Random Distribution Shift in Refugee Placement: Strategies for Building Robust Models

Kirk Bansak*
Department of Political Science
Univeristy of California, Berkeley
kbansak@berkeley.edu

Elisabeth Paulson*
Technology and Operations Management Unit
Harvard Business School
epaulson@hbs.edu

Dominik Rothenhäusler*
Department of Statistics
Stanford University
rdominik@stanford.edu

Abstract

Algorithmic assignment of refugees and asylum seekers to locations within host countries has gained attention in recent years, with implementations in the US and Switzerland. These approaches use data on past arrivals to generate machine learning models that can be used (along with assignment algorithms) to match families to locations, with the goal of maximizing a policy-relevant integration outcome such as employment status after a certain duration. Existing implementations and research train models to predict the policy outcome directly, and use these predictions in the assignment procedure. However, the merits of this approach, particularly in non-stationary settings, has not been previously explored. This study proposes and compares three different modeling strategies: the standard approach described above, an approach that uses newer data and proxy outcomes, and a hybrid approach. We show that the hybrid approach is robust to both distribution shift and weak proxy relationships—the failure points of the other two methods, respectively. We compare these approaches empirically using data on asylum seekers in the Netherlands. Surprisingly, we find that both the proxy and hybrid approaches out-perform the standard approach in practice. These insights support the development of a real-world recommendation tool currently used by NGOs and government agencies.

1 Introduction

Under the status quo refugee resettlement and/or asylum procedures in many countries—including the United States, Switzerland, Netherlands, Sweden, and others—newly arrived refugees and/or asylum seekers are centrally assigned to different localities across the country as they arrive, subject to capacity and other constraints. The objectives of national resettlement and asylum programs often include integration goals, such as helping newcomers achieve economic self-sufficiency. In support of this objective, Bansak et al. [2018] introduced outcome-based refugee assignment, whereby refugees are matched to locations in a manner that seeks to maximize a chosen metric of integration success, such as employment. To accomplish this, machine learning (ML) models, trained on historical data, are used to generate counterfactual outcome predictions for every refugee–location combination upon arrival, which are then used in an assignment procedure. Further research has built upon this initial proposal, incorporating various other features and constraints [Gölz and Procaccia, 2019, Acharya

*Faculty Affiliate, Immigration Policy Lab, Stanford University and ETH Zurich.

et al., 2022, Ahani et al., 2021a,b, Bansak and Paulson, 2022], and algorithmic refugee assignment has been implemented on a limited basis in Switzerland and the United States.

The goal of the algorithm designer is to maximize the outcome of interest to the policy makers—what we refer to as the *policy outcome*—which can be thought of as exogenously given. However, there remains a question as to what outcome(s) and estimation method should be used to produce the predictions that will guide the assignment decisions. Existing research has largely overlooked this question, operating under the assumption that assignments are made by simply generating direct predictions of the policy outcome, using models that are estimated with all available labeled historical data. We refer to this method, currently deployed in Switzerland and the US, as the *standard approach*. While this approach may be suitable in some scenarios (e.g., with stationary data), its implications are less clear in real-world settings characterized by distribution shifts.

In this study we consider how to choose and model the outcome(s) amidst random distribution shifts over time. Understandably, policy outcomes are often long-term, as longer-term outcomes better capture the experience and welfare of refugees than short-term outcomes that may reflect transient dynamics. Yet due to the potentially lengthy duration of policy outcomes, distribution shift can pose a problem, causing models fit on old data to be biased with respect to the true expected outcomes for today's newly arrived cases. This paper addresses the following questions: Given the goal of maximizing a specific policy outcome, what outcome(s) should actually be used—and how should they be used—in estimating the models that will guide assignment decisions? How and under what conditions would using shorter-term outcomes in the estimation process be beneficial?

1.1 Contributions

This study makes several contributions. On the theoretical front, we contribute to the research agenda on algorithmic decision-making under distribution shifts (see related work below) by providing the first comparison of models under random distribution shifts in a nonparametric setting. We formalize and characterize the benefits and drawbacks of several alternative prediction strategies under such random distribution shifts: the *standard approach* described above, an approach that uses newer data to predict a shorter-term but related proxy outcome (the *proxy approach*), and a *hybrid approach* that utilizes both older and newer data, and both the policy and proxy outcome. Theorem 1 provides an exact characterization of the weaknesses of each approach: the standard approach is ineffective when distribution shift is large, and the proxy approach falters if the proxy relationship is not strong enough. The hybrid approach, on the other hand, is robust to both of these failure points.

Moreover, we establish that this issue is not simply a theoretical novelty but a matter of real-world, practical significance by presenting empirical evidence from an IRB-approved study on asylum seekers in the Netherlands. Our findings indicate that among the three proposed strategies, we expect the standard approach to result in significantly lower employment outcomes compared to the other two approaches, and the hybrid and proxy approaches perform similarly, indicating a strong proxy relationship and substantial distribution shift in the data (which is further supported by additional analyses). Consequently, we underscore the non-obvious fact that maximizing a longer-term outcome may, in some cases, be best achieved by making decisions based on a related shorter-term outcome, even when ample historical data is available. These insights support the development of a real-world recommendation tool currently employed by NGOs and government agencies to determine resettlement locations for newly arriving refugees and asylum seekers.

2 Related work

Self-training. A popular method for combining labeled data with unlabeled data is self-training [Scudder, 1965, Chapelle et al., 2006]. In its most basic form, one starts by learning a prediction mechanism only for the labeled data. Then, based on imputations on the unlabelled data, the prediction mechanism is re-trained on its own predictions for the unlabelled data and the original labeled data. In the past few years, there has been an increasing interest in studying the theoretical properties of self-training, in particular, its robustness under distribution shift [Carmon et al., 2019, Chen et al., 2020, Raghunathan et al., 2020, Kumar et al., 2020]. Self-training is similar to the approach τ^C below. Due to the distribution shift structure, we discard the outdated data during the re-training stage. We add to this literature by showing that a variant of self-training is more robust under random, symmetric distribution shift than alternative procedures.

Pre-training. In the transfer learning literature, it is common to learn a feature representation by regressing auxiliary data on the covariates on a large data set and then using this as a feature vector on the smaller (target) data set, either by updating the feature vector or by regressing the final outcome on the feature representation [Caruana, 1997, Weiss et al., 2016, Hendrycks et al., 2019]. The first step is similar to the approach τ^B below, i.e. using the more abundant short-term outcomes to train a first prediction model. One important property of our setting is that predictions of the short-term outcome are already a reasonable optimization target.

Adversarial machine learning. There is a rich literature that models distributional perturbations as adversarial [Huber, 1964, Ben-Tal et al., 2009, Maronna et al., 2019, Biggio and Roli, 2018]. The intuition is that a small change in input at the training or prediction stage should not change the output. Adversarial inputs are somewhat pessimistic for our problem setup, where the changes occur due to natural shifts over time.

Invariance-based approaches. Instead of considering small shifts, another line of work considers distribution shifts that lie on subspaces and relies on the assumption that certain conditional probabilities or representations stay invariant across settings. In other words, these works study representations and procedures that use invariances to transfer across settings [e.g., Ganin and Lempitsky, 2015, Rojas-Carulla et al., 2018, Arjovsky et al., 2019, Rothenhäusler et al., 2021]. For an overview, see Chen and Bühlmann [2021]. These methods have shown some success, but do not consistently beat empirical risk minimization [Gulrajani and Lopez-Paz, 2020, Koh et al., 2021].

Surrogate outcomes. When policy outcomes are long-term, surrogate outcomes are a common tool to guide policy decisions on a faster time scale [Prentice, 1989]. In the traditional setting, the policy maker may have access to historical data containing the long-term outcome but not the treatment decision, and experimental data containing the treatment decision but not the long-term outcome [Yang et al., 2020, Athey et al., 2019]. Thus, the problem can be viewed as a missing data issue. By contrast, in this paper the historical data contains both treatment decisions and the long-term outcome of interest, and the motivation for using a shorter-term version of the outcome is distribution shift.

3 Setup

3.1 Preliminaries

Consider a finite time horizon with three distinct periods: period 0 (the present), period -1 (one period prior), and period -2 (two periods prior). In each period, refugees arrive and must be assigned to a single location within the set $\mathcal{L} = \{1, 2, ...|\mathcal{L}|\}$. Let $J \in \mathcal{L}$ denote the location assignment of a refugee that arrives in the current period, and let X denote a vector of background characteristics.

Let Y_2 be the policy outcome of interest, measured two periods after arrival, and Y_1 be a related but shorter-term outcome measured one period after arrival[1]. Following the potential outcomes framework of Neyman [1923] and Rubin [1974], let $Y_1(j)$ and $Y_2(j)$ denote the one-period and two-period potential outcomes under assignment to location $j \in \mathcal{L}$. We assume that the stable unit treatment assumption (SUTVA) holds, so we observe $Y_1 = Y_1(J)$ and $Y_2 = Y_2(J)$ [Rubin, 1980].

For refugees who arrive in any period t and the variables defined above, we posit the existence of a tuple generated according to a probability distribution P_t:

$$(Y_1(1), ..., Y_1(|\mathcal{L}|), Y_2(1), ..., Y_2(|\mathcal{L}|), X, J)_t \sim P_t.$$

Further, we assume that within any period, the assignment mechanism is such that

$$Y_1(j) \perp\!\!\!\perp J \mid X, \;\; Y_2(j) \perp\!\!\!\perp J \mid X, \;\; \text{and} \;\; P(J = j|X = x) > 0 \;\; \forall \;\; j, x.$$

This is often referred to as the ignorability assumption [Rosenbaum and Rubin, 1983]. It is satisfied here when location assignments are made on the basis of the observed characteristics, and when those characteristics do not preclude refugees from being assigned to any particular location, as in our application to the Dutch asylum system. Therefore, $\mathbb{E}_t[Y_k(j)|X] = \mathbb{E}_t[Y_k|X, J = j]$ for $k = 1, 2$. This paper is focused on the identifiable quantities $\mathbb{E}_t[Y_k|X, J = j]$.

[1]Appendix B discusses how to extend the proposed methods to more than two time periods and outcomes

3.2 Problem

The policy objective in the present period 0 is to assign the refugees arriving in this period to locations in order to maximize the sum of Y_2 among period 0 arrivals. To maximize Y_2, we would ideally choose location assignments for the refugees arriving in period 0 by first estimating $\tau(x, j) = \mathbb{E}_0[Y_2|X = x, J = j]$, and using these estimates within a static or dynamic matching procedure to optimally assign each refugee to a particular location subject to any necessary constraints, such as capacity constraints and the need to keep family members together, as in the algorithmic assignment procedures proposed by Bansak et al. [2018], Ahani et al. [2021a,b] and Bansak and Paulson [2022]. However, we do not observe Y_2 (nor do we observe Y_1) for the cohort that arrives in period 0 at the time of their assignment. Instead, at the time of assignment, we observe the data $(Y_2, Y_1, X, J)_{-2} \sim P_{-2}$, $(Y_1, X, J)_{-1} \sim P_{-1}$, and $(X)_0 \sim P_0$.

