Tree-based Estimation of Heterogeneous Dynamic Policy Effects

Steve Miller[a,*]

[a]*Department of Applied Economics, University of Minnesota, St. Paul, MN, USA*

April 5, 2018

Abstract

Understanding variation in policy effects in both cross-sectional and temporal dimensions can help improve policy targeting, set expectations regarding timing of effects, and assess whether policy implementation has improved over time. To that end, this paper extends recent work adapting regression trees and random forests for identification of heterogeneous treatment effects. Specifically, I adapt those methods to a panel setting, illustrating how dynamic selection assumptions and estimators can be combined with causal forest approaches to jointly investigate not only cross-sectional heterogeneity, but also dynamics of policy effects and changes in implementation effectiveness. To illustrate and evaluate the utility of the approach, I reanalyze how property rights impact the probability of natural resource collapse, finding substantial heterogeneity and nonlinear time dependencies not identified in prior work.

JEL Codes: C23; C22; Q22

*Email: s-miller@umn.edu. Phone: 1-612-625-3212. I am grateful for generous input from many others, including seminar participants at the University of Minnesota, Utah State University, Iowa State University, Arizona State University, the University of Alaska Anchorage, the University of Colorado, conference participants at AERE 2017, Kelly Bishop, Corbett Grainger, and Richard Startz.

1. Introduction

Understanding how the impacts of public policies vary is crucial to evaluating past efforts, setting expectations for the performance of current policies, and targeting future interventions. Variability in policy and program effects is pervasive: labor market interventions (Lechner 2002; Gerfin and Lechner 2002; Crump et al. 2008), changes to class size in schools (Bandiera et al. 2010), and transfer programs (Becker et al. 2013) all may impact different groups in different ways. Similarly, environmental policies may induce a range of emissions abatement across firms (Fowlie 2010), changes in ozone concentration across space (Auffhammer and Kellogg 2011), or modified consumption in response to informational programs (Shimshack et al. 2007; Allcott 2011). Ignoring this variability can lead to misleading conclusions; a policy may have no effect on average but strong impacts on specific people, firms, or locations. (Auffhammer et al. 2009).

One key challenge with many commonly used methods for studying heterogeneity in policy effects is that they require prespecification of its form. Two standard approaches are (1) to interact a treatment variable with one or more cross-sectional characteristics in a parametric model, or (2) to partition the data into groups and estimate a single average effect (parametrically or not) separately for each subset. Both routes require important assumptions about the way in which policy effects vary, codified either by choices of variables with which to interact a treatment variable or by groups of observations on which to estimate separate policy effects. Robustness of estimates can be probed by estimating multiple models under different assumptions, but the range of potential forms of heterogeneity is large.

Two recent contributions by Athey and Imbens (2016) and Wager and Athey (2017) offer a potential way forward, adapting tree-based methods from machine learning to uncover the form of heterogeneity while striving to maintain a causal interpretation of estimates. In line with the second approach described above, the resulting methods, which those authors call 'Causal Trees' and 'Causal Forests,' estimate effects on subsamples, but modify automatic partitioning procedures (regression trees and random forests) to identify the 'best' subsam-

ples for that purpose. The key insight from those studies is that partitions can be compared by the mean squared error between estimated and true treatment effects, and crucially, that error can be estimated without knowledge of the true treatment effect. The main benefit of the resulting method is that the researcher no longer has to prespecify the form of heterogeneity, though she still must choose the set of variables with which the estimated treatment effect is allowed to vary. Identification in these approaches is achieved via an assumption of conditional unconfoundedness, though related methods offer alternatives (Athey et al. 2016).

The principal goal of this paper is to adapt these tree-based methods to study how policy effects vary not only cross-sectionally, but also through time. Effect strength may vary with policy duration for a variety of reasons, e.g., information diffusion, learning, repeated exposure, or the growth of a resource (Hamilton 1995; Lalive et al. 2008; Allcott and Rogers 2014; Costello et al. 2008). In addition, if policy exposure is staggered through time, individuals or firms subjected to the policy starting at a later date may experience different impacts if the effectiveness of the intervention has changed. Further, interactions among these dimensions are possible: the way in which policy effects vary through time may depend on cross-sectional characteristics. As with cross-sectional heterogeneity, a model could be fit separately for each possible treatment duration and policy start year, or for particular time ranges identified by the researcher. The approach proposed here automates that process, identifying temporal ranges over which policy effects can be well approximated by an average.

The proposed approach extends the tree-based methods from Athey and Imbens (2016) and Wager and Athey (2017) in four key ways. The first two build on the suggestion by Athey and Imbens (2016) that their methods can be adapted to study non-randomized treatments. To this end, I first account for non-random and dynamic selection into treatment by adopting the Sequential Conditional Independence Assumption (SCIA) (Robins et al. 2000; Lechner and Miquel 2010), which is a dynamic analog of the unconfoundedness assumption employed in Athey and Imbens (2016). Second, policy effects are estimated using inverse probability of treatment weighting (IPTW) for each group in a given partition, which constructs a

weighted difference between treated and control outcomes with weights inversely proportional to the probability that an entity experienced its entire observed sequence of policy exposure (Robins et al. 2000; Lechner and Miquel 2010). The combination of SCIA and IPTW yields a weighting estimator appropriate for panel settings with dynamic selection into treatment; see Azoulay et al. (2009) for an example application to patenting. Third, I modify the partitioning criterion from Athey and Imbens (2016) to reflect the variance of the IPTW estimator. Finally, to allow the algorithm to partition based on policy start date and duration in addition to cross-sectional characteristics, I introduce a mechanical data pre-processing step that both replicates and labels control observations with different *potential* treatment sequences.

To evaluate the potential benefits of the resulting approach, I revisit whether a common form of rights-based management helps avoid resource collapse in fisheries around the world. Costello et al. (2008) originally examined this question on a global scale using standard panel methods and propensity score adjustments, finding that individual quota (IQ) programs helped reduce the probability of collapse with effect sizes increasing with policy duration. Because the tree-based methods presented here rely primarily on a similar identification strategy, any differences in findings between the studies are plausibly attributable to the methods themselves.

Applying the tree-based methods to a newly compiled and updated global dataset paints a much richer picture of heterogeneity in policy effectiveness. Point estimates of the effect of IQs on the probability of collapse range from a reduction of 17% to an increase of 25%, with an overall mean of 5% reduction. The effectiveness of IQs varies with natural resource characteristics and pre-policy status, policy duration, and year of policy introduction. The most beneficical policy effects accrue in fisheries with moderate catch levels just prior to the policy change: fisheries with already high, stable levels of resource use are likely well managed before the policy change, while previously depleted resources may be optimally closed for several years under the new policy (Reed 1979). Further, as is to be expected

from bioeconomic models (Clark and Munro 1975), effectiveness need not evolve linearly through time: a significant impact of IQs on the probability of collapse takes several years to materialize. Finally, evidence of policy effects on resource collapse is weaker for more recent implementations. In summary, individual quota programs do appear to reduce the probability of fishery collapse, but in a manner that is far from uniform.

This paper complements a growing body of work on the use of machine learning methods for studying heterogeneity in program or policy effects. Applications of tree-based methods to study cross-sectional heterogeneity include Davis and Heller (2017), Handel and Kolstad (2017), and Bertrand et al. (2017). Asher et al. (2016) extend the idea of automated partitioning to generalized method of moments estimation in each subsample, but maintain a focus on cross-sectional heterogeneity. A few studies examine effects at different intervals after program implementation, allowing for cross-sectional heterogeneity within each interval. Bertrand et al. (2017) look at short-run and long-run program effects, but estimate those effects in separate models. With many potential post-treatment effects to be estimated (e.g. one per potential treatment start year and duration), estimating a separate model per effect could become cumbersome or force the researcher to choose duration cutoffs in order to estimate a smaller set of models. Knaus et al. (2017) estimate cross-sectionally averaged effects for different durations, and hand-select a few durations for which to separately estimate cross-sectional heterogeneity, but stop short of automatically exploring potential duration splits or interactions between temporal and cross-sectional heterogeneity.[1]

Given that regression trees and random forests are nonparametric methods, it is also worth noting how the proposed method differs from existing nonparametric approaches to estimating heterogeneous treatment effects. In particular, under the identifying assumption maintained here, kernel regression or other nonparametric techniques can be used directly to

[1] An additional difference is that Knaus et al. (2017) also estimate heterogeneity using the least absolute shrinkage and selection operator (LASSO) rather than tree-based methods. Tree-based methods start with a single average effect and gradually increase heterogeneity ("top down"), while LASSO begins with a fully heterogeneous model and then selects a subset of interactions ("bottom up").

estimate conditional means of counterfactual outcomes, with the difference at a particular covariate values serving as an estimate of the heterogeneous treatment effect. With appropriate covariate-specific and locally-adaptive bandwidths (Racine and Li 2004; Li and Racine 2004; Hall et al. 2007), such a kernel regression can produce counterfactual estimates based on local neighborhoods in a manner similar to regression trees.[2] A defining characteristic and potential advantage of the tree-based methods presented here is that neighborhoods are chosen to adapt to the shape of the estimated treatment effect function. In contrast, bandwidth selection during estimation of counterfactuals via nonparametric methods will typically identify neighborhoods based on the shape of the counterfactual functions themselves. Since the treatment effect may vary when one counterfactual does not, and the treatment effect may be constant when both counterfactuals are highly variable, the adaptivity of the tree-based methods may offer finite-sample performance gains.

The remainder of this paper is as follows. In the next section, I briefly review the causal tree method proposed by Athey and Imbens (2016) and an extension to random forests suggested by Wager and Athey (2017). In the third section, I formally adapt these proposals to estimating treatment effects which may also vary with policy start time and duration. The fourth section applies the resulting method to study the effectiveness of individual quota systems in fisheries, and the final section concludes.

2. Causal Trees and Forests

One way to understand how a policy's effects vary is to estimate its average impact on different subsamples. With an appropriate set of assumptions and a corresponding estimator, the resulting set of estimates can be interpreted as a step function approximation to the true policy effect. A key challenge, of course, is how to select a set of subsamples (locations of steps) so that the approximation of the policy effect function is a good one.

A recent proposal by Athey and Imbens (2016) cleverly automates the selection of cross-

[2] A regression tree identifies disjoint neighborhoods defined by covariate values, and within each neighborhood uses a uniform kernel.

sectional subsamples by adapting a regression tree algorithm. Regression trees are predictive models that make a single prediction for each subsample in a dataset, with the set of subsamples chosen via an algorithm that seeks to minimize squared prediction error.[3] The algorithm begins by estimating a quantity of interest and associated error for the entire sample. Next, the sample is split in two based on the values of a single covariate at a time, estimates are computed for each subsample, and the improvement in estimation error is noted. This calculation is repeated for many possible splits and the split yielding the largest improvement in estimation error is selected. Because the number of potential splits is large (and infinite if at least one covariate is continuous), not all possible splits can be practically considered. The algorithm considers only threshold splits for continuous variables, with possible thresholds located between values in the sample itself. For categorical predictors, treatment effects are estimated per category, categories are sorted according to their estimated treatment effect, and splits can then be considered as if the predictor were continuous.[4] After the first split of the data, each subsample is then recursively split until further divisions give a reduction in prediction error that is sufficiently small. The resulting partition can be viewed as a tree with the final subsamples as leaves, giving the algorithm its name.

