Estimation and Inference of Heterogeneous Treatment Effects with Causal Random Forest

Bachelor Thesis
Department of Economics
University of Zurich

Emanuel Nussli

Supervisors: Prof. Michael Wolf, Ph.D. and
Simon Hediger

Course of studies:	Economics
Student ID:	18-704-205
Address:	Geiselweidstrasse 52
	8400 Winterthur
E-Mail:	emanuel.nussli@uzh.ch
Date:	January 19, 2022

Abstract

Contents

1 Introduction **4**

2 Estimation of Treatment Effects **5**

2.1 Potential Outcomes Framework 5

3 Causal Forests **6**

3.1 Causal Forests 6

3.2 Inference for Causal Forests 10

3.3 Causal Forests and Panel Data 11

4 Monte Carlo Simulation **12**

4.1 Data Generating Process 12

4.2 Delay of Treatment Effects 15

4.3 Performance Evaluation 16

4.4 Results 16

4.4.1 Design 1: The Value of Clustering 17

4.4.2 Design 2: The Influence of Unobserved Confounding . . . 17

5 Empirical Application of Causal Forests: Effectiveness of Facial Mask Policies in Switzerland on Containing the COVID-19 Pandemic **20**

5.1 Introduction 20

5.2 Related Literature 21

5.3 Data 21

5.3.1 Epidemological Data 21

5.3.2 Baseline Cantonal Data 22

5.3.3 Behavioral Data 24

5.3.4 Policy Data 24

5.4 Results 25

5.4.1 Assumptions 26

5.4.2 Average Treatment Effect 26

5.4.3 Treatment Heterogeneity 28

6 Conclusion **31**

References **32**

A Appendix **37**

A.1 Monte Carlo Simulation 37

A.2 Empirical Application 41

A.2.1 Tables 41

A.2.2 Figures 43

A.2.3 Robustness Checks 43

A.2.4 Robustness-Checks 50

1 Introduction

Understanding how individuals react to being assigned a certain treatment.

2 Estimation of Treatment Effects

To lay the foundation of estimating treatment effects, one has to concisely define what a treatment effect is. The potential outcomes framework from [SDS90] and [Rub74] allows for a proper definition of treatment effects and a mathematical framework to think about causality. We describe the potential outcomes framework in section 2.1 while elaborating on the estimands of interest and specifying the assumptions made.

2.1 Potential Outcomes Framework

Assume there is a population of n units indexed $1, \ldots, n$. For each of the units, we observe a feature vector $X_i \in [0,1]^d$ where d is the dimension of the feature space $\mathcal{X}$. We additionally have access to a response $Y_i \in \mathbb{R}$ and a binary treatment indicator $W_i \in \{0,1\}$. In the spirit of [ATW19], we summarize the data into $(X_i, O_i) \in \mathcal{X} \times \mathcal{O}$ where $O_i = \{Y_i, W_i\}$. The data is regarded as an i.i.d sample drawn from a large population [AI16].

[SDS90] and [Rub74] propose the existence of potential outcomes $\{Y_i^{(1)}, Y_i^{(0)}\}$ where $Y_i^{(1)}$ represents the response of the ith unit had it received treatment and $Y_i^{(0)}$ had it not received it. Herein lies the fundamental issue of causality as we only ever observe either $Y_i^{(1)}$ or $Y_i^{(1)}$ making causality inherently a problem of missing data [DL18]. If we observed $\{Y_i^{(1)}, Y_i^{(0)}, W_i, X_i\}$, treatment effect estimation at x would be given by

$$\tau(x) = \mathbb{E}\left[Y_i^{(1)} - Y_i^{(0)} \mid X_i = x\right] \tag{2.1}$$

We denote $\tau(x)$ as the heterogeneous treatment effect function. It is also called the conditional average treatment function for estimating conditional average treatment effects (CATE). The average treatment effect (ATE) on the other hand is defined as

$$\tau = \mathbb{E}\left[Y_i^{(1)} - Y_i^{(0)}\right] = \mathbb{E}_X\left[\mathbb{E}\left[Y_i^{(1)} - Y_i^{(0)} \mid X = x\right]\right] \tag{2.2}$$

where $\mathbb{E}_X$ denotes the expectation over X.

As estimating $\tau(x)$ or τ from observed data (X_i, Y_i, W_i) is generally not

possible, [WA18] assume unconfoundedness, which is standard practice. Unconfoundedness translates as treatment W_i being independent of the potential outcomes conditional on X_i, meaning

$$\left\{Y_i^{(1)}, Y_i^{(0)}\right\} \perp\!\!\!\perp W_i \mid X_i \tag{2.3}$$

Unconfoundedness entails that nearby observations in x-space can be treated as having come from a randomized experiment [WA18].

3 Causal Forests

We firstly describe the Causal Forest algorithm as developed in the sequence of work [AI16], [WA18] and [ATW19] in section 3.1. In section 3.2, we describe the results that enable statistical inference for the estimated treatment effects. Lastly, we discuss panel data and heuristically present why and how Causal Forests can be adapted to suit treatment effect estimation from panel data in section 3.3.

3.1 Causal Forests

Aiming to structure the problem of heterogeneous treatment effects, [Rob88] studied a class of semiparametric problems where we have a model of $\tau(x)$ given by

$$Y_i^{(w)} = f(X_i) + w\tau(X_i) + \epsilon_i(w) \tag{3.1}$$

He restricted the analysis by imposing that $\tau(x)$ is parametrized by $\beta \in \mathbb{R}^d$ via $\tau(x) = \psi(x)\beta$ with $\psi(x)$ being some set of basis functions $\psi : \mathcal{X} \to \mathbb{R}^d$. That setting allows for non-parametric relationships between X_i, W_i and Y_i but parametrizes the treatment effect function by β. This drawback of restricting linearity on a potentially complex treatment function $\tau(x)$ meant that this approach collected some dust until [Rob04] found a way to rewrite the heterogeneous treatment effect function $\tau(x)$ from [Rob88] as a loss minimizer. Under unconfoundedness as defined in section 2.1 and writing the conditional

surface responses as $\mu_w(x) = \mathbb{E}[Y_i^{(w)} \mid X_i = x]$ for $w \in \{0, 1\}$, we have

$$\mathbb{E}\left[\epsilon_i(w) \mid X_i, W_i\right] = 0 \text{ with } \epsilon_i(w) \coloneqq Y_i^{(w)} - \left(\mu_{(0)}(X_i) + w\tau(X_i)\right) \tag{3.2}$$

If we then follow [Rob88] and rewrite equation 3.1 using $m(x) = \mathbb{E}[Y \mid X = x] = \mu_{(0)}(X_i) + e(X_i)\tau(X_i)$ and $\epsilon_i \coloneqq \epsilon_i(W_i)$, we obtain

$$Y_i - m(X_i) = \left(W_i - e(X_i)\right)\tau(X_i) + \epsilon_i \tag{3.3}$$

We can rewrite this equation further yielding

$$\tau^\star(\cdot) = \text{argmin}_\tau \left\{ \mathbb{E}\left(\left[(Y_i - m^\star(X_i)) - (W_i - e^\star(X_i))\tau(X_i) \right]^2 \right) \right\} \tag{3.4}$$

This transformation is central as an oracle who knew $m^\star(x)$ and $e^\star(x)$ could estimate the heterogeneous treatement function $\tau^\star(x)$ via empirical loss minimization of equation 3.4. This is the starting position for machine learning algorithms for causal inference such as the R-learner that estimate $\hat{e}(x)$ and $\hat{m}(x)$ separately in a first step and minimize the empirical loss motivated by equation 3.4 via cross-fitting of $\hat{e}(x)$ and $\hat{m}(x)$ in a second step [NW21]. Causal Forests on the other hand take a more indirect approach that is motivated by a miss-specified partial linear model of equation 3.1. Suppose that the treatment effects are constant, meaning that $Y_i^{(w)} = f(X_i) + w\tau + \epsilon_i(w)$. Robinson's transformation yields

$$Y_i - m(X_i) = \left(W_i - e(X_i)\right)\tau + \epsilon_i \tag{3.5}$$

This allows for consistent estimation of the treatment effect parameter τ given that $\hat{e}(x)$ and $\hat{m}(x)$ are $o(n^{-1/4})$-consistent for m and e in root-mean-squared error. Furthermore the data has to be independent and identically distributed and there needs to be overlap as defined in section 2.1 [Rob88]. We then have

$$\hat{\tau} = \frac{\sum_{i=1}^n \left(Y_i - \hat{m}(X_i)\right)\left(W_i - \hat{e}(X_i)\right)}{\sum_{i=1}^n \left(W_i - \hat{e}(X_i)\right)^2} \tag{3.6}$$

Note that $\hat{\tau}$ is the estimated coefficient of $\{W_i - \hat{e}(X_i)\}$ in the simple regression with $Y_i - \hat{m}(X_i)$ as the response. This property lies behind the idea of

Causal Forests. On a high level, Causal Forests aim to create a partition of the covariate space $\mathcal{X}$ such that the assumption of constant treatment effects across observations in the resulting subspaces is sensible. Having created these subspaces via forest-based methods, Causal Forests aim to use equation 3.6 for treatment effect estimation.

Concretely speaking, for each $b = 1, \ldots, B$ a subsample $\mathcal{S}_b \subseteq \{1, \ldots, n\}$ is drawn. Then, B trees are grown via recursive partitioning, one on each subsample $\mathcal{S}_b$. We define $L_b(x)$ as the set of training samples falling in the same leaf as x. We can construct weights $\alpha_i(x)$ that quantify how often the ith training sample falls into the same leaf as the test point x as

$$\alpha_i(x) = B^{-1} \sum_{b=1}^{B} \frac{\mathbb{1}\{X_i \in L_b(x), i \in \mathcal{S}_b\}}{|\{i : X_i \in L_b(x), i \in \mathcal{S}_b\}|} \tag{3.7}$$

Lastly, we combine that set of weights $\{\alpha_i(x)\}_{i=1}^n$ and the consistent estimator of the constant treatment effects $\hat{\tau}$ from equation 3.6. That allows for heterogeneous treatment effect estimation which gives training samples that are similar to the test point x more weight. We receive the heterogeneous treatment effects from

$$\hat{\tau}(x) = \frac{\sum_{i=1}^n \alpha_i(x) \left(Y_i - \hat{m}^{(-i)}(X_i)\right) \left(W_i - \hat{e}^{(-i)}(X_i)\right)}{\sum_{i=1}^n \alpha_i(x) \left(W_i - \hat{e}^{(-i)}(X_i)\right)^2} \tag{3.8}$$

where the superscript $^{(-i)}$ denotes out-of-bag predictions, meaning that Y_i was not used to compute $\hat{m}^{(-i)}, \hat{e}^{(-i)}$. $\hat{m}(x)$ and $\hat{e}(x)$ are grown in a first step via separate regression and classification forests. The second step consists of building the Causal Forest using equation 3.8 with the set of weights $\{\alpha_i(x)\}_{i=1}^n$ whose construction shall be explained now.

What is left is to specify how exactly the B trees are grown via recursive partitioning. It is obvious that standard random forests should not be used. The partition of $\mathcal{X}$ induced by recursively splitting such that the sum-of-squared in-sample prediction errors are minimized does not create a partition where the treatment effects $\tau(x)$ can be expected to be constant across observations within a leaf. [AI16] first introduced a new splitting rule for heterogeneous treatment effects that was further developed in [WA18]. They proposed two approaches to recursively split the covariate space where one

approach was capable of handling treatment heterogeneity and one was suitable for situations with confounding present. Considering these drawbacks, [ATW19] developed what is now considered the best option for estimating treatment heterogeneity with the method described before.

