Unfairness Detection within Power Systems through Transfer Counterfactual Learning

Song Wei[†], **Xiangrui Kong**[‡], **Sarah A Huestis-Mitchell**[†],
song.wei@gatech.edu xiangruk@andrew.cmu.edu shuestis3@gatech.edu

Shixiang Zhu[‡], **Yao Xie**[†], **Alinson Santos Xavier**[⋆], **Feng Qiu**[⋆]
shixianz@andrew.cmu.edu yao.xie@isye.gatech.edu axavier@anl.gov fqiu@anl.gov

[†]Georgia Institute of Technology, [‡]Carnegie Mellon University, [⋆]Argonne National Laboratory.

Abstract

Energy justice is a growing area of interest in interdisciplinary energy research. However, identifying systematic biases in the energy sector remains challenging due to confounding variables, intricate heterogeneity in treatment effects, and limited data availability. To address these challenges, we introduce a novel approach for counterfactual causal analysis centered on energy justice. We use subgroup analysis to manage diverse factors and leverage the idea of transfer learning to mitigate data scarcity in each subgroup. In our numerical analysis, we apply our method to a large-scale customer-level power outage data set and investigate the counterfactual effect of demographic factors, such as income and age of the population, on power outage durations. Our results indicate that low-income and elderly-populated areas consistently experience longer power outages, regardless of weather conditions. This points to existing biases in the power system and highlights the need for focused improvements in areas with economic challenges.

1 Introduction

Energy justice, an emerging concept in the energy research [20, 12, 10], refers to the equitable distribution of energy's benefits and burdens across society[1]. One of the major objectives in advancing energy justice is to detect and rectify systematic biases in the energy sector. These biases, which might be inherent in policy-making, infrastructure planning, and resource allocation, can intensify discrepancies in energy accessibility and cost for various societal groups.

However, identifying such biases is impeded by three major challenges. Firstly, confounding variables present a significant obstacle. These are external factors that can influence both the treatment (*e.g.*, policy decisions) and the effect (*e.g.*, service reliability), making it difficult to establish clear causal relationships. Secondly, there is the challenge of intricate heterogeneity in treatment effects. This means that the impact of a particular policy or infrastructure change might vary widely across different regions or populations, making it challenging to generalize findings or draw overarching conclusions. Lastly, limited data availability further complicates matters. In the energy section, granular data, in many cases, is either not collected, not shared, or is inaccessible due to various reasons, ranging from proprietary concerns to logistical challenges. This data scarcity can lead to incomplete or skewed analyses, potentially overlooking critical nuances or biases.

[1] https://www.ncsl.org/energy/energy-justice-and-the-energy-transition

NeurIPS 2023 Workshop on Causal Representation Learning.

To tackle the above challenges, we propose a novel approach for counterfactual causal analysis, termed Transfer Counterfactual Learning (TCL). We employ this method to study the counterfactual impact of certain demographic factors on power outage durations [29]. We conduct subgroup analysis to mitigate the counter the effects of model misspecification stemming from heterogeneous treatment effects across distinct subgroups. Additionally, we incorporate transfer learning [25] within the inverse probability weighting (IPW) [11] estimation to rectify inherent model biases and enhance the precision of causal effect estimations. The numerical experiments are conducted on large-scale customer level power outage data, collected from a northeastern region of the United States in March 2018. [8, 22, 21]. The results suggest that low-income and elderly-populated areas consistently experience longer power outages, regardless of weather conditions. This points to existing biases in the power system and highlights the need for focused improvements in areas with economic challenges.

Literature. While there has been increased interest in applying data-integrative TL techniques to causal inference in the presence of heterogeneous covariate spaces [33, 32, 9, 4], these methods typically fail to handle the same covariate space setting, known as the inductive multi-task transfer learning according to [25]. This limitation arises from their algorithm designs, which mostly rely on domain-specific covariate spaces. To the best of our knowledge, the first and only work studying data-integrative TL for causal effect estimation under the inductive multi-task setting is [13]. They proposed to transfer knowledge by using neural network (NN) weights estimated from the source domain as the warm start of the subsequent target domain NN training. Despite its improved empirical performance, the theoretically grounded approach is still largely missing, and such a NN-based TL approach fails to generalize to parametric methods such as (generalized) linear models.

2 Methodology

In this section, we will introduce the connection between energy justice and counterfactual causal analysis. In particular, we will briefly review the causal inference problem setup under Rubin's Causal Model [29], identify the challenges, and introduce our proposed transfer counterfactual learning.

2.1 Energy justice through causal inference

We study energy justice by asking "what if" questions, such as "what would the duration of the power outage have been if this family were a rich one". We answer those questions by studying the causal effect from treatment variable Z, such as wealthiness, density, and so on, to outcome variable Y, indicating the power outage duration. We can then detect unfairness if we can conclude from observational data that being in a wealthier neighborhood does reduce the duration of power outages.

Formally, consider the tuple $(\boldsymbol{X}, Z, Y)$, where random vector $\boldsymbol{X} \in \mathcal{X} \subset \mathbb{R}^d$ denotes pre-treatment covariates (such as demographic information), random variable (r.v.) $Z \in \{0, 1\}$ is the indicator of the treatment (for example, $Z = 1$ if the median household income is greater than a pre-defined threshold and 0 otherwise) and r.v. Y is the outcome, i.e., the power outage duration measurement. Under Rubin's Causal Model, $Y = Y_{\text{trt}} Z + (1 - Z) Y_{\text{ctrl}}$ is referred to as the *observed outcome*, whereas Y_{ctrl} and Y_{trt} are called *potential outcomes* — they are the values of the outcome that would be seen if the subject were to receive control or treatment. One commonly considered causal estimand is the average treatment effect (ATE), which is formally defined as:

$$\text{Average Treatment Effect: } \tau = \mathbb{E}[Y_{\text{trt}}] - \mathbb{E}[Y_{\text{ctrl}}].$$

Challenge 1: selection bias. "When observations in social research are selected so that they are not independent of the outcome variables in a study, sample selection leads to biased inferences about social processes" [31]. To help understand the selection bias, we present a graphical illustration in Figure 1. Unlike experimental studies where we can manually force every subject to receive treatment or randomly assign treatments, it is impossible to manipulate the household income in our observational study. As a result, the outcomes in the selected (or observed) treatment cohort may be influenced by other pre-treatment covariates instead of the treatment itself. Please refer to Appendix A.1 for how to use inverse probability weighting to mitigate the selection bias.