The key insight from Athey and Imbens (2016) is that this procedure can be adapted to estimate policy effects, even though the estimation error cannot be evaluated directly because true policy effects are not observed. Those authors illustrate that under a selection-on-observables assumption, the expected mean squared error (EMSE) of an unbiased policy effect estimator can be estimated, so that the regression tree procedure sketched above can be based on estimates of the policy effect alone. Specifically, if an unbiased estimator $\hat{\tau}(x)$ of a conditional average treatment effect $\tau(x)$ is available (conditioning on a vector of covariates x), Athey and Imbens (2016) show that minimizing EMSE is equivalent to maximizing the

[3]Variants of the algorithm allow for non-uniform estimates/predictions within subsamples, while other objective functions are designed for classification tasks.

[4]If treatment effects are continuous, this procedure is justified by the results of Fisher (1958); see also (Breiman et al. 1984; Chou 1991).

criterion

$$Q \equiv E[\tau^2(x)] - E[Var(\hat{\tau}(x))], \tag{1}$$

which can be estimated by

$$\hat{Q} \equiv \frac{1}{N_{tr}} \sum_i \hat{\tau}^2(x_i) - \left(\frac{1}{N_{tr}} + \frac{1}{N_{est}} \right) \sum_i \widehat{Var}(\hat{\tau}(x_i)), \tag{2}$$

where i indexes observations and N_{tr} and N_{est} are the number of observations in training and estimation datasets, respectively. The training dataset is used to select a partition, while the estimation dataset is used to estimate treatment effects given a fixed partition. This split-sample approach, which Athey and Imbens (2016) term "honest estimation", partly addresses overfitting concerns in an adaptive procedure like the regression tree algorithm.

When evaluating potential splits in the regression tree procedure, the best split is defined as that which causes $\hat{Q}$ to increase by the largest amount.[5] There are two components to this criterion. The first term rewards identification of treatment effects which vary across subpopulations. The second term in the criterion penalizes variance of the treatment effect estimator within a subpopulation. This helps control too aggressively seeking heterogeneity through overfitting. Together, these components have intuitive appeal. The criterion will be highest when the treatment effect is well approximated by a step function: constant within but varying across subpopulations. The resulting modified algorithm maximizing this criterion is referred to as a causal tree.

Wager and Athey (2017) extend this idea by building many such step-function approximations to the policy effect function: a forest of trees. While a single step-function approximation (tree) is highly variable, averaging over many different approximations can provide more stable results. The random forest algorithm (Breiman 2001) does exactly that, building each tree from a dataset resampled from the original data, considering a random subset of

[5]The regression tree algorithm is greedy, seeking the largest local improvement to $\hat{Q}$. It does not guarantee that the final partition will yield the global maximum of $\hat{Q}$.

variables each time the data is split, and averaging over the set of tree estimates to generate a single forest estimate. Wager and Athey (2017) adapt the random forest procedure to use causal trees and establish key asymptotic properties of the resulting causal forest. Aside from the modified splitting criteria, Wager and Athey (2017) also construct tree-specific datasets by subsampling without replacement rather than sampling with replacement as in the original random forest algorithm from Breiman (2001).

In what follows, I build on these ideas to examine how policy effects vary across time in a panel data setting. Both Athey and Imbens (2016) and Wager and Athey (2017) consider cross-sectional heterogeneity; as shown below, these ideas can be readily extended to study how policy effects depend on both when a policy started and how long it has been in place.

3. Estimating policy effects which vary both cross-sectionally and temporally

In this section, I detail how causal trees can be adapted to estimate policy effects that may vary not only with cross-sectional characteristics, but also with both calendar time and policy duration. The proposed method retains the broad structure of a causal forest: a set of causal trees is constructed from resampled datasets, and the conditional average policy effect is estimated as the average of relevant estimates across the trees. However, four main modifications are needed to deal with the dynamic nature of treatment. The first two modifications borrow from existing work on dynamic treatment regimes. First, the assumptions necessary to identify a conditional average policy effect must be changed to reflect a policy adoption (selection) process that occurs over and may vary with time. Second, the estimation method used on any subsample must be updated to reflect the dynamic nature of both treatment and the corresponding effects. Third, the splitting criterion used must be updated to reflect the variance of this modified estimator. Finally, the handling of control observations during the splitting procedure requires elaboration. In a dynamic treatment setting, a control observation can potentially aid in estimation of a counterfactual for treated units first subjected to a policy at different times. As such, when splitting the data on either

policy duration or policy start time, it may be useful to have a control observation appear in both subsamples.[6] These modifications are all relatively straightforward, and I address them below in turn after setting up the estimation objective.

Estimation target

The goal of estimation is to approximate the effect of a policy on an outcome of interest, where that effect may vary with the characteristics of an affected entity, the time the policy went into effect ("start time"), and how long an entity has been subject to the policy ("duration"). The true policy effect will be approximated with a step function, where breaks in the step function may occur with cross sectional characteristics, start time, or duration. Each step corresponds to a constant average policy effect estimated on a subsample.

To formalize the problem, suppose we observe outcomes Y_{it}, characteristics X_{it}, and a binary treatment status T_{it} for entity i at time t. The treatment indicator T_{it} is equal to 1 if entity i is treated at time t and is 0 otherwise. For brevity, let the history of any variable be denoted by its boldface version, e.g., $\mathbf{X}_{it} = \{X_{i1}, ..., X_{it}\}$. To simplify analysis, in what follows I focus on cases in which treatment is permanent – once a policy goes into effect, it remains in effect. This allows representation of a sequence of treatments by a duration $D_{it} = \sum_{t'=1}^{t} T_{it'}$, which simplifies both notation and interpretation of results. Further, to simplify explanation I refer to units of time as years.

With this notation the estimation target can be defined. In the language of potential outcomes popularized by Rubin (1974), each step is a conditional average treatment effect (CATE), with conditioning ensuring an observation falls within a particular step. Denoting a *potential* start year by s and a *potential* policy duration by d, the CATE is defined by

$$\tau(\mathcal{X}, \mathcal{D}, \mathcal{S}) \equiv E[Y_{it}(d, s) - Y_{it}(0, s) | \mathbf{X_{is-1}} \in \mathcal{X}, d \in \mathcal{D}, s \in \mathcal{S}, t = s + d - 1], \tag{3}$$

[6] In a working paper, Prest (2017) independently proposes a related replication strategy for multi-valued treatments, but does so in a setting of randomized treatment introduced simultaneously for all treated units and does not appear to use the method to investigate dynamics.

where $Y_{it}(d, s)$ is the potential outcome for entity i in year t if subjected to d years of treatment beginning in start year s. The CATE is simply the average difference in outcomes for an entity in year t if it had been subjected to d years of a policy starting in year s compared to no treatment in that same timeframe. The sets $\mathcal{X}$, $\mathcal{D}$, and $\mathcal{S}$ define which pre-treatment characteristics, potential treatments, and potential treatment start years the CATE applies to. Note that the characteristics $\mathbf{X_{is-1}}$ used for conditioning are restricted to those appearing prior to the start of the potential treatment.[7] In contrast, selection into treatment is dynamic and allowed to depend on characteristics occurring after year $s-1$. I next make this identifying assumption explicit.

Identifying assumption

Since is impossible to observe i in year t under two different treatment histories, further assumptions are needed to estimate (3). Because of the dynamic, non-random assignment to treatment, a suitable choice is the Sequential Conditional Independence Assumption (SCIA) (Robins et al. 2000; Lechner and Miquel 2010). Like conditional independence in a static context, SCIA essentially states that after conditioning on enough information, whether or not an entity is treated in the current period is as good as random. The conditioning information includes current and past exogenous characteristics as well as past endogenous variables. Letting X_{it}^{EX} and X_{it}^{EN} denote the exogenous and potentially endogenous subsets of covariates, respectively, the SCIA states:

$$Y_{it}(d, s) \perp\!\!\!\perp T_{it} | \mathbf{T}_{it-1}, \mathbf{X}_{it}^{EX}, \mathbf{X}_{it-1}^{EN}. \tag{4}$$

Importantly, the conditioning set includes past treatment and lagged endogenous variables (possibly including lagged outcomes), since both may affect current treatment and outcomes

[7] Characteristics occurring after the initial exposure to treatment could be used for conditioning if we were interested in, e.g. comparing the effect of d years of treatment to that of $d-1$ years of treatment. However, the focus here is on the total effect of d years of treatment as compared to a counterfactual of no treatment.

and thus confound estimation.

Estimation

Under the SCIA, the average effect of a treatment sequence compared to the absence of treatment in any subsample can be estimated several ways, including using inverse probability of treatment weighting (IPTW). The intuition behind IPTW is identical to that for weighting estimators in cross-sectional settings: selection bias overweights observations selected into a particular treatment sequence, and the IPTW estimator attempts to recover a sample that reflects the underlying population by inverting that weighting. The key difference in a dynamic setting is that the weights correspond to the inverse probability that an entity experienced its entire sequence of treatment exposure.

The probability of an entity experiencing its actual sequence of treatment is the product of conditional probabilities of observed treatment in each year. Formally,

$$p(\mathbf{T}_{it}) = \prod_{t'=1}^{t} P(T_{it'}|\mathbf{T}_{it'-1}, \mathbf{X}_{t'}^{EX}, \mathbf{X}_{t'-1}^{EN}). \tag{5}$$

The aforementioned focus on permanent treatment regimes, in which an entity subjected to treatment will be treated in all following years, allows for simplification of these probabilities. Specifically, under permanent policy change the probability of remaining in treatment is one, and the probabilities of interest can be rewritten as functions of duration and start year:

$$p(d,s,t) = \begin{cases} p(T_{is}=1|\mathbf{T}_{is-1}=\mathbf{0}, \mathbf{X}_{is}^{EX}, \mathbf{X}_{is-1}^{EN}) \prod_{t'=1}^{s-1} p(T_{it'}=0|\mathbf{T}_{it'-1}=\mathbf{0}, \mathbf{X}_{it'}^{EX}, \mathbf{X}_{it'-1}^{EN}) & \text{if } d>0 \\ \prod_{t'=1}^{t} p(T_{it'}=0|\mathbf{T}_{it'-1}=\mathbf{0}, \mathbf{X}_{it'}^{EX}, \mathbf{X}_{it'-1}^{EN}) & \text{if } d=0 \end{cases}. \tag{6}$$

Intuitively, the probability of nonzero treatment duration is the probability of beginning treatment in the specified start year and not before, while the probability of zero duration

is simply the probability of never getting treated in any year.

Each conditional probability can be estimated as the probability of starting treatment in a single year conditioned on observed information, and the overall probability of a treatment sequence can be constructed from those estimates. For example, in the empirical application, I estimate the conditional probabilities using a pooled logit model with year effects, though more flexible estimation approaches are possible.