To get into the details of the splitting procedure proposed by [ATW19], we have to start with a general description of Generalized Random Forests. Suppose we have access to data $(X_i, O_i) \in \mathcal{X} \times \mathcal{O}$ with $O_i = \{Y_i, W_i\}$ in the case of estimating heterogeneous treatment effects. We are interested in a quantity $\theta(x)$ that is identified via the local estimating equation which is given by

$$\mathbb{E}\Big[\psi_{\theta(x),\nu(x)}(O_i) \mid X_i = X\Big] = 0 \;\forall\; x \in \mathcal{X} \tag{3.9}$$

where $\psi(\cdot)$ is a scoring function and $\nu(\cdot)$ is a nuisance parameter[1]. The approach of [ATW19] aims to estimate solutions to equation 3.9 by minimizing its empirical similarity-weighted counterpart. Similarity is to measure the relevance of the ith training sample to fitting $\theta(\cdot)$ at x. Note that these similarities called $\alpha_i(x)$ are what we desire to estimate the heterogeneous treatment effects via equation 3.8. We obtain the similarity weights $\alpha_i(x)$ via the splitting procedure described later. In the framework of Generalized Random Forests, they allow to estimate the targets of interest via

$$\Big(\hat{\theta}(x), \hat{\nu}(x)\Big) \in \text{argmin}_{\theta,\nu} \left\{ \left\| \sum_{i=1}^{n} \alpha_i(x) \psi_{\theta(x),\nu(x)}(O_i) \right\|_2 \right\} \tag{3.10}$$

Define the following notation to describe the splitting procedure: let every split start at a parent node $\mathcal{P} \subseteq X$. Given a subsample of data $\mathcal{S}$, we denote $(\hat{\theta}_P, \hat{\nu}_P)(\mathcal{S})$ as the solution to the estimating equation, meaning

$$\Big(\hat{\theta}_P, \hat{\nu}_P\Big)(\mathcal{S}) \in \text{argmin}_{\theta,\nu} \left\{ \left\| \sum_{\{i \in \mathcal{S}: X_i \in P\}} \psi_{\theta,\nu}(O_i) \right\|_2 \right\} \tag{3.11}$$

The algorithm proceeds greedily as proposed by [Bre01] and divides P into

[1] This formulation accommodates a multitude of statistical problems. Let $f_{\theta(x),\nu(x)}$ be the distribution of O_i conditional on X_i. Then, equation 3.9 with $\psi_{\theta(x),\nu(x)}(O_i) = \nabla \ln(f_{\theta(x),\nu(x)}(O_i))$ identifies the local maximum likelihood parameters [ATW19].

two children C_1, C_2 using an axis-aligned split that maximizes the following criterion

$$\Delta(C_1, C_2) := \frac{n_{C_1} n_{C_2}}{n_P^2} \left(\hat{\theta}_{C_1}(\mathcal{S}) - \hat{\theta}_{C_2}(\mathcal{S}) \right)^2 \tag{3.12}$$

given $n_L = |\{i \in \mathcal{S} : X_i \in L\}|,\ L \in \{P, C_1, C_2\}$ [ATW19]. This essentially corresponds to making the heterogeneity of in-sample $\hat{\theta}$ as large as possible. As the computational burden of optimizing $\Delta(C_1, C_2)$ over all candidate splits while solving equation 3.11 for all candidate tuples C_1, C_2 is high, [ATW19] propose an approximate criterion $\tilde{\Delta}(C_1, C_2)$ that uses a gradient-based approximation for solving equation 3.11. Once the trees $b = 1, \ldots, B$ have reached a standard stopping criterion, one can compute the forest-based similarity weights via aggregation of the tree-based weights as given in equation 3.7.

3.2 Inference for Causal Forests

Being able to conduct statistical inference on the estimated treatment effects as described in section 3.1 is crucial. [WA18] provide asymptotic theory for predictors resulting from averages over trees based on U-statistics [Hoe48]. They require their trees to fulfill four conditions to make the asymptotic theory work. Most importantly, the trees need to be honest, meaning that a tree grown on a training sample $((X_1, Y_1), \ldots, (X_s, Y_s))$ does not use the responses $Y_1, \ldots, Y_s$ when choosing where to place the splits while partitioning. The other constrains are more technical and can be found in [WA18]. Under relatively weak assumptions and writing $\mu(x) = \mathbb{E}[Y \mid X = x]$, they show

$$\frac{\mu_n(x) - \mu(x)}{\sigma_n(x)} \xrightarrow{d} \mathcal{N}(0, 1) \text{ for a sequence } \sigma_n \longrightarrow 0 \tag{3.13}$$

where $\xrightarrow{d}$ stands for convergence in distribution. [ATW19] impose the same restrictions on their trees as [WA18]. Further, they manage to restate their approach as a pseudo-forest to make estimates $\hat{\theta}(x)$ from generalized random forest an average of estimates over different trees enabling them to use the asymptotic theory established by [WA18]. In doing so, they show that equation 3.13 holds for $\hat{\theta}(x)$ as well.

Further, they develop a method of constructing asymptotically valid confidence intervals for $\theta(x)$ centered around $\hat{\theta}(x)$ showing that $\lim_{n\to\infty} \mathbb{E}[\theta(x) \in (\hat{\theta}(x) \pm \Phi^{-1}(1-\frac{\alpha}{2})\hat{\sigma}_n(x))] = \alpha$. Deriving a consistent estimator for $\sigma_n(x)$ using the Delta method completes the construction of the confidence intervals.

3.3 Causal Forests and Panel Data

In light of the application part from section 5 where we estimate treatment effects from panel data, we shortly describe the nature of panel data and its emerging challenges. Further, we present the properties of Causal Forests that make them suitable for such as setting. Note that we examine these claims carefully in the following simulation study in section 4.

Going with the notation of [ATW19], suppose we have access to $(X_{it}, O_{it}) \in \mathcal{X} \times \mathcal{O}$ with $O_{it} = \{Y_{it}, W_{it}\}$, $i = 1, \ldots, n$ and $t = 1, \ldots, T$. Put simply, there are observations of the same n entities over T time periods. Causal Forests are however developed for independent and identically distributed data. Yet they exhibit properties that enable integration of the correlation-structures induced within each entity and over time into the estimation of treatment effects.

Firstly, we have the possibility of adjusting for within-entity correlations by using cluster-robust Causal Forests, which [ATW19] provide in the R-package `grf`. They explain that the possibility of clustering Causal Forests allows for a non-parametric random effects modeling [BFJ19]. To achieve this, the algorithm is adapted as follows. Assuming that there are J clusters, draw a subsample of clusters $\mathcal{J}_b \subseteq \{1, \ldots, J\}$ for each tree $b = 1, \ldots, B$ and construct $\mathcal{S}_b$ by drawing k samples at random from each cluster $j \in \mathcal{J}_b$. The splitting scheme remains unchanged. Lastly, an observation i is only considered to be out-of-bag if its cluster was not drawn when subsampling the clusters.

Secondly, we need to address the correlation-structure that is introduced through time. Random forests have been extensively used in panel data settings. [GKX20] provide an extensive analysis showing that tree-based methods are well-suited for forecasting stock returns from panel data. Further, we can include variables that seeks to capture the effect of time such as lagged variables and time-indicators.

4 Monte Carlo Simulation

In order to get an understanding for the performance of Causal Forests in the setting of panel data, I run an extensive simulation study. The specific aim of this section is to understand under which circumstances Causal Forests work well and under which circumstances work less reliably. To do this, I impose different variations on the data generating process. Further, I experiment with cluster-robust Causal Forests and study their advantage over standard Causal Forests. The Causal Forest algorithm as described in [ATW19] is used which is implemented in the R-package `grf`.

4.1 Data Generating Process

The focus of the simulation lies on three distinct difficulties concerning the estimation of treatment effects from panel data. We describe the data as $(X_{it}, Y_{it}, W_{it}), i = 1, \dots, n$ and $t = 1, \dots, T$ as observations of the same n entities over T time periods. Given this structure, we want to examine how well Causal Forests handle correlation-structures within entities. Additionally, we are interested in the algorithms ability to deal with correlation-structures induced over time. Lastly, the interest lies on understanding how well Causal Forests can withstand unobserved confounding variables.

Define the following functions used in the data generating process:

$$\begin{aligned} &\text{main effect} : m(x) = 2^{-1}\, \mathbb{E}\Big[Y^{(0)} + Y^{(1)}\Big] \\ &\text{treatment effect} : \tau(x) = \mathbb{E}\Big[Y^{(1)} - Y^{(0)} \mid X = x\Big] \\ &\text{treatment propensity} : e(x) = \mathbb{P}\Big[W = 1 \mid X = x\Big] \end{aligned} \tag{4.1}$$

The baseline data generating process is inspired by [WA18] which in turn draws from [Rob88] where the outcome Y_{it} can be decomposed into a main effect $m(X_{it})$ and a treatment effect $\tau(X_{it})W_{it}$ as well as an error term ϵ_{it}. We dispense with using the subscripts $(_{it})$ for the remainder of this section to improve readability whenever possible. The features are drawn from a uniform distribution meaning $X \sim \text{Uniform}([0,1]^d)$. The treatment assignment follows a binomial distribution with treatment propensity $e(X)$ mean-

ing $W \sim \text{Bern}(e(X))$. The treatment propensity $e(X)$ is set according to [WA18]. They use the β-density with shape parameters $\{2, 4\}$ to generate $e(X) = \frac{1}{4}(1 + \beta_{2,4}(X_1))$. To ensure that the baseline assumption of overlap holds, meaning that for some $\epsilon > 0$ and all $x \in [0,1]^d$, we have $\epsilon < \mathbb{P}(W = 1 \mid X = x) < 1 - \epsilon$, the treatment propensity $e(X)$ is multiplied by some constant $\kappa = 0.05$. This is not problematic as the assumption of overlap is testable in all applications. We additionally generate a persistent treatment vector per entity i meaning that if $W_{it} = 1 \implies W_{it+p} = 1 \; \forall p \geq 1$. We implement the treatment variable in this fashion as this represents the treatment assignment of introducing the stricter facial mask policy in section 5. The within entity correlation-structure is introduced via entity fixed effects $\alpha_i \sim \mathcal{N}(0, \sigma_\alpha^2)$ while the correlation structure over time is generated by the error term u_{it} following an AR(1) process. The specification of the main effect $m(X) = X_1 + X_3$ is chosen to be simple and is inspired by [WA18] who use $m(X) = 2X_1 - 1$. The treatment effect function $\tau(X)$ is a smooth function given by $\tau(X) = 1 + \frac{1}{1+\exp(-20(X_1 - 1/3))}$ which also roots in [WA18]. They use a slightly more complex treatment function $\tau(X)$ as they are investigating the comparative performance of Causal Forests against nearest-neighbor approaches and they thus want to understand which method reigns superior in learning a complex estimand. Putting everything together, we get:

$$
\begin{aligned}
&Y_{it} = m(X_{it}) + \tau(X_{it}) W_{it} + \epsilon_{it} \\
&X_{it} \sim \text{Uniform}([0,1]^d), \text{ where } d \text{ is the dimension of the feature space } \mathcal{X} \\
&m(X_{it}) = X_{1it} + X_{3it} \\
&\tau(X_{it}) = 1 + \frac{1}{1 + \exp(-20(X_{1it} - 1/3))} \\
&e(X_{it}) = \frac{1}{4}\left(1 + \beta_{2,4}(X_{1it})\right)\kappa, \; \kappa = 0.05 \\
&W_{it} \sim \text{Bern}\left(e(X_{it})\right) \text{ with } W_{it} = 1 \implies W_{it+p} = 1 \; \forall p \geq 1 \\
&\epsilon_{it} = u_{it} + \alpha_i \text{ with } u_{it} = \rho u_{it-1} + \nu_{it} \text{ and } \{\alpha_i, \nu_{it}\} \sim \mathcal{N}(0, \sigma_x^2), x \in \{\alpha, \nu\}
\end{aligned}
\tag{4.2}
$$

As discussed before, we want to test the performance of Causal Forests in the presence of confounding variables. There are two distinguished channels of confounding concerning treatment effect estimation in observational data being

observed and unobserved confounding variables [GRP99]. We will experiment with both types of confounding. Given access to the causal Bayesian network of the data generating process, we define confounding as follows [DAm20]

Definition 4.1 *Let (G, P) with $G = (\boldsymbol{V}, \boldsymbol{E})$ be a causal Bayesian network with $(i, k) \in \boldsymbol{V}, i \neq k$ and there is a directed path from i to k. Then the causal effect from i to k is confounded if $p(x_k \mid x_i) \neq p(x_k \mid do(x_i))$*

The first form of confounding that needs to be present in the simulated data is observed confounding. We achieve this via the interaction of $m(X)$ and $e(X)$ as they both depend on the feature X_1. It follows that the treatment assignment and the potential outcomes $\{Y^{(1)}, Y^{(0)}\}$ are dependent. This structure is present in all our experiments as observed confounding is certainly present in the data from section 5 and we thus want to understand how well Causal Forests control for these confounders.