Challenge 2: heterogeneity. One focus of this work is the (potential) heterogeneity, as the treatment assignment rule and/or the true causal effect could vary among different subgroups within a population

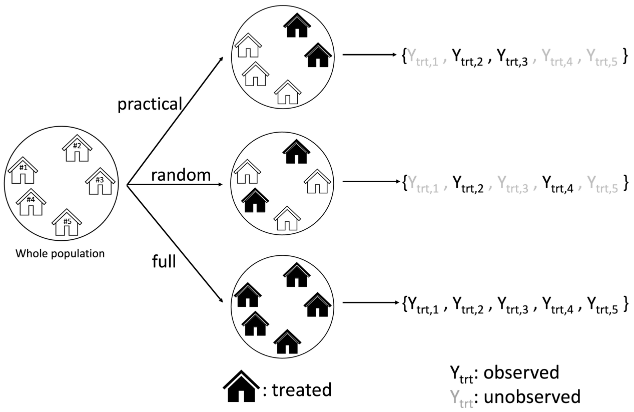
practical
random
full
Whole population
#1
#2
#3
#4
#5
$\{Y_{trt,1}, Y_{trt,2}, Y_{trt,3}, Y_{trt,4}, Y_{trt,5}\}$
$\{Y_{trt,1}, Y_{trt,2}, Y_{trt,3}, Y_{trt,4}, Y_{trt,5}\}$
$\{Y_{trt,1}, Y_{trt,2}, Y_{trt,3}, Y_{trt,4}, Y_{trt,5}\}$
: treated
Y_{trt}: observed
Y_{trt}: unobserved

Figure 1: Illustration of selection bias: In practice (top), the treatment assignment is typically dependent on pre-treatment covariates $\boldsymbol{X}$, making the selected (or observed) treatment cohort NOT independent of the outcome variable. As a result, the selected cohort is not "representative" of the whole population, and inference based on such a selected cohort will typically be biased.

due to diverse individual characteristics and responses. The latter case is referred to as *Heterogeneous Causal Effect (HCE)* problem, under which the ATE for the whole population is less meaningful. One common method to mitigate the impact of HCE is through studying the conditional average treatment effect (CATE) instead of ATE within each subgroup, i.e.,

$$\text{Conditional ATE: } \tau_S = \mathbb{E}[Y_{\text{trt}}|\boldsymbol{X} \in S] - \mathbb{E}[Y_{\text{ctrl}}|\boldsymbol{X} \in S].$$

In our study, the subgroup is obtained as "covariates lying in a target subset of the covariate space", i.e., $\boldsymbol{X} \in S \subset \mathcal{X}$. Even though there exists a data-driven approach to partition the population into subgroups, such as [2], we focus on a more interpretable approach that is based on the label of a binary (or categorical) covariate; Specifically, in this work, we partition the data into normal and extreme weather subgroups based on the whether condition (such as wind and rainfall). Without loss of generality, we consider that the first element in the pre-treatment covariate vector is binary, i.e., $\boldsymbol{X} = (X_1, X_2, \dots), X_1 \in \{0, 1\}$. This yields a natural partition that

$$\mathcal{X} = S_0 \cup S_1, \text{ where } S_\ell = \{X_1 = \ell\}, \ \ell \in \{0, 1\}. \tag{1}$$

In the HCE problem, we typically have heterogeneous causal effect: $\tau_{S_0} \neq \tau_{S_1}$. We are interested in estimating CATE τ_{S_ℓ}, which is typically done by estimating ATE using the subgroup of the observations whose covariates lie in S_ℓ.

Challenge 3: insufficient data. The subgroup analysis above typically suffers from insufficient sample size. To handle this problem, we consider transfer learning techniques by leveraging knowledge obtained from one subgroup (called *source domain*) to help the estimation in the other subgroup (called *target domain*). Existing approaches for HCE estimation largely focus on model-ensemble, such as meta-learning [5], and heterogeneous transfer learning (i.e., TL under the heterogeneous covariate space setting) [4], but is not applicable to our problem set-up.

2.2 Transfer counterfactual learning

As previously mentioned, we conduct subgroup analysis and call S_1, S_0 (1) the target domain and source domain, respectively; Due to the potential heterogeneity, we denote $(\boldsymbol{X}_{\mathtt{t}}, Z_{\mathtt{t}}, Y_{\mathtt{t}})$ and $(\boldsymbol{X}_{\mathtt{s}}, Z_{\mathtt{s}}, Y_{\mathtt{s}})$ as the random vector under the corresponding domains, and our observations are

$$\begin{aligned}\text{Target Domain: } \mathcal{D}_{i,\mathtt{t}} &= (\boldsymbol{x}_{i,\mathtt{t}}, z_{i,\mathtt{t}}, y_{i,\mathtt{t}}), \ i = 1, \dots, n_{\mathtt{t}}, \text{ where } \boldsymbol{x}_{i,\mathtt{t}} \in S_1, \\ \text{Source Domain: } \mathcal{D}_{i,\mathtt{s}} &= (\boldsymbol{x}_{i,\mathtt{s}}, z_{i,\mathtt{s}}, y_{i,\mathtt{s}}), \ i = 1, \dots, n_{\mathtt{s}}, \text{ where } \boldsymbol{x}_{i,\mathtt{s}} \in S_0.\end{aligned}$$

To mitigate the selection bias, we apply the IPW estimator to estimate the CATE, which requires estimating the propensity score. Following [30], we consider a simple yet popular generalized linear form [24] for propensity score models:

$$\mathbb{P}(Z_{\mathtt{t}} = 1|\boldsymbol{X}_{\mathtt{t}}) = g({\boldsymbol{X}_{\mathtt{t}}}^{\mathrm{T}}\beta_{\mathtt{t}}), \quad \mathbb{P}(Z_{\mathtt{s}} = 1|\boldsymbol{X}_{\mathtt{s}}) = g({\boldsymbol{X}_{\mathtt{s}}}^{\mathrm{T}}\beta_{\mathtt{s}}),$$

where superscript $^{\mathrm{T}}$ denotes vector or matrix transpose, and $g(\cdot)$, known as the (inverse) link function, can be either linear, i.e., $g(x) = x,\ x \in [0, 1]$, or nonlinear, such as sigmoid link function $g(x) = 1/(1 + e^{-x}),\ x \in \mathbb{R}$, and exponential link function $g(x) = 1 - e^{-x},\ x \in [0, \infty)$. The IPW estimator for CATE (or ATE in the subgroup) is defined as:

$$\widehat{\tau}_{\mathtt{t}} = \frac{1}{n_{\mathtt{t}}} \sum_{i=1}^{n_{\mathtt{t}}} \frac{z_{i,\mathtt{t}} y_{i,\mathtt{t}}}{g(\boldsymbol{x}_{i,\mathtt{t}}^{\mathrm{T}} \widehat{\beta}_{\mathtt{t}})} - \frac{(1 - z_{i,\mathtt{t}}) y_{i,\mathtt{t}}}{1 - g(\boldsymbol{x}_{i,\mathtt{t}}^{\mathrm{T}} \widehat{\beta}_{\mathtt{t}})}, \quad \widehat{\tau}_{\mathtt{s}} = \frac{1}{n_{\mathtt{s}}} \sum_{i=1}^{n_{\mathtt{s}}} \frac{z_{i,\mathtt{s}} y_{i,\mathtt{s}}}{g(\boldsymbol{x}_{i,\mathtt{s}}^{\mathrm{T}} \widehat{\beta}_{\mathtt{s}})} - \frac{(1 - z_{i,\mathtt{s}}) y_{i,\mathtt{s}}}{1 - g(\boldsymbol{x}_{i,\mathtt{s}}^{\mathrm{T}} \widehat{\beta}_{\mathtt{s}})}, \tag{2}$$

where $\widehat{\beta}_{\mathtt{t}}, \widehat{\beta}_{\mathtt{s}}$ are the estimated nuisance parameters in the target and source domains, respectively.

In the above generalized linear model (GLM) formulation, we consider the same link function $g(\cdot)$ but different model parameters $\beta_{\mathtt{t}} \neq \beta_{\mathtt{s}}$, which accounts for related yet different domains; That is to say, our formulation can account for not only HCE but also heterogeneous treatment assignment rules.

Remark 1. Our TL problem set-up falls into the category of "inductive multi-task transfer learning" according to [25]. To be precise, we consider homogeneous feature space but (potentially) heterogeneous feature distribution, treatment assignment rule, and causal effect across the source and target domains.

ℓ_1-regularized transfer learning for IPW estimation. Since most transfer learning problems fall in the category of supervised learning, it is natural to consider knowledge transfer for the propensity score estimation stage in the IPW estimation. In our setting, we consider improving the estimation accuracy of $\beta_{\mathtt{t}}$ with the knowledge gained on $\beta_{\mathtt{s}}$, which helps resolve the data insufficiency issue due to partition. The key assumption for successful, i.e., theoretically guaranteed, knowledge transfer is the sparsity of the nuisance parameter difference Δ_β [3, 30], defined as:

$$\Delta_\beta = \beta_{\mathtt{t}} - \beta_{\mathtt{s}}. \tag{3}$$

Assumption 1. The difference of nuisance parameters Δ_β (3) is s-sparse, i.e., for $0 \leq s \leq d$,

$$\|\Delta_\beta\|_0 \leq s.$$

This assumption states that treatment assignment mechanisms are very similar across both domains.

To estimate $\beta_{\mathtt{t}}$, we first leverage source domain data to estimate $\beta_{\mathtt{s}}$, which serves as a rough estimator of $\beta_{\mathtt{t}}$ due to Assumption 1. Next, we correct the bias of the rough estimator by using ℓ_1 regularization to learn the difference Δ_β from target domain data. The idea is that we can accurately estimate $\beta_{\mathtt{s}}$ using abundant source domain data and faithfully capture the sparse Δ_β from limited target domain data with the help of ℓ_1 regularization. To be precise:

Rough estimation: $\widehat{\beta}_{\mathtt{t}}^{\mathrm{rough}} = \arg\min_b \frac{1}{n_{\mathtt{s}}} \sum_{i=1}^{n_{\mathtt{s}}} -z_{i,\mathtt{s}} \boldsymbol{x}_{i,\mathtt{s}}^{\mathrm{T}} b + G\left(\boldsymbol{x}_{i,\mathtt{s}}^{\mathrm{T}} b\right),$
Bias correction: $\widehat{\beta}_{\mathtt{t}}^{\mathrm{TL}} = \arg\min_b \frac{1}{n_{\mathtt{t}}} \sum_{i=1}^{n_{\mathtt{t}}} -z_{i,\mathtt{t}} \boldsymbol{x}_{i,\mathtt{t}}^{\mathrm{T}} b + G\left(\boldsymbol{x}_{i,\mathtt{t}}^{\mathrm{T}} b\right) + \lambda \|b - \widehat{\beta}_{\mathtt{t}}^{\mathrm{rough}}\|_1,$

where $\lambda > 0$ is a tunable regularization strength hyperparameter, and function G satisfies $G' = g$; please refer to [30] for a detailed definition and estimation of the generalized linear model. Lastly, the proposed ℓ_1-TCL estimation of $\tau_{\mathtt{t}}$ can be done by plugging $\widehat{\beta}_{\mathtt{t}} = \widehat{\beta}_{\mathtt{t}}^{\mathrm{TL}}$ into (2).

Challenge 4: hyperparameter selection. One practical issue is the selection of the ℓ_1 regularization hyperparameter λ; In fact, this is a rather important topic as proper hyperparameter tuning can sometimes close the performance gap among different causal estimators [15]. The challenge is that a "golden standard", such as prediction accuracy in classic supervised learning, is largely missing since we cannot evaluate the causal effect estimation accuracy from data due to unobserved counterfactual outcomes. Unfortunately, there is no good solution beyond the recent empirical analysis of some nuisance model prediction performance metrics as the selection criteria [2, 6, 15]; Here, we propose one additional solution by studying the covariate distribution balanceness, and we compare the empirical performance of those criteria.

One straightforward approach is to use the performance of the supervised learning of the nuisance model as the selection criterion. As first studied in [2], the generalization performance by performing a train-validation-test split is of vital importance; Here, we apply cross-validation (CV) since the sample size in our real application is relatively small. In our setting, the nuisance model is the propensity score model which learns the binary treatment assignment, and we choose several binary prediction metrics to quantify the performance of the propensity score model — ℓ_2 norm of the prediction error (ℓ_2 err.), Cross Entropy error (CE err.), and the area under the receiver operating characteristic curve (AUC).