As with cross-sectional analyses involving estimated probability of treatment, both overlap and balance assessments should be conducted before probabilities are used to construct weights. While overlap could be assessed for the probability of entire treatment sequences, since selection happens on an annual basis, it may be more practical to assess overlap for the conditional probability of transitioning into treatment. That conditional probability is the model that underlies the treatment sequence probabilities and can be assessed using standard methods. Further, as in cross-sectional IPTW estimation, it may be prudent to use trimming to cap the influence of observations with extremely low estimated probability of observed treatment sequence.

Because the conditioning set for the SCIA and treatment probabilities will differ across observations from different points in time, some care must be taken in estimating a treatment effect using a subsample with observations from multiple years. In particular, conditioning on $\mathbf{X_{is-1}} \in \mathcal{X}$ as in (3) restricts attention to observations with similar pre-treatment characteristics, but the pre-treatment period may differ across observations if $\mathcal{S}$ contains more than one start year. Similarly, even if $\mathcal{S}$ has a single element so that conditioning on $\mathbf{X_{is-1}}$ refers to a uniquely defined pre-treatment period, if $\mathcal{D}$ has multiple elements, the observations y_{it} used to estimate (3) may come from different years. One way forward is to estimate $\tau(\mathcal{X}, \mathcal{D}, \mathcal{S})$ by first estimating $\tau(\mathcal{X}, \{d\}, \{s\})$ for each $(d, s) \in \mathcal{D} \times \mathcal{S}$, then averaging across those estimates. Choosing a specific duration d and start date s implies a specific observation year t, which in turn ensures that the conditioning set used is coherent across treated and control observations.

With the preceding groundwork in place, the IPTW estimator can be defined. The average treatment effect within a sub-population is estimated as follows:

$$\hat{\tau}(\mathcal{X}, \mathcal{D}, \mathcal{S}) = \frac{1}{|\mathcal{D} \times \mathcal{S}|} \sum_{d \in \mathcal{D}} \sum_{s \in \mathcal{S}} \hat{\tau}(\mathcal{X}, d, s), \tag{7}$$

where

$$\hat{\tau}(\mathcal{X}, d, s) = \sum_{\substack{s=t-d+1, \\ i,t:\mathbf{X_{is-1}} \in \mathcal{X}, \\ D_{it}=d}} w_{it}^{d}(d, s) y_{it} - \sum_{\substack{s=t-d+1, \\ i,t:\mathbf{X_{is-1}} \in \mathcal{X}, \\ D_{it}=0}} w_{it}^{0}(d, s) y_{it}, \tag{8}$$

$$w_{it}^{a}(d, s) = \frac{\frac{1(D_{it}=a)}{\hat{p}(a,s,t)}}{\sum_{\substack{s=t-d+1, \\ i,t:\mathbf{X_{is-1}} \in \mathcal{X}, \\ D_{it}=a}} \frac{1(D_{it}=a)}{\hat{p}(a,s,t)}} \text{ for } a \in \{0, d\}. \tag{9}$$

The estimated treatment effect for a set of potential start years and durations is the average across estimated treatment effects for each pair of a potential duration and potential start date. Each duration- and start year-specific estimated effect is estimated via IPTW, which is a weighted difference in mean outcomes between entities experiencing the specified duration of treatment beginning in the specified year and those experiencing no treatment. Weights are equal to the inverse estimated probability that the entity received its observed sequence of treatment, with weights normalized to one separately for the treated and control groups.

The criterion (2) used to evaluate partitions depends not only on the estimated treatment effect $\hat{\tau}$ in a subsample but also on an estimate of its variance. In Athey and Imbens (2016), under the assumption of randomization the estimated variance reduces to a simple weighted sum of the estimated variances in the treated and control observations in the subsample. That simplification does not hold for (7). However, an estimate of the variance can be based on the linear representation (8). In particular, Imbens (2015) suggests that for linear estimators such as $\hat{\tau}(\mathcal{X}, d, s)$, the variance can be estimated as

$$\widehat{Var}(\hat{\tau}(\mathcal{X}, d, s)) = \sum_{i:\mathbf{X_{is-1}} \in \mathcal{X}} (w_{it}^{D_{it}}(d, s))^2 \hat{\sigma}^2(\mathbf{X_{is-1}}, D_{it}). \tag{10}$$

Here $\hat{\sigma}^2(\mathbf{X_{is-1}}, D_{it})$ is an estimator of the error variance conditional upon pre-treatment

characteristics and whether an observation belongs to the treated or control group. This conditional variance can be estimated many ways, but I adopt the matching approach from Imbens (2015). In that approach, an observation is matched based on $\mathbf{X_{is-1}}$ to the closest observation with the same treatment history, and $\hat{\sigma}^2(\mathbf{X_{is-1}}, D_{it})$ is estimated as one half the squared difference between those observations' outcomes.

With the estimator defined, the splitting procedure defined in Athey and Imbens (2016) can be applied to splits on s or d with one final modification. Because s and d represent *potential* start year and *potential* duration, they do not exist in observed data. Assigning d and s is simple for treated observations: set $d = D_{it}$ and $s = t - d + 1$. In contrast, since control observations have $D_{it} = 0$, it is not obvious which values of d and s should be assigned. One option, which I adopt, is to replicate control observations for any values of s and d for which the control observation could plausibly aid in counterfactual estimation. For example, and untreated observation in 2017 could act as a control for $d = 1$ and $s = 2017$, or as a control for $d = 2$ and $s = 2016$. This approach is similar to matching with replacement, where the same control can contribute to multiple counterfactual estimates. Mechanically, this is accomplished as a data pre-processing step before the tree procedure starts.[8]

Panel unit subsampling

In a panel setting, unobserved determinants of an outcome of interest are likely to be correlated across time. As a result, the forest building procedure proposed here uses a block subsampling approach in which all observations from a given panel unit are either included or excluded from the sample used to construct each tree. This approach differs from the i.i.d. subsampling used in Wager and Athey (2017) but is suitable for resampling panel data (Kapetanios 2008).

[8]To save on computation, each untreated observation could be replicated for a random subset of the (d, s) pairs for which it could serve as a potential control.

Inference

Some guidance on inference for causal trees and forests is provided in Athey and Imbens (2016) and Wager and Athey (2017). If a forest rather than a single tree is used, the inference approach suggested in Athey and Imbens (2016), which applies to individual leaves in a single tree, is not directly applicable. The infinitesimal jacknknife approach proposed in Wager and Athey (2017) constructs a variance estimate from the set of point estimates provided by the trees in the forest. I use that method with two modifications. First, I adopt the finite sample bias correction from Wager et al. (2014), which accounts for Monte Carlo variability in the random forest procedure. Second, because the bootstrapping procedure used here resamples at the panel unit level, the formula for the variance uses the (smaller) number of panel units in a sample rather than the number of observations. Doing so inflates estimated variances and confidence intervals to account for potential intertemporal correlation in observations from a panel unit.[9]

Given the focus here on heterogeneity in policy effects, it is worth noting that the same set of estimates provided by the trees in the forest can be used to conduct inference on differences in treatment effects between any two groups $(\mathcal{X}, \mathcal{D}, \mathcal{S})$ and $(\mathcal{X}', \mathcal{D}', \mathcal{S}')$. Each tree provides an estimate of the difference in treatment effects evaluated at two covariate vectors, and so an infinitesimal jackknife approach can be applied to those estimated differences. Accounting for covariance in estimates across subpopulations in this way is likely to be especially important since each tree in a forest generates a constant estimate for all elements in a given partition.

Handling unobservables

A clear threat to identification under SCIA or its cross-sectional analog is the presence of confounding unobservable factors, such as unit-specific effects or time effects. Time-varying

[9] An alternative, but computationally intensive method is to bootstrap the entire forest-building procedure and resample entire forests. See, e.g., Asher et al. (2016) and Knaus et al. (2017) for related approaches.

but cross-sectionally uniform unobservables are of less concern given that (7) first constructs year-specific estimates and then averages over those estimates in $\mathcal{D} \times \mathcal{S}$. A complementary approach makes use of double differencing as a data pre-processing step. Let ϕ_t and μ_i represent additive year and unit effects, respectively. Given their additive nature, a natural approach is to eliminate these effects via double differencing. Any differencing must account for the staggered nature of treatment: unit-level means depend on the unit-specific frequency of treatment, while year-level means depend on the year-specific incidence of treatment, and there is no global pre-treatment period. To this end, I make use of never-treated units in the following way. First, I average outcomes within year t across never-treated units, and subtract those means to remove the effect of ϕ_t. Then, the pre-treatment, within-unit mean of that intermediate outcome (with year effects removed) can be subtracted out to remove unit-level effects without making the demeaning process dependent on the frequency of treatment for a unit. The resulting transformed outcome can be used as the outcome of interest in the algorithm described above.[10] An alternative to this double differencing approach would be to estimate a regression with year and unit dummies in each subsample. However, data pre-processing via double differencing as just described maintains the broad structure of the algorithm developed in Athey and Imbens (2016) and retains the computational advantages of a simple comparison of weighted means in each candidate partition.

4. Application: Individual Transferable Quota Markets in Fisheries

Beginning in the mid 1970s, a number of commercial fisheries around the world have shifted to management based on individual quotas (IQs)[11] that allocate the rights to catch fish to specific entities (individuals or groups). Such a management approach has potential

[10] Specifically, assume $Y_{it}(d,s) = \mu_i + \phi_t + \tau(d,s)$. Estimate $\hat{\phi}_t = \frac{1}{|i:T_{it}=0|}\sum_{i:T_{it}=0} Y_{it}$ and define $\tilde{Y}_{it} = Y_{it} - \hat{\phi}_t$. Estimate $\hat{\mu}_i = \frac{1}{|t:T_{it}=0|}\sum_{t:T_{it}=0} \tilde{Y}_{it}$ and define $\ddot{Y}_{it} = \tilde{Y}_{it} - \hat{\mu}_i$. Use $\ddot{Y}_{it}$ as the outcome of interest.

[11] The analysis here does not distinguish between tradable and non-tradable forms of harvest rights, focusing broadly on any policy that assigns an individual right to a flow benefit.

to improve both economic and biological outcomes by addressing dynamic externalities, reducing incentives to overinvest in extraction capital, and increasing the value of landings by lengthening harvest seasons (Grafton et al. 2000). However, both biological and economic effects of IQ adoption are likely to vary across fisheries and through time: optimal management and the rents that flow from it will depend upon initial resource conditions and the biology governing system dynamics, among other factors.