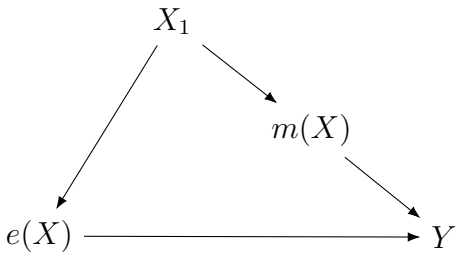

Figure 1: Graph of causal bayesian network with observed confounding variable X_1. Note that there are two directed paths being $\{X_1 \to e(X) \to Y\}$ and $\{X_1 \to m(X) \to Y\}$ from X_1 to Y which means that $p(y \mid x_1) \neq p(y \mid do(x_1))$

The second issue concerning confounding in trying to estimate treatment effects from observational panel data is unobserved confounding. Unobserved confounding is in principle the same as observed confounding but with the confounding variable not being measured. Note that the presence of unmeasured confounding generally impedes identification and we do not claim that Causal Forests bypass that restriction. However, we are interested in gauging the robustness of the performance of Causal Forests under unmeasured confounding [VWB19].

4.2 Delay of Treatment Effects

It is unreasonable to assume that the assignment of treatment has a contemporaneous effect on the response variable. Proposing that the assignment of treatment of entity i at time t which is given by $W_{it} = 1$ if the canton decides to enforce the stricter facial mask policy at time t has a direct influence of the response $Y_{it} = \ln \mathrm{R}_{e,i,t}$ is inaccurate. A delay has to be included as changes in the dynamics of COVID-19 occur with delay due to the incubation period and due to delay of behavioral alternations [TB19]. There is a growing body of literature that is concerned with estimating these dynamic treatment effects [SA21]. The most common approach can be written as

$$Y_{it} = \alpha_i + \lambda_t + \sum_l \mu_l \mathbb{1}\{t - E_i = l\} + \nu_{it} \tag{4.3}$$

where α_i and λ_t are the entity and time fixed-effects while l denotes the choice of lag and E_i is the time period when entity i initially receives the binary treatment. As Causal Forests aim to identify the treatment effects by partitioning the covariate space $\mathcal{X}$ such that observations within a terminal leaf can be assumed to exhibit constant treatment effects, this approach is not suitable. Adding the indicators associated with $\mathbb{1}\{t-E_i = l\}$ to the design matrix would not work as one treatment vector has to be specified for Causal Forests. We approach the issue in the fashion of [CKS21] who propose lagging the data allowing the covariates and the treatment to enter the response variable as follows

$$Y_{it} = m(X_{it-l}) + \tau(X_{it-l})W_{it-l} + \epsilon_{it} \tag{4.4}$$

With the right choice of lag l, this approach might work very well in an application where the RHS of equation 4.4 influences the response through some complicated dynamical system as is the case in the empirical application of section 5. Plainly lagging $\{X_{it}, W_{it}\}$ in the simulation study is however going to produce

4.3 Performance Evaluation

The performance of the algorithm is evaluated in terms of the MSE and the Bias for estimating $\tau(X)$ as well as the expected coverage of $\tau(X)$ with a targeted coverage rate of 0.95. We report both the MSE and Bias as it allows us to understand the composition of the MSE via the Bias-Variance decomposition[2] [Dom00]. Let S be the number of simulated datasets and n the number of observations. Further, let $\mathbb{1}_i = 1$ if the 95% confidence interval for observation i contains the true parameter τ_i and $\mathbb{1}_i = 0$ otherwise. For each of the S datasets, we compute the following measures

$$\text{MSE}_s(\hat{\tau}, \tau) = n^{-1} \sum_{i=1}^{n} (\hat{\tau}_i - \tau_i)^2$$

$$\text{Bias}_s(\hat{\tau}, \tau) = n^{-1} \sum_{i=1}^{n} \hat{\tau}_i - \tau_i$$

$$\text{Coverage}_s(\hat{\tau}, \tau) = n^{-1} \sum_{i=1}^{n} \mathbb{1}_i$$

We aggregate the results over S datasets by averaging. For the MSE for example, we compute $\text{MSE}(\hat{\tau}, \tau) = S^{-1} \sum_{s=1}^{S} \text{MSE}_s(\hat{\tau}, \tau)$. The same applies for the Bias and the Coverage. We compute and report these measures for both the CATE as well as the ATE.

4.4 Results

After constructing the data, we lag the main effect and treatment effect by a lag length of $l = 7$ days[3]. In doing so, we mimic the properties of the data from the empirical application from section 5 as the response of COVID-19 in terms of new infections is delayed due to the incubation period and delay in behavioral changes [Che+21]. The dimension of the feature space d varies for all simulation setups with $d \in \{3, 4, 5, 6, 7, 8, 24\}$ while $d = 24$ represents the dimension of the design matrix from the application part. The number of observations varies while we hold the number of entities $\mathcal{I} = |i \in \{1, \ldots, I\}|$ constant at 10 and change the number of time periods $\mathcal{T} = |t \in \{1, \ldots, T\}|$

[2] The Bias-Variance decomposition of an estimator is given by $\text{MSE} = \text{Bias}^2 + \text{Variance}$

[3] Concretely, we run Causal Forests on $Y_{it} = m(X_{it-l}) + \tau(X_{it-l})W_{it-l} + \epsilon_{it}$

with $T \in \{200, 1000\}$. The coefficient of the autocorrelated error term $\{u_{it}\}$ is $\rho = 0.2$. We also implement $\{\alpha_i, \nu_{it}\} \sim \mathcal{N}(0, 1)$. Lastly, the number of simulations per configuration is $S = 100$.

In the following, we describe the two simulation designs and discuss the results.

4.4.1 Design 1: The Value of Clustering

The data is generated according to the description in equation 4.2. Note that the error term of the simulated data $\epsilon_{it} = u_{it} + \alpha_i$ is comprised of an autocorrelated error process u_{it} and an entity fixed effect α_i that captures the effect of the entity i on the outcome Y_i. To gauge the ability of Causal Forests of capturing these entity fixed effects α_i, we employ the algorithm on the data described once while using cluster-robust Causal Forests as described in section 3.3 and once while not taking these clusters into account. Comparing the performances allows us to judge the value of clustering.

We observe comparable performances regarding the Bias as well as the MSE for both the CATE and the ATE but cluster-robust Causal Forests reign supreme on coverage. This lies in accordance with our expectations as the ability to cluster allows for cluster-robust inference where the within-cluster correlations are taken into account when estimating the standard errors for the heterogeneous treatment effects [AW19]. We also note that the overall performance of clustered Causal Forests is very satisfactory reaching an equally-weighted average coverage rate over all combinations of $\{d, T\}$ of 89% for the CATE and 86% for the ATE. For detailed results of the comparison, check out table 2 in appendix A.1.

As clustering for entities improves the performance considerably, we work with cluster-robust Causal Forests for the remainder of the simulation study as well as for the empirical application.

4.4.2 Design 2: The Influence of Unobserved Confounding

Given that Causal Forests seem to handle data with observed confounding present adequately, we want to test the performance when there is an unobserved confounding variable present in the data.

We generate the data as described in equation 4.1 while deviating in regard to two aspects. Firstly, we make the treatment propensity $e(X)$ dependent on a variable that is not used in the estimation. Using $\bar{X}_{12it} = 2^{-1}(X_{1it} + X_{2it})$, we compute the treatment propensities as

$$e(X_{it}) = \frac{1}{4}(1 + \beta_{2,4}(\bar{X}_{12it}))\kappa, \ \kappa = 0.05 \tag{4.5}$$

This means that the treatment propensity is a function of the average of X_{1it} and X_{2it}. Secondly, we make the variables X_1 and X_2 dependent. This is achieved by drawing $X \sim \text{Uniform}([0,1]^d)$ as usual and subsequently setting $\tilde{X}_1 = X_1 + 0.5X_2$. Once the data is simulated as described, X_2 is deleted reducing the number of features by 1 resulting in the design matrix $\boldsymbol{X}_{-X_2}$. Furthermore, X_1 is replaced by $\tilde{X}_1$. In generating the data in this fashion, we create a confounding variable in X_2 as defined in definition 4.1.

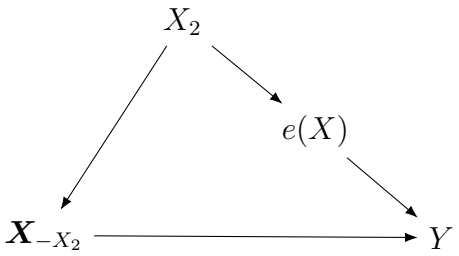

Figure 2: Graph of causal bayesian network with unobserved X_2. X_2 is a classical confounder in this structure.

We now present the results for this design in more detail as it resembles the data from the empirical application most precisely. We first start by noting that $\tau(X)$ is bounded from below by 1 and bounded from above by 2 which allows us to relate the criteria MSE and Bias to the treatment effect $\tau(X)$[4].

The performance regarding MSE and Bias is sufficient which can be seen in table 3. We observe that the performance does not seem to vary considerably along d and T. The coverage rate falls short of the targeted coverage rate of 0.95. We reach an equally-weighted average coverage rate over all combinations of $\{d, T\}$ of 89% for the CATE and 88% for the ATE. Considering that we

[4]Remember that $X_1 = X_1 + 0.5X_2$ where $X_1, X_2 \in [0,1]$. We see that $\tau(0) = 1$ and $\tau(1.5) = 2$. Noticing that $\tau(X)$ is an increasing function yields the desired upper and lower bounds

simulate data containing entity fixed effects, an autocorrelated error term, a complex treatment function $\tau(X)$, unobserved and observed confounding and we estimate heterogeneous treatment effects on data lagged by $l = 7$ days, we consider the performance a big success.

Apart from the classical performance measures as introduced in section 4.3, we additionally investigate the performance using methods specifically designed for Causal Forests as described in sections 5.4.1 and 5.4.2. To achieve this, we use data generated from design 2 while choosing $d = 24$, $T = 200$ and 10 entities. We chose this combination as it most closely represents the data from the empirical application. We generate $S = 100$ datasets. The calibration regression explained in section 5.4.2 is estimated for all $S = 100$ runs. In figure 6 we see that $\hat{\alpha}_s$ are centered around 0.999 with 95% of the mass being between 0.886 and 1.172. As all the p-values of the one-sided hypothesis tests of $\hat{\alpha}_s < 0$ for the $S = 100$ runs lie within $[0, 0.0041]$, we can conclude that the ATE is well estimated. The heterogeneous treatment effects on the other side are not well calibrated with the $\hat{\beta}_s$ being far from 1. That either stems from the design of the treatment effect function $\tau(X)$ that is constant for all $X_1 \in [1, 1.5]$ or from the inability of Causal Forests to accurately estimate heterogeneous treatment effects in panel data. To distinguish these two channels, we run the proposed procedure with $X_1 \in [0, 1]$ such that we do not have a largely constant treatment effect function. In doing so, we observe the same pattern as before. According to the calibration regressions, Causal Forests accurately estimate the average treatment effect but fail to pick up the treatment heterogeneity. We keep that in mind for the empirical application in section 5 as it provides evidence that the calibration regressions misrepresent the Causal Forests ability to estimate heterogeneous treatment effects from from panel data. This conclusion can be drawn as there are large discrepancies when comparing the performance of estimating heterogeneous treatment effects using the criteria from section 4.3 and the calibration regressions. The detailed results can be found in appendix A.1.

5 Empirical Application of Causal Forests: Effectiveness of Facial Mask Policies in Switzerland on Containing the COVID-19 Pandemic

The great uncertainty caused by COVID-19 poses large difficulties for all societies around the world. In close collaboration with science, politicians are trying to balance containing the spread of COVID-19 and returning back to life as it was before COVID-19. In doing so, they have to make decisions based on scarce information. Quantifiying the effect of containment measures is therefore central as it allows for a better understanding of the effectiveness of policies which is crucial for the remainder of the pandemic.

