Estimating the effects of universal transfers: new ML approach and application to labor supply reaction to child benefits

How Do the Results of Forest-Based Estimation Depend on Hyper-Parameters of the Forest?

Filip Premik*

Abstract

With the development of statistical theory behind the machine learning algorithms, they are becoming an important tool in the empirical economists' toolbox. By construction, they rely on a set of pre-specified hyper parameters governing the architecture of the algorithm chosen arbitrarily by a researcher. In this note I consider a model of female labor supply by Premik (2021). I show that the economic interpretation of the estimates (obtained via Generalized Random Forest (Athey et al., 2019)) is robust to different choices of the hyper parameters. This is an encouraging result suggesting that despite their complexity, the machine learning algorithms are likely to become a part of applied econometricians' toolbox.

*The author gratefully acknowledges the support of the National Science Center (grant #2016/23/N/HS4/03637).

Introduction

Machine learning methods have gained remarkable attention over the span of last 20 years. They have been developed primarily as tools aiming for improving accuracy of prediction tasks. Initially, little was known about statistical properties of estimators based on these approaches. It created an obstacle preventing their broader use within empirical economics. This obstacle seems to be overcome by recent advances in the literature that establish statistical theory for estimation based on various machine learning algorithms. As a result, machine learning algorithms have increasingly been present in the econometric toolbox of empirical economists.

Contrary to the traditional econometric methods, estimators based on machine learning algorithms often require specifying so called hyper (or nuisance) parameters. They serve for algorithm calibration and are not subject of estimation. The task of setting up values of these parameters most frequently reflects a well known trade-off between low variance and bias of an estimator. Traditionally, data scientists choose values for these parameters from a pre-specified grid in a way to optimize prediction accuracy (an exercise frequently referred to as cross-validation). Researchers usually lack this obvious heuristic when it comes to estimation of economic models. Similarly, statistical theory does not always suggest a right choice of values for the hyper-parameters.

In this note I analyze the model developed and estimated in Premik (2021) and examine how much interpretation of the estimation results depends on the method's hyper-parameters. Premik (2021) essentially uses the Generalized Random Forest estimator by Athey et al. (2019) to obtain consistent (and asymptotically normal) conditional probabilities of choice. I focus on estimating the quarterly decomposition of the labor force flows, as given by equation (8) in the paper. There are two reasons for this choice. First, quarterly estimates are obtained on smaller subsamples of the data compared to the pre-post approach. By construction that is likely to elevate their variance and hence make it more vulnerable to the choice of nuisance parameters. Second, the main result in Premik (2021) is based on the quarterly models. Evaluating their performance is hence a test of reliability of paper's findings.

I re-estimate the quarterly version of the model using different calibrations for nuisance parameters to obtain time series of the decomposition given by equation (8) in Premik (2021). In order to evaluate their impact on estimation results, I compare the variance resulting from varying nuisance parameters with the estimator's sampling error. In other words, I verify whether the estimated time paths under various hyper parametrizations fall within the bounds of 95% confidence interval of the main result reported in the paper. If that is the case, I draw a conclusion that the choice of hyper parameters does not affect significantly interpretation of the paper's results.

In the remainder of this note, I describe the original nuisance parametrization in Premik (2021). Then I provide a brief overview of the particular random forest hyper parameters that need to be set in order to perform GRF estimation and discuss their impact on interpretation of the economic result in the paper.

Original Parametrization

To estimate the conditional choice probabilities of being in the labor force I use the `regression.forest` function from the `GRF` package in `R` (Tibshirani et al., 2020), developed as an implementation of the method proposed by Athey et al. (2019). Premik (2021) uses the following calibration of the nuisance parameters:

- `seed` = 7335
- `num.trees` = 500
- `sample.fraction` = 0.5
- `mtry` = 40
- `alpha` = 0.05
- `honesty.fraction` = 0.5

If not specified differently, parameters of the any forest discussed in this note are specified as above.

Forest Hyper-Parameters

Initial Seed and Size of a Forest

The random forest approach is a randomized ensemble of trees. Each tree is grown on a random subsample of the data. In addition, for a single tree the data splits in each consecutive step are performed based on a randomly selected subset of covariates. Therefore, depending on the initial seed set for pseudo-random number generation, the draws of subsamples and subsets of covariates vary, leading to differences in the final estimates of the conditional choice probabilities. This additional variation may affect interpretation of the final results. A way to decrease it (relative to the sampling error) is to increase the number of trees in the forest. The intuition is that with a number of trees large enough the same observations and covariates will be drawn sufficiently often for the forest to capture their impact on the conditional choice probabilities.

Figures (1)-(4) show time series of parameters of decomposition (8) in the paper. The main result is compared to estimates obtained through forests based on 20 different initial `seed` values and with 100, 500, 1000, 2000 trees (parameter `num.trees`) respectively. The choice of the initial seed does not introduce significant variance to the estimates, as paths estimated for different seed never leave the original confidence interval. They are also numerically very close to the main parametrization time series, regardless of the seed and number of trees - particularly for forests with 500 and more trees.

Sample Fraction for Growing a Tree

Before running a random-forest based estimation, one needs to decide about the fraction of a sample that is used to grow a single tree (`sample.fraction` option in the `GRF` package). Small values of this parameter may increase variance of the estimator as it effectively restricts the tree estimation sample size. Large values may in turn increase the finite sample bias, as some potential heterogenous effects among a smaller sub-populations may be overwhelmed by the sample average effects.

Figure (5) shows the decomposition (8) estimated with `sample.fraction` varying among 0.1, 0.2, 0.3, 0.4 and 0.5[1]. The estimated paths are numerically very close to each other are contained in the original 95% confidence interval, suggesting that the choice of this hyper-parameter does not affect interpretation of the final result.