Using the methods described above, I examine how adopting an individual quota policy impacts the probability of resource collapse, with the primary goal of understanding what new insights these methods might provide as compared to more conventional approaches. This empirical context offers a useful test case for two reasons. First, a high-profile benchmark study (Costello et al. 2008) used a mix of propensity score and fixed effects methods to examine whether IQs reduce the probability of fishery collapse, with collapse defined as catch falling below 10% of its historical maximum (Worm et al. 2006). Results indicated that IQ adoption reduces probability of collapse on average, but the specifications allowed for only limited heterogeneity in policy effects: a linear trend with policy duration. Because the tree-based methods used here also rely on propensity score methods, a comparison of results should offer illustration of what new insights tree-based methods might provide.[12] Second, IQ effects are likely heterogeneous: prior work has identified dependency on, e.g., social factors (Gutiérrez et al. 2011), scientific robustness of quotas set (Mora et al. 2009), biological characteristics of targeted species (Pinkerton and Edwards 2009), and country development (Ban et al. 2009).[13] More recent work also suggests that quota stringency itself (and hence outcomes) depends upon the strength of property rights in a fishery (Costello and Grainger 2018). Figure 1 presents suggestive evidence of heterogeneous effects, depict-

[12]Regression trees and random forests have been applied before to examine factors associated with improved fisheries outcomes, though in a predictive fashion only. Gutiérrez et al. (2011) use random forests to predict outcomes for co-managed fisheries. Melnychuk et al. (2016) use random forests to examine factors associated with desired outcomes for fisheries under individual quota management. In both cases, the authors attach a causal interpretation to estimates, but do not directly tackle the issue of causality.

[13]For a broader review of potential limitations of IQs see Copes et al. (1986).

ing trends in collapse among fisheries not under IQ management, and two groups under IQ management: fisheries with low catch prior to policy implementation and those with high catch. Differences between the IQ groups and the non-IQ group vary through time, but there are also clear differences in both levels and trends across the two IQ groups.

In this setting, i indexes a fishery, t a year, and D_{it} the number of years that fishery i was subjected to IQ management as of year t. The outcome of interest, Y_{it}, is a binary variable taking value one if and only if the catch c_{it} for fishery i in year t is less than 10% of the maximum catch observed in fishery i in all years $t' < t$, i.e., $Y_{it} \equiv \mathbb{1}(c_{it} \leq 0.1 \cdot c_{it}^{max})$ with $c_{it}^{max} \equiv \max_{t':t'<t} c_{it'}$.[14] The covariates X_{is} considered are one year lagged maximum historical catch (c_{s-1}^{max}), one year lagged absolute catch (c_{is-1}) and relative catch (c_{is-1}/c_{is-1}^{max}), one year lagged absolute catch trend ($c_{is-1} - c_{is-2}$) and relative catch trend ($(c_{is-1} - c_{is-2})/c_{s-1}^{max}$), and the Von-Bertalanffy growth rate K_i of the species.

The probability of transitioning to IQ management is estimated as a pooled logit model using the set of observations that are not already under IQ management in a given year. Explanatory variables include the lagged catch (both absolute and relative), lagged catch trend (absolute), the maximum historical catch, year effects, and a linear time trend in the effect of the lagged relative catch. The logic behind such a model is that catch history is likely to influence adoption of a new management policy.

I apply this method to a global dataset of fisheries catch, IQ management status, and species characteristics ranging from 1950-2014. For consistency with Costello et al. (2008), a fishery is defined by a species and Large Marine Ecosystem (LME) pair. Catch records (tons per fishery per year) come from the Sea Around Us project, IQ policy adoption indicators

[14]There are clear drawbacks to the use of catch-based measures of collapse, including mathematical reasons (Wilberg and Miller 2007) and differences from management-based measures of collapse (de Mutsert et al. 2008; Branch et al. 2011). Still, 1) catch data are available for far more fisheries than biomass estimates, 2) the purpose of the application is to illustrate what new methods offer, motivating use of the same outcome measure as prior work, and 3) some of the mathematical concerns with the chosen collapse metric (Wilberg and Miller 2007) are mitigated by the estimator used here. Those concerns state that even without a trend in true rates of collapse, the random nature of catch statistics make it more likely for a fishery to experience an extremely low catch year as time passes. However, since the estimator outlined above first compares IQ and non-IQ fisheries in a given year (see (8)), trends in collapse metrics are of less concern.

are based on a database complied by the Environmental Defense Fund, and Von Bertalanffy growth rate (year^{-1}), where available, is collected from FishBase. Where a species is subjected to multiple forms of management within a single LME, the earliest date of IQ implementation for a species in an LME is used to determine IQ status for the fishery. Records are matched across data sources using species, location, and year. The dataset is constructed with the intent of being an updated and expanded version of the original data used in Costello et al. (2008). See the Appendix for more details on dataset assembly.

Prior to estimation, the dataset is cleaned and filtered in two ways to make it comparable with that used in Costello et al. (2008). First, attention is restricted to fisheries occurring in LMEs which contain at least one fishery under IQ management during the study period, which should improve similarity between IQ fisheries and control fisheries. Second, the data are filtered to complete records containing all outcomes and explanatory variables.

The resulting dataset, summarized in Table 1, has 141,538 observations from 4,317 fisheries, with 320 fisheries managed via an IQ system at some point during the study window. For models including species growth rate, data availability limits the sample to 120,332 observations from 3,586 fisheries, 305 of which enter into IQ management during the study window. Many of the observations without available growth rates correspond to shellfish fisheries for which Von Bertalanffy growth is not applicable.

During the course of estimation, after propensity scores are estimated, records are eliminated in which estimated propensity scores are numerically equivalent to zero or one in order to satisfy overlap requirements. In addition, trimming is used to eliminate records with probability of observed treatment sequence below 0.001.

4.1. Assessing propensity score methods

The internal validity of the causal forest method used here rests on the ability of the weighting scheme to render treatment as good as random. To this end, Figure 2 presents standardized differences in means of covariates before and after reweighting; all standardized

differences are smaller in absolute value than 0.1 (Austin and Stuart 2015; Stuart et al. 2014). Similarly, Figures 3 and 4 present balance plots for the covariates used in the propensity score model before and after reweighting by the inverse probability of selection into an IQ program. The 25th, 50th, and 75th quantiles are all similar between treated and control observations after reweighting. Note that the weights used here differ from the weights ultimately used in the forest procedure, since the latter correspond to the probability that an observation experienced its entire treatment history. Because treatment histories are of different lengths in panel data, assessing balance is far simpler using weights based on a single probability of selection into treatment. Moreover, concern over endogeneity in this empirical context is limited to initial program adoption since IQ programs are never removed in the dataset, so weighting by the inverse probability of selection into treatment is suitable for assessing balance.

In addition to these standard balance assessments, section 4.2.1 presents parametric falsification tests examining whether IQ adoption impacts outcomes which should not be affected under the assumptions outlined earlier.

4.2. Results

4.2.1. Standard methods and heterogeneity

As a baseline for the tree-based approach, I first provide results using conventional parametric methods. Doing so has three main benefits. First, estimating a model analogous to a primary specification used in Costello et al. (2008) provides suggestive evidence that any differences in results obtained from the new methods presented here are not likely due to differences between the original and updated data. Second, I conduct falsification tests using those established methods to further probe the identification offered by a propensity score approach. To reiterate, my goal is not to improve on the identification from Costello et al. (2008) but instead to examine what forest-based methods offer under similar identifying assumptions; still, the credibility of the comparison and identification are linked. Finally,

examining results from alternate conventional specifications that encode different types of treatment effect heterogeneity highlights the challenge of correctly specifying the form of heterogeneity.

One of the primary results from Costello et al. (2008) suggests that while the probability of fishery collapse increases through time, each additional year under IQ management can provide a counteracting and potentially completely offsetting effect on the probability of collapse. While those authors investigate many models, a basic specification they explore is a model with a constant and three regressors: whether a fishery is ever an IQ, a year effect, and how many years the fishery has been under IQ management. The last regressor is simply an interaction between a dummy variable indicating that a fishery is under IQ management and a continuous year variable, thereby capturing treatment effect heterogeneity.

I estimate the same model using an analog to their dataset produced by subsetting the updated data described earlier. Specifically, I filter to include only data up to 2003 and, for the purposes of this comparison, I consider only fisheries established as IQs in 2003 or before to be IQs. I estimate linear probability models, but logit specifications provide similar results. The marginal effects from the model described above on this dataset are presented in column 1 (unweighted) and column 2 (inverse probability of treatment weighted) of Table 2. The key findings are as follows: while the probability of collapse increases around half a percent a year, an additional year of IQ management reduces the probability of collapse by around the same amount. These results are consistent with the main results presented Costello et al. (2008), suggesting the datasets are sufficiently similar up to 2003 that substantive differences in results from the tree-based methods are likely due to the methods themselves.

Estimating variants of this basic model with alternate outcomes provides no evidence of identification concerns. Specifically, I estimate model variants with 1) the outcome variable in the year prior to IQ implementation, 2) outcome variables randomized across fisheries within a year, and 3) outcome variables randomized across fisheries within a region (LME) in a year. None of these models should display significant effects of treatment duration on

the outcome of interest; the first is an assessment of pre-treatment differences among units, while the second and third evaluate potential spillovers across fisheries. As indicated in Table 3, there is no significant evidence of policy duration affecting any of these variables.

One concern with the conventional parametric approach is that the way in which the policy effect varies may be misspecified; policy effects may vary in ways not accounted for in the model. To provide just one example, the effectiveness of IQ management may depend upon the (relative) catch in a fishery in the year just prior to IQ implementation. If a fishery is already collapsed or nearly so, management under IQs may optimally impose very low harvest levels in order to rebuild the stock, while an already healthy fishery may enjoy continued high landings (see, e.g., Reed 1979). The former case may show up as collapse in the dataset, while the latter will not. To explore this possibility, I estimate an alternate specification which interacts the linear trend in IQ impact with a dummy variable indicating whether the fishery was considered collapse in the year prior to the policy change.

The results of this exercise are presented in the final column of Table 2. While the main effect of policy duration has a similar sign and magnitude, the net policy duration effect has opposite sign for fisheries that were collapsed prior to policy adoption. These findings highlight the both the challenges inherent in and consequences of choosing a specification for treatment effect heterogeneity.

4.3. Causal forest methods

To investigate the potential for the modified honest forest procedure outlined above to aid in identifying heterogeneous effects of IQs, I apply the method to two samples. The first sample omits species growth rate in order to allow for the use of more data. For that sample, heterogeneity in four dimensions is investigated: pre-treatment relative catch, pre-treatment relative catch trend, policy duration, and policy start year. The second sample and set of results allow for heterogeneity in effects by species growth rate, using only observations for which that information is also available.

Estimation excluding biological parameters produces compelling results. First, beginning by allowing splitting only on policy duration, results are broadly consistent with those from Costello et al. (2008): over the course of 30 years of policy implementation, the probability of collapse declines at roughly a half percent per year of implementation (Figure 5). However, even without allowing for heterogeneity in other dimensions, the method already reveals substantial nonlinearities. Much of the decline in the probability of collapse with continued IQ implementation appears to accrue between years 5-12 and 25-27, with a relatively flat profile and even decay of effects in other periods. This suggests that expectations of immediate benefits may need to be tempered, and that reductions in the probability of collapse may not be steady. Note that while the confidence intervals for the many estimated policy effects overlap, due to correlations among the estimates, many differences between policy effects are, in fact, significant. As an example, Figure 6 illustrates differences between the estimated policy effect after one year and effects for all other durations along with confidence intervals for those differences; approximately 40 percent of the differences are significant at the 5% level.