5.1 Introduction

After the country-wide lockdown in Switzerland that ended on June 19, 2020 ,each of the 26 cantons that make up Switzerland was given political autonomy concerning the introduction of COVID-19 containment measures. As a consequence, we observe strong cantonal heterogeneity in how the situation was handled. We use this heterogeneity to infer the effect of making facial masks mandatory in all public or shared spaces on slowing down the spread of COVID-19 in Switzerland.

The contribution of this paper lies firstly in estimating the effect of facial mask policies on reducing the spread of COVID-19 which we measure through the effective reproductive number denoted by $R_{e,i,t}$. Secondly, we apply Causal Forests in a scenario where methods such as Differences in Differences [BDM04] or Synthetic Control Methods [AG03] are usually employed. We show that Causal Forests provide a viable alternative providing convincing estimates.

The results show that the introduction of making facial masks mandatory in all public or shared spaces has a significant average treatment effect of ATE $= -0.044$ on $\ln R_{e,i,t}$. This corresponds to lowering the effective reproductive number by 5% which is explained in more detail in section 5.4.2. Causal Forests enable the estimation of a treatment effect per canton and day. We do not find significant treatment heterogeneity. That means that neither the canton in which the policy was introduced nor the time when the policy was

introduced significantly influences the treatment effect.

The time period of interest for the analysis is restricted to the period from August 21, 2020 to October 19, 2020. Starting on June 20, 2020, the cantons were allowed to tighten the policies that were put in place country-wide. On August 21, 2020, canton Neuenburg was the first canton to enforce a stricter facial mask policy. Over the course of the next two months, 11 out of the 26 cantons implemented a stricter facial mask policy than the lower bound determined by the federal government. On October 19, 2020, the federal government established multiple country-wide policies, including a tighter facial mask policy. See appendix A.2 for more details.

In the next section, we present related literature. The data is presented in section 5.3 and the results in section 5.4.

5.2 Related Literature

5.3 Data

The identification of heterogeneous treatment effects relies heavily on the forests ability to create a sensible partition of the feature space $\mathcal{X}$. This is crucial, as we want to interpret nearby observations in x-space as having come from a randomized experiment [WA18]. Therefore, we construct a rich data set containing different types of information. The data can be classified into four categories being epidemological data 5.3.1, baseline cantonal data 5.3.2, behaviorial data 5.3.3 and policy data 5.3.4. We present all types of data and document where they can be accessed. A concise overview of the data used and where it can be accessed can be found in appendix A.2. We also include a complete composition of the variables used in appendix A.2. Note that most of our data is at the cantonal level and at a daily frequency to suit the analysis.

5.3.1 Epidemological Data

To accurately represent the epidemological situation, we construct a dataset containing information concering COVID-19 cases, capacity and utilization of the health case system as well as $R_{e,i,t}$ and related measures[5].

[5]The data is published by the Federal Office of Public Health FOPH: https://opendata.swiss/en/dataset/covid-19-schweiz

The logarithm of the effective reproductive number $\mathrm{R}_{e,i,t}$ is used as the response variable. Using NINF_t as the number of newly infected people on day t, we write

$$R_e(t) = R_{e,t} = \frac{\mathrm{NINF}_t}{\sum_{s=1}^{t} \mathrm{NINF}_{t-s} w_s} \tag{5.1}$$

with w_s being the value of the infectivity profile s days after infection which measures how infective a individual is relative to the beginning of their symptoms [PSS21] [Ash+20]. [PSS21] show that the growth rate of new infections can be approximated by $\ln \mathrm{R}_{e,i,t}$ which is why its very natural to use it as the response.

In order to capture the delay from the time of infection to the time of reporting of the case, we work with a lag of $l = 7$ days for the main effect as well as the treatment effect, meaning

$$Y_{it} = m(X_{it-l}) + \tau(X_{it-l}) W_{it-l} + \epsilon_{it} \tag{5.2}$$

Using $l = 7$ days lies in accordance with estimates of the incubation period and further with the delay of political measures on behavior changes of people [Che+21]. We work with $l \in \{5, 7, 8, 10, 14\}$ and show that the results are robust to the choice of l in appendix A.2.

Furthermore, we add a feature that indicates the week an observation at time t lies in to capture a time trend that might otherwise be unaccounted for.

5.3.2 Baseline Cantonal Data

We gather data on baseline differences between the cantons that are not directly linked to COVID-19. Information on the population and demographics, weather data and information about public holidays are assembled[6] [7] [8].

We collect the percentage of people over aged over 80 years old, the settlement area as well as the population density. Further, the data contains

[6] The data on population and demographics is from https://github.com/daenuprobst/covid19-cases-switzerland

[7] The weather data is published by the Federal Office of Meteorology and Climatology MeteoSwiss: https://opendata.swiss/en/dataset/klimamessnetz-tageswerte

[8] The holiday data is gathered from https://www.edk.ch/en/education-system/websites-of-the-cantons

information about the number of people over 65 years old as well as the total population. These features are important as evidence shows that infection growth is strongly linked to measures of residential density [WK20].

Further, [Zor+20] suggest that weather conditions are closely linked to infection growth of COVID-19. They find in particular that dry air supports the transmission of the COVID-19 virus[9]. To seize these relationships, we collect weather data from 14 weather stations. We solely use weather stations that represent the cantons weather accurately meaning we exclude stations situated on mountains. Next, we match each of the 26 cantons to the next weather station to get an accurate characterization of each cantons weather.

The holiday indicator is constructed to have value 1 if the majority of public schools are on holiday and 0 otherwise. There are some small variations within the cantons and mostly between primary schools and secondary schools but the discrepancies in the time period of interest are small.

Another important factor in the composition of $\ln \mathrm{R}_{e,i,t}$ is geographical proximity to an area of large growth of COVID-19 cases. [Kap20] find that there are distinct spatial patterns in the dynamics of COVID-19 which means that geographical proximity of cantons has to be modeled. As we need a cantonal indicator to run cluster-robust Causal Forests, we seek to build an indicator with 26 distinct values that represent spatial proximity of the cantons. To do this, we employ classical MDS. First, the euclidian distance matrix $\boldsymbol{D} = \{d_{ij}\}_{i,j=1}^{26}$ between all the cantons main cities is computed[10]. We then apply classical MDS on $\boldsymbol{D}$ to project $\boldsymbol{D}$ from $\mathbb{R}^2 \to \mathbb{R}$ while preserving the euclidian distances as best as possible. Lastly, we sort the elements of the indicator vector $\boldsymbol{d}$ and replace the elements $\{d_i\}_{i=1}^{26}$ with the sequence of natural numbers from 1 to 26 preserving the previous sorting.

[9] Their findings find a lot of support: [Zhu+20], [FR20]

[10] Note that the distance between two points on a sphere is more accurately represented by the length of the connecting geodesic but euclidian distances are sufficient here due to short distances [Rap91]

5.3.3 Behavioral Data

We collect data on household spending as well as mobility data to quantify the behavior of the population during the pandemic[11] [12].

To obtain a measurement of household spending we use the transactions from credit and debit cards as well as from mobiles. We include transactions where the origin of the cardholder can either be domestic or foreign. As we are working with cantonal data and absolute numbers do not interest us but rather growth rates, we compute the daily growth rate of the number of transactions as well as the daily growth rate of the amount spent in CHF per canton. [PSS21] argue that consumption can be seen as a proxy for the level of social distancing by the population as we do not include e-commerce transactions.

Changes in behavior can also be identified trough changes in mobility. The data is geographically structured along language giving us three distinct time series. We match each canton to the corresponding geographical region and compute the daily growth rate of the median distance traveled per person and day in km.

5.3.4 Policy Data

We use the policy data collected by the KOF Swiss Economic Institute to account for the different COVID-19 policies put in place by the cantonal governments[13].

The approach that is implemented by the KOF closely follows the Oxford Stringency index [PSS21] [Hal+21]. The idea of the stringency index is to combine a set of COVID-19 containment policies into an index that represents the stringency of a governments COVID-19 policies. The Oxford Stringency index is only computed at the country level. The KOF Stringency-Plus index

[11] The data on household spending is published by Monitoring Consumption Switzerland: https://monitoringconsumption.com/

[12] The data on mobility is published by KOF Swiss Economic Institue: https://kofdata.netlify.app/#/

[13] The data on the policies is published by KOF Swiss Economic Institute: https://kofdata.netlify.app/#/

(KSI$^+$) is constructed at the cantonal level and is computed as

$$\text{KSI}^+ = 10^{-1} \sum_{j=1}^{10} \left(100 \frac{v_{j,t}}{N_j} \right) \in [0, 100] \tag{5.3}$$

where $v_{j,t}$ is the policy value of indicator j on day t while N_j is the maximum value of policy indicator j. The 10 indicators that make up the KSI$^+$ are facial coverings, school closing, workplace closing, cancellation of public events, restrictions on gatherings, closure of public transport, stay-at-home requirements, restrictions on internal movement, international travel controls and public info campaign [PSS21]. Compared to the Oxford Stringency index, they add the 10^{th} indicator facial coverings and make a slight modification to the coding of workplace closing[14]. We exclude indicators from the analysis that neither vary across cantons over the time period of analysis being $t_1 =$ August 19, 2020 to $T =$ October 19, 2020 nor change at the national level during that time period. That leaves us with the three indicators being cancellation of public events, restrictions on gatherings as well as facial coverings.

5.4 Results

All our models are estimated using Causal Forests from the R package `grf` while clustering on canton. We use the design matrix $\boldsymbol{X}$ with $\dim(\boldsymbol{X}) = 1612 \times 24$ described in appendix A.2. Given that there are 26 cantons, we end up with data for $T = 60$ days. We dispense with preselection of variables by first running Causal Forests to determine feature importance as described by [AW19] as it unnecessarily reduces the number of features while not improving performance in our case[15].

In the following, we first evaluate the estimated average treatment effect. Causal Forests allow for simultaneous estimation of the ATE, ATT as well as for an overlap-weighted ATE. In section 5.4.3 we search for heterogeneity in the treatment effects and examine if there is a feature that is predictive of heterogeneity.

[14] The coding of the indicators is explained here: https://github.com/OxCGRT/covid-policy-tracker/blob/master/documentation/codebook.md

[15] We show in the appendix A.2 that our results are robust to preselection

5.4.1 Assumptions

The identifying assumption of overlap as discussed in section 4.1 can be graphically evaluated. Remeber that overlap states that for some $\epsilon > 0$ and all $x \in [0,1]^d$, we have

$$\epsilon < \mathbb{P}(W = 1 \mid X = x) < 1 - \epsilon \tag{5.4}$$

The two-stage approach of Causal Forests as discussed in section 3.1 means that we have access to an estimate of $e(X) = \mathbb{P}(W = 1 \mid X = x)$. By looking at the histogram of $\hat{e}(X)$ in figure 9, we can check if $\hat{e}(X)$ is sufficiently bounded away from 0 and 1. The estimated treatment propensities are not centered around 0.5 as one would for example expect from data coming from a randomized control trial. But given $\min(\hat{e}(X)) = 0.045$ and $\max(\hat{e}(X)) = 0.880$, we can conclude that the assumption of overlap is met.

5.4.2 Average Treatment Effect

Understanding the average effect of making facial masks mandatory in all public or shared spaces in Switzerland in the time period of analysis on the growth rate of new infections is of central interest. We hence first abstract from the ability to estimate heterogeneous treatment effects and estimate the average treatment effect defined as τ. [AW19] define the aggregation of the estimated ATE $\hat{\tau}$ under clustering as

$$\begin{aligned}
\hat{\Gamma}_i &= \hat{\tau}^{(-i)}(X_i) + \frac{W_i - \hat{e}^{(-i)}(X_i)}{\hat{e}^{(-i)}(X_i)(1 - \hat{e}^{(-i)}(X_i))} \\
&\quad \times \left(Y_i - \hat{m}^{(-i)}(X_i) - \left(W_i - \hat{e}^{(-i)}(X_i) \right) \hat{\tau}^{(-i)}(X_i) \right) \\
\hat{\tau}_j &= n_j^{-1} \sum_{\{i:A_i=j\}} \hat{\Gamma}_i \\
\hat{\tau} &= J^{-1} \sum_{j=1}^{J} \hat{\tau}_j \text{ and } \hat{\sigma}^2 = \frac{1}{J(J-1)} \sum_{j=1}^{J} (\hat{\tau}_j - \hat{\tau})^2
\end{aligned} \tag{5.5}$$

where we categorize observations into cantons $A_i \in \{1, \ldots, J\}$ with $J = 26$. This is an augmented inverse-propensity weighted average treatment effect

where $\hat{\Gamma}_i$ are the augmented inverse-propensity weighted treatment effects per canton and day. $\hat{\tau}_j$ is the average treatment effect for canton j [Far15]. Running everything as described above, we obtain $\hat{\tau} = -0.044$ with $\hat{\sigma}^2_{\hat{\tau}} = 0.021$. Using the theory set forth in section 3.1 which states that predictions from Causal Forests are asymptotically Gaussian and unbiased allows us to perform hypothesis testing. We can reject the null hypothesis $H_0 : \hat{\tau} = 0$ in a two-sided hypothesis test at the 5% confidence level[16]. This suggests that the introduction of the stricter facial mask policy compared to the lower bound defined by the national government in Switzerland was effective in containing the spread of COVID-19 in the period of analysis.