As the fitted propensity scores are used to balance the covariate distribution, one popular method to access the goodness-of-fit in causal inference is to directly examine the "balanceness" of the covariate distribution of the re-weighted samples. One commonly used measure of the covariate distribution "balanceness" is the standardized mean difference (SMD) [34], which is essentially Cohen's d and can measure the discrepancy between two probability distributions; alternatively, one can use the distance between two probability distributions to measure the balanceness and here we choose Maximum Mean Discrepancy (MMD). For a definition of our selection criteria, one can see Appendix A.2.

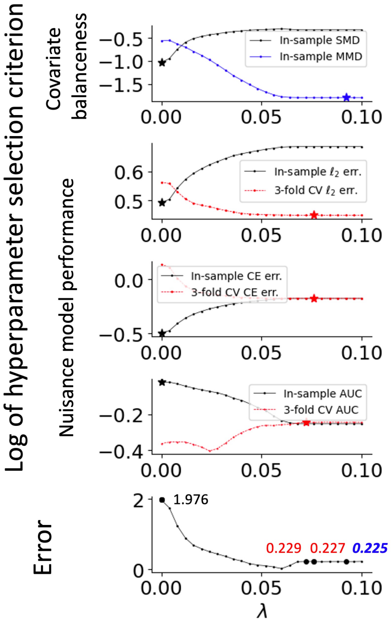

Figure 2: Comparison of different hyperparameter selection criteria. We can observe that in-sample MMD and all cross validation nuisance model performance metrics can select λ's that output similar and accurate causal effect estimates.

Simulation results. As the counterfactuals are unobservable in practice, we conduct numerical simulation, in which we know the ground truth causal effect, to compare the effectiveness of the aforementioned selection criteria. We consider a $d = 50, s = 2$ example where we have $n_{\mathtt{t}} = 100$ target domain samples and $n_{\mathtt{s}} = 2000$ source domain samples; Please refer to Appendix B.1 for detailed experimental configurations. If we estimate the causal effect using the target domain data only, the IPW estimate will have a very large error (estimation error is 1.45 whereas the ground truth causal effect is $\tau_{\mathtt{t}} = 3$). We perform the aforementioned ℓ_1-regularized transfer learning and plot how the resulting ℓ_1-TCL estimation error varies with different λ in Figure 2. Additionally, we plot how the aforementioned criteria vary with different λ and use "$\star$" to denote the selected hyperparameter.

Surprisingly, the SMD considered in [34] does not perform well in our set-up, and it may require more sophisticated aggregation (than the simple average) of the Cohen's d's. The observation that in-sample nuisance model performances are rather poor agrees with existing literature [2, 6, 15]. Importantly, we find that the MMD criterion has a similar performance to the CV nuisance model performance criteria; Indeed, it performs the best in this specific example.

Remark 2. One drawback of all aforementioned criteria is that they do not involve outcome variable, and therefore they cannot be used to test whether or not the heterogeneous causal effect assumption hold. As a result, performing subgroup analysis and transfer learning among subgroups might be the most robust approach that we can take. In our numerical example, IPW estimation on the entire dataset (i.e., both domains combined) yields a point estimate of 4.91, which is very close to the source domain true causal effect $\tau_{\mathtt{s}} = 5$.

3 Real-Data Experiment

Dataset. Our data was collected from multiple sources: The outage data was collected for all cities in a state located in the northeastern United States from the local government, spanning 2018-03-01 00:00:00 to 2018-03-31 23:00:00. The weather data was collected from High-Resolution Rapid Refresh (HRRR) model, and the demographic factors are collected from the American Community Survey (ACS) 5-Year survey from US Census Bureau. We obtained land cover and land use data from the Massachusetts government's MassGIS Data: 2016 Land Cover/Land Use (LCLU) dataset.

Our real-data experiment examined the causal effects from several factors (i.e., treatment variables, including median income, elderly percentage, population density, and total population) to the power outage duration (i.e., the outcome variable) with demographics, land cover, land use, and weather data as pre-treatment covariates. The power outage is characterized by the System Average Interruption Duration Index (SAIDI), which is usually used to measure the average outage duration for each customer served in a certain period of time. The two weather factors considered here are wind speed (m/s) and precipitable water (kg/m^2). To account for potential heterogeneity, the outage events are split into two groups: severe weather events and normal weather events. The outage event is considered to happen during severe weather if the maximum wind speed or the maximum precipitation passes the preset thresholds. The thresholds here are 19 m/s for wind speed and 16 kg/m^2 for precipitable water. SAIDI would be calculated for severe weather events and normal weather events separately. With this setting, we are able to get two sets of data to conduct the transfer counterfactual learning. Details about SAIDI definition, demographics, land cover, and land use factors considered in this study can be found in Appendix B.2.

Results. The treatment variables are binarized via the criterion "whether the variable **exceeds** a threshold (80% quantile of the whole population)"; We report our results in Table 1, where we use "↑" to indicate such a criterion. For instance, in the case of median income, a household is categorized as poor or wealthy by comparing its income with the 80% quantile of all cities' household median income; From Table 1, we can observe that, without transfer learning, the estimated causal effect was positive for the normal weather subgroup, indicating that being wealthy led to more power outages, which contradicts common sense. Specifically, the causal effect estimate without transfer learning was -460.79 for the severe weather group and 68.03 for the normal weather group. However, with transfer learning, the causal effect estimates aligned with expectations, showing that being wealthy led to fewer power outages — the causal effect estimate was -762.25 for the severe weather group and -881.88 for the normal weather group. Additionally, the observations that vanilla causal inference approach (i.e., IPW without TL) fails to output reasonable results can be made for other demographic factors as the treatment variable: For example, it is counter-intuitive that there are less power outage in the elderly-populated areas. Our ℓ_1-TCL approach in Table 1 uses covariate balanceness (i.e., in-sample MMD) criterion for hyperparameter selection; See Figure 3 in Appendix B.2 for results of ℓ_1-TCL with nuisance model predictive performance criteria.

Table 1: Causal effect estimates. We quantize the estimated value $\widehat{\tau}$ as: "− −" if $\widehat{\tau} \leq -500$, "−" if $-500 < \widehat{\tau} \leq -100$, "⋆" if $-100 < \widehat{\tau} \leq 100$, "+" if $\widehat{\tau} > 100$. We can see TL can help yield very different results which agree more with common sense, compared with estimation without the help of source domain knowledge (i.e., no TL).