Allowing policy effects to depend on more than just duration reveals a richer picture. Figures 7, 8, and 9 show similar plots of estimated policy effects against duration, but depict those profiles for groups of fisheries that vary in some other dimension. For example, Figure 7 depicts partial effects of policy duration at three different levels of pre-treatment relative catch (1%, 50%, and 99%). When catch is already low prior to IQ implementation, the policy change has a *positive* but insignificant estimated effect on the probability of collapse regardless of policy duration. Even if significant, positive estimates do not necessarily imply perverse effects of IQs: responsible management of an already depleted stock may entail reducing catch to very low or zero levels, which is collapse by the definition stated earlier. The strongest benefits of the policy accrue when pre-treatment catch is at moderate levels, with steadily improving reductions in the probability of collapse through time. Finally, when pre-treatment catch is already high, the policy yields insignificant reductions in the

probability of collapse. In short, duration matters for policy effects, but in a way that depends upon the history of the resource. This interaction highlights the benefit of using tree-based methods to examine temporal and cross-sectional heterogeneity simultaneously.

The causal forest estimates also suggest that IQ policies have had a less beneficial impact in more recent implementations. Figure 8 indicates that the estimated reduction in the probability of collapse shifts upward, flattens in time profile, and loses significance for implementations that began in more recent years. While this apparent lack of effectiveness in more recent years could be concerning, not all IQ programs are introduced for (exclusively) biological reasons. More recent adoptions may target by economic rather than biological objectives, with better biological outcomes achieved even in the absence of IQs. Unfortunately, assessing impacts on those economic outcomes is precluded by data availability at the scale of this analysis.

Allowing policy effects to depend on resource growth rates yields limited but sensible patterns of heterogeneity. Figure 9 shows the same time profile of policy effects, grouped now according to the 25th, 50th, and 75th quantiles of growth rates in the dataset. The fastest growing species display reductions in the probability of collapse sooner, though that comparative advantage disappears after roughly 15 years.

On a final note, a caveat on interpretation of results is in order. In some fisheries the introduction of individual quotas coincides with the first introduction of any kind of binding harvest quota. For those cases any improved outcomes reflect the composite effect of both changes, which could differ from the effect of introducing individual-level rights alone. Nevertheless, the preceding analysis highlights important heterogeneity in how the probability of collapse under IQ management differs from other forms of management.

5. Conclusion

Public policies are likely to impact people, firms, or jurisdictions in different ways, and those effects may vary depending on when a policy first took effect and for how long it has

been in place. Learning more about heterogeneity, dynamics, and policy improvement has many potential benefits. At a minimum, we can set expectations as to when policy effects might start to arise for a new policy and who stands to benefit the most from it. Knowledge of heterogeneity in policy effects might also help better target a particular type of policy to settings in which it is likely to have the most beneficial effect.

This paper has shown how to adapt a recently proposed suite of methods – causal trees (Athey and Imbens 2016) and causal forests (Wager and Athey 2017) – to examine cross-sectional heterogeneity, dynamics, and changes in policy effectiveness with the time of policy introduction. The estimation approach uses a modified regression tree procedure to identify subpopulations in which the treatment effect can is best approximated by a constant, and makes use of inverse probability of treatment weighting to construct estimates in each subsample. The key benefit of this method is that it requires weak assumptions about how policy effects vary. The researcher need only specify which variables might influence policy effects, but not the functional form of that influence. This allows data to potentially reveal a richer pattern of heterogeneity and time dependence than what may be hypothesized.

Applying these methods to examine the effect of a market-based natural resource management policy reveals heterogeneity across fisheries and through time. Individual Quota programs substantively reduce the probability of fishery collapse for some fisheries, while having no or even detrimental estimated effects for others. Moreover, the strongest effects take several years to materialize and do not always follow a linear pattern as assumed in prior work. These findings suggest that application of these methods to other policy questions could reveal heterogeneity that has been previously ignored or assumed away.

As discussed in the introduction, tree-based methods are not the only tools that permit estimation of heterogeneous policy effects under weaker functional form assumptions. Examination of cross-sectional heterogeneity could be done via, e.g., a variant of support vector regression (Imai and Ratkovic 2013) least absolute selection and shrinkage operator (LASSO) estimation (Tian et al. 2014), Bayesian nonparametric approaches (Taddy et al. 2016), or

via kernel regression with inverse propensity score weights determined by a separate, initial kernel regression (Abrevaya et al. 2015). A comparative analysis of these methods and their relative performance across a range of empirical contexts seems a promising area for future research.

6. Appendix

Dataset construction

The list of fisheries subjected to IQ management was based on the database of catch share programs provided by the Environmental Defense Fund. Since that database provides a single start year for an entire management program, which may involve multiple species, start dates for each species in each LME were subsequently compiled from published government sources or academic research papers. In some cases, a single species may be subjected to several forms of management in a single LME. Some LMEs span national jurisdictions, and even within a jurisdiction and LME, different groups of vessels (e.g. those using different fishing gear) may be subjected to different management. The start date used for IQs in such cases was the earliest year in which the species was subjected to IQ management in that LME. A more refined definition of a fishery could be used to avoid this issue, but defining a fishery as an LME x species pair allows for a clean comparison with Costello et al. (2008).

As explained in the main text, species name and LME were used to match management information with catch records. Catch data from the Sea Around Us project was downloaded per LME, and records were restricted to reported landings data (excluding estimates of unreported catch and fish which were discarded prior to returning to port). Records from the management database without an exact (species x LME) match in the catch records were manually reviewed. Where possible, exact matches were supplemented with manual matches, correcting differences in spelling or the use of non-standard species synonyms in one source or the other. Catch records were matched to trait data from FishBase based on species names alone. If no exact match was found, several additional steps were used to increase coverage. First, the species name from the catch data was matched against genus

names in FishBase. In some cases, catch data was actually reported at the genus level, so that the 'species' name actually represented a genus. Genus traits were computed as means and modes of species-level traits for species in that genus. Supplementary steps included attempting matches based on synonyms for species names used in the catch data, as well as manually mapping taxonomic family names used in catch data to a representative genus for matching to the genus-level traits just described.

References

Abrevaya, J., Hsu, Y.-C., Lieli, R. P., 2015. Estimating conditional average treatment effects. Journal of Business & Economic Statistics 33 (4), 485–505.

Allcott, H., 2011. Social norms and energy conservation. Journal of Public Economics 95 (9), 1082–1095.

Allcott, H., Rogers, T., 2014. The short-run and long-run effects of behavioral interventions: Experimental evidence from energy conservation. The American Economic Review 104 (10), 3003–3037.

Asher, S., Nekipelov, D., Novosad, P., Ryan, S. P., December 2016. Classification trees for heterogeneous moment-based models. Working Paper 22976, National Bureau of Economic Research.

Athey, S., Imbens, G., 2016. Recursive partitioning for heterogeneous causal effects. Proceedings of the National Academy of Sciences 113 (27), 7353–7360.

Athey, S., Tibshirani, J., Wager, S., 2016. Generalized random forests. arXiv preprint arXiv:1610.01271.

Auffhammer, M., Bento, A. M., Lowe, S. E., 2009. Measuring the effects of the clean air act amendments on ambient pm10 concentrations: The critical importance of a spatially

disaggregated analysis. Journal of Environmental Economics and Management 58 (1), 15–26.

Auffhammer, M., Kellogg, R., 2011. Clearing the air? the effects of gasoline content regulation on air quality. The American Economic Review, 2687–2722.

Austin, P. C., Stuart, E. A., 2015. Moving towards best practice when using inverse probability of treatment weighting (iptw) using the propensity score to estimate causal treatment effects in observational studies. Statistics in Medicine 34 (28), 3661–3679.

Azoulay, P., Ding, W., Stuart, T., 2009. The impact of academic patenting on the rate, quality and direction of (public) research output. The Journal of Industrial Economics 57 (4), 637–676.

Ban, N. C., Caldwell, I. R., Green, T. L., Morgan, S. K., O'Donnell, K., Selgrath, J. C., 2009. Diverse fisheries require diverse solutions. Science 323 (5912), 338–339.

Bandiera, O., Larcinese, V., Rasul, I., 2010. Heterogeneous class size effects: New evidence from a panel of university students. The Economic Journal 120 (549), 1365–1398.

Becker, S. O., Egger, P. H., Von Ehrlich, M., 2013. Absorptive capacity and the growth and investment effects of regional transfers: A regression discontinuity design with heterogeneous treatment effects. American Economic Journal: Economic Policy 5 (4), 29–77.

Bertrand, M., Crépon, B., Marguerie, A., Premand, P., 2017. Contemporaneous and post-program impacts of a public works program: Evidence from côte d'ivoire. Working paper.

Branch, T. A., Jensen, O. P., Ricard, D., Ye, Y., Hilborn, R., 2011. Contrasting global trends in marine fishery status obtained from catches and from stock assessments. Conservation Biology 25 (4), 777–786.

Breiman, L., 2001. Random forests. Machine learning 45 (1), 5–32.

Breiman, L., Friedman, J., Stone, C. J., Olshen, R. A., 1984. Classification and regression trees. CRC press.

Chou, P. A., 1991. Optimal partitioning for classification and regression trees. IEEE Transactions on Pattern Analysis & Machine Intelligence (4), 340–354.

Clark, C. W., Munro, G. R., 1975. The economics of fishing and modern capital theory: a simplified approach. Journal of Environmental Economics and Management 2 (2), 92–106.

Copes, P., et al., 1986. A critical review of the individual quota as a device in fisheries management. Land Economics 62 (3), 278–291.

Costello, C., Gaines, S. D., Lynham, J., 2008. Can catch shares prevent fisheries collapse? Science 321 (5896), 1678–1681.

Costello, C., Grainger, C. A., 2018. Property rights, regulatory capture, and exploitation of natural resources. Journal of the Association of Environmental and Resource Economists 5 (2), 441–479.

Crump, R. K., Hotz, V. J., Imbens, G. W., Mitnik, O. A., 2008. Nonparametric tests for treatment effect heterogeneity. The Review of Economics and Statistics 90 (3), 389–405.

Davis, J. M., Heller, S. B., 2017. Using causal forests to predict treatment heterogeneity: An application to summer jobs. American Economic Review: Papers and Proceedings 107 (5), 546–550.

de Mutsert, K., Cowan, J. H., Essington, T. E., Hilborn, R., 2008. Reanalyses of gulf of mexico fisheries data: landings can be misleading in assessments of fisheries and fisheries ecosystems. Proceedings of the National Academy of Sciences 105 (7), 2740–2744.

Fisher, W. D., 1958. On grouping for maximum homogeneity. Journal of the American Statistical Association 53 (284), 789–798.

Fowlie, M., 2010. Emissions trading, electricity restructing, and investment in pollution abatement. The American Economic Review, 837–869.

Gerfin, M., Lechner, M., 2002. A microeconometric evaluation of the active labour market policy in switzerland. The Economic Journal 112 (482), 854–893.

Grafton, R. Q., Squires, D., Fox, K. J., 2000. Private property and economic efficiency: a study of a common-pool resource. The Journal of Law and Economics 43 (2), 679–714.

Gutiérrez, N. L., Hilborn, R., Defeo, O., 2011. Leadership, social capital and incentives promote successful fisheries. Nature 470 (7334), 386–389.