3.3 Alternative approaches

In the presence of distribution shift, there are (at least) three strategies one could pursue:

1. $\tau^A(x, j) = \mathbb{E}_{-2}[Y_2|X = x, J = j]$
2. $\tau^B(x, j) = \mathbb{E}_{-1}[Y_1|X = x, J = j]$
3. $\tau^C(x, j) = \mathbb{E}_{-1}[\mathbb{E}_{-2}[Y_2|Y_1, X, J]|X = x, J = j]$

The first approach estimates the policy outcome using the data from period -2.[2] Using the resulting model(s) for $\tau^A(x, j)$, expected outcomes in each possible location j can be estimated for the newly arriving refugees in period 0 given their background characteristics x, and optimal assignments then made using those estimates. This is the default approach in the literature. Unfortunately, its performance may suffer under distribution shift due to its reliance on period -2 data.

The second approach estimates the shorter-term outcome Y_1 using data from period -1, and uses these estimates to inform the location assignments. Relative to $\tau^A(x, j)$, the advantage of this strategy is that it employs more recent data, and hence is less susceptible to the effects of distribution shift. However, this approach is only reasonable if optimizing the outcome Y_2 results in similar assignments as optimizing the proxy outcome Y_1. When $\mathbb{E}_t[Y_2|X = x, J = j] = \beta_1\mathbb{E}_t[Y_1|X = x, J = j] + \beta_0$, where $\beta_1 > 0$ and β_0 are unknown constants and $t \in \{-2, 1, 0\}$, optimizing on Y_1 will yield identical assignments as optimizing with respect to Y_2 (due to the linearity of assignment algorithms). Even if perfect linearly does not hold, the assignments may still remain close.

The third approach again estimates Y_2, but uses the data from both periods -2 and -1. This strategy proceeds by first regressing Y_2 on X and Y_1 using the period -2 data, and then regressing that fitted model on X using the period -1 data. The general idea behind the third strategy is that we want to devise the best available approximation of the density $P_0(y_2, y_1, x, j) = P_0(y_2|y_1, x, j)P_0(y_1, x, j)$. We cannot know $P_0(y_2|y_1, x, j)$ and so we use the best guess $P_{-2}(y_2|y_1, x, j)$. And we cannot know $P_0(y_1, x, j)$ so we use the best guess $P_{-1}(y_1, x, j)$, since P_{-1} is likely closer to P_0 than is P_{-2}.

Relationships between strategies in the absence of distribution shift. In the special case of no distribution shift, $\tau^A = \tau$. Hence, the first strategy is the natural strategy of choice in reasonably stationary data-generating environments. Furthermore, by the tower property of conditional expectations, it is also the case that $\tau = \tau^A = \tau^C$. From an infinite data perspective, then, the first and third strategies are interchangeable in the absence of distribution shift. In the following, we will study these procedures in more detail from a distribution-shift perspective.

4 Predictions under distribution shift

We model distributional changes as random [Jeong and Rothenhäusler, 2022], and introduce a simplified model that captures natural shifts and allows us to study tradeoffs between the three methods. To avoid measurability issues and simplify the discussion, we will focus on discrete distributions, that is, the random variable takes value in a finite alphabet $(X, Y_1, Y_2, J) \in \mathcal{X} \times \mathcal{Y}_1 \times \mathcal{Y}_2 \times \mathcal{L}$. Note that this focus does not impose a serious practical constraint, as any continuous

[2] In practice, data from earlier periods could also be used. This is also true for the second and third approaches.

variables of interest can be arbitrarily coarsened to meet this requirement. We consider the distribution $P_{-2}(y_1, y_2|x, j)$ as fixed, and assume that $P_{-1}(y_1, y_2|x, j)$ differs randomly from $P_{-2}(y_1, y_2|x, j)$ and that $P_0(y_1, y_2|x, j)$ differs randomly from $P_{-1}(y_1, y_2|x, j)$. To be more specific, we write

$$S_t(y_1, y_2|x, j) = P_{t+1}(y_1, y_2|x, j) - P_t(y_1, y_2|x, j) \text{ for } t \in \{-2, -1\},$$

where we call $S_t(y_1, y_2|x, j)$ a distributional shift variable. That is, between time 0 and -1 and between time -1 and -2 the probabilities of events get shifted by a random amount that can depend on y_1, y_2, x, and j. There are two constraints on the shift variables to make P_{-1} and P_{-2} well-defined probability measures. First, the shift variables must satisfy $0 \leq P_t(y_1, y_2|x, j) + S_t(y_1, y_2|x, j) \leq 1$. Furthermore, since $\sum_{y_1,y_2} P_0(y_1, y_2|x, j) = 1 = \sum_{y_1,y_2} P_{-1}(y_1, y_2|x, j)$ the shift variable S_t has to satisfy $\sum_{y_1,y_2} S_t(y_1, y_2|x, j) = 0$.

We will not discuss how to construct such distribution shifts. Such constructions can be found in Jeong and Rothenhäusler [2022], where for fixed x and j the authors introduce a perturbation process, which we employ here, that satisfies

$$\begin{aligned} \text{Var}(S_{-2}(\bullet|x, j)) &= \kappa_{-2} P_{-2}(\bullet|x, j)(1 - P_{-2}(\bullet|x, j)), \text{ and} \\ \text{Var}(S_{-1}(\bullet|x, j)|S_{-2}) &= \kappa_{-1} P_{-1}(\bullet|x, j)(1 - P_{-1}(\bullet|x, j)), \end{aligned} \tag{1}$$

for some potentially unknown $\kappa_t > 0$ for all events $\bullet \subseteq \mathcal{Y}_1 \times \mathcal{Y}_2$. In equation 1, the event $\bullet$, and the variables x and j are considered fixed, and the variance is over the randomness in the shift S_{-2} (top equation) or S_{-1} (bottom equation).

Intuitively, this perturbation process captures "well-behaved" or "symmetric" distribution shift [Jeong and Rothenhäusler, 2022], where the perturbation of an event depends on its initial probability: events that have a small probability are only perturbed by a small amount. In the concrete context of modeling employment, this captures gradual shifts in labor market dynamics. One can think about the unknown scaling factor κ_t as a measure of the strength of distribution shift between P_{t+1} and P_t.

In addition to the structure above, we allow the distribution to shift arbitrarily in (X, J); however, to avoid issues of identifiability, we require the support of (X, J) to be time invariant. Note that we also allow the shifts to be correlated across different x, that is we allow

$$\text{Cov}(S_t(y_1, y_2|x, j), S_t(y_1, y_2|x', j')) \neq 0$$

for $x \neq x'$ or $j \neq j'$. This allows for shifts that affect many units in a similar way (for example, an economic boom). We also assume that the perturbation process has (conditional) mean zero, that is $\mathbb{E}[S_{-2}] = 0$ and $\mathbb{E}[S_{-1}|S_{-2}] = 0$. Intuitively, this means that the perturbation has no momentum: the perturbation that shifts P_{-2} to P_{-1} is uncorrelated with the perturbation that shifts P_{-1} to P_0. This does not rule out memory/momentum driven by systematic factors (e.g. systematic or secular changes in the labor market) that can be captured by time and other covariates contained in X.

We now present a theorem that compares the predictive performance of the three approaches under distribution shift, in the infinite-data case. Finite-sample considerations are discussed in Section 4.3. The proof can be found in Appendix D. For clarity, we will assume that Y_1 and Y_2 are on the same scale. That is, a positive linear transformation has been applied to Y_1 in a pre-processing step such that $\mathbb{E}_{-2}[(\mathbb{E}_{-2}[Y_1|X, J] - \mathbb{E}_{-2}[Y_2|X, J])^2]$ is as low as possible. Strictly speaking, the mathematical results below do not require this scaling, but the interpretation is clearer with scaling. When Y_1 is rescaled, we will rename τ^B as $\tilde{\tau}^B$ and Y_1 as $\tilde{Y}_1$.

Theorem 1 (Comparison of methods)**.** *Under the assumptions made above,*

$$\begin{aligned} \text{MSE}(\tau^C) &= \underbrace{\kappa_{-2}\mathbb{E}_{-2}[(Y_2 - \mathbb{E}_{-2}[Y_2|Y_1, X, J])^2]}_{\textit{shift between } -2 \textit{ and } -1} + K_1 + O(\kappa_{-1}\kappa_{-2}) \\ \text{MSE}(\tau^A) &= \underbrace{\kappa_{-2}\mathbb{E}_{-2}[(Y_2 - \mathbb{E}_{-2}[Y_2|X, J])^2]}_{\textit{shift between } -2 \textit{ and } -1} + K_1 + O(\kappa_{-1}\kappa_{-2}) \\ \text{MSE}(\tilde{\tau}^B) &= \underbrace{\mathbb{E}_{-2}[(\mathbb{E}_{-2}[Y_2 - \tilde{Y}_1|X, J])^2]}_{\textit{error due to using proxy outcome}} + \underbrace{\kappa_{-2}\mathbb{E}_{-2}[(Y_2 - \tilde{Y}_1 - \mathbb{E}_{-2}[Y_2 - \tilde{Y}_1|X, J])^2]}_{\textit{inflation of proxy error under shift}} + K_1 \\ &\quad + O(\kappa_{-1}\kappa_{-2}) \end{aligned}$$

where $K_1 := \kappa_{-1}\mathbb{E}_{-2}[(Y_2 - \mathbb{E}_{-2}[Y_2|X, J])^2]$ *is the shift between* -1 *and* 0 *that impacts all methods equally. Here the MSE is over the randomness in distribution shifts and the randomness in* X

and J, that is MSE$(\tau^\bullet) = \mathbb{E}[(\tau^\bullet(X,J) - \tau(X,J))^2]$, where the outer expectation is both over the randomness in distribution shift and the randomness in X and J.

Let us discuss how to interpret the results for the case of τ^C. On a high level, one can think about the error as

$$\text{MSE}(\tau^C) = \sum_{\text{shifts}} \text{strength of shift} \cdot \text{ sensitivity}$$

One part of the error depends on the shift between -2 and -1, and one part depends on the shift between -1 and 0. The strength of the shift enters the equation via κ_{-1} and κ_{-2}, and is multiplied by the sensitivity of the $\tau^C(x,j)$ with respect to that particular distributional change. The sensitivity of $\tau^C(x,j)$ depends on the noise level $\mathbb{E}_{-2}[(Y_2 - \mathbb{E}_{-2}[Y_2|Y_1,X,J])^2]$. This makes sense intuitively: if the noise is high that means there are important unobserved variables besides X and J that determine the value of Y_2. Since those unobserved variables can shift in distribution over time, generalization may suffer. The estimator τ^A has a similar decomposition and interpretation as τ^C.