	Median income$_{(\uparrow)}$		Elderly percentage$_{(\uparrow)}$		Population density$_{(\uparrow)}$		Total population$_{(\uparrow)}$	
Weather	Severe	Normal	Severe	Normal	Severe	Normal	Severe	Normal
ℓ_1-TCL	− − (-762.25)	− − (-881.88)	+ (420.74)	+ (255.76)	− (-394.05)	⋆ (0.82)	− − (-616.26)	⋆ (34.55)
No TL	− (-460.79)	⋆ (68.03)	− (-253.89)	− (-146.62)	− − (-616.02)	− (-233.17)	− − (-557.00)	− (-141.18)

Findings. As shown in Table 1, our ℓ_1-TCL provides more reasonable results compared to the vanilla causal inference approach: It highlights prolonged power outages in areas characterized by both low income and a higher percentage of elderly residents. Moreover, the elderly population area may have increased vulnerability during severe weather events. Interestingly, extended power outages in less densely populated areas are only observed under severe weather conditions. This observation aligns with the notion that power companies may prioritize supplying electricity to densely populated areas during severe weather events to minimize the impact on a larger number of residents.

One particularly intriguing observation is the consistency in prolonged power outages in areas with lower median income. This phenomenon occurs under both severe and normal weather conditions, indicating that these areas may face persistent challenges in maintaining power supply reliability. This suggests a need for targeted infrastructure improvements and support in economically disadvantaged regions to mitigate power outage durations.

4 Discussion

This work presents a homogeneous transfer learning approach within the inverse probability weighting estimation to tackle the data sacristy and improve estimation accuracy of the causal effect. The proposed method is subsequently deployed to the unfairness detection within power systems, suggesting that low-income and elderly-populated areas may consistently experience prolonged power outages, and highlighting the necessity for focused improvements in areas with economic challenges.

Although the real-data results appear reasonable, there does not exist ground truth to externally validate our findings. Therefore, it would be beneficial to conduct experiments on benchmark datasets with known causal relationships or to develop a goodness-of-fit score to quantify the reliability of the real results. We leave those topics for future study.

References

[1] Assimilation and Verification Innovation Division (AVID). The High-Resolution Rapid Refresh (HRRR), 2018.

[2] Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

[3] Hamsa Bastani. Predicting with proxies: Transfer learning in high dimension. *Management Science*, 67(5):2964–2984, 2021.

[4] Ioana Bica and Mihaela van der Schaar. Transfer learning on heterogeneous feature spaces for treatment effects estimation. *arXiv preprint arXiv:2210.06183*, 2022.

[5] Alicia Curth and Mihaela van der Schaar. Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms. In *International Conference on Artificial Intelligence and Statistics*, pages 1810–1818. PMLR, 2021.

[6] Alicia Curth and Mihaela van der Schaar. In search of insights, not magic bullets: Towards demystification of the model selection dilemma in heterogeneous treatment effect estimation. *arXiv preprint arXiv:2302.02923*, 2023.

[7] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

[8] Seung-Ryong Han, Seth D Guikema, and Steven M Quiring. Improving the predictive accuracy of hurricane power outage forecasts using generalized additive models. *Risk Analysis: An International Journal*, 29(10):1443–1453, 2009.

[9] Tobias Hatt, Jeroen Berrevoets, Alicia Curth, Stefan Feuerriegel, and Mihaela van der Schaar. Combining observational and randomized data for estimating heterogeneous treatment effects. *arXiv preprint arXiv:2202.12891*, 2022.

[10] Raphael J Heffron. Applying energy justice into the energy transition. *Renewable and Sustainable Energy Reviews*, 156:111936, 2022.

[11] Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685, 1952.

[12] Walter Keady, Bindu Panikkar, Ingrid L Nelson, and Asim Zia. Energy justice gaps in renewable energy transition policy initiatives in vermont. *Energy Policy*, 159:112608, 2021.

[13] Sören R Künzel, Bradly C Stadie, Nikita Vemuri, Varsha Ramakrishnan, Jasjeet S Sekhon, and Pieter Abbeel. Transfer learning for estimating causal effects using neural networks. *arXiv preprint arXiv:1808.07804*, 2018.

[14] Yanling Lin, Jianhui Wang, and Meng Yue. Equity-based grid resilience: How do we get there? *The Electricity Journal*, 35(5):107–135, 2022. Behind the meter strategies for enhancing the electricity grid resilience, reliability, economics, sustainability, and security.

[15] Damian Machlanski, Spyridon Samothrakis, and Paul Clarke. Hyperparameter tuning and model evaluation in causal effect estimation. *arXiv preprint arXiv:2303.01412*, 2023.

[16] MassGIS and NOAA Office of Coastal Management (OCM). Full documentation for MassGIS' 2016 Land Use/Land Cover Data, 2019.

[17] Sara Meerow, Pani Pajouhesh, and Thaddeus R. Miller. Social equity in urban resilience planning. *Local Environment*, 24(9):793–808, 2019.

[18] Sayanti Mukherjee, Roshanak Nateghi, and Makarand Hastak. A multi-hazard approach to assess severe weather-induced major power outage risks in the u.s. *Reliability Engineering & System Safety*, 175:283–305, 2018.

[19] Mukhopadhyay and Sayanti. *Towards a Resilient Grid: A Risk-Based Decision Analysis Incorporating the Impacts of Severe Weather-Induced Power Outages*. PhD thesis, Purdue University, 2017.

[20] Luis Mundaca, Henner Busch, and Sophie Schwer. 'successful'low-carbon energy transitions at the community level? an energy justice perspective. *Applied Energy*, 218:292–303, 2018.

[21] Roshanak Nateghi, Seth Guikema, and Steven M Quiring. Power outage estimation for tropical cyclones: Improved accuracy with simpler models. *Risk analysis*, 34(6):1069–1078, 2014.

[22] Roshanak Nateghi, Seth D Guikema, and Steven M Quiring. Comparison and validation of statistical methods for predicting power outage durations in the event of hurricanes. *Risk Analysis: An International Journal*, 31(12):1897–1906, 2011.

[23] Roshanak Nateghi, Seth D Guikema, and Steven M Quiring. Comparison and validation of statistical methods for predicting power outage durations in the event of hurricanes. *Risk analysis : an official publication of the Society for Risk Analysis.*, 31(12), 2011-12.

[24] John Ashworth Nelder and Robert WM Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384, 1972.

[25] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.

[26] Steven M. Quiring, Laiyin Zhu, and Seth D. Guikema. Importance of soil and elevation characteristics for modeling hurricane-induced power outages. *Natural Hazards*, 58(1):365–390, 2011.