Hall, P., Li, Q., Racine, J. S., 2007. Nonparametric estimation of regression functions in the presence of irrelevant regressors. The Review of Economics and Statistics 89 (4), 784–789.

Hamilton, J. T., 1995. Pollution as news: media and stock market reactions to the toxics release inventory data. Journal of Environmental Economics and Management 28 (1), 98–113.

Handel, B., Kolstad, J., 2017. Wearable technologies and health behaviors: New data and new methods to understand population health. American Economic Review 107 (5), 481–85.

Imai, K., Ratkovic, M., 2013. Estimating treatment effect heterogeneity in randomized program evaluation. The Annals of Applied Statistics 7 (1), 443–470.

Imbens, G. W., 2015. Matching methods in practice: Three examples. Journal of Human Resources 50 (2), 373–419.

Kapetanios, G., 2008. A bootstrap procedure for panel data sets with many cross-sectional units. The Econometrics Journal 11 (2), 377–395.

Knaus, M., Lechner, M., Strittmatter, A., 2017. Heterogeneous employment effects of job search programmes: A machine learning approach. Working Paper 23326, IZA Institute of Labor Economics.

Lalive, R., Van Ours, J. C., Zweimüller, J., 2008. The impact of active labour market programmes on the duration of unemployment in switzerland. The Economic Journal 118 (525), 235–257.

Lechner, M., 2002. Program heterogeneity and propensity score matching: An application to the evaluation of active labor market policies. Review of Economics and Statistics 84 (2), 205–220.

Lechner, M., Miquel, R., 2010. Identification of the effects of dynamic treatments by sequential conditional independence assumptions. Empirical Economics 39 (1), 111–137.

Li, Q., Racine, J., 2004. Cross-validated local linear nonparametric regression. Statistica Sinica, 485–512.

Melnychuk, M. C., Essington, T. E., Branch, T. A., Heppell, S. S., Jensen, O. P., Link, J. S., Martell, S. J., Parma, A. M., Smith, A. D., 2016. Which design elements of individual quota fisheries help to achieve management objectives? Fish and Fisheries 17 (1), 126–142.

Mora, C., Myers, R. A., Coll, M., Libralato, S., Pitcher, T. J., Sumaila, R. U., Zeller, D., Watson, R., Gaston, K. J., Worm, B., 06 2009. Management effectiveness of the world's marine fisheries. PLOS Biology 7 (6), 1–11.

Pinkerton, E., Edwards, D. N., 2009. The elephant in the room: the hidden costs of leasing individual transferable fishing quotas. Marine Policy 33 (4), 707–713.

Prest, B. C., 2017. Peaking interest: How awareness drives the effectiveness of time-of-use electricity pricing.

Racine, J., Li, Q., 2004. Nonparametric estimation of regression functions with both categorical and continuous data. Journal of Econometrics 119 (1), 99–130.

Reed, W. J., 1979. Optimal escapement levels in stochastic and deterministic harvesting models. Journal of Environmental Economics and Management 6 (4), 350–363.

Robins, J. M., Hernán, M. Á., Brumback, B., 2000. Marginal structural models and causal inference in epidemiology. Epidemiology, 550–560.

Rubin, D. B., 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology 66 (5), 688.

Shimshack, J. P., Ward, M. B., Beatty, T. K., 2007. Mercury advisories: information, education, and fish consumption. Journal of Environmental Economics and Management 53 (2), 158–179.

Stuart, E. A., Huskamp, H. A., Duckworth, K., Simmons, J., Song, Z., Chernew, M. E., Barry, C. L., 2014. Using propensity scores in difference-in-differences models to estimate the effects of a policy change. Health Services and Outcomes Research Methodology 14 (4), 166–182.

Taddy, M., Gardner, M., Chen, L., Draper, D., 2016. A nonparametric bayesian analysis of heterogenous treatment effects in digital experimentation. Journal of Business & Economic Statistics 34 (4), 661–672.

Tian, L., Alizadeh, A. A., Gentles, A. J., Tibshirani, R., 2014. A simple method for estimating interactions between a treatment and a large number of covariates. Journal of the American Statistical Association 109 (508), 1517–1532, pMID: 25729117.

Wager, S., Athey, S., 2017. Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association.

Wager, S., Hastie, T., Efron, B., 2014. Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. The Journal of Machine Learning Research 15 (1), 1625–1651.

Wilberg, M. J., Miller, T. J., 2007. Comment on "impacts of biodiversity loss on ocean ecosystem services". Science 316 (5829), 1285–1285.

Worm, B., Barbier, E. B., Beaumont, N., Duffy, J. E., Folke, C., Halpern, B. S., Jackson, J. B., Lotze, H. K., Micheli, F., Palumbi, S. R., et al., 2006. Impacts of biodiversity loss on ocean ecosystem services. Science 314 (5800), 787–790.

Tables

	All fisheries		IQ fisheries	
	mean	sd	mean	sd
Collapsed	0.23	0.42	0.13	0.33
Collapsed Under IQ			0.17	0.38
Relative Catch	0.49	0.38	0.58	0.35
Relative Catch Under IQ			0.45	0.32
Pre-IQ Relative Catch			0.53	0.35
Pre-IQ Relative Catch Trend			-0.02	0.17
Von-Bertalanffy Growth (K)	0.27	0.29	0.19	0.16
Policy Start Year			1996	9
Policy Duration			12	8
Fisheries	4,317		320	
Observations	141,538		15,295	

Table 1: Summary statistics

	(1)	(2)	(3)
Ever IQ	-0.176^{***}	-0.157^{***}	-0.172^{***}
	(0.020)	(0.023)	(0.020)
Year	0.006^{***}	0.007^{***}	0.006^{***}
	(0.000)	(0.001)	(0.001)
Is IQ x Years in IQ	-0.005^{*}	-0.005^{*}	-0.005^{*}
	(0.002)	(0.002)	(0.003)
Is IQ x Years in IQ x Pre-Policy Catch$\leq$0.1			0.083^{***}
			(0.023)

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 2: Linear probability model estimation of collapse using (1) model in Costello, Gaines, and Lynham (2008) unweighted, (2) the same model weighted by inverse probability of observed treatment sequence, and (3) adding an interaction between treatment duration and an indicator of whether the fishery was collapsed in the year prior to IQ implementation.

	(1)	(2)	(3)
Ever IQ	0.029	0.002	0.017
	[−0.043; 0.102]	[−0.043; 0.050]	[−0.033; 0.068]
Year	−0.003	0.001	0.000
	[−0.011; 0.005]	[−0.002; 0.004]	[−0.002; 0.002]
Is IQ x Years in IQ	0.002	−0.000	0.000
	[−0.008; 0.012]	[−0.005; 0.004]	[−0.004; 0.005]

* 0 outside the confidence interval

Table 3: Falsification tests: (1) outcome equal to relative catch in year prior to IQ implementation (2) outcomes randomized across fisheries within a year (3) outcomes randomized across fisheries in the same LME and same year.

Figures

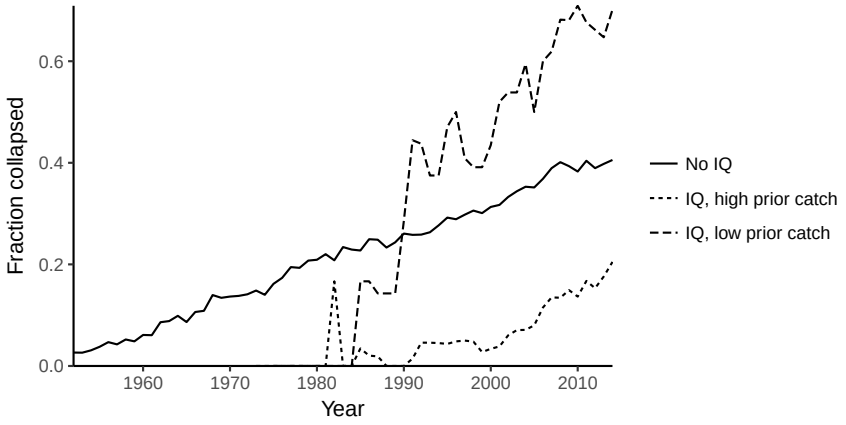

Figure 1: Fraction of fisheries collapsed in a given year in three groups: those not under IQ management (solid), those under IQ management with low catch ($\leq 20\%$ of historical max) just prior to IQ implementation (long dash), and those under IQ management with high catch ($> 20\%$ of historical max) just prior to IQ implementation (short dash).

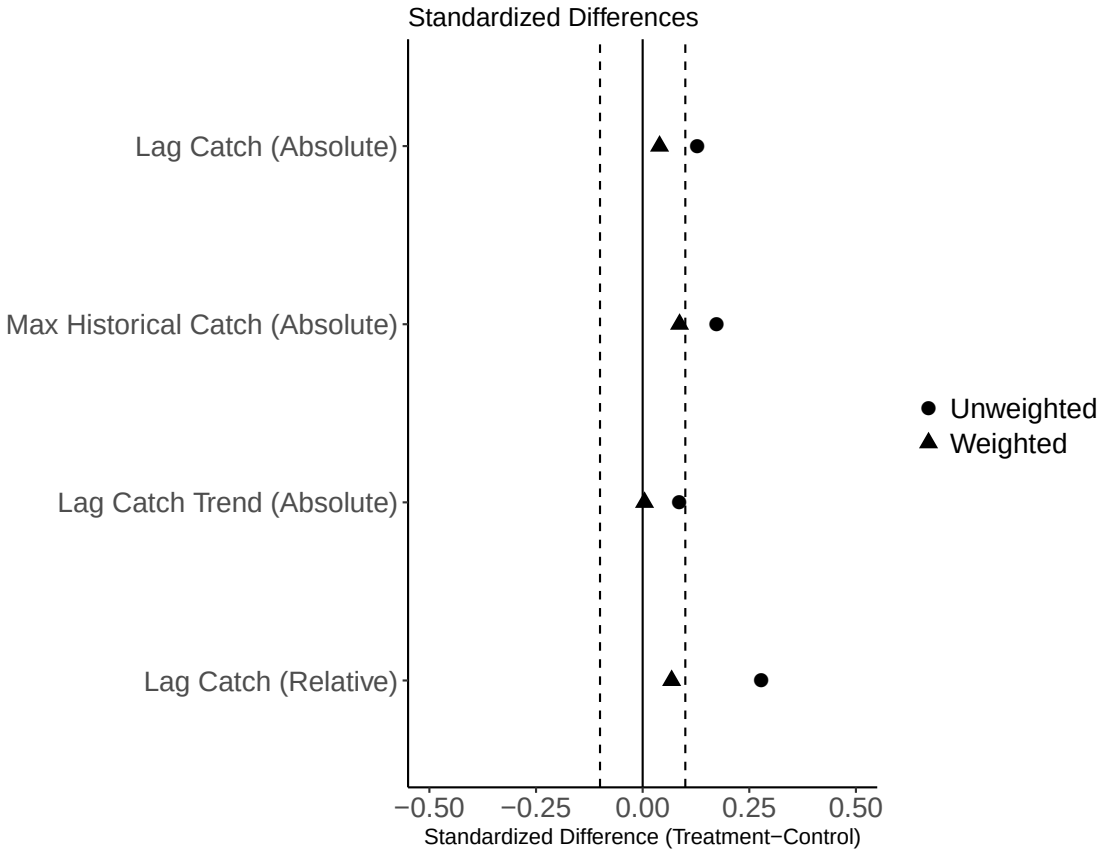

Figure 2: Standardized differences of covariates across non-IQ and IQ observations before and after reweighting by stabilized inverse probability of transition from non-IQ to IQ management..