Number of Covariates to Grow a Tree

At each split of a tree, a subset of variables is drawn from a set of all available covariates to perform a split. Hyper parameter `mtry` controls the size of this subset. The `GRF` package suggest a default value of

$$\min\{\sqrt{p} + 20, p\}$$

to mimic the Poisson process used in the theoretical part of their paper (where p is the total number of available covariates). With $p = 379$ (generating default `mtry` = 40 used in the paper), I consider `mtry` values of 5,10,20,40,60,80,100,150,200,300. Figure (6) shows the estimated paths. For nearly all parameters of the decomposition (8), the choice of `mtry` is irrelevant for their economic interpretation. The only exception concerns the selection parameter for outflows (figure 6e). However, in this case the only path that goes out of the 95% confidence interval is the one for `mtry` = 5. With such a small number of covariates to perform splits, trees – and in consequence the whole forest – are likely to suffer accuracy deficiency and poor ability to reflect complex patterns in the data. Hence, I conclude that `mtry` does not affect the interpretation of the results for reasonably parametrized `mtry`.

Split Imbalance

Parameter `alpha` controls degree of imbalance that is allowed at each tree split. As Tibshirani et al. (2020) explain in the user manual at their webpage, *when splitting a parent node, the size of each child node is not allowed to be less than size(parent) * alpha.*. I checked values 0, 0.05, 0.1, 0.15, 0.2 and 0.25. Similarly as `mtry`, the only parameter of the decomposition (8) for which it may matter is the selection parameter for outflows (figure (7e). In particular, for

[1] 0.5 is maximal level feasible using `GRF` package

`alpha` larger or equal than 0.15 the estimates are visibly lower than the main paper's result in 2019Q1 and subsequent quarters.

Large values of `alpha` impose that the tree splits create subsamples of similar sizes. That restricts away the forest ability of capturing the heterogenous impact on smaller subpopulations. For example, if a certain value of covariate x makes 10% of females in the sample changing their labor force participation decisions and `alpha` is at 0.25, the effect will not be captured leading to biased estimates. Therefore, paths' deviations at large values of `alpha` should be treated with a caution. As a result, the divergence in time series after 2019Q1 at figure (7e) is likely to be generated by too restrictive parametrization of the forest.

Honesty Fraction

Honesty in random forest estimation is a crucial concept for establishing the desired statistical properties (see Wager and Athey (2018)). It aims to reduce bias of tree predictions by using different subsamples for performing the splits (i.e. growing a tree) and making predictions (i.e. populating the final leaves). The hyper parameter `honesty.fraction` $\in (0, 1)$ decides which proportion of the subsample drawn for estimating a single tree (of a size of `sample.fraction` times the original sample size) to be used for performing splits. The remaining part is then used to populate the final leaves. I consider the following values for `honesty.fraction`: 0.4, 0.45, 0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85 and 0.9. As figure (8) shows, the choice of `honesty.fraction` does not change the economic interpretation of the main result in the paper.

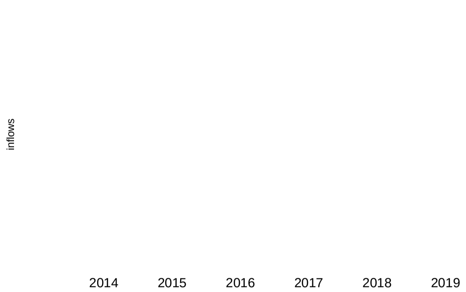
Conclusion

In this note I analyze how the economic interpretation of estimation based on a forest-based algorithm depends on pre-specified hyper parameters. Using an example of a recent study by Premik (2021), I show that the nuisance parameters does not introduce significant variation in the economic interpretation of the estimates, particularly for reasonable parametrizations.

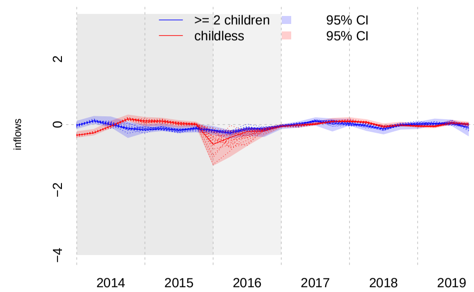
This is an encouraging result, suggesting that machine learning techniques can be used successfully by economists without necessity to rely on essentially arbitrary assumptions regarding the architecture of an algorithm.

FIGURES

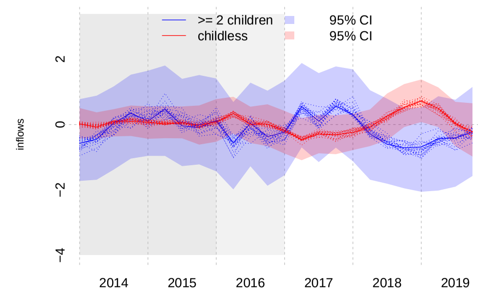
Figure 1: Estimated Parameters of Decomposition (**??**), Time Series Approach, 100 forests. The sold lines with 95% confidence interval present the main result from Premik (2021). The dotted lines present paths for forests obtained using various initial seeds.