The magnitude of the average treatment effect is considerable as well. An estimated average treatment effect of $\hat{\tau} = -0.044$ on the response variable $\ln \mathrm{R}_{e,i,t}$ implies $\ln \mathrm{R}_{e,i,t,Treated} - \ln \mathrm{R}_{e,i,t,Control} = -0.044 \iff \frac{\mathrm{R}_{e,i,t,Treated}}{\mathrm{R}_{e,i,t,Control}} = \exp(-0.044) = 0.95$. That implies that we estimate that the daily effective reproductive number of cantons that introduce the stricter facial mask policy is 5% lower compared to cantons that do not introduce the policy. The effective reproductive number is closely related to the basic reproductive number R_0 via $\mathrm{R}_e = \frac{S(t)}{N}\mathrm{R}_0$ where $S(t)$ is the number of people susceptible at time t while N is the total population. Values of R_0 correspond to exponential spreading of infectious diseases while values of $\mathrm{R}_0 < 1$ are associated a decline in the spreading of the disease [TB19]. All cantons apart from Schaffhausen display effective reproductive numbers $\mathrm{R}_e < 1$ for some period and $\mathrm{R}_e > 1$ for some other period during the period of analysis which can be seen in figure 10. Going with the prediction from epidemiological models, the proposed 5% reduction of the effective reproductive number has important consequences as this reduction can delay or impede exponential spreading of COVID-19.

Having presented the estimated average treatment effect and its implications, we henceforth investigate the credibility of the estimate. Firstly, we use an approach motivated by [Che+18]. They propose a best linear predictor approach allowing to judge the accuracy of the estimated average treatment effect in models where CATE are estimated. For Causal Forests, this translates

[16] The analysis for the ATT and overlap-weighted ATE can be found in the appendix A.2

into running the following calibration regression

$$Y_i - \hat{m}^{(-i)}(X_i) = \alpha \underbrace{\bar{\tau}(W_i - \hat{e}^{(-i)}(X_i))}_{C_i} + \beta \underbrace{(\hat{\tau}^{(-i)}(X_i) - \bar{\tau})(W_i - \hat{e}^{(-i)}(X_i))}_{D_i} + \epsilon_i \tag{5.6}$$

where $\bar{\tau}$ is the equally-weighted estimated average treatment obtained from the full set of CATE. This calibration procedure is implemented in the R-package `grf` and estimated using heteroskedastic-robust (HC3) standard errors. A coefficient of $\hat{\alpha} = 1$ implies that the mean forest prediction is accurate while $\hat{\beta} = 1$ suggests that heterogeneity in the treatment effect was captured by the Causal Forest [AW19]. Obtaining a coefficient of $\hat{\alpha} = 0.978$ with a standard error of 0.638, we can reject the null hypothesis $H_0 : \hat{\alpha} < 0$ in a one sided test at the 10% confidence level. This proposes that the average prediction is correct. This is of course very promising. Further, [Ath+17] derive an expression for the Bias of the estimated average treatment effect given by

$$\text{Bias}(\hat{\tau}, \tau) = \Big(\mathbb{E}[Y_i \mid W_i = 1] - \mathbb{E}[Y_i \mid W_i = 0]\Big) - \tau = \frac{1}{p(1-p)}\mathbb{E}[b(X_i)] \tag{5.7}$$

where $b(X_i)$ is called the bias function which we can evaluate from what we have estimated previously. $b(x)$ measures the contribution of observations with $X_i = x$ to the overall $\text{Bias}(\hat{\tau}, \tau)$. Writing $p = \mathbb{E}[X_i]$ and $\mu_w = \mathbb{E}[Y_i^{(w)}]$, the bias function is given as $b(x) = (e(x) - p) \cdot p(\mu_0(x) - \mu_0) + (1-p)(\mu_1(x) - \mu_1)$. In figure 11, we can see that the Bias is centered around 0 with 95% of the mass being between -0.05 and 0.05 which is an indication that the Bias in estimating the average treatment effect is very small.

Considering the great performance in the calibration regression regarding the estimation of the average treatment effect and the small Bias found, we consider the estimation of a significant treatment effect of $\hat{\tau} = -0.044$ to be trustworthy.

5.4.3 Treatment Heterogeneity

As Causal Forests are capable of estimating heterogeneous treatment effects, we want to examine if there was heterogeneity found and subsequently how well the heterogeneity was estimated.

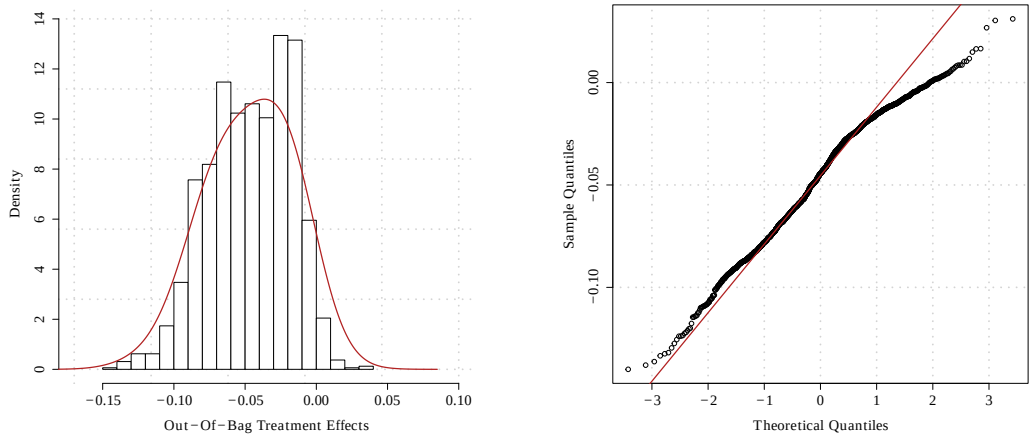

Figure 3: Histogram of out-of-bag heterogeneous treatment effects from the Causal forest on the left. QQ-plot against normal distribution of the out-of-bag heterogeneous treatment effects displaying some light tails on the right

Looking at figure 3 might lead one to believe that there is treatment heterogeneity as there is clearly variation in the out-of-bag heterogeneous treatment effects. This however does not mean that $\hat{\tau}^{(-i)}(X_i)$ provides a better estimate of the true treatment effect function $\tau(X_i)$ than the average treatment effect $\hat{\tau}$ [AW19].

Understanding if a variable of our design matrix $\boldsymbol{X}$ is predictive of the heterogeneous treatment effects $\tau(x) = \mathbb{E}[Y_i^{(1)} - Y_i^{(0)} \mid X_i = x]$ helps to grasp through which channels the introduction of the stricter facial mask policy affects the spread of COVID-19. To test this, we compute the best linear projection of the heterogeneous treatment effects $\tau(x)$ onto $\boldsymbol{X}$. Writing $\hat{\boldsymbol{\tau}}$ for the vector containing the estimated heterogeneous treatment effects with $\dim(\hat{\boldsymbol{\tau}}) = 1612$, we estimate the linear model $\hat{\boldsymbol{\tau}} = \boldsymbol{\beta}_0 + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ using heteroskedastic-robust (HC3) standard errors. Apart from the coefficient on government response index (`gov_resp`), we cannot reject the null-hypothesis of $\hat{\beta}_j = 0$ in the two-sided hypothesis tests. We consider the coefficient on the government response index $\hat{\beta}_{\texttt{gov\_resp}} = 2.67 \times 10^{-2}$ not to be of particular interest as the magnitude is very small. Note that the weekly indicator is contained in the design matrix indicating that the time in which the stricter facial mask policy was introduced does not affect the treatment effect.

After concluding that none of the variables seem to have predictive power on the heterogeneous treatment effects, we seek to investigate cantonal-level

heterogeneity. To achieve this, we compute the best linear projection of the cantonal average treatment effects $\hat{\tau}_j$ from equation 5.5 on $\boldsymbol{X}_{Cantonal} \subset \boldsymbol{X}$. The subset contains the variables that are constant within a canton being settlement area, density, people over 65 years, beds per capita, percentage of people over 80 years and population. Using heteroskedastic-robust (HC3) standard errors, we again find no substantial relationships. This is little surprising as the cantonal average treatment effects $\hat{\tau}_j$ exhibit very little variation with a standard error of $\hat{\sigma}_{\hat{\tau}_j} = 0.023$. We have to conclude that we do not find evidence for treatment heterogeneity.

Judging the accuracy of the estimates of treatment heterogeneity is enabled using the calibration procedure described in section 5.4.2. Estimating equation 5.6 yields a coefficient of $\hat{\beta} = -0.67018$ with a standard error of 0.97403. A coefficient on D_i of 1 indicates well calibrated treatment heterogeneity estimates while a positive and significant coefficient provides evidence of an association between $\hat{\tau}^{(-i)}(X_i)$ and the true treatment function $\tau(X_i)$ [AW19]. Distinguishing between bad estimation accuracy of the heterogeneous treatment effects and a true underlying constant treatment function $\tau(X_i) = \tau$ is generally not possible from observational data using Causal Forests. With this scenario in mind, we run calibration regressions on estimates from Causal Forests on simulated data where treatment heterogeneity is present in section 4.4.2. We find that the estimated coefficients $\hat{\beta}$ are far from 1 which corresponds to failing the calibration regression approach.

To summarize, we do not find any strong association of a covariate with the estimated treatment effects. This might partly be driven by constant treatment effects and partly by the Causal Forests limited ability to accurately estimate heterogeneous treatment effect from panel data. Hence, the estimated average treatment effect is more trustworthy in our opinion.

6 Conclusion

References

[AG03] Alberto Abadie and Javier Gardeazabal. "The economic costs of conflict: A case study of the Basque country." In: *American Economic Review* 93.1 (2003). ISSN: 00028282. DOI: 10.1257/000282803321455188.

[AI16] Susan Athey and Guido Imbens. "Recursive partitioning for heterogeneous causal effects." In: *Proceedings of the National Academy of Sciences of the United States of America* 113.27 (2016). ISSN: 10916490. DOI: 10.1073/pnas.1510489113.

[Ash+20] Peter Ashcroft et al. *COVID-19 infectivity profile correction*. 2020. DOI: 10.4414/smw.2020.20336.

[Ath+17] Susan Athey et al. "Estimating average treatment effects: Supplementary analyses and remaining challenges." In: *American Economic Review*. Vol. 107. 5. 2017. DOI: 10.1257/aer.p20171042.

[ATW19] Susan Athey, Julie Tibshirani, and Stefan Wager. "Generalized random forests." In: *Annals of Statistics* 47.2 (2019). ISSN: 00905364. DOI: 10.1214/18-AOS1709.

[AW19] Susan Athey and Stefan Wager. "Estimating Treatment Effects with Causal Forests: An Application." In: *Observational Studies* 5.2 (2019). DOI: 10.1353/obs.2019.0001.

[BDM04] Marianne Bertrand, Esther Duflo, and Sendhil Mullainathan. *How much should we trust differences-in-differences estimates?* 2004. DOI: 10.1162/003355304772839588.

[BFJ19] Andrew Bell, Malcolm Fairbrother, and Kelvyn Jones. "Fixed and random effects models: making an informed choice." In: *Quality and Quantity* 53.2 (2019). ISSN: 15737845. DOI: 10.1007/s11135-018-0802-x.

[Bre01] Leo Breiman. "Random forests." In: *Machine Learning* 45.1 (2001). ISSN: 08856125. DOI: 10.1023/A:1010933404324.