Now let us turn to the interpretation of the error of $\tilde{\tau}^B$. There are three terms. The first term captures whether $\tilde{Y}_1$ is a reasonable proxy for Y_2 in the absence of distribution shift. Even if $\tilde{Y}_1$ is a good proxy under P_{-2}, due to distribution shift this might change between P_{-2} and P_{-1}. This is captured in the second term. The second term may be zero, for example in the case where $\tilde{Y}_1 \overset{\text{a.s.}}{=} Y_2$. Finally, the last term corresponds to the shift between P_{-1} and P_0, which is the same for τ^C, τ^A, and $\tilde{\tau}^B$. It can be thought of as an irreducible error since all procedures are equally affected by the shift between P_{-1} and P_0. One takeaway from the error terms of $\tilde{\tau}^B$ is that ideally, a proxy satisfies $\mathbb{E}_{-2}[\tilde{Y}_1|X,J] \approx \mathbb{E}_{-2}[Y_2|X,J]$. In addition, it is also important that $\text{Var}(Y_2 - \tilde{Y}_1|X,J)$ is low, since otherwise the proxy error can inflate drastically between P_{-2} and P_0.

4.1 Comparison of the three approaches

In the following, we build further intuition for the strengths and weaknesses of the approaches.

τ^A versus τ^C. On a high level, τ^A is more affected by shifts between -2 and -1 than τ^C. In the theorem this is reflected by the fact that κ_{-2} is multiplied by sensitivity factor $\mathbb{E}_{-2}[(Y_2 - \mathbb{E}_{-2}[Y_2|Y_1,X,J])^2]$, which is always smaller, and can be much smaller, than τ^A's sensitivity factor $\mathbb{E}_{-2}[(Y_2 - \mathbb{E}_{-2}[Y_2|X,J])^2]$. Using the theorem,

$$\text{MSE}(\tau^C) \ll \text{MSE}(\tau^A) + O(\kappa_{-1}\kappa_{-2}).$$

By this argument, τ^C is generally preferable over τ^A (and is always preferable under the assumptions made above). That being said, there may be real-world situations where one would prefer τ^A over τ^C. For example, if P_{-1} is the time period of a pandemic, then the optimal prediction mechanism might change drastically for this particular year (P_{-1}), and then switch back to the original mechanism, leading to $P_{-2} \approx P_0$. In this case, the (random) distribution shifts S_{-1} and S_{-2} are negatively correlated, which we assumed not to be the case.

$\tilde{\tau}^B$ versus τ^C. The comparison between $\tilde{\tau}^B$ and τ^C is slightly more technical. Please note that both have the same dependence on κ_{-1}, the shift between P_0 and P_{-1}. However, they have a different sensitivity with respect to the shift between P_{-2} and P_{-1}. It turns out that this sensitivity is lower for τ^C:

$$\begin{aligned}\mathbb{E}_{-2}[(Y_2 - \mathbb{E}_{-2}[Y_2|Y_1,X,J])^2] &= \min_{g(Y_1,X,J)} \mathbb{E}_{-2}[(Y_2 - g(Y_1,X,J))^2] \\ &\leq \mathbb{E}_{-2}[(Y_2 - \tilde{Y}_1 - \mathbb{E}_{-2}[Y_2 - \tilde{Y}_1|X,J])^2].\end{aligned}$$

Thus, τ^C also outperforms $\tilde{\tau}^B$ (up to lower-order terms), and the gap will generally grow as the correlation of Y_2 and $\tilde{Y}_1$ decreases.

4.2 Robust predictions under distribution shift.

$\tilde{\tau}^B$ and τ^A have very different failure points. It turns out that τ^C can be seen as an interpolation estimator that behaves similarly to τ^A and $\tilde{\tau}^B$, in the following extreme cases.

Example 1 (Proxy violated)**.** *Let us consider the case where Y_1 does not help when predicting Y_2, that is $\mathbb{E}_{-2}[Y_2|Y_1, X, J] \approx \mathbb{E}_{-2}[Y_2|X, J]$. According to Theorem 1, $\tilde{\tau}^B$ may have a large mean-squared error due to a large proxy error. On the other hand, τ^A will have small mean-squared error if κ_{-2} is small. Using the tower property,*

$$\begin{aligned}\tau^C(x,j) &= \mathbb{E}_{-1}[\mathbb{E}_{-2}[Y_2|Y_1, X, J]|X = x, J = j] \\ &\approx \mathbb{E}_{-1}[\mathbb{E}_{-2}[Y_2|X, J]|X = x, J = j] \\ &= \mathbb{E}_{-2}[Y_2|X = x, J = j] \\ &= \tau^A(x,j).\end{aligned}$$

Thus, in this case the performance of τ^C is comparable to τ^A. Now let us turn to a different extreme case.

Example 2 ($\tilde{Y}_1 \approx Y_2$)**.** *Consider the case where $\mathbb{E}_{-2}[(Y_2 - \tilde{Y}_1)^2]$ is small and κ_{-2} is large, that is, where the shift between P_{-2} and P_{-1} is large. Using Theorem 1, we see that τ^A will have a large mean-squared error, while $\tilde{\tau}^B$ should be much less affected by the shift between P_{-2} and P_{-1}. Applying Jensen's inequality, we get $\mathbb{E}_{-2}[(\mathbb{E}_{-2}[Y_2|Y_1, X, J] - \tilde{Y}_1)^2] \leq \mathbb{E}_{-2}[(Y_2 - \tilde{Y}_1)^2]$. Thus, we have $\tilde{Y}_1 \approx \mathbb{E}_{-2}[Y_2|Y_1, X, J]$. Hence,*

$$\begin{aligned}\tilde{\tau}^B(x,j) &= \mathbb{E}_{-1}[\tilde{Y}_1|X = x, J = j] \\ &\approx \mathbb{E}_{-1}[\mathbb{E}_{-2}[Y_2|Y_1, X, J]|X = x, J = j] \\ &= \tau^C(x,j).\end{aligned}$$

Thus, under this extreme scenario, τ^C will perform similarly to $\tilde{\tau}^B$. In either case, τ^C will perform roughly as well as the best of both methods. We see this interpolation property as a major advantage of τ^C.

4.3 Finite-sample considerations

These results provide intuition for the relative strengths and weaknesses of each approach, which are further demonstrated using real-world data in the following section. Recall that these results apply to an infinite data setting. Since we expect the distribution shift to be of larger order than sampling uncertainty, in a finite data setting the intuition on the strengths and weaknesses of each estimator should remain consistent. However, if the distribution shift is very small there are cases where τ^A or $\tilde{\tau}^B$ may outperform τ^C asymptotically. See Appendix C for more details.

5 Empirical Assessment

In this section, we demonstrate the performance of the proposed approaches on asylum seeker data from the Netherlands. The purpose of this assessment is to empirically evaluate the implications of distribution shift on algorithmic decision-making and the relative merits of our proposed approaches on real-world data. This research was approved by the Institutional Review Boards at UC Berkeley, Harvard, and Stanford.

5.1 Data and Estimation

The data consists of background characteristics, assigned location within the Netherlands, and employment outcomes for adult asylum seekers in the Netherlands. The modeling approach used to estimate τ^A, τ^B, and τ^C is based on the methodology developed in Bansak et al. [2018]. In this approach, separate stochastic gradient boosted tree models are fit for each individual assignment location. In this context, the assignment locations are comprised of the 35 labor market regions in the Netherlands. More details on the data and estimation, including the covariates that comprise X, can be found in Appendix A. Note that year and month are also included in X to allow for systematic time trends to be captured by the models.

5.2 Counterfactual Impact Assessment

For this assessment, we consider the following two outcomes: Y_2 is the proportion of time worked in an asylum seeker's first two years after assignment (the policy outcome), and Y_1 is the proportion of

time worked in the first year after assignment. Based on preliminary analyses of the data, we find strong evidence of distribution shift in the data and a strong proxy relationship between Y_1 and Y_2 (see Appendix A.5).

We designate as a "test cohort" the n_0 asylum seekers who were assigned to labor market regions between 2018-06-01 and 2018-08-31 (the latest available data at the time of analysis) and employ the proposed approaches as if these asylum seekers were arriving (and needing to be assigned) in the present time, with the goal of maximizing the average two-year employment of the cohort. Specifically, define $\boldsymbol{W}^A$, $\boldsymbol{W}^B$, and $\boldsymbol{W}^C$, which are the analogous matrices that correspond to the predictions of τ^A, τ^B, and τ^C onto the test cohort. Each row of each matrix corresponds to an individual asylum seeker in the test cohort, with the columns corresponding to labor market regions. These matrices are estimated following the procedures described in the previous sections, with additional details provided in Appendix A. Given the test cohort's start date of 2018-06-01 (along with the fact that we treat this as the cohort needing to presently be assigned for this evaluation), estimation of τ^A, τ^B, and τ^C can only utilize data on Y_2 prior to 2016-06-01 and Y_1 prior to 2017-06-01. With these matrices in hand, let $\boldsymbol{Z^A}$, $\boldsymbol{Z^B}$, and $\boldsymbol{Z^C}$ be defined as:

$$\boldsymbol{Z^k} = \arg\max_{\mathbf{z}\in\mathcal{Z}} \sum_{i=1}^{n_0}\sum_{j=1}^{|\mathcal{L}|} w_{ij}^k z_{ij}^k \text{ for } k = A, B, C \tag{2}$$

where $\mathcal{Z}$ is the set of feasible assignments of asylum seekers to labor market regions[3] and w_{ij}^k denotes the entry in the ith row and jth column of $\boldsymbol{W}^k$. Our goal is to understand how these alternative assignment strategies perform in terms of improving two-year employment for the test cohort. Thus, we would like to compare the following quantities:

$$E^A = \sum_{i=1}^{n_0}\sum_{j=1}^{|\mathcal{L}|} w_{ij}^* z_{ij}^A,\ E^B = \sum_{i=1}^{n_0}\sum_{j=1}^{|\mathcal{L}|} w_{ij}^* z_{ij}^B,\ E^C = \sum_{i=1}^{n_0}\sum_{j=1}^{|\mathcal{L}|} w_{ij}^* z_{ij}^C \tag{3}$$

where $\boldsymbol{W}^*$ is the matrix containing the true counterfactual employment probabilities corresponding to $\tau(x, j)$. That is, for an arrival i with covariates x, $w_{ij}^* = \mathbb{E}_0[Y_2|X = x, J = j]$. Using the benefit of hindsight, we estimate a model for $\boldsymbol{W}^*$ that utilizes all available data (including the contemporaneous data that would not be available in real time). With slight abuse of notation, we will refer to this estimate as $\boldsymbol{W}^*$ and treat it as the ground truth.

We can compare each quantity E^A, E^B, and E^C to the actual average employment score achieved by the test cohort given their actual status quo assignments (14.26%) to perform a counterfactual assessment of the extent to which two-year employment could have been improved had assignments instead been made algorithmically. Finally, by performing this assessment for each of these three quantities, we can assess the relative performance of the approaches.