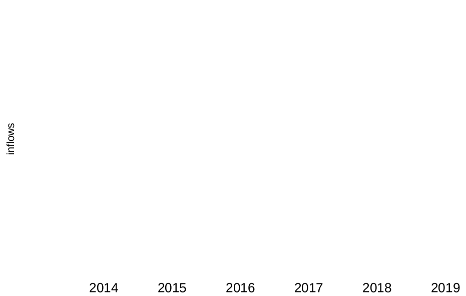
(a) inflows: *treatment* parameters: $\hat{\beta}(s_{t-1}) \equiv \hat{\varrho}_t(s_{t-1}) - \hat{\varrho}_{t-1}(s_{t-1})$



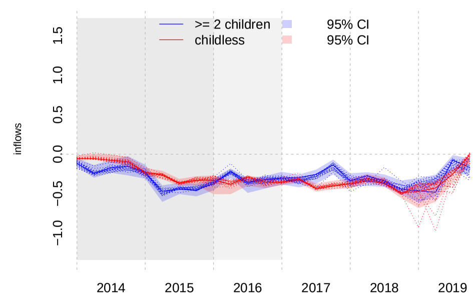
(b) inflows: *selection* parameters: $\hat{\gamma}_t(s_t, s_{t-1}) \equiv \hat{\varrho}_t(s_t) - \hat{\varrho}_t(s_{t-1})$



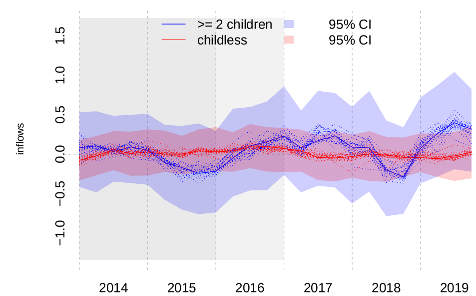
(c) inflows: *idiosyncratic* parameters: $\hat{\xi}_t(s_t, s_{t-1}) \equiv \big(\overline{y}_1 - \overline{y}_0\big) - \big(\hat{\varrho}_1(s_1) - \hat{\varrho}_0(s_0)\big)$



(d) outflows: *treatment* parameters: $\hat{\beta}(s_{t-1}) \equiv \hat{\varrho}_t(s_{t-1}) - \hat{\varrho}_{t-1}(s_{t-1})$

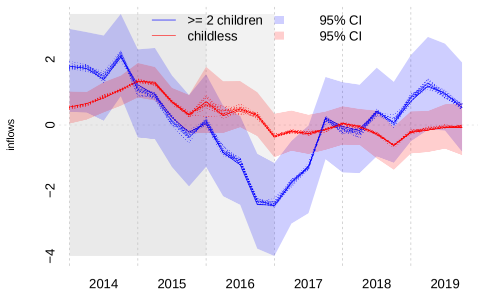


(e) outflows: *selection* parameters: $\hat{\gamma}_t(s_t, s_{t-1}) \equiv \hat{\varrho}_t(s_t) - \hat{\varrho}_t(s_{t-1})$

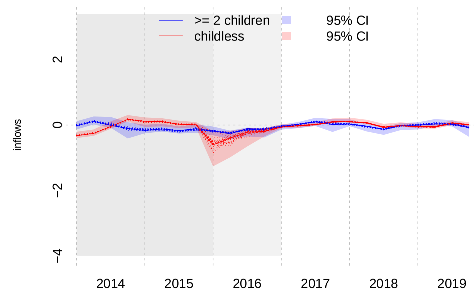


(f) outflows: *idiosyncratic* parameters: $\hat{xi}_t(s_t, s_{t-1}) \equiv \big(\overline{y}_1 - \overline{y}_0\big) - \big(\hat{\varrho}_1(s_1) - \hat{\varrho}_0(s_0)\big)$

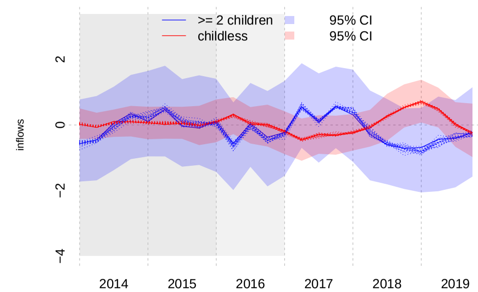
Figure 2: Estimated Parameters of Decomposition (**??**), Time Series Approach, 500 forests. The sold lines with 95% confidence interval present the main result from Premik (2021). The dotted lines present paths for forests obtained using various initial seeds.



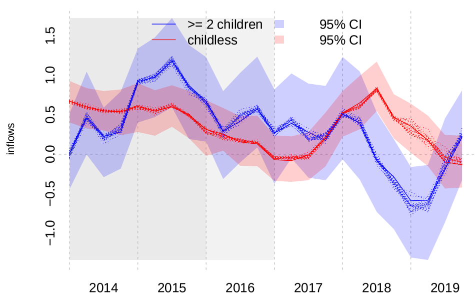
(a) inflows: *treatment* parameters: $\hat{\beta}(s_{t-1}) \equiv \hat{\varrho}_t(s_{t-1}) - \hat{\varrho}_{t-1}(s_{t-1})$



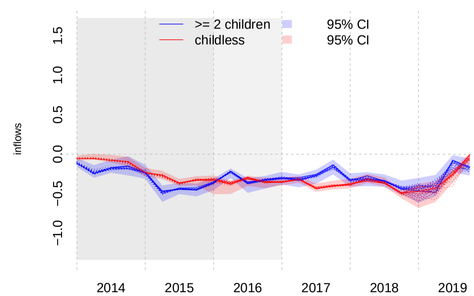
(b) inflows: *selection* parameters: $\hat{\gamma}_t(s_t, s_{t-1}) \equiv \hat{\varrho}_t(s_t) - \hat{\varrho}_t(s_{t-1})$



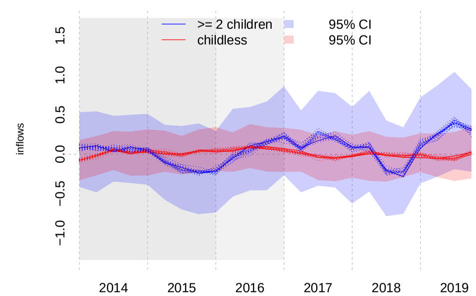
(c) inflows: *idiosyncratic* parameters: $\hat{\xi}_t(s_t, s_{t-1}) \equiv (\overline{y}_1 - \overline{y}_0) - (\hat{\varrho}_1(s_1) - \hat{\varrho}_0(s_0))$