[Che+18] Victor Chernozhukov et al. *Generic Machine Learning Inference on Heterogeneous Treatment Effects in Randomized Experiments, with an Application to Immunization in India*. Tech. rep. 2. 2018.

[Che+20] Simiao Chen et al. "Buying Time for an Effective Epidemic Response: The Impact of a Public Holiday for Outbreak Control on COVID-19 Epidemic Spread." In: *Engineering* 6.10 (2020). ISSN: 20958099. DOI: 10.1016/j.eng.2020.07.018.

[Che+21] Cheng Cheng et al. *The incubation period of COVID-19: a global meta-analysis of 53 studies and a Chinese observation study of 11 545 patients*. 2021. DOI: 10.1186/s40249-021-00901-9.

[CKS21] Victor Chernozhukov, Hiroyuki Kasahara, and Paul Schrimpf. "Causal impact of masks, policies, behavior on early covid-19 pandemic in the U.S." In: *Journal of Econometrics* 220.1 (2021). ISSN: 18726895. DOI: 10.1016/j.jeconom.2020.09.003.

[Cri20] Eileen M Crimmins. "Age-Related Vulnerability to Coronavirus Disease 2019 (COVID-19): Biological, Contextual, and Policy-Related Factors." In: *Public Policy & Aging Report* 30.4 (2020). ISSN: 1055-3037. DOI: 10.1093/ppar/praa023.

[DAm20] Alexander D'Amour. "On multi-cause causal inference with unobserved confounding: Counterexamples, impossibility, and alternatives." In: *AISTATS 2019 - 22nd International Conference on Artificial Intelligence and Statistics*. 2020.

[DL18] Peng Ding and Fan Li. "Causal inference: A missing data perspective." In: *Statistical Science* 33.2 (2018). ISSN: 08834237. DOI: 10.1214/18-STS645.

[Dom00] Pedro Domingos. "A Unified Bias-Variance Decomposition." In: *Aaai/Iaai* (2000).

[Far15] Max H. Farrell. "Robust inference on average treatment effects with possibly more covariates than observations." In: *Journal of Econometrics* 189.1 (2015). ISSN: 18726895. DOI: 10.1016/j.jeconom.2015.06.017.

[FR20] Daniele Fattorini and Francesco Regoli. *Role of the chronic air pollution levels in the Covid-19 outbreak risk in Italy*. 2020. DOI: 10.1016/j.envpol.2020.114732.

[GKX20] Shihao Gu, Bryan Kelly, and Dacheng Xiu. "Empirical Asset Pricing via Machine Learning." In: *Review of Financial Studies* 33.5 (2020). ISSN: 14657368. DOI: 10.1093/rfs/hhaa009.

[GRP99] Sander Greenland, James M. Robins, and Judea Pearl. "Confounding and collapsibility in causal inference." In: *Statistical Science* 14.1 (1999). ISSN: 08834237. DOI: 10.1214/ss/1009211805.

[Hal+21] Thomas Hale et al. "A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker)." In: *Nature Human Behaviour* 5.4 (2021). ISSN: 23973374. DOI: 10.1038/s41562-021-01079-8.

[Hoe48] Wassily Hoeffding. "A Class of Statistics with Asymptotically Normal Distribution." In: *The Annals of Mathematical Statistics* 19.3 (1948). ISSN: 0003-4851. DOI: 10.1214/aoms/1177730196.

[Hui+20] Jana S Huisman et al. "Estimation and worldwide monitoring of the effective reproductive number of SARS-CoV-2." In: *medRxiv* (2020).

[Kap20] Nikos Kapitsinis. "The underlying factors of the COVID-19 spatially uneven spread. Initial evidence from regions in nine EU countries." In: *Regional Science Policy and Practice* 12.6 (2020). ISSN: 17577802. DOI: 10.1111/rsp3.12340.

[KL51] S. Kullback and R. A. Leibler. "On Information and Sufficiency." In: *The Annals of Mathematical Statistics* 22.1 (1951). ISSN: 0003-4851. DOI: 10.1214/aoms/1177729694.

[LMZ18] Fan Li, Kari Lock Morgan, and Alan M. Zaslavsky. "Balancing Covariates via Propensity Score Weighting." In: *Journal of the American Statistical Association* 113.521 (2018). ISSN: 1537274X. DOI: 10.1080/01621459.2016.1260466.

[NW21] X Nie and S Wager. "Quasi-oracle estimation of heterogeneous treatment effects." In: *Biometrika* 108.2 (2021). ISSN: 0006-3444. DOI: 10.1093/biomet/asaa076.

[PSS21] Regina Pleninger, Sina Streicher, and Jan-Egbert Sturm. "Do COVID-19 Containment Measures Work? Evidence from Switzerland." Zürich, June 2021.

[Rap91] Rh Rapp. *Geometric Geodesy: Part I.* 1991.

[Rob04] James M. Robins. "Optimal Structural Nested Models for Optimal Sequential Decisions." In: 2004. DOI: 10.1007/978-1-4419-9076-1{_}11.

[Rob88] P. M. Robinson. "Root-N-Consistent Semiparametric Regression." In: *Econometrica* 56.4 (1988). ISSN: 00129682. DOI: 10.2307/1912705.

[Rub74] Donald B. Rubin. "Estimating causal effects of treatments in randomized and nonrandomized studies." In: *Journal of Educational Psychology* 66.5 (1974). ISSN: 00220663. DOI: 10.1037/h0037350.

[SA21] Liyang Sun and Sarah Abraham. "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects." In: *Journal of Econometrics* 225.2 (2021). ISSN: 18726895. DOI: 10.1016/j.jeconom.2020.09.006.

[SDS90] Jerzy Splawa-Neyman, D M Dabrowska, and T P Speed. "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9." In: *Statistical Science* 5.4 (1990), pp. 465–472. ISSN: 08834237. URL: http://www.jstor.org/stable/2245382.

[TB19] R. N. Thompson and Ellen Brooks-Pollock. *Preface to theme issue 'Modelling infectious disease outbreaks in humans, animals and plants: Epidemic forecasting and control'.* 2019. DOI: 10.1098/rstb.2019.0375.

[VWB19] Victor Veitch, Yixin Wang, and David M. Blei. "Using embeddings to correct for unobserved confounding in networks." In: *Advances in Neural Information Processing Systems.* Vol. 32. 2019.

[WA18] Stefan Wager and Susan Athey. "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests." In: *Journal of the American Statistical Association* 113.523 (2018). ISSN: 1537274X. DOI: 10.1080/01621459.2017.1319839.

[WK20] William C. Wheaton and Anne Kinsella Thompson. "The Geography of COVID-19 growth in the US: Counties and Metropolitan Areas." In: *SSRN Electronic Journal* (2020). ISSN: 1556-5068. DOI: 10.2139/ssrn.3570540.

[Zhu+20] Yongjian Zhu et al. "Association between short-term exposure to air pollution and COVID-19 infection: Evidence from China." In: *Science of the Total Environment* 727 (2020). ISSN: 18791026. DOI: 10.1016/j.scitotenv.2020.138704.

[Zor+20] Maria A. Zoran et al. "Assessing the relationship between surface levels of PM2.5 and PM10 particulate matter impact on COVID-19 in Milan, Italy." In: *Science of the Total Environment* 738 (2020). ISSN: 18791026. DOI: 10.1016/j.scitotenv.2020.139825.

A Appendix

A.1 Monte Carlo Simulation

The Value of Clustering: Results for CATE							
		MSE		**Bias^2**		**Coverage**	
d	T	Clustered	Standard	Clustered	Standard	Clustered	Standard
3	200	0.190	0.142	0.123	0.091	0.891	0.584
4	200	0.193	0.146	0.127	0.097	0.878	0.558
5	200	0.150	0.112	0.093	0.068	0.904	0.622
6	200	0.144	0.106	0.090	0.067	0.933	0.624
7	200	0.124	0.093	0.079	0.058	0.926	0.610
8	200	0.220	0.166	0.131	0.098	0.860	0.509
24	200	0.205	0.153	0.128	0.094	0.867	0.495
3	1000	0.135	0.091	0.089	0.060	0.892	0.650
4	1000	0.168	0.114	0.102	0.070	0.874	0.614
5	1000	0.154	0.104	0.096	0.064	0.868	0.601
6	1000	0.130	0.089	0.078	0.053	0.894	0.630
7	1000	0.135	0.091	0.078	0.054	0.894	0.605
8	1000	0.130	0.088	0.080	0.054	0.883	0.576
24	1000	0.143	0.097	0.088	0.059	0.903	0.510

Table 1: Results in terms of the CATE for the comparison of cluster-robust and standard Causal Forests. The data is generated according to design 1 with $d \in \{3, 4, 5, 6, 7, 8, 24\}$ and $\in \{200, 1000\}$ and 10 entities. The standard Causal Forests slightly outperform the cluster-robust Causal Forests in terms of MSE and Bias^2. However, they clearly fall short in terms of coverage as they fail to incorporate the entity fixed-effects when estimating the standard errors of the CATE. The cluster-robust option is clearly preferred as accurate estimation of the standard errors is crucial for inference.

The Value of Clustering: Results for ATE							
		MSE		**Bias^2**		**Coverage**	
d	T	Clustered	Standard	Clustered	Standard	Clustered	Standard
3	200	0.192	0.143	0.124	0.091	0.870	0.370
4	200	0.193	0.147	0.129	0.097	0.830	0.300
5	200	0.152	0.112	0.094	0.068	0.880	0.420
6	200	0.142	0.107	0.089	0.067	0.930	0.370
7	200	0.123	0.094	0.079	0.058	0.920	0.390
8	200	0.218	0.165	0.130	0.099	0.830	0.330
24	200	0.206	0.152	0.128	0.093	0.860	0.320
3	1000	0.132	0.091	0.087	0.060	0.870	0.350
4	1000	0.165	0.114	0.102	0.070	0.820	0.370
5	1000	0.154	0.104	0.097	0.064	0.850	0.350
6	1000	0.127	0.088	0.076	0.052	0.860	0.440
7	1000	0.134	0.092	0.078	0.054	0.880	0.400
8	1000	0.129	0.088	0.080	0.054	0.820	0.420
24	1000	0.144	0.097	0.088	0.059	0.880	0.370

Table 2: Results in terms of the ATE for the comparison of cluster-robust and standard Causal Forests. The data is generated according to design 1 with $d \in \{3, 4, 5, 6, 7, 8, 24\}$ and $T \in \{200, 1000\}$ and 10 entities. As in table 2, the standard Causal Forests slightly outperform the cluster-robust Causal Forests in terms of MSE and Bias^2 but clearly fall short in terms of coverage as they fail to incorporate the entity fixed-effects when estimating the standard errors of the CATE and therefore the standard errors of the ATE which can be seen from equation 5.5. We observe that the results for the MSE as well as the Bias^2 are almost identical to those from the CATE but the coverage differs considerably when comparing to table 2 which suggests that standard Causal Forests considerably underestimate the standard error of the ATE when there are natural clusters present in the data. Once again, the cluster-robust option is clearly preferred as accurate estimation of the standard errors is crucial for inference on the ATE.

True Unobserved Confounding: Results							
		MSE	**Bias**^{2}	**Coverage**	**MSE**	**Bias**^{2}	**Coverage**
d	T	CATE	CATE	CATE	ATE	ATE	ATE
3	200	0.177	0.112	0.918	0.175	0.111	0.930
4	200	0.168	0.104	0.844	0.166	0.103	0.860
5	200	0.190	0.117	0.885	0.194	0.119	0.880
6	200	0.158	0.089	0.916	0.159	0.090	0.920
7	200	0.174	0.107	0.888	0.175	0.107	0.870
8	200	0.178	0.115	0.903	0.174	0.114	0.890
24	200	0.171	0.108	0.907	0.169	0.107	0.880
3	1000	0.149	0.095	0.888	0.147	0.094	0.890
4	1000	0.160	0.096	0.881	0.160	0.096	0.880
5	1000	0.134	0.089	0.891	0.134	0.090	0.920
6	1000	0.147	0.092	0.876	0.148	0.091	0.860
7	1000	0.149	0.084	0.894	0.148	0.082	0.920
8	1000	0.111	0.070	0.889	0.111	0.069	0.860
24	1000	0.124	0.081	0.913	0.125	0.081	0.880

Table 3: Results in terms of the CATE and ATE for the design 2. We simulate data with $d \in \{3, 4, 5, 6, 7, 8, 24\}$ and $T \in \{200, 1000\}$ with 10 entities. The results are very promising. We again notice that the performances regarding the CATE and the ATE are very similar. Further, we observe that the performance regarding all three measures does not decrease compared to design 1 when adding hidden confounding variables as described for design 2. We are cautious about generalizing these findings as its generally very hard to test an estimators robustness to the presence of unobserved confounding variables in the data as there are many ways hidden confounders can enter the data. It is nevertheless astonishing that the performance remains this good.