Figure 1 (left panel) shows the resulting gain in two-year employment using each method. The counterfactual gains in two-year employment are largest using the τ^C estimator, closely followed by τ^B, followed by τ^A. These results suggest that both the distribution shift between period -2 and 0 is relatively large, and one-year employment is a good proxy for two-year employment. Unsurprisingly, the estimated mean squared error between τ^k and τ for $k = 1, 2, 3$ is smallest for τ^C (0.0075), followed by $\tilde{\tau}^B$ (0.00939), followed by τ^A (0.01008). Thus, not only does the τ^C estimator result in the best estimation of τ, but these improved predictions translate into meaningful gains in employment ($\sim$4 percentage points above the 14.26% baseline) when used in an algorithmic matching procedure.

5.3 Additional Tests

To further assess the relative performance of the alternative methods under different scenarios and to highlight the underlying theory, we undertake a set of additional tests in which we induce (a) a violation of the proxy relationship, and (b) additional distribution shift. For the proxy violation test, Y_1 is randomly permuted for 50% of the full data set, thereby weakening the relationship between Y_1 and Y_2. For the additional distribution shift test, the pair (Y_1, Y_2) is randomly permuted for 50% of the data prior to 2016-06-01. Importantly, the two-year outcomes (Y_2) for the test cohort are

[3] At a minimum, z_{ij} must be binary with each asylum seeker assigned to exactly one location, and $\mathbf{Z}$ must satisfy capacity constraints at each location.

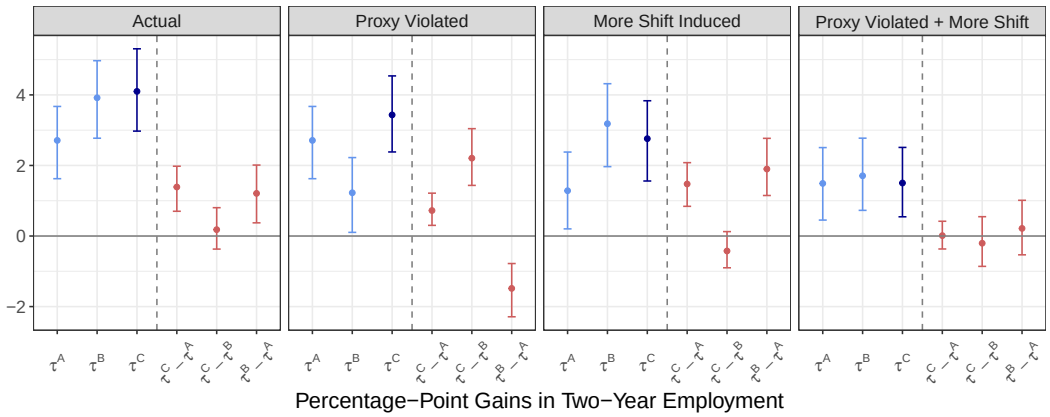

Figure 1: Counterfactual gains in two-year employment (compared to status quo) with actual data (left) and perturbed data using the alternative strategies τ_A, τ_B, and τ_C. Confidence intervals are 95% bootstrapped. For details see Appendix A.

kept intact across the tests, allowing for meaningful comparison of the results across the different test specifications. For each test, the permutations are applied prior to the training of the models corresponding to τ^A, τ^B, and τ^C, while the "ground truth" matrix $\boldsymbol{W}^*$ is fixed across tests.

The impact on gains in two-year employment for each test is shown in Figure 1. When the proxy relationship is violated, the performance of τ^B dramatically suffers, as does the performance of τ^C to a lesser extent. However, τ^C still outperforms τ^A, thus demonstrating its robustness to proxy violations. When more distribution shift is present in the data, the performance of all three predictors suffers, and τ^B and τ^C perform similarly well (with a slight edge to τ^B, although their difference is not statistically significant). When both a proxy violation and more distribution are introduced, all three estimators perform similarly. These additional tests highlight the robustness of τ^C, which performs at least as well as the other estimators, and better than in certain conditions.

6 Conclusions, Limitations, and Broader Impact

In this research, we proposed three distinct estimation strategies aimed at enhancing algorithmic location recommendations for refugees and asylum seekers, given the presence of random distribution shifts. We characterized the strengths and weaknesses of the three approaches, both theoretically and empirically, using real-world asylum-seeker data from the Netherlands. The standard approach in outcome-based refugee assignment involves training an ML model to predict an outcome of interest (e.g., employment status after two years), and using these predictions in an algorithm that recommends locations [Bansak et al., 2018, Ahani et al., 2021b,a, Bansak and Paulson, 2022]. However, when the policy outcome of interest is long-term, as is often the case in refugee resettlement, this approach will suffer under the presence of distribution shift. An alternative approach is therefore to leverage information on a shorter-term but related outcome using more recent data.

Both in theory and empirically, we show that an estimator that predicts the long-term policy outcome by combining both older data in which the policy outcome is observed, with more recent data in which only a short-term proxy is observed, is robust to both random distribution shift and the strength of the proxy relationship. Thus, in practice, using this estimator to inform the location recommendations results in higher total employment than the standard method where only the older data is used.

This research has potential for real-world impact through its integration into a recommendation software tool, and can help various stakeholders including a) the algorithms designers, b) the agencies/NGOs in charge of resettlement, and c) the refugee and asylum-seeker families. While this paper does not specifically address issues like biases and fairness, future work could include, for example, understanding whether different estimation methods could be used for different subgroups to achieve better and more equitable outcomes.

The model considered in this paper has several limitations. First, we consider a specific type of distribution shift, characterized by Equation 1. The results of our empirical analyses, which track closely with our theory, show how this can be a reasonable model in certain contexts. However, it does prohibit certain dynamics; for example, we do not address situations where $P_{-2} \approx P_0$, but $P_{-1} \neq P_0$. Such a phenomenon could occur if an isolated event impacts period -1 but period 0 reverts back to P_{-2}. Furthermore, this paper only considers random shifts. There could be large, unpredictable shifts that make all historical data useless (for example a new technology that disrupts an economy).

Acknowledgments

The authors are grateful for the guidance and data access provided by the Central Agency for the Reception of Asylum Seekers (COA) in the Netherlands and Statistics Netherlands (CBS). This work was supported by Stanford University's Human-Centered Artificial Intelligence (HAI) Hoffman-Yee Grant and J-PAL's European Social Inclusion Initiative (ESII).

References

Avidit Acharya, Kirk Bansak, and Jens Hainmueller. Combining outcome-based and preference-based matching: A constrained priority mechanism. *Political Analysis*, 30(1):89–112, 2022.

Narges Ahani, Tommy Andersson, Alessandro Martinello, Alexander Teytelboym, and Andrew C Trapp. Placement optimization in refugee resettlement. *Operations Research*, 69(5):1468–1486, 2021a.

Narges Ahani, Paul Gölz, Ariel D Procaccia, Alexander Teytelboym, and Andrew C Trapp. Dynamic placement in refugee resettlement. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, page 5, 2021b.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Susan Athey, Raj Chetty, Guido W Imbens, and Hyunseung Kang. The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely. Technical report, National Bureau of Economic Research (NBER), Working Paper No. 26463, 2019.

Kirk Bansak and Elisabeth Paulson. Outcome-driven dynamic refugee assignment with allocation balancing. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pages 1182–1183, 2022.

Kirk Bansak, Jeremy Ferwerda, Jens Hainmueller, Andrea Dillon, Dominik Hangartner, Duncan Lawrence, and Jeremy Weinstein. Improving refugee integration through data-driven algorithmic assignment. *Science*, 359(6373):325–329, 2018.

Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust Optimization*, volume 28. Princeton University Press, 2009.

Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pages 2154–2156, 2018.

Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C Duchi, and Percy S Liang. Unlabeled data improves adversarial robustness. *Advances in Neural Information Processing Systems*, 32, 2019.

Rich Caruana. Multitask learning. *Machine Learning*, 28:41–75, 1997.

O. Chapelle, A. Zien, and B. Scholkopf. *Semi-Supervised Learning*. MIT Press, 2006.

Yining Chen, Colin Wei, Ananya Kumar, and Tengyu Ma. Self-training avoids using spurious features under domain shift. *Advances in Neural Information Processing Systems*, 33:21061–21071, 2020.

Yuansi Chen and Peter Bühlmann. Domain adaptation under structural causal models. *The Journal of Machine Learning Research*, 22(1):11856–11935, 2021.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*, pages 1180–1189, 2015.

Paul Gölz and Ariel D Procaccia. Migration as submodular optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 549–556, 2019.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty. In *International Conference on Machine Learning*, pages 2712–2721, 2019.

P Huber. Robust location of a location parameter. *Annals of Mathematical Statistics*, 35(1):73–101, 1964.

Yujin Jeong and Dominik Rothenhäusler. Calibrated inference: statistical inference that accounts for both sampling uncertainty and distributional uncertainty. *arXiv preprint arXiv:2202.11886*, 2022.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664, 2021.

Ananya Kumar, Tengyu Ma, and Percy Liang. Understanding self-training for gradual domain adaptation. In *International Conference on Machine Learning*, pages 5468–5479, 2020.

Ricardo A Maronna, R Douglas Martin, Victor J Yohai, and Matías Salibián-Barrera. *Robust Statistics: Theory and Methods (with R)*. John Wiley & Sons, 2019.

Jersey Neyman. Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, 10:1–51, 1923.

Ross L Prentice. Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in Medicine*, 8(4):431–440, 1989.

Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John Duchi, and Percy Liang. Understanding and mitigating the tradeoff between robustness and accuracy. *arXiv preprint arXiv:2002.10716*, 2020.

Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *The Journal of Machine Learning Research*, 19(1):1309–1342, 2018.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Dominik Rothenhäusler, Nicolai Meinshausen, Peter Bühlmann, and Jonas Peters. Anchor regression: Heterogeneous data meet causality. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(2):215–246, 2021.

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

Donald B. Rubin. Comment: Randomization analysis of experimental data: The Fisher randomization test. *Journal of the American Statistical Association*, 75(371):591–593, 1980.

Henry Scudder. Probability of error of some adaptive pattern-recognition machines. *IEEE Transactions on Information Theory*, 11(3):363–371, 1965.

Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big data*, 3(1):1–40, 2016.

David H Wolpert. The lack of a priori distinctions between learning algorithms. *Neural Computation*, 8(7):1341–1390, 1996.

Jeremy Yang, Dean Eckles, Paramveer Dhillon, and Sinan Aral. Targeting for long-term outcomes. *arXiv preprint arXiv:2010.15835*, 2020.

A Appendix for Empirical Assessment

A.1 Data

We use data on asylum seekers in the Netherlands from two data sources. The first is the administrative data on individuals granted temporary asylum residence permit ("permit holders") from the Central Agency for the Reception of Asylum Seekers (COA), which is the government agency in charge of the asylum system in the Netherlands. This is a comprehensive data set covering all permit holders in the Netherlands containing their background characteristics, procedural information, and location data. The second is the Asielcohort microdata compiled by Statistics Netherlands (CBS). The Asielcohort microdata is comprised of merged data from a number of administrative databases, several of which track various measures of permit holders in the Netherlands after arrival. These include economic, educational, and other indicators of integration and well-being that can be evaluated as downstream outcomes.