[27] James M Robins, Miguel Angel Hernan, and Babette Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, pages 550–560, 2000.

[28] 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.

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

[30] Song Wei, Ronald Moore, Hanyu Zhang, Yao Xie, and Rishikesan Kamaleswaran. Transfer causal learning: Causal effect estimation with knowledge transfer. *arXiv preprint arXiv:2305.09126*, 2023.

[31] Christopher Winship and Robert D Mare. Models for sample selection bias. *Annual review of sociology*, 18(1):327–350, 1992.

[32] Lili Wu and Shu Yang. Transfer learning of individualized treatment rules from experimental to real-world data. *Journal of Computational and Graphical Statistics*, pages 1–10, 2022.

[33] Shu Yang and Peng Ding. Combining multiple observational data sources to estimate causal effects. *Journal of the American Statistical Association*, 115(531):1540–1554, 2020.

[34] Zhongheng Zhang, Hwa Jung Kim, Guillaume Lonjon, Yibing Zhu, et al. Balance diagnostics after propensity score matching. *Annals of translational medicine*, 7(1), 2019.

Appendix of Unfairness Detection within Power Systems through Transfer Counterfactual Learning

Table of Contents

A Additional Technical Details

A.1 Selection bias and inverse probability weighting estimation

In observational study, the pre-treatment covariates $\boldsymbol{X}$ that determine the treatment may also be correlated, or “confounded”, with the outcome. As a result, selecting treated (or control) subjects and averaging the selected outcomes will yield a biased estimate of $\mathbb{E}[Y_{\text{trt}}]$ (or $\mathbb{E}[Y_{\text{ctrl}}]$). Although such statistically independence typically does not hold in practice, in causal inference literature, a common practice to handle such a problem is to assume there are ‘no unmeasured confounders” (also known as the Ignorability Assumption) [27]:

$$(Y_{\text{ctrl}}, Y_{\text{trt}}) \perp\!\!\!\perp Z \mid \boldsymbol{X}.$$

In the following, we shall continue our study under the above assumption.

One way to handle the selection bias is to re-weight each sample in the selected cohort such that the re-weighted sample is “representative” of the whole population, and one popular method is the inverse probability weighting [11]. To be precise, [28] showed that the propensity score $e(\boldsymbol{X}) = \mathbb{P}(Z = 1|\boldsymbol{X})$, i.e., the probability of receiving treatment given covariates, satisfies the following:

$$(Y_{\text{ctrl}}, Y_{\text{trt}}) \perp\!\!\!\perp Z \mid e(\boldsymbol{X}).$$

This leads to the following unbiased estimate of $\mathbb{E}[Y_{\text{trt}}]$:

$$\mathbb{E}\left[\frac{ZY}{e(\boldsymbol{X})}\right] = \mathbb{E}\left\{\mathbb{E}\left[\frac{I(Z=1)Y_{\text{trt}}}{e(\boldsymbol{X})} \,\middle|\, Y_{\text{trt}}, \boldsymbol{X}\right]\right\} = \mathbb{E}\left\{\frac{Y_{\text{trt}}}{e(\boldsymbol{X})}\mathbb{E}\left[I(Z=1)|Y_{\text{trt}}, \boldsymbol{X}\right]\right\} = \mathbb{E}[Y_{\text{trt}}].$$

Similarly, we have $\mathbb{E}\left[\frac{ZY}{1-e(\boldsymbol{X})}\right] = \mathbb{E}[Y_{\text{ctrl}}]$. Then, the famous inverse probability weighting (IPW) estimator of the ATE τ can be obtained by replacing the expectation with sample average in $\mathbb{E}\left[\frac{ZY}{e(\boldsymbol{X})}\right] - \mathbb{E}\left[\frac{ZY}{1-e(\boldsymbol{X})}\right]$; We will introduce the IPW estimator in detail later.

A.2 Hyperparameter selection criteria

Nuisance model performance. In our study AUC is implemented using `sklearn.metrics.auc` in python, and ℓ_2 err. as well as CE err. are defined as follows:

Definition 1 (ℓ_2 err.)**.** Given binary samples $\{z_i,\ i = 1, \ldots, n\}$ their predictions $\{\widehat{z}_i,\ i = 1, \ldots, n\}$, the ℓ_2 norm of the prediction error is given by $\sum_{i=1}^{n}(z_i - \widehat{z}_i)^2/n$.

Definition 2 (CE err.)**.** Given binary samples $\{z_i,\ i = 1, \ldots, n\}$ their predictions $\{\widehat{z}_i,\ i = 1, \ldots, n\}$, the Cross Entropy error is given by $\sum_{i=1}^{n}\left(z_i \log(\widehat{z}_i) + (1 - z_i)\log(1 - \widehat{z}_i)\right)/n$.

Covariate balanceness. One commonly used measure of the covariate distribution "balanceness" is the standardized mean difference (SMD) [34], which is essentially the Cohen's $\mathtt{d}$:

Definition 3 (Cohen's $\mathtt{d}$)**.** Given two sets of samples $\mathcal{A} = \{a_i,\ i = 1, \dots, m\}$ and $\mathcal{B} = \{b_j,\ j = 1, \dots, n\}$, the Cohen's $\mathtt{d}$ is defined as follows:

$$\mathtt{d}_{\text{Cohen}}(\mathcal{A}, \mathcal{B}) = (\bar{a} - \bar{b}) \Bigg/ \sqrt{\frac{S_{\mathcal{A}}^2 + S_{\mathcal{B}}^2}{2}},$$

where $\bar{a} = \sum_{i=1}^{m} a_i/m,\ \bar{b} = \sum_{i=1}^{n} b_j/n$, and the pooled sample variance can be calculated as:

$$S_{\mathcal{A}} = \frac{1}{m}\sum_{i=1}^{m}(a_i - \bar{a})^2, \quad S_{\mathcal{B}} = \frac{1}{n}\sum_{j=1}^{n}\left(b_j - \bar{b}\right)^2.$$