(d) outflows: *treatment* parameters: $\hat{\beta}(s_{t-1}) \equiv \hat{\varrho}_t(s_{t-1}) - \hat{\varrho}_{t-1}(s_{t-1})$

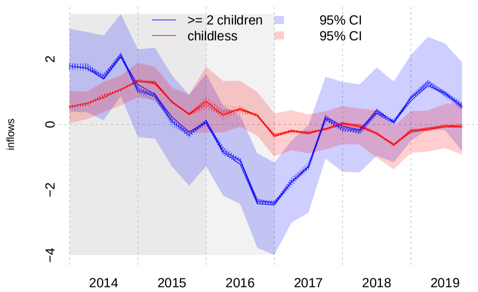


(e) outflows: *selection* parameters: $\hat{\gamma}_t(s_t, s_{t-1}) \equiv \hat{\varrho}_t(s_t) - \hat{\varrho}_t(s_{t-1})$

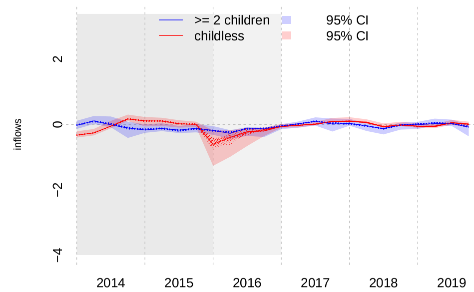


(f) outflows: *idiosyncratic* parameters: $\hat{xi}_t(s_t, s_{t-1}) \equiv (\overline{y}_1 - \overline{y}_0) - (\hat{\varrho}_1(s_1) - \hat{\varrho}_0(s_0))$

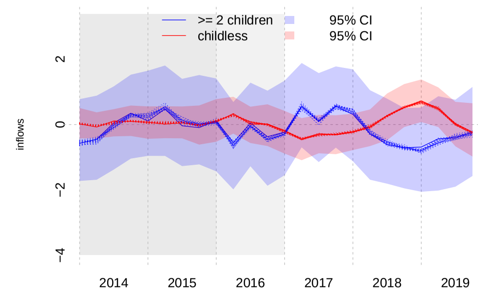
Figure 3: Estimated Parameters of Decomposition (??), Time Series Approach, 1000 forests. The sold lines with 95% confidence interval present the main result from Premik (2021). The dotted lines present paths for forests obtained using various initial seeds.



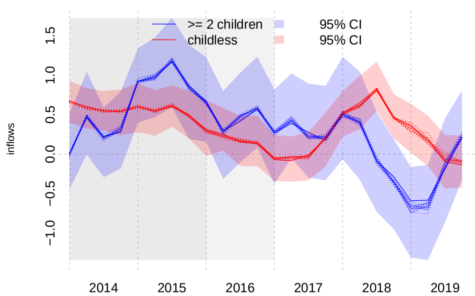
(a) inflows: *treatment* parameters: $\hat{\beta}(s_{t-1}) \equiv \hat{\varrho}_t(s_{t-1}) - \hat{\varrho}_{t-1}(s_{t-1})$



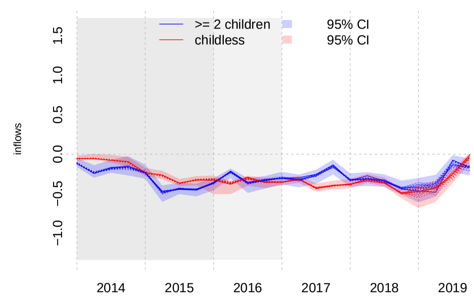
(b) inflows: *selection* parameters: $\hat{\gamma}_t(s_t, s_{t-1}) \equiv \hat{\varrho}_t(s_t) - \hat{\varrho}_t(s_{t-1})$



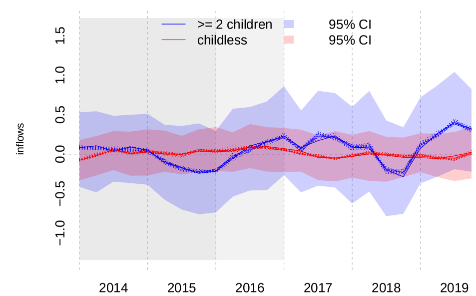
(c) inflows: *idiosyncratic* parameters: $\hat{\xi}_t(s_t, s_{t-1}) \equiv (\overline{y}_1 - \overline{y}_0) - (\hat{\varrho}_1(s_1) - \hat{\varrho}_0(s_0))$



(d) outflows: *treatment* parameters: $\hat{\beta}(s_{t-1}) \equiv \hat{\varrho}_t(s_{t-1}) - \hat{\varrho}_{t-1}(s_{t-1})$

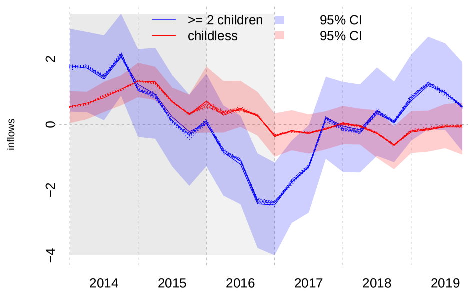


(e) outflows: *selection* parameters: $\hat{\gamma}_t(s_t, s_{t-1}) \equiv \hat{\varrho}_t(s_t) - \hat{\varrho}_t(s_{t-1})$