The target population for our assessment is comprised of permit holders who were geographically assigned in the Netherlands through the regular housing procedure. Further, we only consider data and outcomes for adults (i.e. 18 years or older). We also exclude some subsets of permit holders who fall outside the scope of the objectives of algorithmic assignment or for whom data are unreliable.[4] In addition, the target population for algorithmic assignment further excludes family reunifiers, since such permit holders are automatically reunified with their family in the Netherlands.[5] The permit holders in the available data were assigned between January 2014 and August 2018 ($n \approx 46,000$).

As per rules on usage of these data sets, and in accordance with our data access and use agreements with CBS, our access to and analysis of these data is conducted entirely via a secure Remote Access Environment. All statistics and results that are reported from these data are checked and cleared for export by CBS.

A.2 Modeling and Estimation Procedures

The modeling approach is based on the methodology developed in Bansak et al. [2018]. We first merge the historical data for the target population's background characteristics, employment outcomes, and geographic locations. Using supervised learning on these merged data, we fit separate models across each labor market region (LMR) that predict one- or two-year employment using the predictors described in a section below. To do so, for each model we first subset the training data to permit holders who were assigned to a given LMR, and then use this subset for the model training. The next subsection describes the basis upon which LMRs were targeted as the geographic locations of interest.

To generate our models, we employ stochastic gradient boosted trees with squared error loss, which we implement in `R` using the `gbm` package. We employ cross-validation to determine the optimal values for tuning parameters. Specifically, we cross-validate over the number of boosting iterations (trees) and the interaction depth of the trees. We employ 5 folds in our cross-validation, allowing the `gbm` functionality to determine the random splits. For each model, we cross-validate over tree depths of 3-8 and a number of trees that we ensure are sufficient to be able to identify the number that yields the minimum CV mean squared error (i.e. we ensure that we do not consider too few trees to find the minimum CV error). Parameters were tuned independently for each location-specific model. Preliminary assessments on the bag fraction, learning/shrinkage rate, and minimum number of observations per node demonstrated relatively little to no impact on model performance within conventional value ranges, and so these parameters were held fixed at 0.5, 0.01, and 5, respectively. Where appropriate, 95% confidence intervals are generated using a nonparametric bootstrap.

[4] We exclude permit holders who fell under the 2019 Children's Pardon, resettlers / relocants / asylum seekers covered by the EU-Turkey deal, and permit holders who have "hard criteria" that pre-determined their locations. Beyond family reunification constraints, hard criteria may be motivated by other determinants such as employment contracts and medical issues.

[5] Note, however, that data for family reunifiers is included the data used to estimate the various models comprising τ^A and τ^B, with an family reunifier indicator also included as a variable in the models. Family reunifiers were not, however, included in the "test cohort" defined below since they are outside the scope of algorithmic assignment.

A.3 Target Geography

Selecting the appropriate target unit of geography for algorithmic recommendations requires balancing three goals: (1) generating a large number of geographic options to provide the algorithmic recommendation procedure with as much geographic variation as possible, (2) ensuring that each geographic option is associated with a sufficient amount of historical data such that accurate and effective predictive models can be trained, and (3) identifying levels or regions of geography that are administratively compatible with the underlying goals and procedures.

Based on these criteria, and in consultation with our partners at COA, we determined that the best target unit of geography in the Dutch context is the labor market region (LMR). Hence, as noted above, we generate separate predictive models for each of the 35 LMRs, and the algorithmic assignment procedure determines the optimal LMR for each family of permit holders (i.e. the LMR where they have the highest chance of employment success, subject to all the constraints).

A.4 Predictors

Based on data availability, the pre-arrival characteristics we include as the covariates X are age, gender, martial status, prior education, number of family members, country of origin, religion, native language, ethnicity, and prior work experience and industry. In addition, we also include in X several key variables related to the assignment procedure as predictors, including the month of assignment, year of assignment, whether or not someone is a family reunifier, and what type of processing location their housing interview took place at.

A.5 Evidence of proxy relationship and distribution shift

As a simple analysis of the proxy relationship, we fit a linear regression of Y_2 on Y_1 using the entire dataset and find a large R^2 value of 0.625. The R^2 increases to 0.642 when the covariates are added to the model.

To establish the presence of distribution shift, we perform two analyses making use of the data from 2015-06-01 to 2016-05-31 (analogous to period -2 in our impact assessment) and from 2016-06-01 to 2017-05-31 (analogous to period -1). First, we fit classifiers (using gradient boosted trees according to the procedures described above, but with binomial deviance loss in place of squared error loss) that predict which period each data point belongs to as a function of the outcomes and the covariates (described above). Note that for this procedure, we omit the month and year covariates for obvious reasons. We find solid predictive performance with these models: a classification accuracy of 0.768 (relative to 0.547 under a null/intercept model) and an R^2 of 0.344.

Second, we split the period -1 data into training and test sets, and we estimate models (again using gradient boosted trees according to the procedures described above, and with squared error loss) that predict the two-year outcome as a function of the covariates. We fit models using the period -1 training data, and separately also fit models using a period -2 training set that is randomly sampled to have identical size as the period -1 training set. We then compare the performance of these models in predicting onto the period -1 test set, and we find clearly superior performance of the period -1 training models (a test set R^2 of 0.126) over the period -2 training models (a test set R^2 of 0.081). We find similar results when we use these models to predict onto data that is analogous to period 0.

A.6 Application

$\boldsymbol{W}^*$ is estimated by fitting models for expected two-year employment using all of the available data, including the data for the 2018-06-01 – 2018-08-31 test cohort. We use all of the data for this quantity for the sake of counterfactual evaluation; if one were trying to assign the 2018-06-01 – 2018-08-31 in the real world (i.e. if that was the present), it would obviously not be possible to estimate $\boldsymbol{W}^*$. In contrast, $\boldsymbol{W}^A$, $\boldsymbol{W}^B$, and $\boldsymbol{W}^C$ are all estimated in ways that would be possible, allowing them to be used to determine the counterfactual assignment decisions.

$\boldsymbol{W}^A$ is estimated by fitting models for expected two-year employment using the data prior to 2016-06-01 (i.e. two years earlier than the test cohort), equivalent to period -2, and then applying those models to the test cohort. $\boldsymbol{W}^B$ is estimated by fitting models for expected one-year employment using the data prior to 2017-06-01, equivalent to the union of period -1 and -2, and then applying

those models to the test cohort. $\boldsymbol{W}^C$ is estimated via a two-step process. In the first step, models for expected two-year employment are fit using using the data from prior to 2016-06-01, with one-year employment being included as a regressor. Those models are applied to the data between 2016-06-01 and 2017-05-31. With these predicted/expected values in hand, in the second step, another set of models for two-year employment are fit using all data prior to 2017-06-01—which can be done "feasibly" by using the predicted/expected values for two-year employment as the outcome for these models—with only the variables contained in X as the regressors. In addition, the data from prior to 2016-06-01 are also included in the training of these models, with the true/observed values (rather than predicted/expected values) of two-year employment used for those observations. This methodology effectively combines period -1 and -2 for the τ^B estimator and for the second stage of the τ^C estimator, which is a deviation from the assumptions made in Section 4. One could also estimate τ^B and the second stage of τ^C using only period -1 data (between 2016-06-01 and 2017-05-31). In Section A.9 we show a figure analogous to Figure 2 where τ^B and τ^C were estimated using this approach.

A.7 Assignment and Constraints

Once the models corresponding to τ^A, τ^B, and τ^C have been estimated, the models can then be applied to the test cohort to estimate their expected employment success (based on Y_2 or Y_1, respectively for each method) at each of the possible LMRs. Optimal algorithmic assignment decisions on which specific LMR each permit holder should be assigned to can then be made, based on maximizing the expected average employment subject to the constraints. There are two key types of assignment constraints that we take into account for our assessment. The first are family-related constraints. All permit holders in the same family (or associated with the same case number for other reasons) must be assigned together to the same location. The second are location capacity constraints. That is, permit holders must be assigned across labor market regions according to pre-determined capacity and proportionality guidance.

A.8 Results

The mean squared error for method K is computed by taking the average over locations and refugee strata x, in the sense of the following:

$$\frac{1}{|\mathcal{L}|}\sum_{j\in\mathcal{L}}\mathbb{E}_X\left[\mathbb{E}\left[(\tau^K(x,j)-\tau(x,j))^2\right]\right]$$

This is estimated by computing the mean of the squared differences between the entries in the ground truth matrix $\boldsymbol{W}^*$ and the matrix corresponding to the appropriate method.

The MSE for each test is shown in Table A.8.

	τ^A	$\tilde{\tau}^B$	τ^C
Actual	0.010084	0.009388	0.007520
Proxy Violated	0.010084	0.013649	0.009756
More Shift	0.012559	0.009721	0.008650
Proxy Violated + More Shift	0.012407	0.013527	0.012243

Table 1: MSE Results

A.9 Results where old data is dropped for τ^B and τ^C

Figure 2 is analogous to Figure 1 but shows the results where τ^B and the second stage of τ^C are estimated only using data from between 2016-06-01 and 2017-05-31. This method more closely aligns with the theoretical results, but practically turns out to be less desirable because the sample size becomes much smaller. As can be seen by comparing Figure 2 and 1, both τ^B and τ^C perform slightly worse using this methodology.

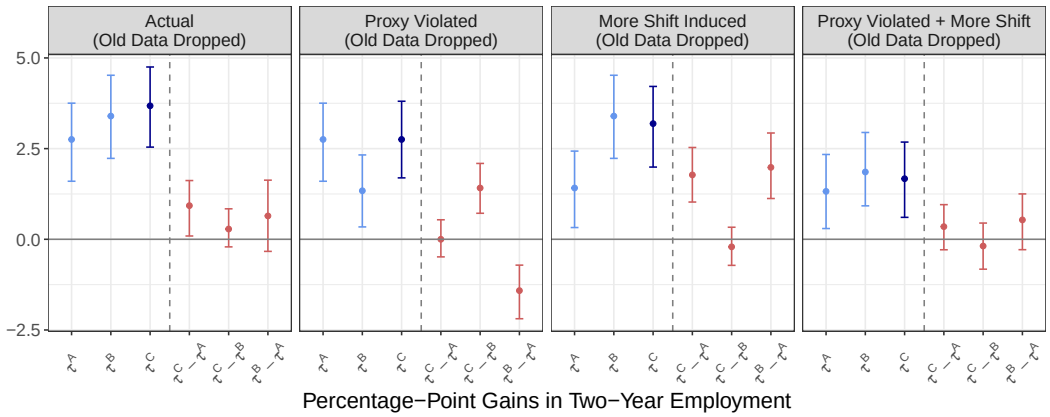

Figure 2: Counterfactual gains in two-year employment (compared to status quo) with actual data (left) and perturbed data using the alternative strategies τ_A, τ_B, and τ_C. Confidence intervals are 95% bootstrapped.