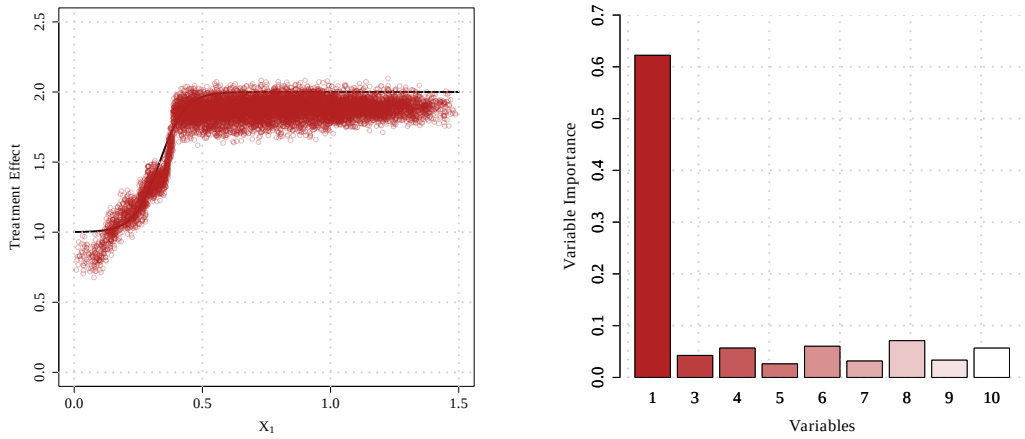

Figure 4: True treatment function $\tau(X)$ and Causal Forest estimates on the left and feature importance of features $\{X_1, X_j\}_{j=3}^{10}$ on the right. Note that X_2 is missing for the feature importance as we delete X_2 to introduce unobserved confounding

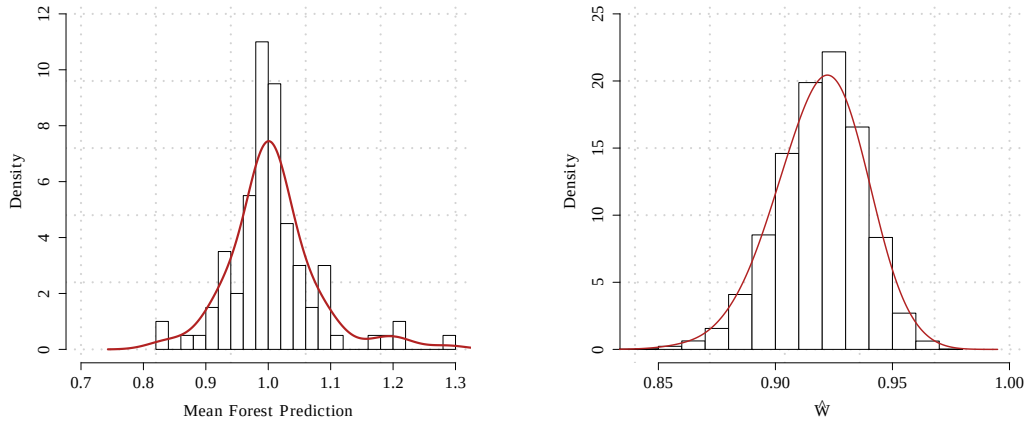

Figure 5: Histogram of $\hat{\alpha}_s$ over the $S = 100$ runs on the left and histogram of $\hat{W}$ from one randomly chosen run on the right

A.2 Empirical Application

A.2.1 Tables

Nr.	Acronym	Description	Frequency & Unit	Literature
1	`r_median`	median R_e	Daily & Cantonal	[PSS21]
2	`deaths`	cum. deaths	Daily & Cantonal	[TB19]
3	`recovered`	cum. recovered	Daily & Cantonal	[TB19]
4	`tests`	cum. tests	Daily & Cantonal	[TB19]
5	`hosp`	curr. hospitalized	Daily & Cantonal	[TB19]
6	`perc_age`	age ≥ 80 years in %	Constant & Cantonal	[Cri20]
7	`abs_age`	age ≥ 65 years	Constant & Cantonal	[Cri20]
8	`area`	area in ha	Constant & Cantonal	[WK20]
9	`density`	people per km^2	Constant & Cantonal	[WK20]
10	`pop`	population	Constant & Cantonal	[WK20]
11	`beds`	hospital beds per capita	Constant & Cantonal	[TB19]
12	`holidays`	official school holidays	Daily & Cantonal	[Che+20]
13	`sunshine`	sunshine in minutes	Daily & Cantonal	[Zor+20]
14	`temp`	mean air-temperature in C°	Daily & Cantonal	[Zor+20]
15	`humidity`	relative humidity in %	Daily & Cantonal	[Zor+20]
16	`amount_spent`	growth CHF spent debit, credit and mobiles	Daily & Cantonal	[PSS21]
17	`transactions`	growth transactions debit, credit and mobiles	Daily & Cantonal	[PSS21]
18	`mobility`	median distance in km	Daily & National	[PSS21]
19	`eco_supp`	economic support index	Daily & National	[CKS21]
20	`gov_resp`	government response index	Daily & National	[CKS21]
21	`canc_events`	cancellation of events indicator	Daily & Cantonal	[PSS21]
22	`rest_gatherings`	restrictions on gatherings indicator	Daily & Cantonal	[PSS21]
23	`facial_covering`	mask-policy indicator	Daily & Cantonal	[PSS21]
24	`kof_stringency`	KOF stringency index	Daily & Cantonal	[PSS21]
25	`week`	weekly indicator	Weekly & National	
26	`canton`	cantonal indicator	Constant & Cantonal	

Table 4: This table lists the $d + 2 = 26$ variables used. The variables are described and the literature justifying their use is listed. The variables are separated along their type as described in section 5.3 which is represented by the horizontal lines. Note that cum. stands for cumulative. Check out the respective web-pages referenced in 5.3 for a detailed description of the policy-indicators and policy-indices.

Nr.	Dataset	URL	Literature
1	COVID19Re_geoRegion	https://www.covid19.admin.ch/en/overview	
2	COVID19Cases_geoRegion	https://www.covid19.admin.ch/en/overview	
3	COVID19Cases_geoRegion	https://www.covid19.admin.ch/en/overview	

Table 5: hallo

A.2.2 Figures

A.2.3 Robustness Checks

In the following, we examine our data more closely. We first investigate the response $Y_{it} = \ln \mathrm{R}_{e,i,t}$, then the treatment variable $W_{i,t}$ which is given by `facial_covering` and lastly the variables that make up the design matrix $\boldsymbol{X}$.

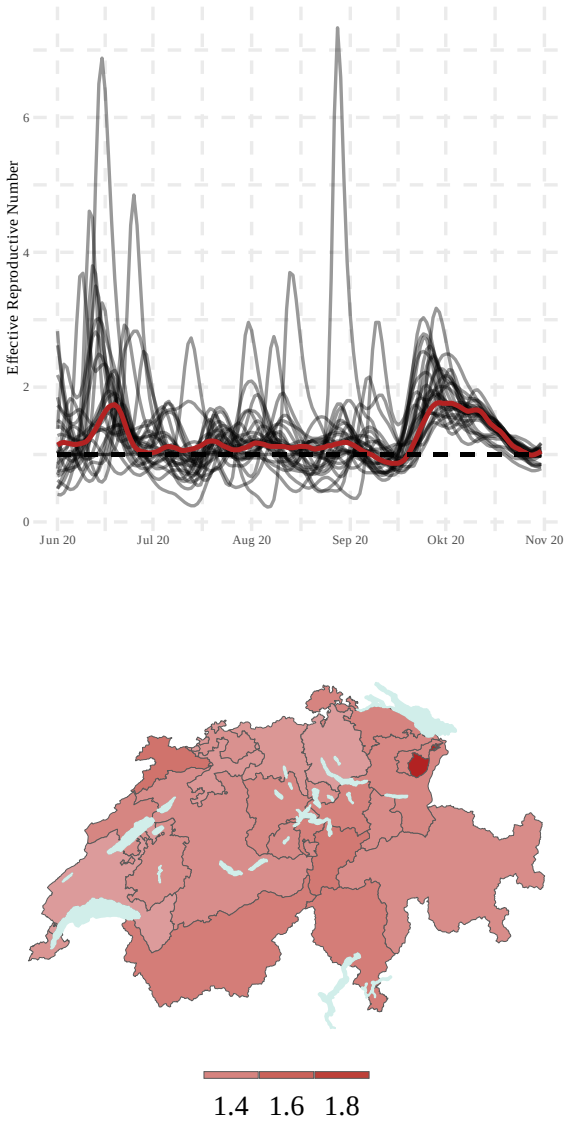

Figure 6: We observe strong heterogeneity in R_e across cantons which can easily be seen from the plot on the left. The red line represents the effective reproductive number on the national level while the dashed line visualizes $R_e = 1$. On the right plot, we can observe the average R_e per canton over the period of analysis.

We conclude that there is sufficient heterogeneity in R_e across cantons. See [Hui+20] for more details on how the effective reproductive number is estimated for Switzerland.

Further, we want to justify using $\ln R_{e,i,t}$ as an approximation for the growth rate of new infections $\text{NINF}_{i,t}$. Remember that $R_e(t) = \frac{\text{NINF}_t}{\sum_{s=1}^{t} \text{NINF}_{t-s} w_s}$ with w_s

being the value of the infectivity profile s days after infection which measures how infective a individual is relative to the beginning of their symptoms. Note that we have $\sum_s w_s = 1$ being modeled via the serial interval distribution [Hui+20]. We proceed as [PSS21] assuming that $w_s = \mathbb{1}\{s = \tau\}$ implying that the transmission of COVID-19 only takes places exactly τ days after infection. Using this, we rewrite

$$\mathrm{R}_e(t) = \frac{\mathrm{NINF}_t}{\mathrm{NINF}_{t-\tau}} \tag{A.1}$$

Using a first order Taylor approximation, one can show that $\ln \mathrm{NINF}_t - \ln \mathrm{NINF}_{t-\tau}$ is a good approximation of the τ-day growth rate of new infections, given that the growth rate is sufficiently small.

Furthermore, understanding if there is enough variation in the mask-policies across cantons and time to credibly estimate treatment effects is crucial. The variable is coded as an ordinal variable with four levels. A value of 0 stands for no policy, a value of 1 represents instances where wearing a mask is recommended. A value of 2 means that wearing a mask is required in some shared or public spaces when social distancing is not possible. In Switzerland, that coincides with the national introduction of mandatory mask-wearing on public transport on July 6, 2020[17]. This law forms a lower bound concerning mask-policies for the period of analysis from which the cantons had the authority to differ by introducing stricter policies. A value of 3 represents policies where wearing a mask is mandatory in all shared or public spaces where social distancing is not possible. Concretely for Switzerland, that means masks are mandatory in all public indoor spaces as well as in train stations, airports, bus stations and tram stations[18]. A value of 4 was never realized in Switzerland as it corresponds to a mask law that enforces mask wearing whenever not at home. As standard Causal Forests require binary treatments $W_i \in \{0, 1\}$, we have to reduce the facial mask indicator down to 2 levels. This is straight forward as we only observe values of 2 and 3 for the period of analysis. This means that a canton is considered treated if it employs as facial mask policy

[17] https://www.admin.ch/gov/de/start/dokumentation/medienmitteilungen.msg-id-79711.html

[18] https://www.bag.admin.ch/bag/de/home/das-bag/aktuell/news/news-18-10-2020.html

corresponding to a value of 3 and untreated otherwise.