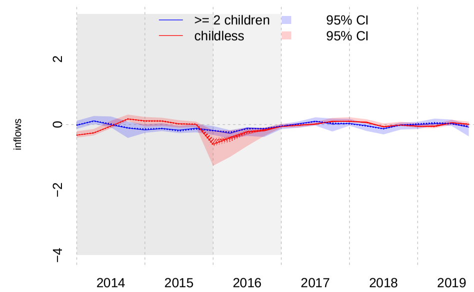


(f) outflows: *idiosyncratic* parameters: $\hat{xi}_t(s_t, s_{t-1}) \equiv (\overline{y}_1 - \overline{y}_0) - (\hat{\varrho}_1(s_1) - \hat{\varrho}_0(s_0))$

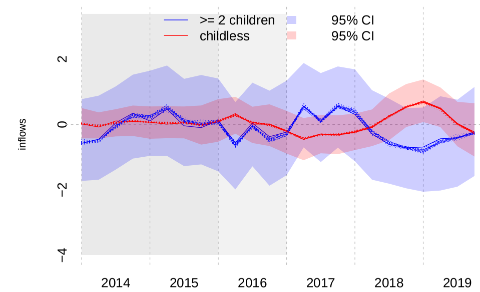
Figure 4: Estimated Parameters of Decomposition (**??**), Time Series Approach, 2000 forests. The sold lines with 95% confidence interval present the main result from Premik (2021). The dotted lines present paths for forests obtained using various initial seeds.



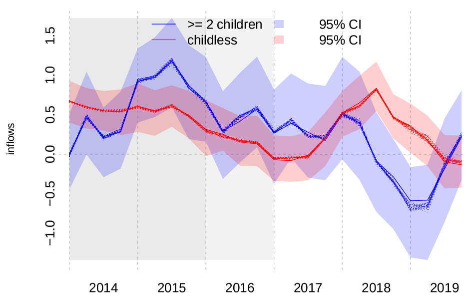
(a) inflows: *treatment* parameters: $\hat{\beta}(s_{t-1}) \equiv \hat{\varrho}_t(s_{t-1}) - \hat{\varrho}_{t-1}(s_{t-1})$



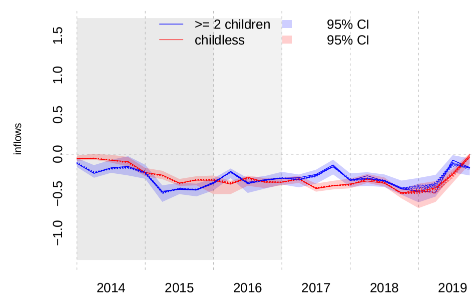
(b) inflows: *selection* parameters: $\hat{\gamma}_t(s_t, s_{t-1}) \equiv \hat{\varrho}_t(s_t) - \hat{\varrho}_t(s_{t-1})$



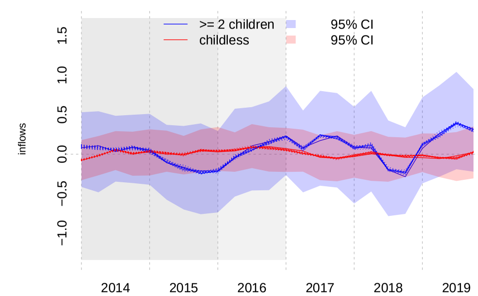
(c) inflows: *idiosyncratic* parameters: $\hat{\xi}_t(s_t, s_{t-1}) \equiv \left(\overline{y}_1 - \overline{y}_0\right) - \left(\hat{\varrho}_1(s_1) - \hat{\varrho}_0(s_0)\right)$



(d) outflows: *treatment* parameters: $\hat{\beta}(s_{t-1}) \equiv \hat{\varrho}_t(s_{t-1}) - \hat{\varrho}_{t-1}(s_{t-1})$

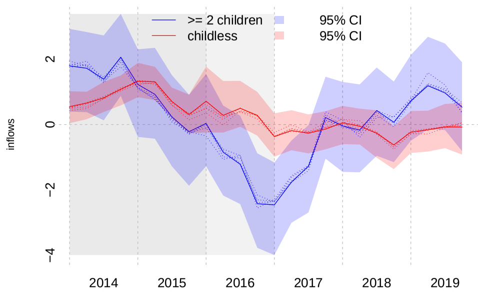


(e) outflows: *selection* parameters: $\hat{\gamma}_t(s_t, s_{t-1}) \equiv \hat{\varrho}_t(s_t) - \hat{\varrho}_t(s_{t-1})$

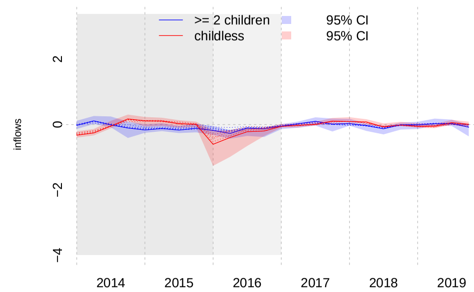


(f) outflows: *idiosyncratic* parameters: $\hat{xi}_t(s_t, s_{t-1}) \equiv \left(\overline{y}_1 - \overline{y}_0\right) - \left(\hat{\varrho}_1(s_1) - \hat{\varrho}_0(s_0)\right)$

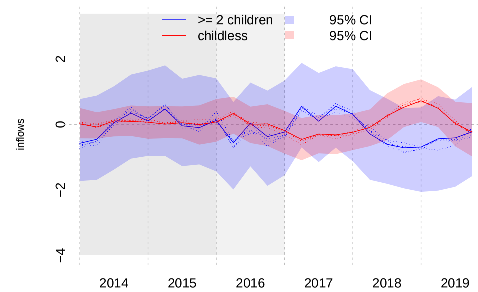
Figure 5: Estimated Parameters of Decomposition (8), Time Series Approach, calibration of `sample.fraction`. The sold lines with 95% confidence interval present the main result from Premik (2021). The dotted lines present paths for forests obtained using various initial seeds.