When the Cohen's $\mathtt{d}$'s absolute value is close to zero, the standardized means of two distributions are similar to each other, i.e., the covariates' distributions are balanced. In our multivariate setting, we can simply use the average of the absolute Cohen's $\mathtt{d}$'s as the selection criterion, i.e.,

$$\text{SMD} = \frac{1}{d}\sum_{j=1}^{d}\left|\mathtt{d}_{\text{Cohen}}\big(\{\boldsymbol{x}(j) : \boldsymbol{x} \in \mathcal{D}_{\boldsymbol{x},\text{trt}}\}, \{\boldsymbol{x}(j) : \boldsymbol{x} \in \mathcal{D}_{\boldsymbol{x},\text{ctrl}}\}\big)\right|,$$

where $\boldsymbol{x}(j)$ is the j-th element of the vector $\boldsymbol{x} \in \mathbb{R}^d$, and

$$\mathcal{D}_{\boldsymbol{x},\text{trt}} = \left\{\frac{\boldsymbol{x}_{i,\mathtt{t}}}{g(\boldsymbol{x}_{i,\mathtt{t}}^{\mathrm{T}}\widehat{\beta}_{\mathtt{t}})} : z_{i,\mathtt{t}} = 1\right\}, \quad \mathcal{D}_{\boldsymbol{x},\text{ctrl}} = \left\{\frac{\boldsymbol{x}_{i,\mathtt{t}}}{1 - g(\boldsymbol{x}_{i,\mathtt{t}}^{\mathrm{T}}\widehat{\beta}_{\mathtt{t}})} : z_{i,\mathtt{t}} = 0\right\}.$$

This above metric tells us the average standardized mean difference after propensity score weighting, and we will refer to it as the SMD is the following analysis.

As one can see, the goal is to quantitatively characterize how similar two empirical distributions are after propensity score weighting, and this is indeed extensively suited in two sample test problem. Here, we introduce one very popular non-parametric distance metric between distributions (i.e., two sample test statistic), Maximum Mean Discrepancy (MMD) [7], as follows:

Definition 4 (Maximum Mean Discrepancy)**.** Given two sets of samples $\mathcal{A} = \{\boldsymbol{a}_i,\ i = 1, \dots, m\}$ and $\mathcal{B} = \{\boldsymbol{b}_j,\ j = 1, \dots, n\}$, an unbiased estimator of MMD can be obtained via U-statistics as follows:

$$\widehat{\text{MMD}}^2(\mathcal{A}, \mathcal{B}) = \frac{1}{m(m-1)}\sum_{i=1}^{m}\sum_{j \neq i}^{m} k(\boldsymbol{a}_i, \boldsymbol{a}_j) - \frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} k(\boldsymbol{a}_i, \boldsymbol{b}_j) + \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j \neq i}^{n} k(\boldsymbol{b}_i, \boldsymbol{b}_j).$$

Our MMD-based selection criterion (referred to as MMD for brevity) is then given as:

$$\text{MMD} = \widehat{\text{MMD}}^2\left(\mathcal{D}_{\boldsymbol{x},\text{trt}}, \mathcal{D}_{\boldsymbol{x},\text{ctrl}}\right).$$

In the above definition, k is the user-specified kernel function, i.e., $k(\cdot, \cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Commonly used kernel functions include Gaussian radial basis function $k(\boldsymbol{x}_1, \boldsymbol{x}_2) = \exp\{-\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2^2/r^2\}$, where $\|\cdot\|_2$ is the vector ℓ_2 norm and $r > 0$ is the bandwidth parameter; the bandwidth is typically chosen using median heuristic.

B Additional Experimental Details

B.1 Numerical simulation

In our numerical simulation, the covariates are the absolute values of normally distributed random numbers with mean 0 and standard deviation 1. The source domain PS model parameters are also absolute values of normally distributed random numbers with mean 0 and standard deviation 0.1. In the data generating process, the "unknown" true PS model is GLM with exponential link function whereas our specified model considers GLM with sigmoid link function, i.e., we consider model mismatch in our simulation. We take absolute values since the link function, i.e., the exponential function $g(x) = 1 - \exp\{-x\}$, is supported on $x \in (0, \infty)$. The s-sparse difference has magnitude 0.2. The ground truth ATE is $\tau = 3$ for in target domain under both settings; for the heterogeneous PS model case (i.e., setting 1), we consider a follow up experiment where the ground truth ATE is increased to 5 in target domain. Our specified model is GLM with sigmoid function and we fit vanilla logistic regression using source domain data for the rough estimation step. Afterwards, we perform gradient descent to learn the sparse difference with ℓ_1 regularization; the total number of iterations is 20000, the initial learning rate (lr) is 0.001 and it decays by 1% (i.e., lr $\leftarrow 0.99$ lr) every 1000 iterations.

In addition to compare different hyperparameter selection criteria, we also compare some naive baseline methods: We naively merge the target and source domain datasets; since there source domain samples significantly outnumber the target domain one, it is no surprising that the IPW estimate is 4.91, which is very close to the source domain ground truth causal effect. Additionally, we consider the inference using only target domain dataset, and the IPW estimate is 4.45, which is still less than satisfactory compared with the ℓ_1-TCL estimate.

B.2 Real-data experiment

SAIDI. Here, we do little modifications to represent the power reliability during the month (quantified in minutes) and calculate SAIDI for each city. Consider one city with M power grid users, it went through N outage events in the month with each outage lasting for $L_n, n = 1, \dots N$ hours. The number of power grid users recorded to be without electricity is C_{ln} during hour l for outage event n. Then SAIDI for the city in the month can be calculated as

$$\text{SAIDI} = \frac{\sum_{n=1}^{N} \sum_{l=1}^{L_n} C_{ln}}{M} \times 60$$

We consider all the power outage events with a minimum outage rate bigger than 0.1% and last for over 2 hours.

Weather. Extreme weather and climate events have emerged as primary catalysts for infrastructure damage, leading to widespread power outages and supply inadequacy risks in the United States [18, 19]. Consequently, analyzing the patterns of these outages can be facilitated by leveraging weather data.

To perform the analysis, we obtained weather data from the High-Resolution Rapid Refresh (HRRR) model [1]. Developed by the National Centers for Environmental Prediction (NCEP), the HRRR model is a numerical weather prediction model that provides high-resolution and frequently updated forecasts for regional weather conditions. By assimilating data from various sources, such as satellites, radars, and weather stations, the HRRR model combines observational information with advanced mathematical equations to simulate atmospheric behavior. This enables the model to generate highly detailed and accurate short-term forecasts, ranging from 1 to 18 hours, with a spatial resolution as fine as 3 kilometers.

For our analysis, we specifically focused on two key weather parameters extracted from the HRRR model: wind speed (m/s) and precipitable water (kg/m^2). Hourly weather data was collected from all available stations, and for each city, we matched the data with the nearest stations to obtain the corresponding weather information.