B Extensions

B.1 More than two outcomes

In a more general version of the problem discussed above, we have $T+1$ time periods: $t \in \{0, -1, ..., -T\}$, where period 0 is the present time, and we have T outcomes, Y_1 through Y_T, with Y_T being the policy outcome of interest. The period t data contains outcomes Y_1 through Y_t. There are now *at least* $T + T(T+1)/2$ strategies one could pursue in choosing an outcome and estimation method to guide the algorithmic recommendations. In particular, there are T estimands of the form: $\tau^t(x, j) = \mathbb{E}_{-t}[Y_t|X = x, J = j]$ for $t = 1, ..., T$. There are also $T(T+1)/2$ nested methods of the form:

$$\tau^{t_1-t_2}(x,j) = \mathbb{E}_{-t_1}[\mathbb{E}_{-t_1-1}[...\mathbb{E}_{-t_2+1}[\mathbb{E}_{-t_2}[Y_{t_2}|Y_{t_1}, ..., Y_{t_2-1}, x, j]|Y_{t_1}, ..., Y_{t_2-2}, x, j]...]|x, j]$$

for $t_2 > t_1$. For example, when $T = 3$,

$$\tau^{1-3}(x,j) = \mathbb{E}_{-1}[\mathbb{E}_{-2}[\mathbb{E}_{-3}[Y_3|Y_2, Y_1, x, j]|Y_1, x, j]|x, j].$$

Similar intuition for the relative merits and drawbacks of each method as in Section 4 will hold. In particular, we expect $\tau^{1-T}(x, j)$ to have similar robustness properties as τ^C.

B.2 Other data fusion problems

In addition to extending the modeling framework to an arbitrary number of time periods, one could also extend the framework to non-temporal shifts. In particular, consider a scenario with three datasets, I, II, and III. Our goal is to predict a particular outcome, outcome A, for individuals in dataset I, where no outcome labels are present. Dataset II consists of a population of individuals and an outcome B, related to outcome A. Finally, dataset III contains yet another population for which outcomes A and B are both present. The question is how to use datasets II and III, along with the outcomes they contain, to best predict outcome A for dataset I. In this paper, the random distribution shift model effectively meant that period -1 data is closer to period 0 than period -2 is to period 0. In this generalization, a key difficulty is measuring the "closeness" of datasets I, II, and III. If dataset II is closer to dataset I than dataset III is to I, one could employ an analogous estimator to estimator $\tau^C(x, j)$.

C Finite-sample considerations

In the following we will study how the procedures compare from a asymptotic viewpoint, in the absence of distribution shift (that is, if $P_{-2} = P_{-1} = P_0$). This setting is particularly important for relatively small shifts, since in this case the error due to sample size will dominate the overall

mean-squared error. For mathematical simplicity, we assume that the ratio of units for the two periods converges to a constant, that is $\frac{n_{-1}}{n_{-2}} \to \rho \in (0, \infty)$. We write $n = n_{-1} + n_{-2}$.

In general, it is difficult to compare the three approaches since the performance depends both on the choice of machine learning algorithms and on the data set. In particular, similar to "no free lunch" theorems in machine learning [Wolpert, 1996], we expect that for any choice of algorithm, any of the strategies can perform best depending on the dataset. That being said, in the following, we will see that in the setting where X and Y_1 is discrete, a clear comparison can be drawn. In this case, in low-dimensional settings the estimator can be based on sample proportions. Let $I_{-2,x,j}$ be the set of indices for units from period $t = -2$ that have $X_i = x$, and $J_i = j$.

$$\hat{\tau}^A(x, j) = \frac{1}{\#I_{-2,x,j}} \sum_{i \in I_{-2,x,j}} Y_{2,i}.$$

We now define an estimate of τ^C. To this end, let $I_{-1,x,j}$ denote the set of indices for units from time period $t = -1$ that have $X_i = x$ and $J_i = j$. Similarly, let $I_{-2,x,j,y}$ be the set of indices for units from time period $t = -2$ that have $X_i = x$, $J_i = j$, and $Y_{1,i} = y$. As above, we can define

$$\hat{\tau}^C(x, j) = \frac{1}{\#I_{-1,x,j}} \sum_{i \in I_{-1,x,j}} \hat{Q}(x, j, Y_{1,i}), \quad \text{where } \hat{Q}(x, j, y) = \frac{1}{\#I_{-2,x,j,y}} \sum_{i \in I_{-2,x,j,y}} Y_{2,i}.$$

In the following, we introduce an estimator for τ^B. Analogously to the above, define

$$\hat{\tau}^B = \frac{1}{\#I_{-2,x,j}} \sum_{i \in I_{-2,x,j}} Y_{1,i}.$$

Let us now compare the asymptotic performance of the three approaches. The estimator $\hat{\tau}^B$ is justified if the relationship between the target Y_2 and the proxy outcome is linear, that is if $\mathbb{E}_{-2}[Y_2|X = x, J = j] = \beta_1 \mathbb{E}_{-2}[Y_1|X = x, J = j] + \beta_0$. This is a very strong assumption. In particular, it assumes that the relationship does not change for different values of $X = x$ and $J = j$. The proof of the following result can be found in Section D.

Example 3 (Invariant distribution; proxy assumption does not hold)**.** *Assume that there is no shift, i.e. that $P_{-2}(\cdot|x, j) = P_{-1}(\cdot|x, j) = P_0(\cdot|x, j)$. Then,*

$$\sqrt{n}(\hat{\tau}^A(x, j) - \tau(x, j)) \rightharpoonup \mathcal{N}(0, \sigma_A^2)$$
$$\sqrt{n}(\hat{\tau}^C(x, j) - \tau(x, j)) \rightharpoonup \mathcal{N}(0, \sigma_C^2),$$

where $\sigma_A^2 < \sigma_C^2$ if and only if $\rho < 1$. Furthermore, if the linear proxy assumption does not hold, we have

$$\beta_1 \hat{\tau}^B(x, j) + \beta_0 - \tau(x, j) \not\to_P 0.$$

Please note that we expect $n_{-2} > n_{-1}$, since for P_{-2} we can pool many observations from previous timepoints. Thus, if there is not distribution shift and the proxy assumption is violated, one would generally prefer $\hat{\tau}^A$. If the linear proxy assumption holds and $n_{-1} > n_{-2}$, then the conclusions change. This will be discussed in the following.

The following example shows that in this case τ^B is preferable over the other procedures. The proof can be found in Section D.

Example 4 (Invariant distribution; proxy assumption holds)**.** *Assume that there is no shift, that is $P_{-2}(\cdot|x, j) = P_{-1}(\cdot|x, j) = P_0(\cdot|x, j)$ and that $\rho > 1$. If $\mathbb{E}_{-2}[Y_2|Y_1, X] = \beta_1 Y_1 + \beta_0$, then*

$$\sqrt{n}(\hat{\tau}^A(x, j) - \tau(x, j)) \rightharpoonup \mathcal{N}(0, \sigma_A^2)$$
$$\sqrt{n}(\hat{\tau}^C(x, j) - \tau(x, j)) \rightharpoonup \mathcal{N}(0, \sigma_C^2)$$
$$\sqrt{n}(\hat{\tau}^B(x, j) - \tau(x, j)) \rightharpoonup \mathcal{N}(0, \sigma_B^2).$$

Furthermore,

$$\sigma_B^2 < \min(\sigma_A^2, \sigma_C^2).$$

D Proofs

D.1 Proof of Theorem 1

Proof. For notational simplicity, without loss of generality, we will assume that Y_1 is already transformed, that is $Y_1 = \tilde{Y}_1$. First, let us prove a few auxiliary results. By assumption, for fixed $y_2, y_1, y_2', y_1', x, j, x', j'$ we have

$$\mathbb{E}[S_t(y_2, y_1|x, j)] = 0 \text{ and } \text{Cov}(S_{-2}(y_2, y_1|x, j), S_{-1}(y_2', y_1'|x', j')) = 0, \tag{4}$$

for $t \in \{-2, -1\}$. For every function f and fixed x, j we have

$$\begin{aligned} &\text{Var}(\sum_{y_1, y_2} f(y_1, y_2, x, j) S_{-2}(y_1, y_2|x, j)) \\ &= \sum_{y_1, y_2} \sum_{y_1', y_2'} f(y_1, y_2, x, j) f(y_1', y_2', x, j) \text{Cov}(S_{-2}(y_1, y_2|x, j), S_{-2}(y_1', y_2'|x, j)) \end{aligned} \tag{5}$$

Here, the variance is over the randomness in the shift S_{-2}. Now if $y_1 = y_1'$ and $y_2 = y_2'$, then by equation (1),

$$\begin{aligned} &\text{Cov}(S_{-2}(y_1, y_2, |x, j), S_{-2}(y_1', y_2'|x, j)) \\ &= \text{Var}(S_{-2}(y_1, y_2|x, j)) \\ &= \kappa_{-2} P_{-2}(y_1, y_2|x, j)(1 - P_{-2}(y_1, y_2|x, j)). \end{aligned}$$

If $y_1 \neq y_1'$ or $y_2 \neq y_2'$, let A be the event that $(y_1', y_2') = (Y_1, Y_2)$ or $(y_1, y_2) = (Y_1, Y_2)$. Then we can use additivity, that is $P_{-1}(A|X = x, J = j) = P_{-1}((y_1', y_2') = (Y_1, Y_2)|X = x, J = j) + P_{-1}((y_1, y_2) = (Y_1, Y_2)|X = x, J = j)$ to obtain

$$\begin{aligned} &\text{Var}(P_{-1}(A|X = x, J = j)) \\ &= \text{Var}(S_{-2}(y_1', y_2'|x, j)) + 2\text{Cov}(S_{-2}(y_1, y_2|x, j), S_{-2}(y_1', y_2'|x, j)) + \text{Var}(S_{-2}(y_1, y_2|x, j)) \end{aligned}$$

Applying equation (1) directly on both sides, we get

$$\begin{aligned} &\kappa_{-2} P_{-2}(A|x, j)(1 - P_{-2}(A|x, j)) \\ &= \kappa_{-2} P_{-2}(y_1, y_2|x, j)(1 - P_{-2}(y_1, y_2|x, j)) + 2\text{Cov}(S_{-2}(y_1, y_2|x, j), S_{-2}(y_1', y_2'|x, j)) \\ &+ \kappa_{-2} P_{-2}(y_1', y_2'|x, j)(1 - P_{-2}(y_1', y_2'|x, j)). \end{aligned}$$

Applying additivity to the left-hand side, we get

$$\begin{aligned} &\kappa_{-2}(P_{-2}(y_1, y_2|x, j) + P_{-2}(y_1', y_2'|x, j))(1 - P_{-2}(y_1, y_2|x, j) - P_{-2}(y_1', y_2'|x, j)) \\ &= \kappa_{-2} P_{-2}(y_1, y_2|x, j)(1 - P_{-2}(y_1, y_2|x, j)) + 2\text{Cov}(S_{-2}(y_1, y_2|x, j), S_{-2}(y_1', y_2'|x, j)) \\ &+ \kappa_{-2} P_{-2}(y_1', y_2'|x, j)(1 - P_{-2}(y_1', y_2'|x, j)). \end{aligned}$$