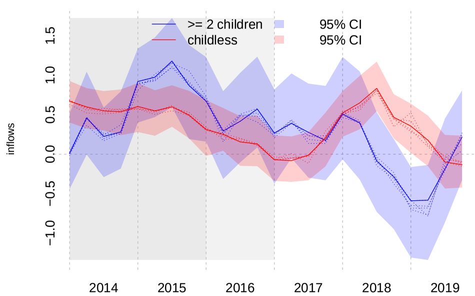
(a) inflows: *treatment* parameters: $\hat{\beta}(s_{t-1}) \equiv \hat{\varrho}_t(s_{t-1}) - \hat{\varrho}_{t-1}(s_{t-1})$



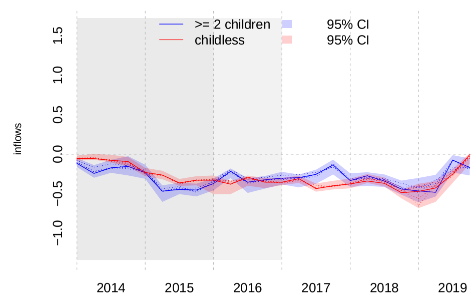
(b) inflows: *selection* parameters: $\hat{\gamma}_t(s_t, s_{t-1}) \equiv \hat{\varrho}_t(s_t) - \hat{\varrho}_t(s_{t-1})$



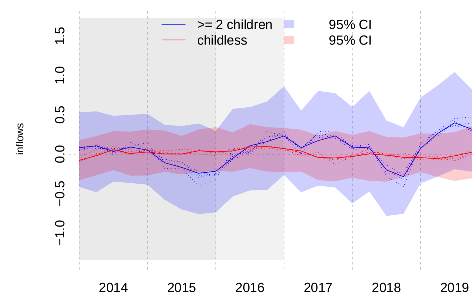
(c) inflows: *idiosyncratic* parameters: $\hat{\xi}_t(s_t, s_{t-1}) \equiv (\overline{y}_1 - \overline{y}_0) - (\hat{\varrho}_1(s_1) - \hat{\varrho}_0(s_0))$



(d) outflows: *treatment* parameters: $\hat{\beta}(s_{t-1}) \equiv \hat{\varrho}_t(s_{t-1}) - \hat{\varrho}_{t-1}(s_{t-1})$

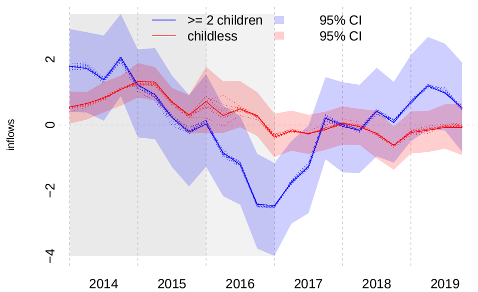


(e) outflows: *selection* parameters: $\hat{\gamma}_t(s_t, s_{t-1}) \equiv \hat{\varrho}_t(s_t) - \hat{\varrho}_t(s_{t-1})$

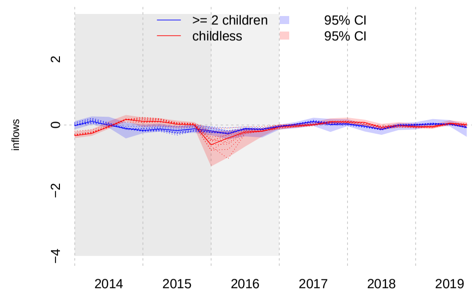


(f) outflows: *idiosyncratic* parameters: $\hat{xi}_t(s_t, s_{t-1}) \equiv (\overline{y}_1 - \overline{y}_0) - (\hat{\varrho}_1(s_1) - \hat{\varrho}_0(s_0))$

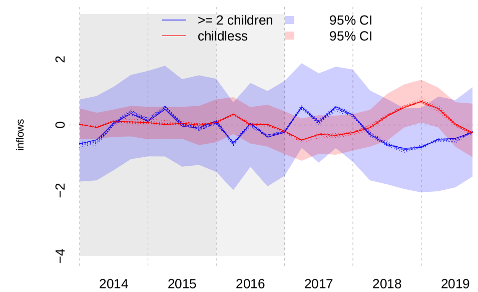
Figure 6: Estimated Parameters of Decomposition (8), Time Series Approach, calibration of `mtry`. The sold lines with 95% confidence interval present the main result from Premik (2021). The dotted lines present paths for forests obtained using various initial seeds.



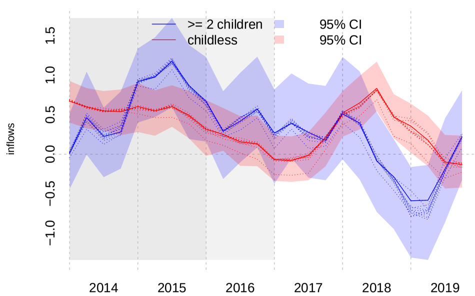
(a) inflows: *treatment* parameters: $\hat{\beta}(s_{t-1}) \equiv \hat{\varrho}_t(s_{t-1}) - \hat{\varrho}_{t-1}(s_{t-1})$



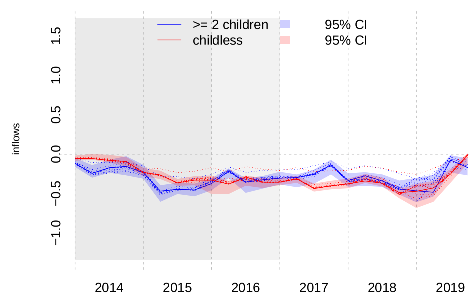
(b) inflows: *selection* parameters: $\hat{\gamma}_t(s_t, s_{t-1}) \equiv \hat{\varrho}_t(s_t) - \hat{\varrho}_t(s_{t-1})$



