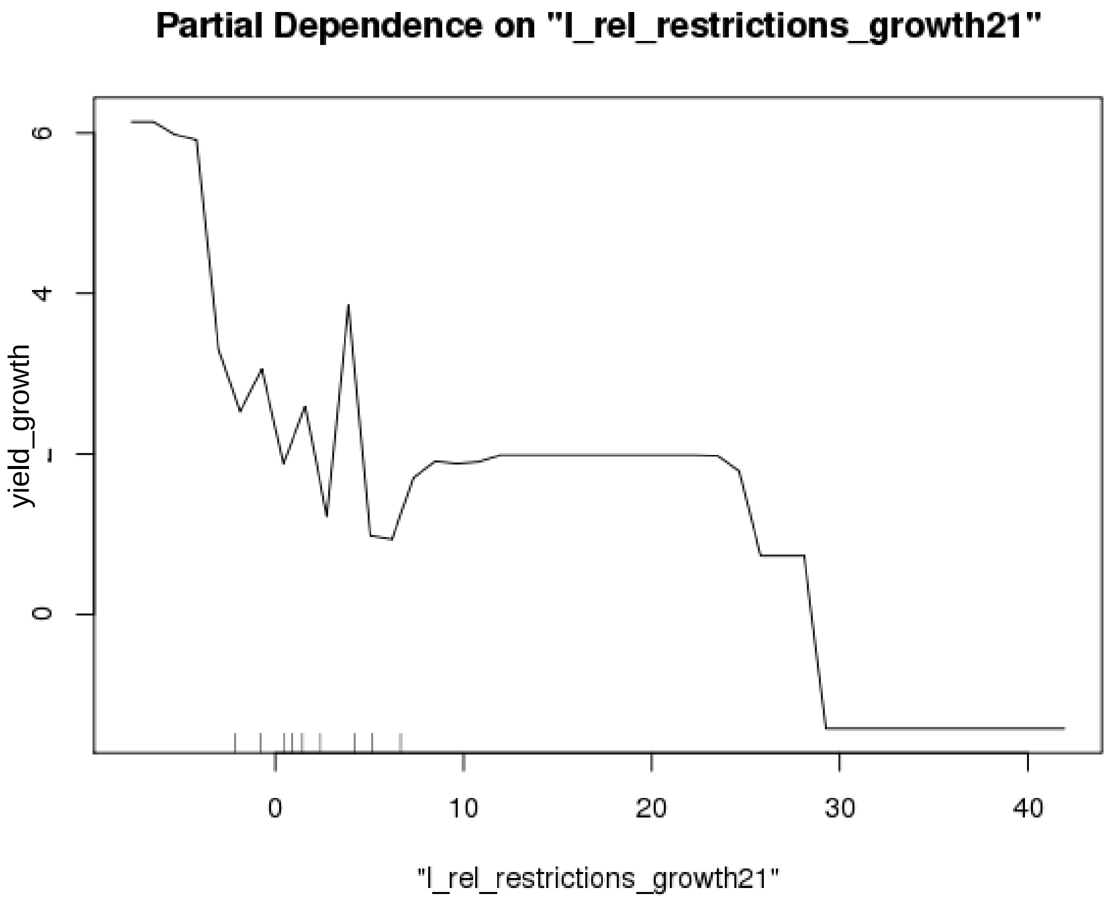
- *Uncertainty bands* for the PDP, illustrated in Figure 2, are derived by discarding the most extreme tails (in Figure 2, the upper 5% and lower 5%) of the frequency distribution of predicted values of the dependent variable for each value of the explanatory variable being varied. The width of the uncertainty band at any point reflects the impact of model uncertainty on predictions: if the correct model were known with certainty, then there would be no spread in these bands, since the correct value could be predicted with certainty by using the correct model.

- *Individual conditional expectation (ICE) cluster plots* search for heterogeneity (e.g., sensitive subpopulations of years in which variations in regulatory growth had exceptionally large or small impacts on yield growth) by clustering predicted responses for each individual case (row) in the data set. Figure 3 shows that, in this application, such clustering did not identify much heterogeneity: the ICE cluster plots are roughly parallel and relatively close to each other, indicating that the approximate size and direction of effects of changes in *l_rel_restrictions_growth21* is on changes in *yield_growth* were relatively uniform throughout the data set.