Simplifying, we get

$$-2\kappa_{-2} P_{-2}(y_1', y_2'|x, j) P_{-2}(y_1, y_2|x, j) = 2\text{Cov}(S_{-2}(y_1, y_2|x, j), S_{-2}(y_1', y_2'|x, j)).$$

Thus,

$$\text{Cov}(S_{-2}(y_1, y_2|x, j), S_{-2}(y_1', y_2'|x, j)) = -\kappa_{-2} P_{-2}(y_1, y_2|x, j) P_{-2}(y_1', y_2'|x, j).$$

To summarize, we have

$$\begin{aligned} &\text{Cov}(S_{-2}(y_1, y_2, |x, j), S_{-2}(y_1', y_2'|x, j)) \\ &= \begin{cases} \kappa_{-2} P_{-2}(y_1, y_2|x, j)(1 - P_{-2}(y_1, y_2|x, j)) & \text{if } (y_1, y_2) = (y_1', y_2'), \\ -\kappa_{-2} P_{-2}(y_1, y_2|x, j) P_{-2}(y_1', y_2'|x, j) & \text{if } (y_1, y_2) \neq (y_1', y_2'). \end{cases} \end{aligned}$$

This allows us to re-write equation (5) as

$$\begin{aligned} &\text{Var}(\sum_{y_1, y_2} f(y_1, y_2, x, j) S_{-2}(y_1, y_2|X = x, J = j)) \\ &= \kappa_{-2}(\sum_{y_1, y_2} f(y_1, y_2, x, j)^2 P_{-2}(y_1, y_2|x, j) - (\sum_{y_1, y_2} f(y_1, y_2, x, j) P_{-2}(y_1, y_2|x, j))^2) \\ &= \kappa_{-2} \text{Var}_{-2}(f(Y_1, Y_2, X, J)|X = x, J = j). \end{aligned}$$

On the left-hand side, the variance is over the randomness in the shift S_t, where on the right-hand side the variance is computed under P_{-2}. Since S_{-2} has mean zero, for any function f with $\mathbb{E}_{-2}[f(Y_1, Y_2, X, J)|X = x, J = j] = 0$ we have

$$\begin{aligned}&\mathbb{E}[(\sum_{y_1,y_2} f(y_1, y_2, x, j)S_{-2}(y_1, y_2|x, j))^2]\\ &= \kappa_{-2}\text{Var}_{-2}(f(Y_1, Y_2, X, J)|X = x, J = j)\\ &= \kappa_{-2}\mathbb{E}_{-2}[f(Y_1, Y_2, X, J)^2|X = x, J = j].\end{aligned} \tag{6}$$

By an analogous argument, for any function f with $\mathbb{E}_{-2}[f(Y_1, Y_2, X, J)|X = x, J = j] = 0$ we have

$$\begin{aligned}&\mathbb{E}[(\sum_{y_1,y_2} f(y_1, y_2, x, j)S_{-1}(y_1, y_2|x, j))^2]\\ &= \kappa_{-1}(1 - \kappa_{-2})\mathbb{E}_{-2}[f(Y_1, Y_2, X, J)^2|X = x, J = j].\end{aligned} \tag{7}$$

(Notice that κ_{-2} is naturally upper bounded by 1: For an event with $P_t(\bullet|x, j) = 1/2$, the right-hand side of equation (1) is $\kappa_t/4$. The left-hand side of (1) is upper bounded by $1/4$, since $P_t + S_t$ is a $[0, 1]$-valued random variable, and the maximum variance of a $[0, 1]$-valued random variable is achieved by $\text{Ber}(1/2)$. Thus, equation (1) implies $\frac{1}{4} \geq \frac{\kappa_t}{4}$. Thus, we have $0 \leq \kappa_t \leq 1$.)

We now have all auxiliary results in place to investigate the MSE of τ^C. For every fixed x, j we have

$$\begin{aligned}\tau(x, j) - \tau^C(x, j) &= \sum_{y_2,y_1} y_2(P_{-2}(y_1, y_2|x, j) + S_{-1}(y_1, y_2|x, j) + S_{-2}(y_1, y_2|x, j))\\ &\quad - \sum_{y_2,y_1} \mathbb{E}_{-2}[Y_2|Y_1 = y_1, X = x, J = j](P_{-2}(y_1, y_2|x, j) + S_{-2}(y_1, y_2|x, j))\\ &= \sum_{y_2,y_1} y_2(S_{-1}(y_1, y_2|x, j) + S_{-2}(y_1, y_2|x, j))\\ &\quad - \sum_{y_2,y_1} \mathbb{E}_{-2}[Y_2|Y_1 = y_1, X = x, J = j]S_{-2}(y_1, y_2|x, j)]\\ &= \sum_{y_2,y_1} (y_2 - \mathbb{E}_{-2}[Y_2|Y_1 = y_1, X = x, J = j])S_{-2}(y_1, y_2|x, j)\\ &\quad + \sum_{y_2,y_1} y_2 S_{-1}(y_1, y_2|x, j)\end{aligned}$$

The first inequality follows by definition. The second inequality follows from the fact that $\sum_{y_2,y_1} y_2 P_{-2}(y_1, y_2|x, j) = \mathbb{E}_{-2}[Y_2|X = x, J = j]$ and $\sum_{y_2,y_1} \mathbb{E}_{-2}[Y_2|Y_1 = y_1, X = x, J = j]P_{-2}(y_1, y_2|x, j) = \mathbb{E}_{-2}[\mathbb{E}_{-2}[Y_2|Y_1, X = x, J = j]|X = x, J = j] = \mathbb{E}_{-2}[Y_2|X = x, J = j]$ and thus these terms cancel out. The third inequality follows by rearranging terms.

Since $\sum_{y_2,y_1} S_{-1}(y_1, y_2|x, j) = 0$, we have $\sum_{y_2,y_1} \mathbb{E}_{-1}[Y_2|X = x, J = j]S_{-1}(y_1, y_2|x, j) = 0$. Combining this with the equation above,

$$\begin{aligned}&\tau(x, j) - \tau^C(x, j)\\ &= \sum_{y_2,y_1} (y_2 - \mathbb{E}_{-2}[Y_2|Y_1 = y_1, X = x, J = j])S_{-2}(y_1, y_2|x, j)\\ &\quad + \sum_{y_2,y_1} (y_2 - \mathbb{E}_{-2}[Y_2|X = x, J = j])S_{-1}(y_1, y_2|x, j).\end{aligned}$$

Squaring and taking expectations, using equation (4), equation (7), and equation (6) yields

$$\begin{aligned}\mathbb{E}[(\tau(x, j) - \tau^C(x, j))^2|X = x, J = j] &= \kappa_{-2}\mathbb{E}_{-2}[(Y_2 - \mathbb{E}_{-2}[Y_2|Y_1, X, J])^2|X = x, J = j]\\ &\quad + \kappa_{-1}\mathbb{E}_{-2}[(Y_2 - \mathbb{E}_{-2}[Y_2|X, J])|X = x, J = j] + O(\kappa_{-1}\kappa_{-2})\end{aligned}$$

Now taking the expectation over X and J yields

$$\begin{aligned}\mathbb{E}[(\tau(x, j) - \tau^C(x, j))^2] &= \kappa_{-2}\mathbb{E}_{-2}[(Y_2 - \mathbb{E}_{-2}[Y_2|Y_1, X, J])^2]\\ &\quad + \kappa_{-1}\mathbb{E}_{-2}[(Y_2 - \mathbb{E}_{-2}[Y_2|X, J])^2] + O(\kappa_{-1}\kappa_{-2}).\end{aligned}$$

This completes the first claim. Similarly,

$$\begin{aligned}\tau(x,j) - \tau^A(x,j) &= \sum_{y_2,y_1} y_2(P_{-2}(y_2,y_1|x,j) + S_{-1}(y_2,y_1|x,j) + S_{-2}(y_2,y_1|x,j)) \\ &\quad - \sum_{y_2,y_1} y_2 P_{-2}(y_2,y_1|x,j) \\ &= \sum_{y_2,y_1} y_2 S_{-1}(y_2,y_1|x,j) + S_{-2}(y_2,y_1|x,j)) \\ &= \sum_{y_2,y_1} (y_2 - \mathbb{E}_{-2}[Y_2|X=x,J=j])(S_{-1}(y_2,y_1|x,j) + S_{-2}(y_2,y_1|x,j))\end{aligned}$$

Analogously as above, squaring and taking expectations (using equation equation (4), equation (7), and equation (6)) yields the claim. Lastly,

$$\begin{aligned}\tau(x,j) - \tau^B(x,j) &= \mathbb{E}_{-2}[Y_2|X=x,J=j] \\ &\quad + \sum_{y_2,y_1} (y_2 - \mathbb{E}_{-2}[Y_2|X=x,J=j])(S_{-1}(y_1,y_2|x,j) + S_{-2}(y_1,y_2|x,j)) \\ &\quad - \mathbb{E}_{-2}[Y_1|X=x,J=j] \\ &\quad - \sum_{y_2,y_1} (y_1 - \mathbb{E}[Y_1|X=x,J=j]) S_{-2}(y_1,y_2|x,j) \\ &= \mathbb{E}_{-2}[Y_2|X=x,J=j] - \mathbb{E}_{-2}[Y_1|X=x,J=j] \\ &\quad + \sum_{y_2,y_1} (y_2 - y_1 - \mathbb{E}_{-2}[Y_2 - Y_1|X=x,J=j]) S_{-2}(y_1,y_2|x,j) \\ &\quad + \sum_{y_2,y_1} (y_2 - \mathbb{E}_{-2}[Y_2|X=x,J=j]) S_{-1}(y_1,y_2|x,j).\end{aligned}$$

As before, squaring and taking expectations yields the claim. □

D.2 Proof of Example 3

Proof. For the second statement, please note that due to the law of large numbers $\hat{\tau}^B(x,j) \to_P \mathbb{E}_{-2}[Y_1|X=x,J=j]$. Since by assumption $\mathbb{E}_{-2}[Y_2|X=x,J=j] \not\equiv \beta_1 \mathbb{E}_{-2}[Y_1|X=x,J=j] + \beta_0$, $\beta_1 \hat{\tau}^B(x,j) + \beta_0 - \tau(x,j) \not\to_P 0$. This proves the second statement.