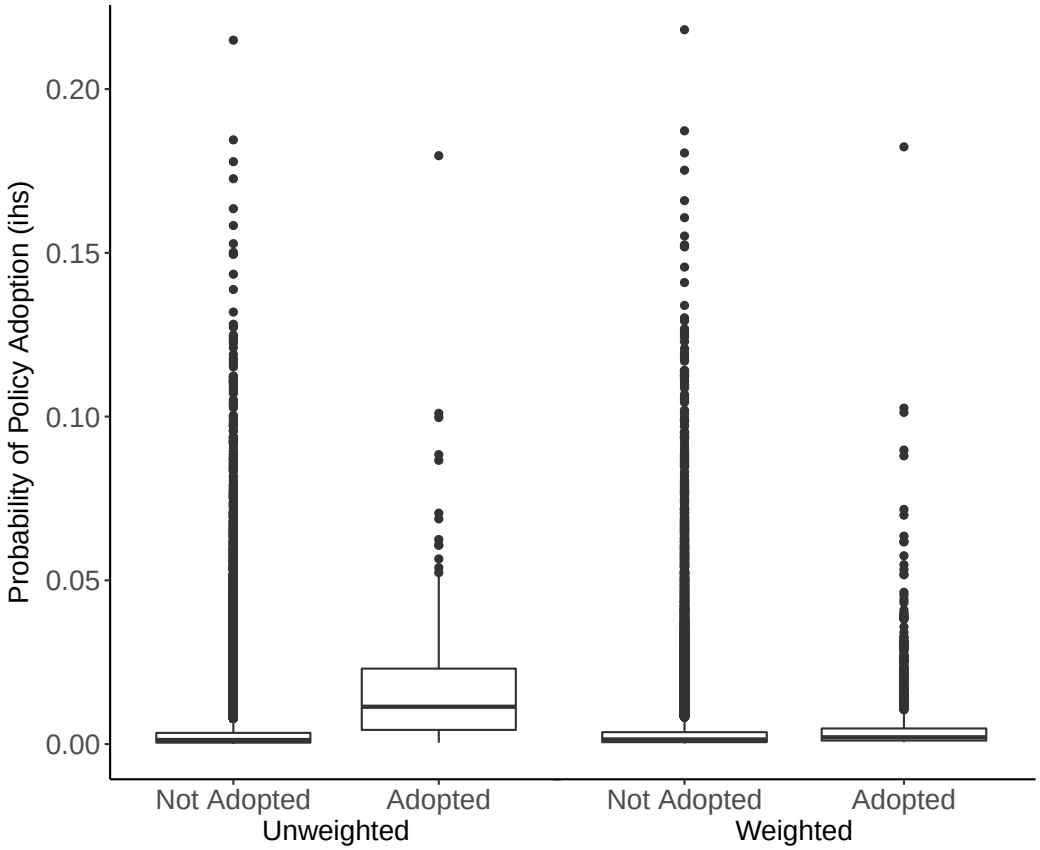

Figure 3: Balance of propensity of policy adoption before and after reweighting by stabilized inverse probability of transition from non-IQ to IQ management..

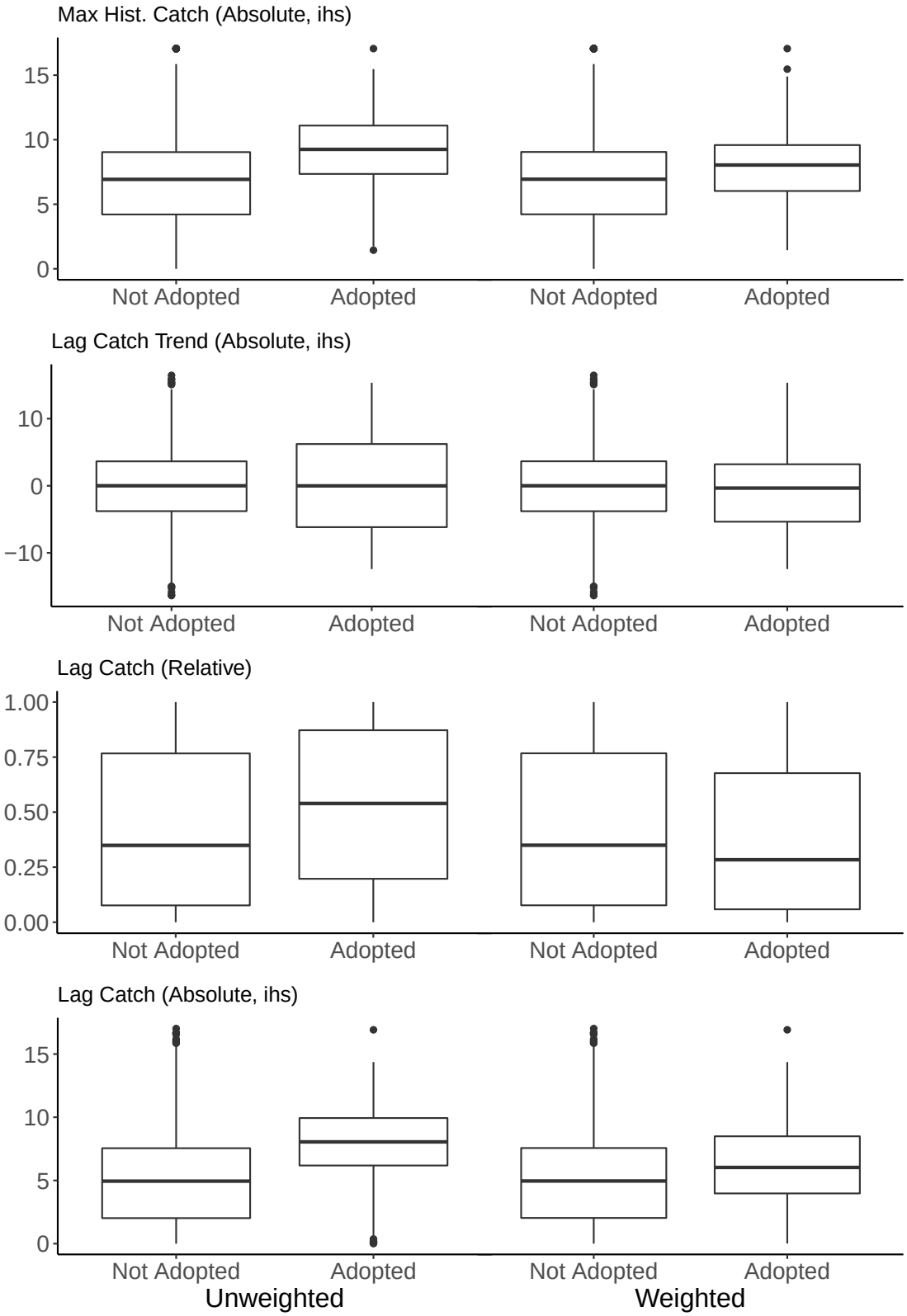

Figure 4: Covariate balance across treatment status before (left column) and after (right column) reweighting by stabilized inverse probability of transition from non-IQ to IQ management.

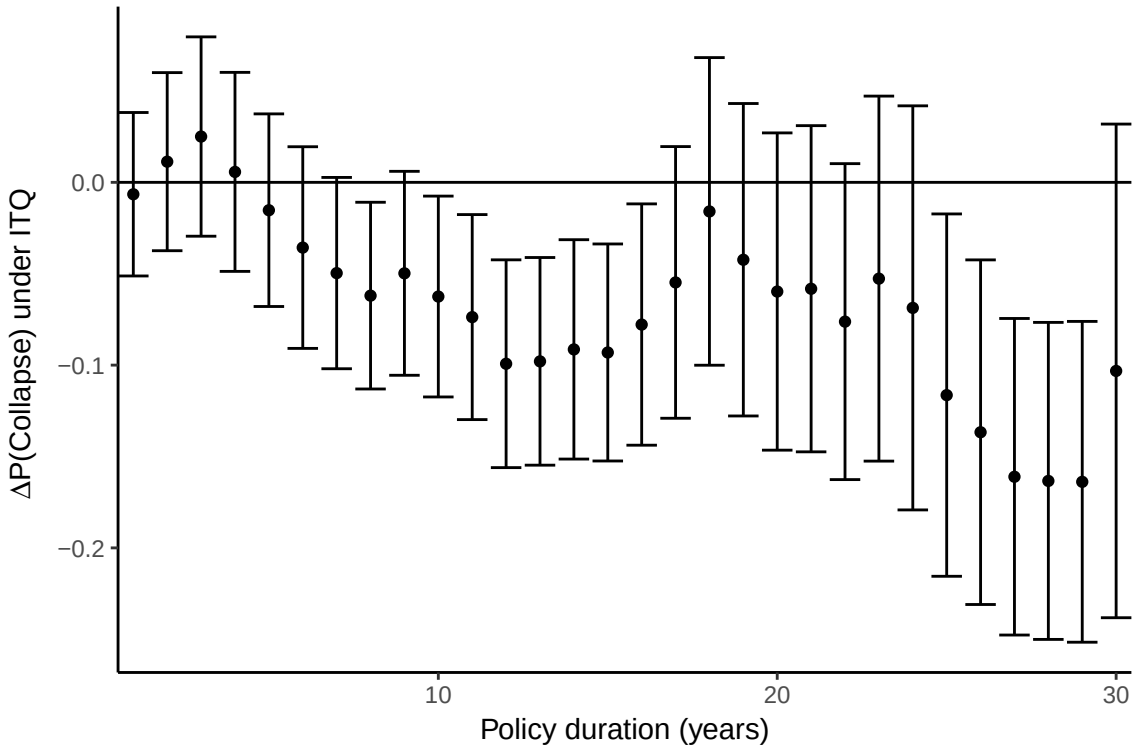

Figure 5: Change in probability of collapse due to IQs as a function of the number of years IQ management has been in effect. Points represent mean estimates and error bars coincide with 95% confidence intervals, with all quantities computed via panel bootstrapping at the fishery level.