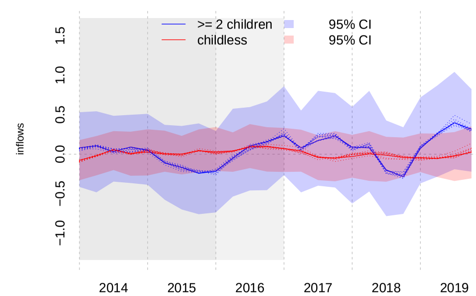
(c) inflows: *idiosyncratic* parameters: $\hat{\xi}_t(s_t, s_{t-1}) \equiv \left(\overline{y}_1 - \overline{y}_0\right) - \left(\hat{\varrho}_1(s_1) - \hat{\varrho}_0(s_0)\right)$



(d) outflows: *treatment* parameters: $\hat{\beta}(s_{t-1}) \equiv \hat{\varrho}_t(s_{t-1}) - \hat{\varrho}_{t-1}(s_{t-1})$

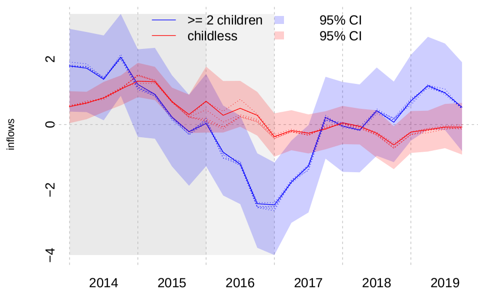


(e) outflows: *selection* parameters: $\hat{\gamma}_t(s_t, s_{t-1}) \equiv \hat{\varrho}_t(s_t) - \hat{\varrho}_t(s_{t-1})$

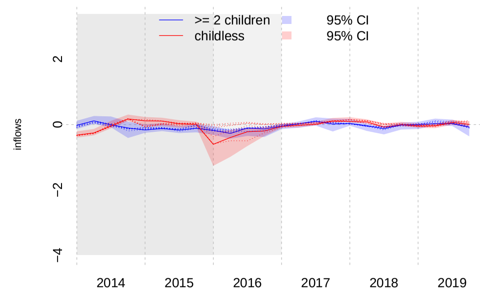


(f) outflows: *idiosyncratic* parameters: $\hat{xi}_t(s_t, s_{t-1}) \equiv \left(\overline{y}_1 - \overline{y}_0\right) - \left(\hat{\varrho}_1(s_1) - \hat{\varrho}_0(s_0)\right)$

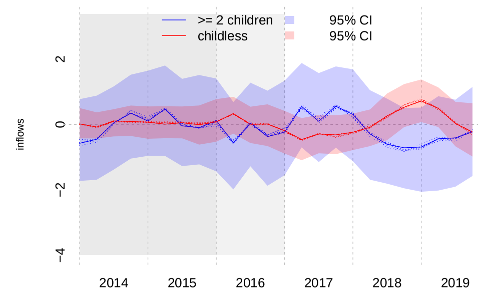
Figure 7: Estimated Parameters of Decomposition (8), Time Series Approach, calibration of `alpha`. The sold lines with 95% confidence interval present the main result from Premik (2021). The dotted lines present paths for forests obtained using various initial seeds.



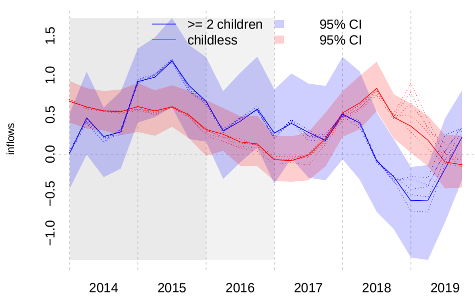
(a) inflows: *treatment* parameters: $\hat{\beta}(s_{t-1}) \equiv \hat{\varrho}_t(s_{t-1}) - \hat{\varrho}_{t-1}(s_{t-1})$



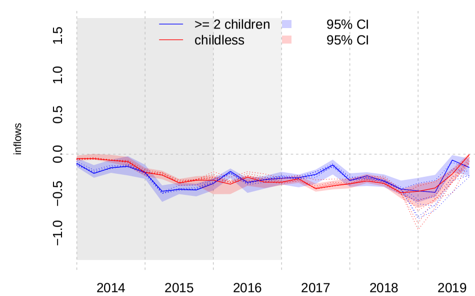
(b) inflows: *selection* parameters: $\hat{\gamma}_t(s_t, s_{t-1}) \equiv \hat{\varrho}_t(s_t) - \hat{\varrho}_t(s_{t-1})$



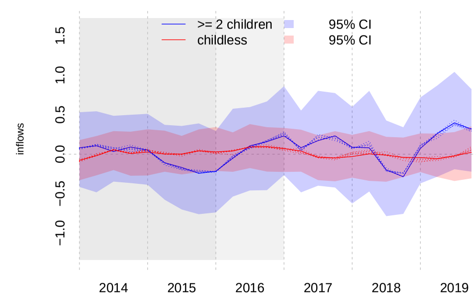
(c) inflows: *idiosyncratic* parameters: $\hat{\xi}_t(s_t, s_{t-1}) \equiv \big(\overline{y}_1 - \overline{y}_0\big) - \big(\hat{\varrho}_1(s_1) - \hat{\varrho}_0(s_0)\big)$