Demographic census. Previous research in the field of power equity has extensively examined the correlation between demographic factors and power outages. Studies have demonstrated that certain disadvantaged communities, such as low-income and minority communities, experience a disproportionate impact from power outages due to their limited resources for recovery [14, 17]

To investigate the demographic factors influencing power outages, we leverage the comprehensive data provided by the American Community Survey (ACS) 5-Year Data. This dataset is a continuous survey conducted by the U.S. Census Bureau, focusing on gathering in-depth information about the demographic, social, economic, and housing characteristics of the U.S. population. The ACS 5-Year Data is particularly valuable as it encompasses information collected over a 5-year period, providing a more robust and accurate representation of the population and its diverse characteristics. To complement the demographic factors, we also obtained the number of power grid users from the local utilities.

Land cover and land use. Several studies [23, 26] have indicated that land cover and land use variables can serve as reasonable proxies for power systems data, such as the number of poles. This suggests that land cover variables have the potential to serve as a valuable tool in developing generalized outage models that can be applied to service areas lacking detailed power system data.For our analysis, we obtained land cover and land use data from the Massachusetts government's MassGIS Data: 2016 Land Cover/Land Use (LCLU) dataset [16]. This statewide dataset combines land cover mapping from 2016 aerial imagery with land use information derived from standardized assessor parcel data specific to Massachusetts. The creation of this dataset was a collaborative effort between MassGIS and the National Oceanic and Atmospheric Administration's (NOAA) Office of Coastal Management (OCM).

In our study, we considered various land cover categories, including forest, grass, and wetland. Additionally, we examined different land use types, such as single-family residential, multi-family residential, and agriculture. Each of these land cover and land use variables was measured as a percentage of area for each city in the dataset.

Additional results. We report how the selection criteria and the estimated causal effect vary with different ℓ_1 regularization strength λ's in Figure 3. It is worthwhile noting that the similar performance of different selection criteria is again observed in our real data example when we consider the median income and elderly percentage as the treatment variables, reaffirming their effectiveness and supporting the enhanced reliability and accuracy of the ℓ_1 TCL results. Meanwhile, when we take the population density or total number as the treatment, we can observe that the in-sample MMD criterion can output very different result compared with CV nuisance model performance: On one hand, the numerical simulation shows in-sample MMD criterion can help yield a bit more accurate causal effect estimate; On the other hand, it makes sense that power company will prioritize the power supply in populated areas under sever conditions, but such a power supply difference should not exist under normal conditions.

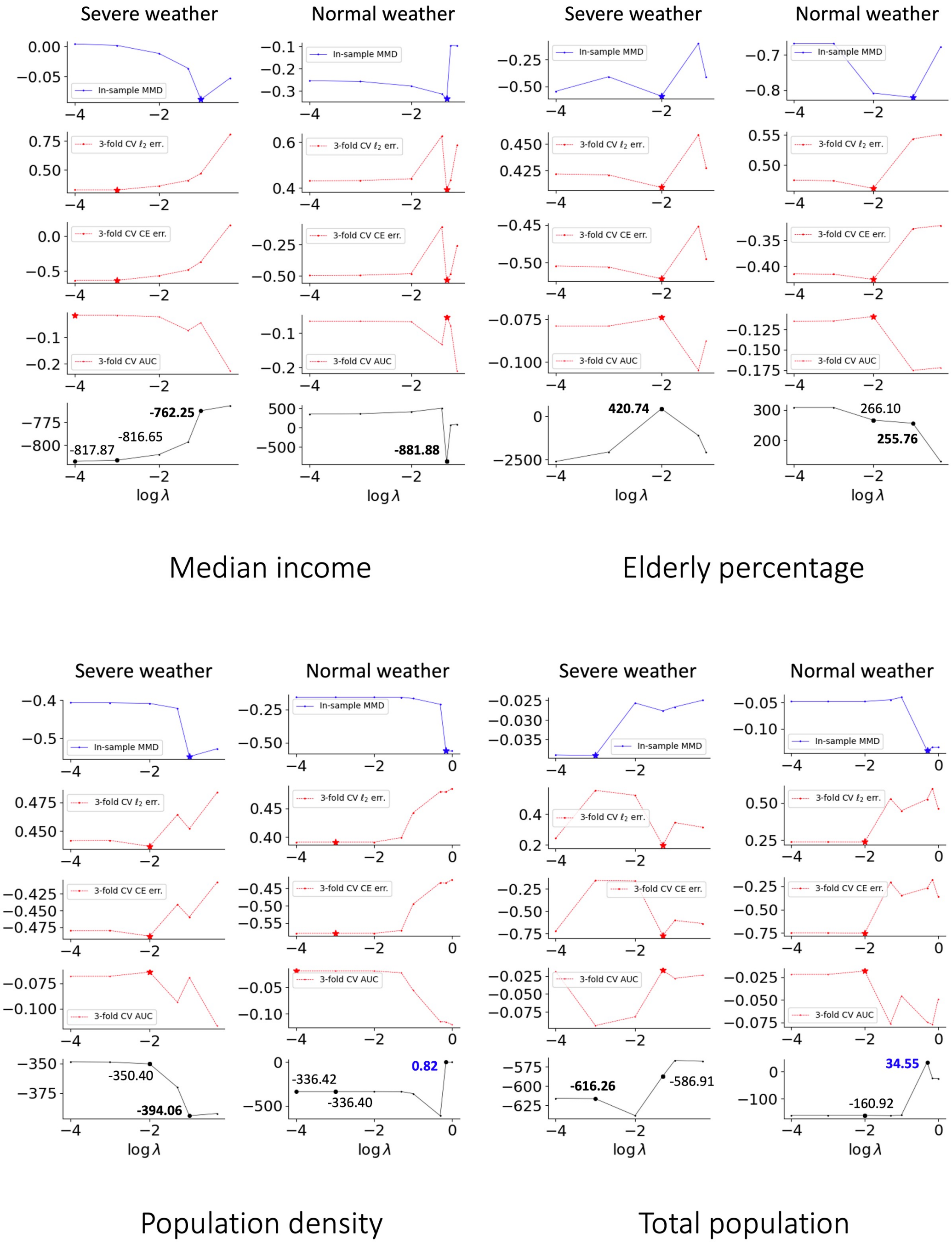

Figure 3: Hyperparameter selection in the real data example. When we take median income or elderly percentage as the treatment variable, ℓ_1-TCL with different selection criteria can output similar results. Our in-sample MMD criterion can help consistently output reasonable results.