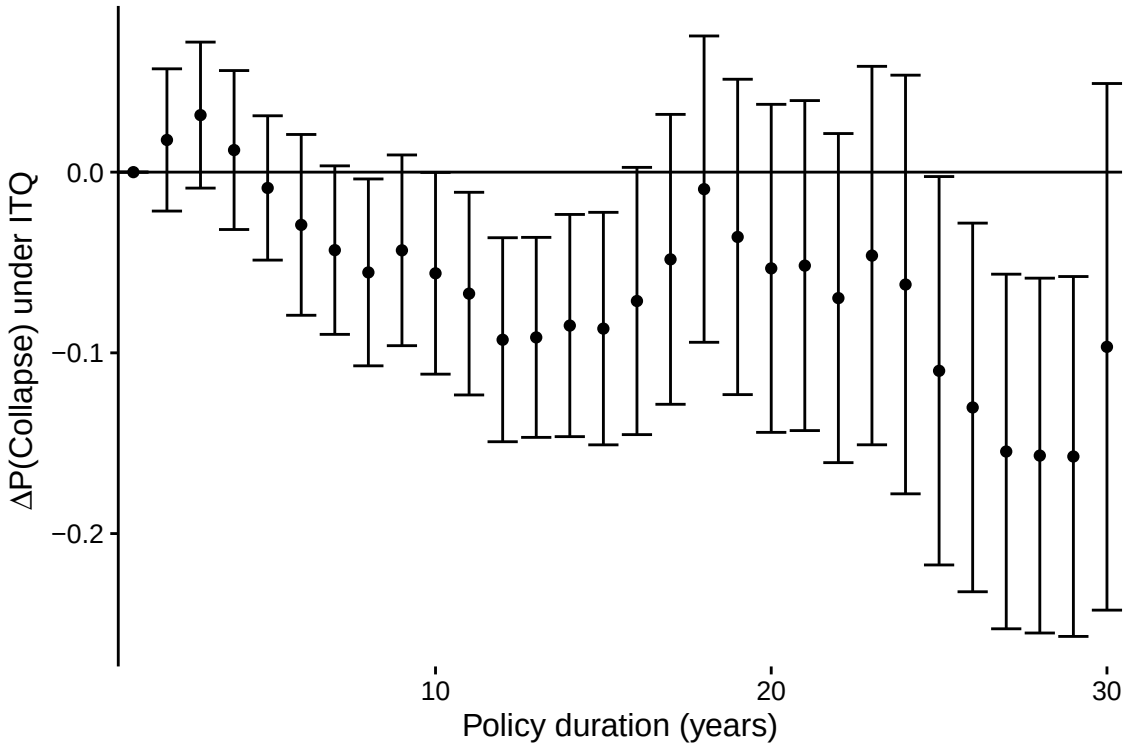

Figure 6: Difference in policy effects from the effect in the first year of implementation. Negative estimated differences indicate IQs reduce the probability of collapse more in that year than in the first year of implementation. Points represent mean estimates and error bars coincide with 95% confidence intervals.

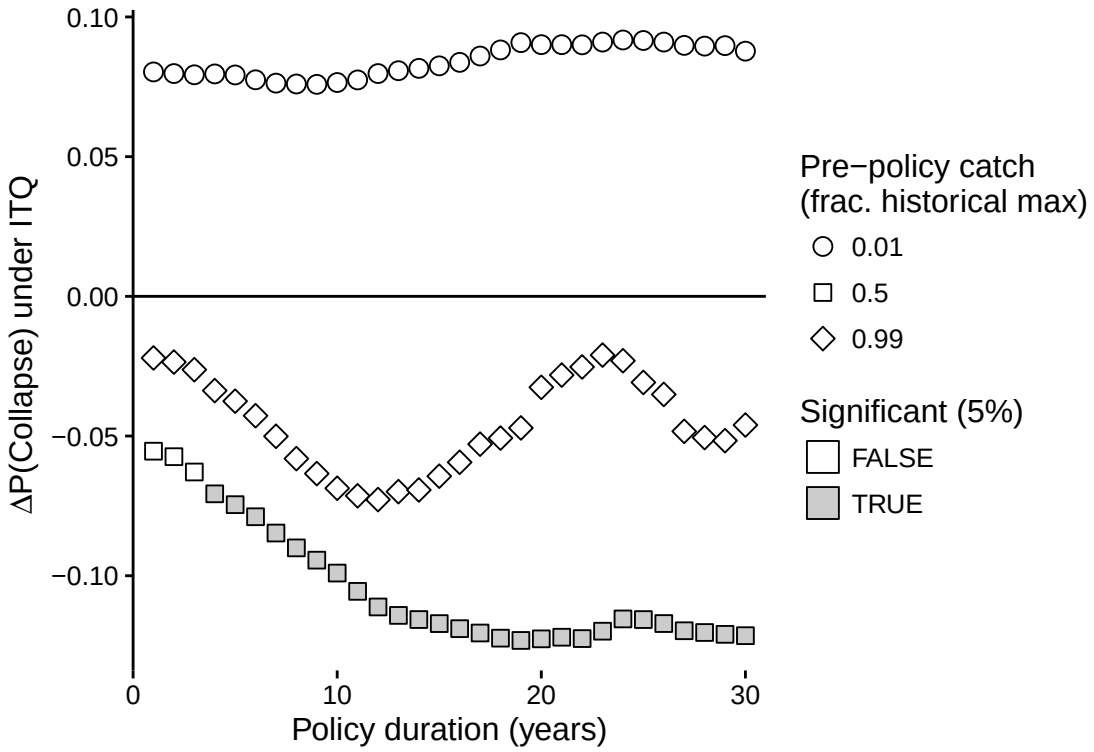

Figure 7: Change in probability of collapse due to IQs as a function of policy duration, grouped by pre-policy catch as a fraction of historical maximum catch (point shape). Filled (unfilled) shapes are significant (not significant) at the 5% level. All estimated policy effects are for a policy start year of 1992 and a sample mean of pre-policy relative catch trends.

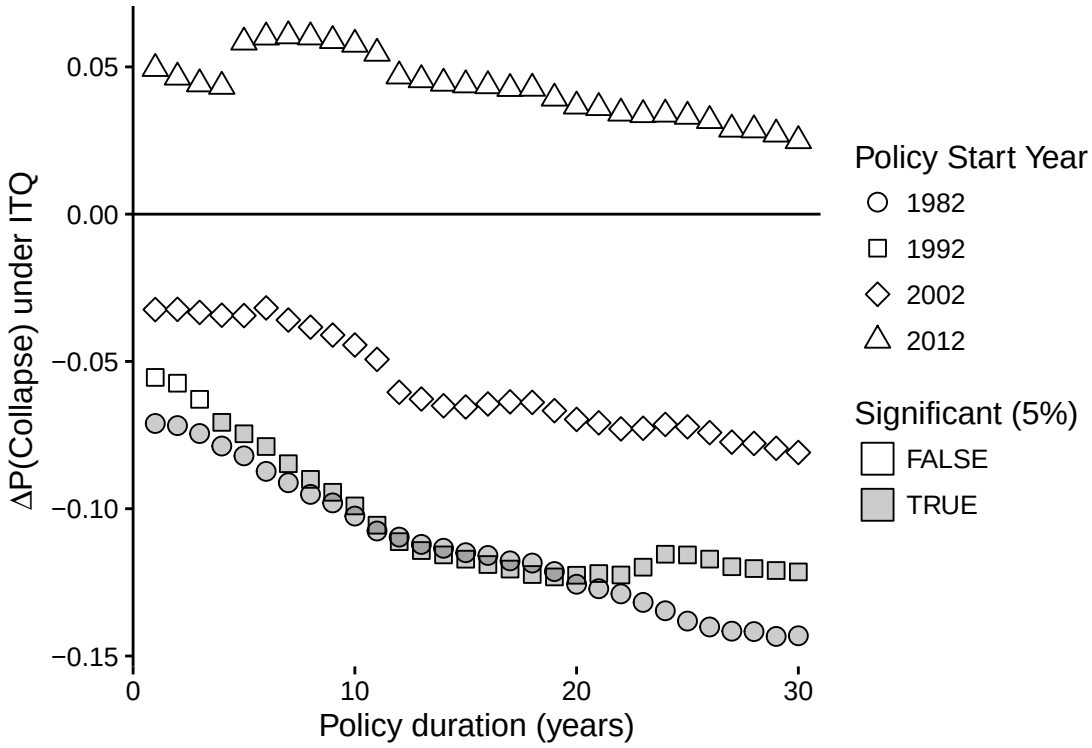

Figure 8: Change in probability of collapse due to IQs as a function of policy duration, grouped by potential policy start year (point shape). Filled (unfilled) shapes are significant (not significant) at the 5% level. All estimated policy effects are for pre-policy relative catch of 0.5 and the sample mean of pre-policy relative catch trends.

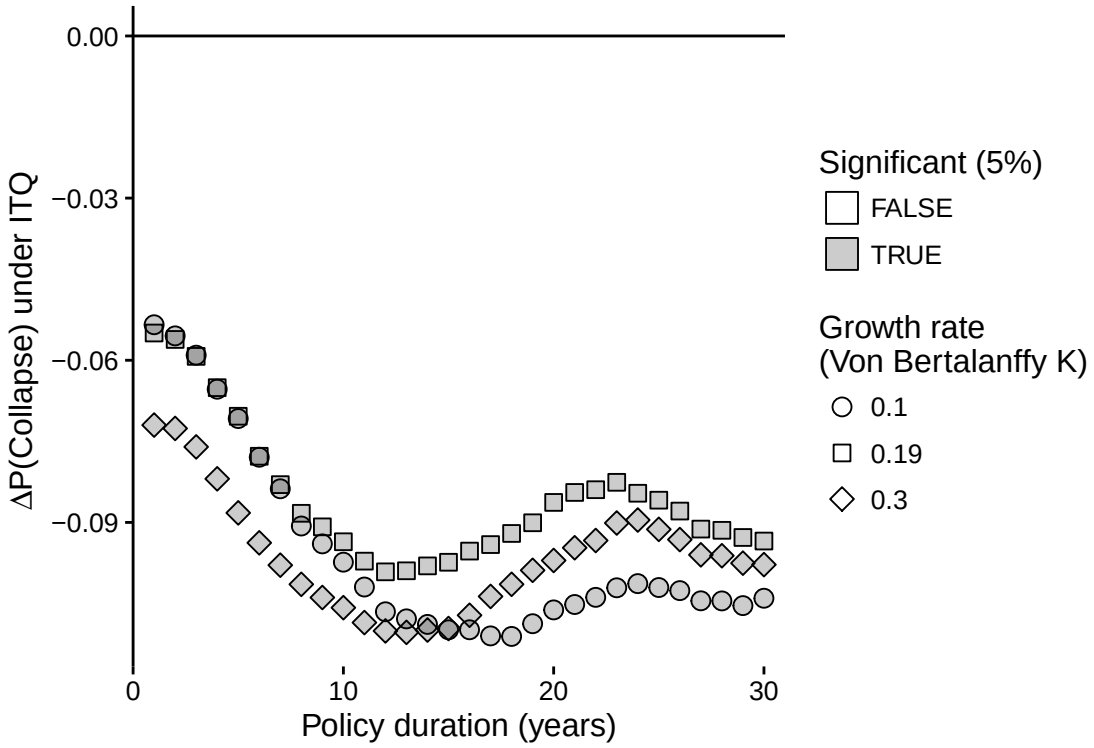

Figure 9: Change in probability of collapse due to IQs as a function of policy duration, grouped by Von-Bertalanffy growth rate (point shape). Growth rates shown are the 25th, 50th and 75th quantiles in the data. Filled (unfilled) shapes are significant (not significant) at the 5% level. All estimated policy effects are for potential policy start year of 1992, pre-policy relative catch of 0.5, and the sample mean of pre-policy relative catch trends.