Overall, these results suggest that our main findings from regression analysis hold up well when more flexible and relatively assumption-free (non-parametric ensemble learning) techniques are used. Although we only explored machine learning methods for a few selected associations, the results illustrated in Figures 1-3 suggest that measures of causal impacts (natural direct effects) that do not depend on the relatively restrictive assumptions of regression modeling still show that measures of growth in some types of regulations are negatively associated with yield growth.

[1] Variables are defined and constructed in the same ways as in Chapter 4.

Figure 1. Example of a PDP generated by CAT software. Yield growth (*yield-growth*) in industry 111110 is significantly negatively associated with growth in regulations (*l_rel_restrictions_growth21*) in non-parametric analysis (Spearman's rack correlation = 0.083, p < 0.00001)



The partial dependence plot shows that the association between < l_rel_restrictions_growth21 > and < yield_growth > is: significantly negative (based on Spearman's rank correlation of -0.836 and p-value 0.00000)

Figure 2. Example of 90% uncertainty bands around the PDP showing the range within which 90% of model predictions for *yield_growth* lie for each value of *l_rel_restrictions_growth21*. Even with model uncertainty, it is clear that the association is negative.

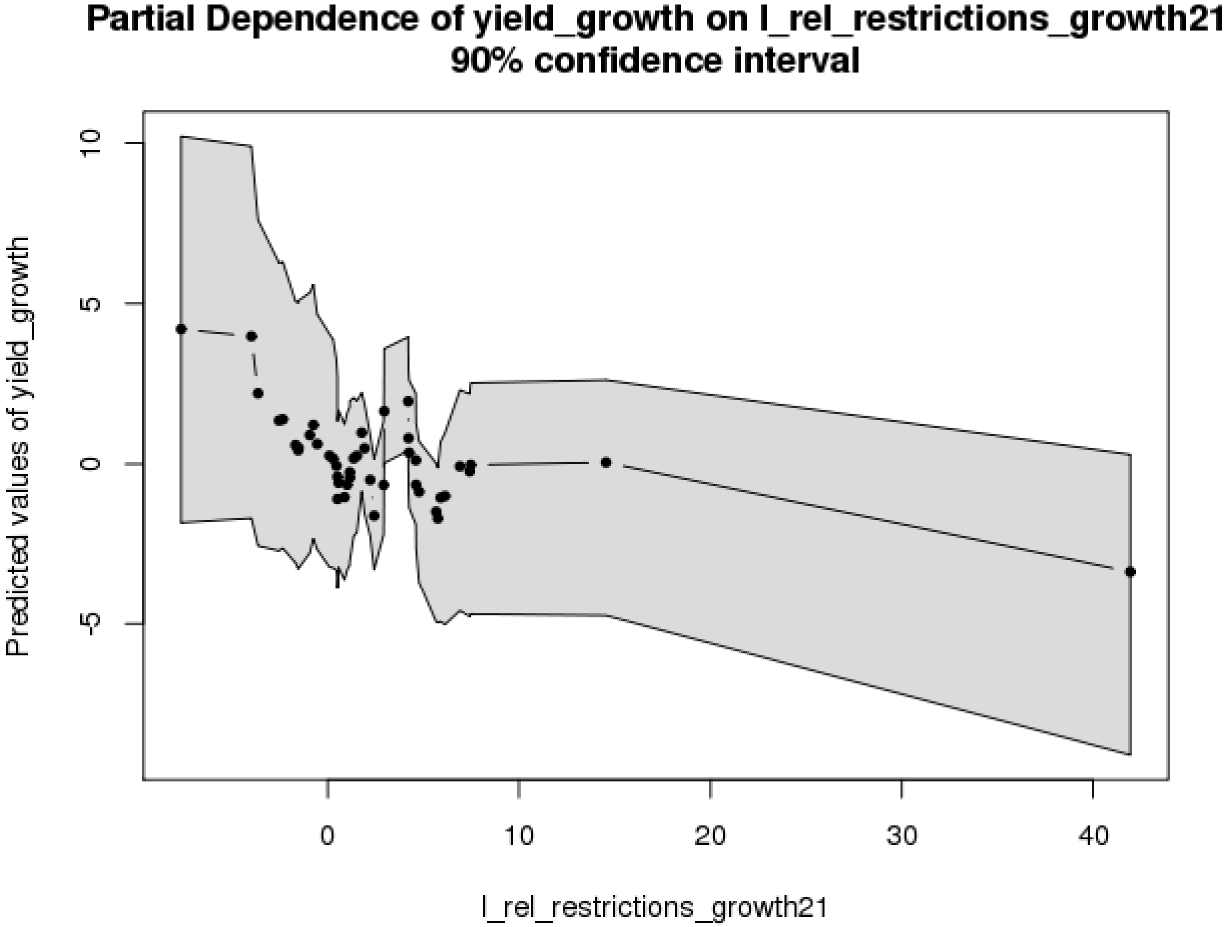
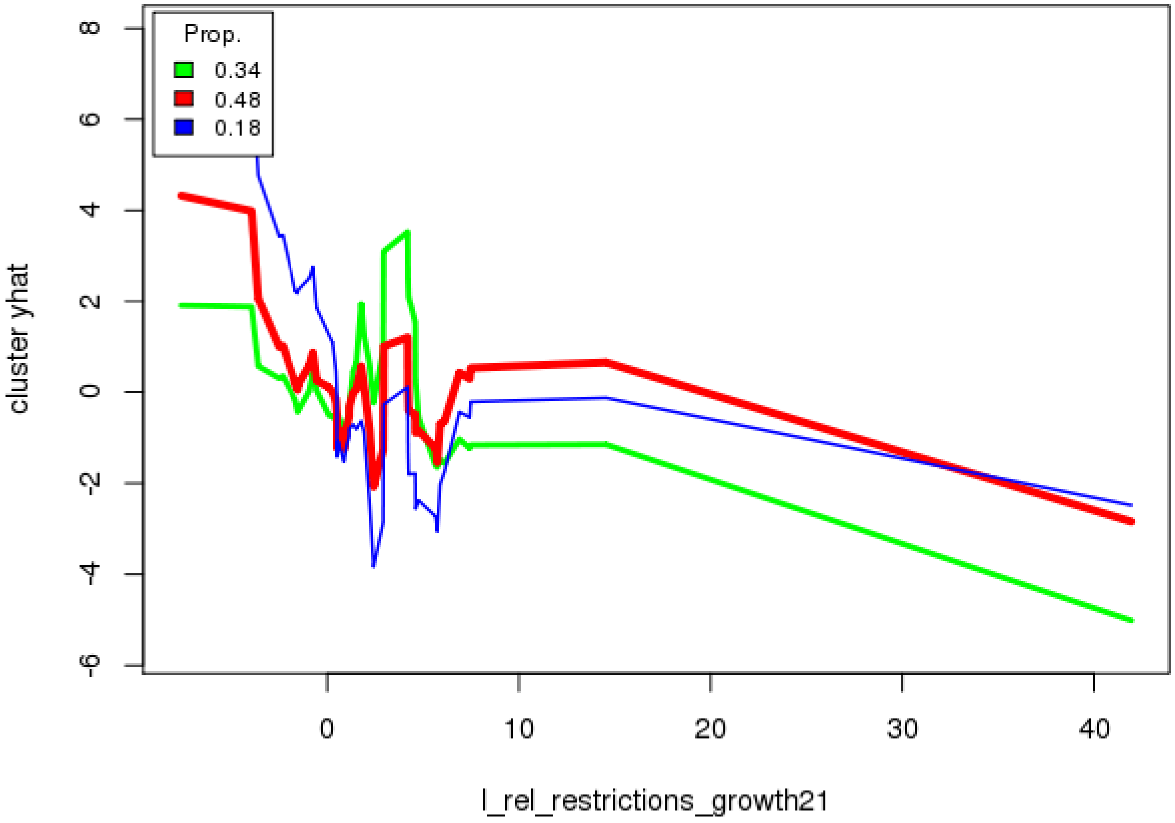


Figure 3. Example of individual conditional expectation (ICE) cluster plot showing the heterogeneity of model predictions for *yield_growth* from *l_rel_restrictions_growth21*. Clustering the model predictions reveals that there is little heterogeneity: the plots are roughly parallel.


Prop.
🟩 0.34
🟥 0.48
🟦 0.18
cluster yhat
l_rel_restrictions_growth21