(d) outflows: *treatment* parameters: $\hat{\beta}(s_{t-1}) \equiv \hat{\varrho}_t(s_{t-1}) - \hat{\varrho}_{t-1}(s_{t-1})$

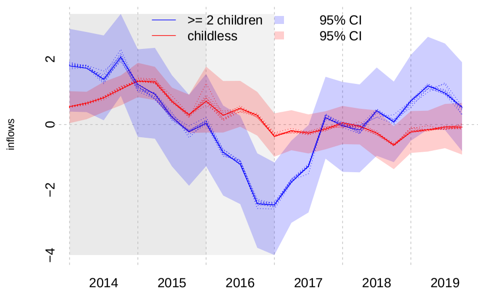


(e) outflows: *selection* parameters: $\hat{\gamma}_t(s_t, s_{t-1}) \equiv \hat{\varrho}_t(s_t) - \hat{\varrho}_t(s_{t-1})$

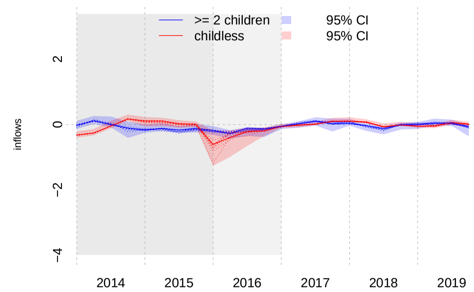


(f) outflows: *idiosyncratic* parameters: $\hat{xi}_t(s_t, s_{t-1}) \equiv \big(\overline{y}_1 - \overline{y}_0\big) - \big(\hat{\varrho}_1(s_1) - \hat{\varrho}_0(s_0)\big)$

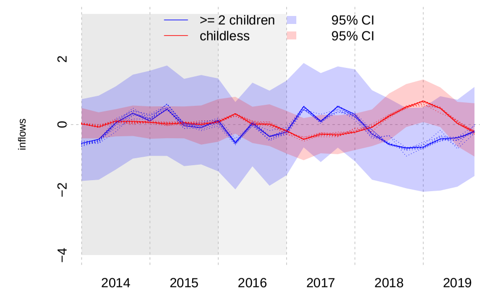
Figure 8: Estimated Parameters of Decomposition (8), Time Series Approach, calibration of `honesty.fraction`. The sold lines with 95% confidence interval present the main result from Premik (2021). The dotted lines present paths for forests obtained using various initial seeds.



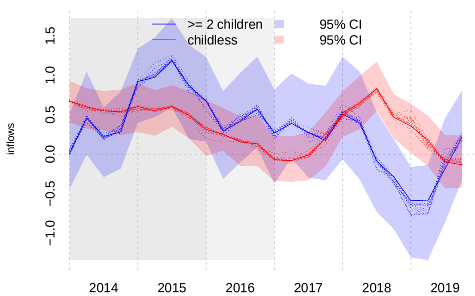
(a) inflows: *treatment* parameters: $\hat{\beta}(s_{t-1}) \equiv \hat{\varrho}_t(s_{t-1}) - \hat{\varrho}_{t-1}(s_{t-1})$



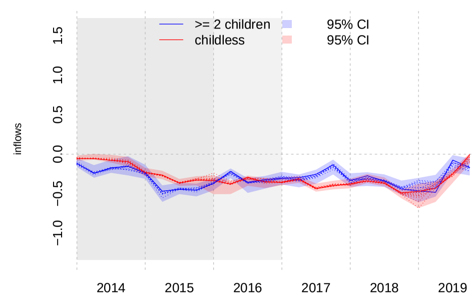
(b) inflows: *selection* parameters: $\hat{\gamma}_t(s_t, s_{t-1}) \equiv \hat{\varrho}_t(s_t) - \hat{\varrho}_t(s_{t-1})$



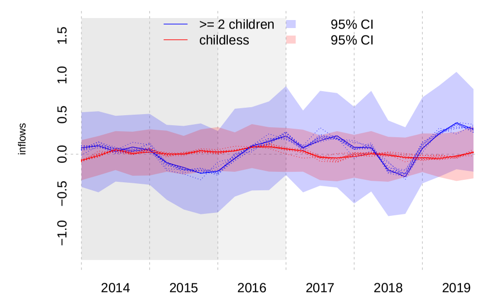
(c) inflows: *idiosyncratic* parameters: $\hat{\xi}_t(s_t, s_{t-1}) \equiv (\overline{y}_1 - \overline{y}_0) - (\hat{\varrho}_1(s_1) - \hat{\varrho}_0(s_0))$



(d) outflows: *treatment* parameters: $\hat{\beta}(s_{t-1}) \equiv \hat{\varrho}_t(s_{t-1}) - \hat{\varrho}_{t-1}(s_{t-1})$



(e) outflows: *selection* parameters: $\hat{\gamma}_t(s_t, s_{t-1}) \equiv \hat{\varrho}_t(s_t) - \hat{\varrho}_t(s_{t-1})$



(f) outflows: *idiosyncratic* parameters: $\hat{xi}_t(s_t, s_{t-1}) \equiv (\overline{y}_1 - \overline{y}_0) - (\hat{\varrho}_1(s_1) - \hat{\varrho}_0(s_0))$

References

Athey, S., J. Tibshirani, S. Wager, et al. (2019): "Generalized random forests," *The Annals of Statistics*, 47, 1148–1178.

Premik, F. (2021): "Estimating the effects of universal transfers: new ML approach and application to labor supply reaction to child benefits," *FAME—GRAPE Working Paper*.

Tibshirani, J., S. Athey, and S. Wager (2020): *grf: Generalized Random Forests*, r package version 1.2.0.

Wager, S. and S. Athey (2018): "Estimation and inference of heterogeneous treatment effects using random forests," *Journal of the American Statistical Association*, 113, 1228–1242.