Let I_{-2} denote the set of indices from time period -2 and I_{-1} denote the set of indices from time period -1. A Taylor expansion reveals that

$$\begin{aligned}\hat{\tau}^A(x,j) - \tau(x,j) &= \frac{1}{\#I_{-2,x,j}} \sum_{i \in I_{-2,x,j}} Y_{2,i} - \mathbb{E}_{-2}[Y_2|X=x,J=j] \\ &= \frac{1}{n_{-2}} \sum_{i \in I_{-2}} \frac{1}{\frac{\#I_{-2,x,j}}{n_{-2}}} 1_{X_i=x,J_i=j}(Y_{2,i} - \mathbb{E}_{-2}[Y_2|X=x,J=j]) \\ &= \frac{1}{n_{-2}} \sum_{i \in I_{-2}} \frac{1_{X_i=x,J_i=j}}{\mathbb{P}_{-2}(X=x,J=j)}(Y_{2,i} - \mathbb{E}_{-2}[Y_2|X=x,J=j]) + o_P(1/\sqrt{n}).\end{aligned}$$

Thus,

$$\sqrt{n_{-2}}(\hat{\tau}^A(x,j) - \tau(x,j)) \rightharpoonup \mathcal{N}(0, \frac{1}{P_{-2}(X=x,J=j)} \text{Var}_{-2}(Y_2|X=x,J=j))$$

As $n_{-1}/n_{-2} \to \rho$ and $n = n_{-1} + n_{-2}$,

$$\sqrt{n}(\hat{\tau}^A(x,j) - \tau(x,j)) \rightharpoonup \mathcal{N}(0, (1+\rho)\frac{1}{P_{-2}(X=x,J=j)} \text{Var}_{-2}(Y_2|X=x,J=j)).$$

For τ^C the proof proceeds analogously but is a bit more technical:

$$
\begin{aligned}
&\hat{\tau}^C(x,j) - \tau(x,j)) \\
&= \frac{1}{\#I_{-1,x,j}} \sum_{i \in I_{-1,x,j}} \frac{1}{\#I_{-2,x,j,Y_{1,i}}} \sum_{i' \in I_{-2,x,j,Y_{1,i}}} Y_{2,i'} - \mathbb{E}_{-2}[Y_2 | X = x, J = j] \\
&= \frac{1}{n_{-1}} \sum_{i \in I_{-1}} \frac{1_{X_i = x, J_i = j}}{\#I_{-1,x,j}/n_{-1}} \frac{1}{n_{-2}} \sum_{i' \in I_{-2}} \frac{1_{X_{i'} = x, J_{i'} = j, Y_{1,i'} = Y_{1,i}}}{\#I_{-2,x,j,Y_{1,i}}/n_{-2}} (Y_{2,i'} - \mathbb{E}_{-2}[Y_2 | X = x, J = j]) \\
&= \frac{1}{n_{-1}} \sum_{i \in I_{-1}} \frac{1_{X_i = x, J_i = j}}{\#I_{-1,x,j}/n_{-1}} \frac{1}{n_{-2}} \sum_{i' \in I_{-2}} \frac{1_{X_{i'} = x, J_{i'} = j, Y_{1,i'} = Y_{1,i}}}{\#I_{-2,x,j,Y_{1,i}}/n_{-2}} \\
&\qquad \cdot (\mathbb{E}_{-2}[Y_2 | Y_1 = Y_{1,i}, X = x, J = j] - \mathbb{E}_{-2}[Y_2 | X = x, J = j]) \\
&+ \frac{1}{n_{-1}} \sum_{i \in I_{-1}} \frac{1_{X_i = x, J_i = j}}{\#I_{-1,x,j}/n_{-1}} \frac{1}{n_{-2}} \sum_{i' \in I_{-2}} \frac{1_{X_{i'} = x, J_{i'} = j, Y_{1,i'} = Y_{1,i}}}{\#I_{-2,x,j,Y_{1,i}}/n_{-2}} (Y_{2,i'} - \mathbb{E}_{-2}[Y_2 | Y_1 = Y_{1,i}, X = x, J = j])
\end{aligned}
$$

The first part can be drastically simplified and we can add an additional sum over y to decouple the second term:

$$
\begin{aligned}
&\hat{\tau}^C(x,j) - \tau(x,j)) \\
&= \frac{1}{n_{-1}} \sum_{i \in I_{-1}} \frac{1_{X_i = x, J_i = j}}{\#I_{-1,x,j}/n_{-1}} (\mathbb{E}_{-2}[Y_2 | Y_1 = Y_{1,i}, X = x, J = j] - \mathbb{E}_{-2}[Y_2 | X = x, J = j]) \\
&+ \frac{1}{n_{-1}} \sum_{i \in I_{-1}} \frac{1_{X_i = x, J_i = j}}{\#I_{-1,x,j}/n_{-1}} \frac{1}{n_{-2}} \sum_{i' \in I_{-2}} \frac{1_{X_{i'} = x, J_{i'} = j, Y_{1,i'} = Y_{1,i}}}{\#I_{-2,x,j,Y_{1,i}}/n_{-2}} (Y_{2,i'} - \mathbb{E}_{-2}[Y_2 | Y_1 = Y_{1,i}, X = x, J = j]) \\
&= \frac{1}{n_{-1}} \sum_{i \in I_{-1}} \frac{1_{X_i = x, J_i = j}}{\#I_{-1,x,j}/n_{-1}} (\mathbb{E}_{-2}[Y_2 | Y_1 = Y_{1,i}, X = x, J = j] - \mathbb{E}_{-2}[Y_2 | X = x, J = j]) \\
&+ \sum_y \frac{1}{n_{-1}} \sum_{i \in I_{-1}} \frac{1_{X_i = x, J_i = j, y_i = y}}{\#I_{-1,x,j}/n_{-1}} \frac{1}{n_{-2}} \sum_{i' \in I_{-2}} \frac{1_{X_{i'} = x, J_{i'} = j, Y_{1,i'} = y}}{\#I_{-2,x,j,y}/n_{-2}} (Y_{2,i'} - \mathbb{E}_{-2}[Y_2 | Y_1 = y, X = x, J = j])
\end{aligned}
$$

Now by the law of large numbers,

$$
\begin{aligned}
&\hat{\tau}^C(x,j) - \tau(x,j)) \\
&= \frac{1}{n_{-1}} \sum_{i \in I_{-1}} \frac{1_{X_i = x, J_i = j}}{P_{-2}(X = x, J = j)} (\mathbb{E}_{-2}[Y_2 | Y_1 = Y_{1,i}, X = x, J = j] - \mathbb{E}_{-2}[Y_2 | X = x, J = j]) \\
&+ \sum_y \frac{P_{-2}(X = x, J = j, Y_1 = y)}{P_{-2}(X = x, J = j)} \frac{1}{n_{-2}} \sum_{i' \in I_{-2}} \frac{1_{X_{i'} = x, J_{i'} = j, Y_{1,i'} = y}}{P_{-2}(X = x, J = j, Y_1 = y)} (Y_{2,i} - \mathbb{E}_{-2}[Y_2 | Y_1 = y, X = x, J = j]) \\
&+ o_P(1/\sqrt{n}) \\
&= \frac{1}{n_{-1}} \sum_{i \in I_{-1}} \frac{1_{X_i = x, J_i = j}}{P_{-2}(X = x, J = j)} (\mathbb{E}_{-2}[Y_2 | Y_1 = Y_{1,i}, X = x, J = j] - \mathbb{E}_{-2}[Y_2 | X = x, J = j]) \\
&\frac{1}{n_{-2}} \sum_{i' \in I_{-2}} \sum_y \frac{1_{X_{i'} = x, J_{i'} = j, Y_{1,i'} = y}}{P_{-2}(X = x, J = j)} (Y_{2,i} - \mathbb{E}_{-2}[Y_2 | Y_1 = y, X = x, J = j]) \\
&+ o_P(1/\sqrt{n}) \\
&= \frac{1}{n_{-1}} \sum_{i \in I_{-1}} \frac{1_{X_i = x, J_i = j}}{P_{-2}(X = x, J = j)} (\mathbb{E}_{-2}[Y_2 | Y_1 = Y_{1,i}, X = x, J = j] - \mathbb{E}_{-2}[Y_2 | X = x, J = j]) \\
&\frac{1}{n_{-2}} \sum_{i' \in I_{-2}} \frac{1_{X_{i'} = x, J_{i'} = j}}{P_{-2}(X = x, J = j)} (Y_{2,i'} - \mathbb{E}_{-2}[Y_2 | Y_1 = Y_{1,i'}, X = x, J = j]) \\
&+ o_P(1/\sqrt{n})
\end{aligned}
$$

Thus, by the CLT, with $\frac{n_{-1}}{n_{-2}} \to \rho$ we have

$$\sqrt{n}(\hat{\tau}^C(x,j) - \tau(x,j)) \rightharpoonup \mathcal{N}(0, \sigma_C^2),$$

where

$$\begin{aligned}\sigma_C^2 = &\frac{1+\rho}{\rho}\frac{1}{P_{-2}(X=x,J=j)}\text{Var}_{-2}(\mathbb{E}_{-2}[Y_2|Y_1,X=x,J=j]|X=x,J=j)\\ &+(1+\rho)\frac{1}{P_{-2}(X=x,J=j)}\text{Var}_{-2}(Y_2-\mathbb{E}_{-2}[Y_2|Y_1,X=x,J=j]|X=x,J=j).\end{aligned}$$

Thus, $\tau_A^2 < \tau_C^2$ if and only if $\rho < 1$.

□

D.3 Proof of Example 4

Proof. Please note that following the proof of Example 3, the estimators $\hat{\tau}^A(x,j)$, $\hat{\tau}^B_{\text{scaled}}(x,j) := \beta_1\tau^B(x,j) + \beta_0$ and $\hat{\tau}^C(x,j)$ are all asymptotically unbiased. Thus, in the following we will compare their respective asymptotic variances. For $\hat{\tau}^B_{\text{scaled}}(x,j)$ we have asymptotic variance

$$\begin{aligned}&\frac{1+\rho}{\rho P_{-2}(X=x,J=j)}\text{Var}(\beta_1 Y_1+\beta_0|X=x,J=j)\\ &=\frac{1+\rho}{\rho P_{-2}(X=x,J=j)}\text{Var}(\mathbb{E}_{-2}[Y_2|Y_1,X=x,J=j]|X=x,J=j)\end{aligned}$$

From the proof of Example 3 we know that the asymptotic variance of $\hat{\tau}^A(x,j)$ is

$$\begin{aligned}&\frac{1+\rho}{P_{-2}(X=x,J=j)}\text{Var}(Y_2-\mathbb{E}_{-2}[Y_2|Y_1,X=x,J=j]|X=x,J=j)\\ &+\frac{1+\rho}{P_{-2}(X=x,J=j)}\text{Var}(\mathbb{E}_{-2}[Y_2|Y_1,X=x,J=j]|X=x,J=j).\end{aligned}$$

Similarly, for $\hat{\tau}^C$ we know that the asymptotic variance is

$$\begin{aligned}&(1+\rho)\frac{1}{P_{-2}(X=x,J=j)}\text{Var}(Y_2-\mathbb{E}_{-2}[Y_2|Y_1,X=x,J=j]|X=x,J=j)\\ &+\frac{1+\rho}{\rho}\frac{1}{P_{-2}(X=x,J=j)}\text{Var}(\mathbb{E}_{-2}[Y_2|Y_1,X=x,J=j]|X=x,J=j).\end{aligned}$$

Since $\rho > 1$, the claim follows. □