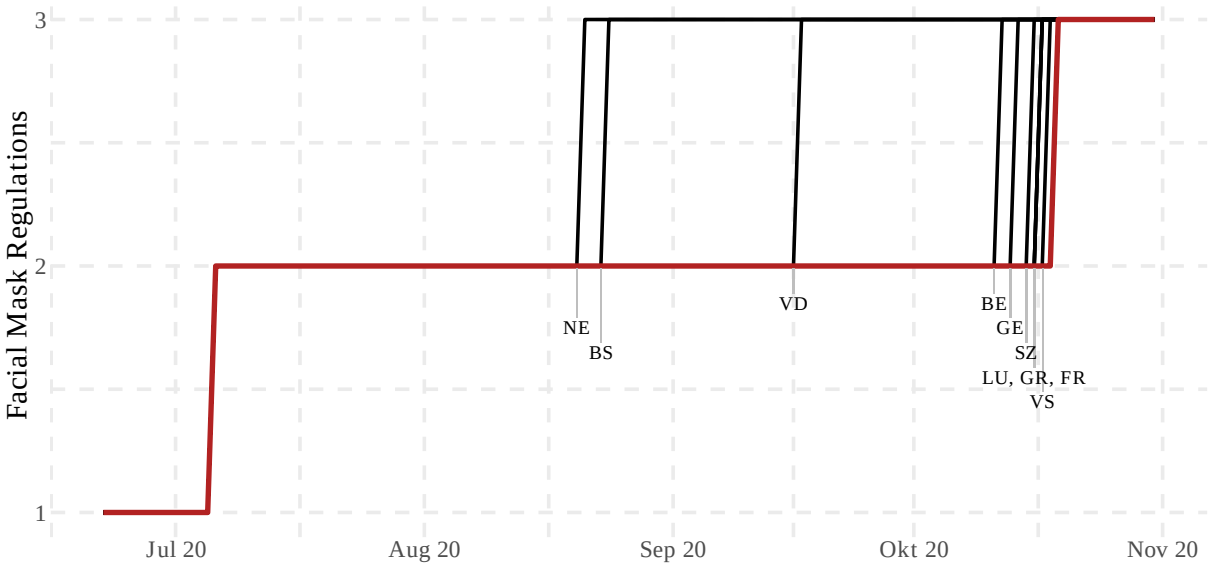

Figure 7: The black line represents the lower bound put in place at the national level. 11 out of the 26 cantons enforced a stricter mask policy than nationally required which is indicated by the lines diverging from the black line. From the $26 \times T = 1612$ data points, we observe $195 = 12\%$ where a stricter policy is in place.

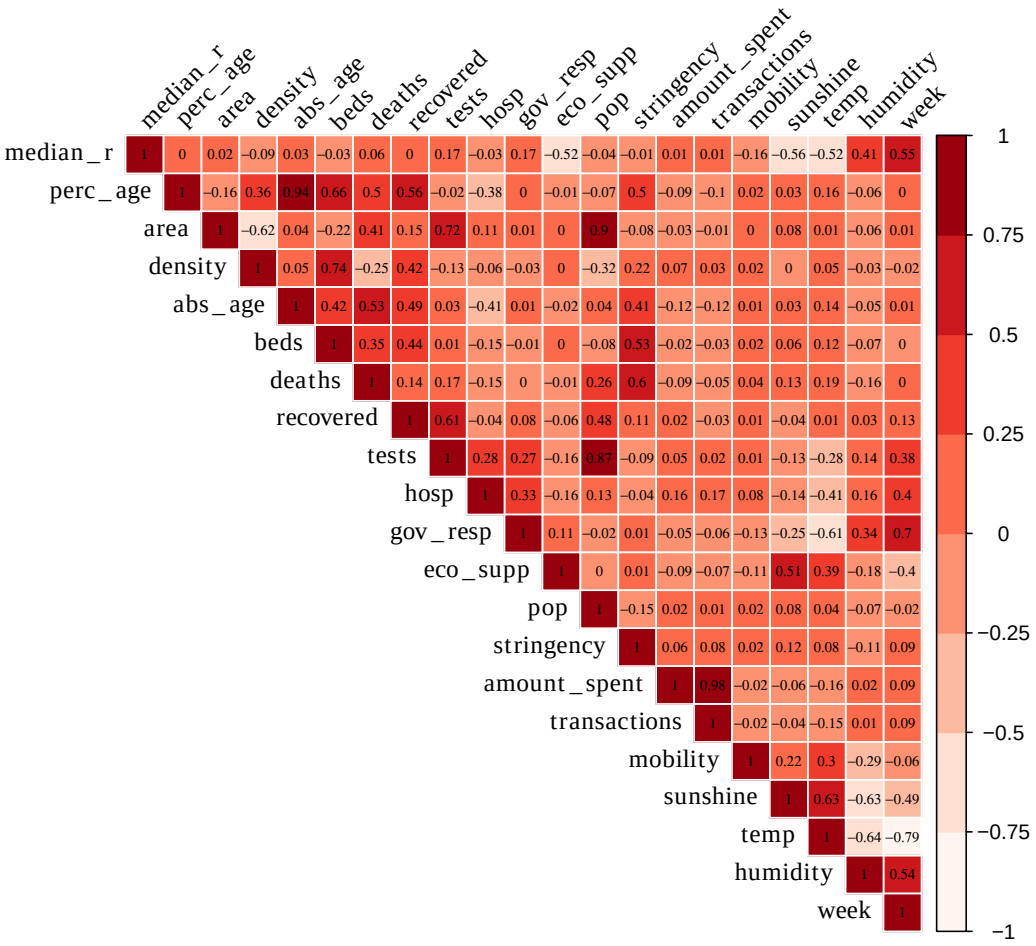

Figure 8: Correlation matrix of all numerical variables including the indicator week as low values of week can be interpreted as early and high values as late with respect to the period of analysis. Particularly interesting are the high correlations of the weather variables with the response variable.

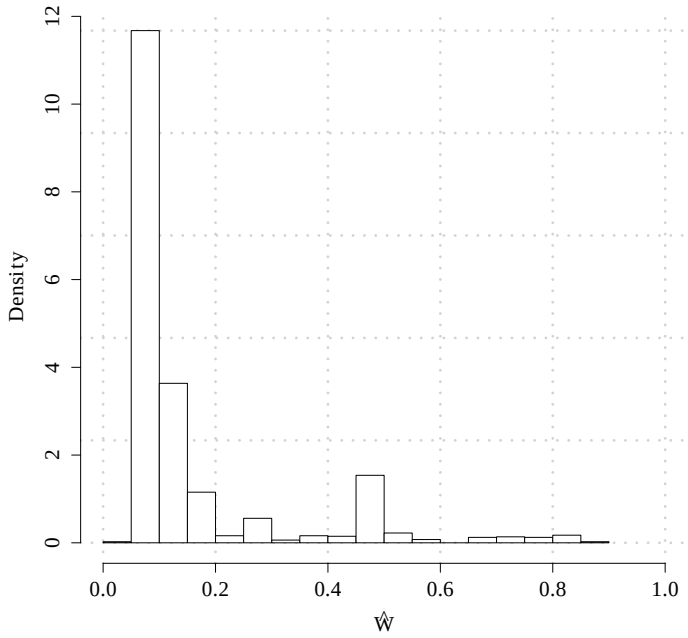

Figure 9: Histogram of $\hat{e}(X)$ from the Causal Forest. The estimated propensity scores are not centered around 0.5 as one would for example expect from data coming from a randomized control trial. The propensity scores are however sufficiently bounded away from 0 and 1 with $\min(\hat{e}(X)) = 0.045$ and $\max(\hat{e}(X)) = 0.880$

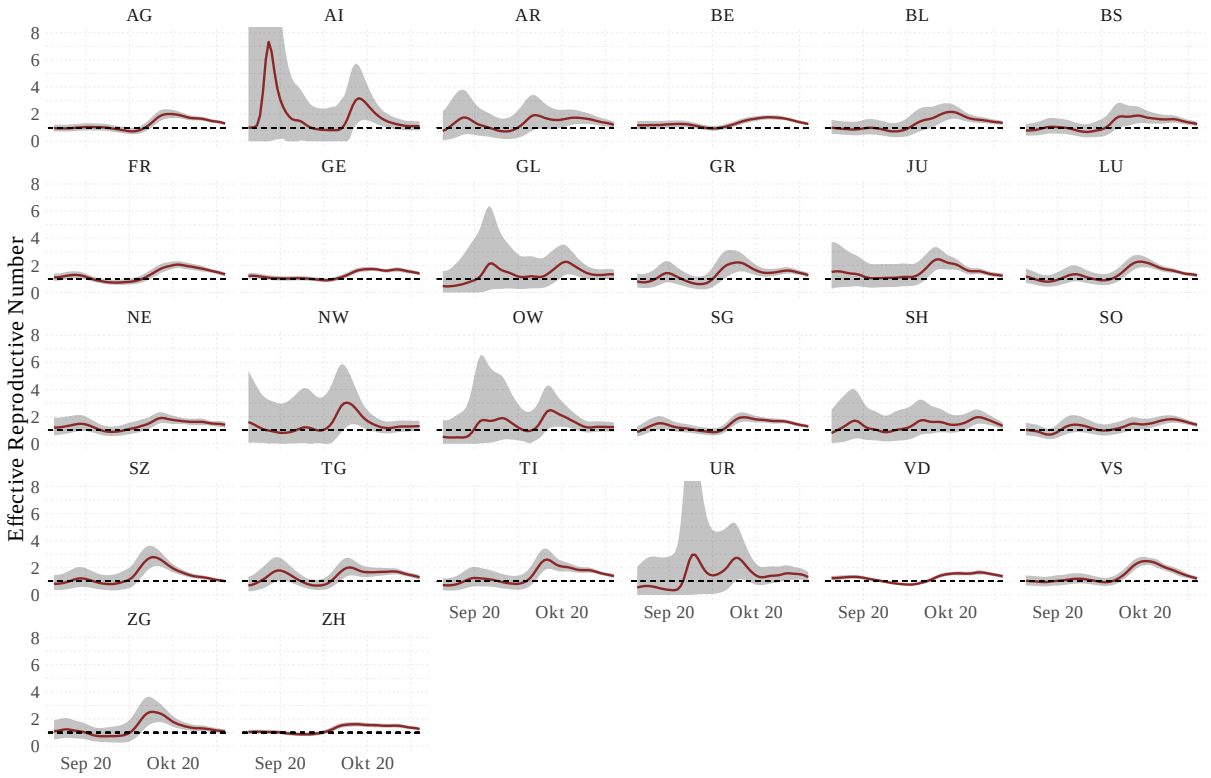

Figure 10: blabla

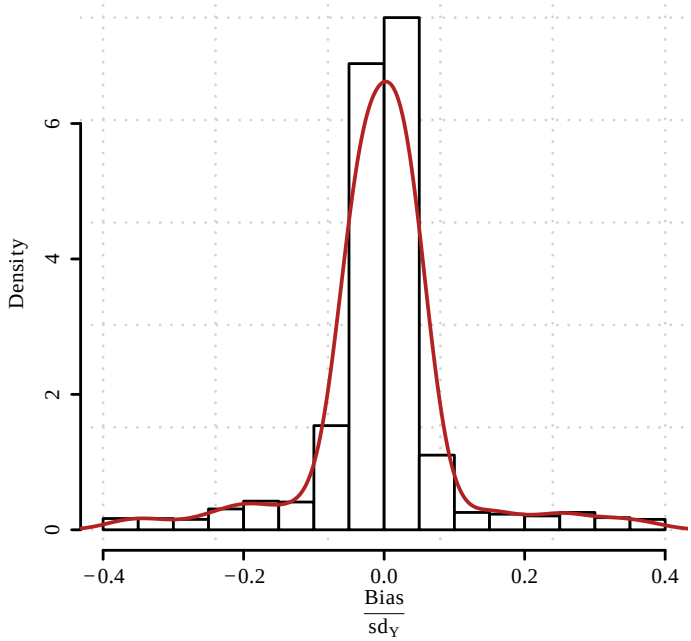

Figure 11: Histogram of $b(X_i)$ from the Causal Forest rescaled by $\text{sd}_Y = \hat{\sigma}_Y$. We observe that $b(X_i)$ is centered around 0 with most of the mass being very close to 0 indicating small bias

A.2.4 Robustness-Checks

We present various robustness checks related to our main results from section 5.4.2 and 5.4.3. To compare the heterogeneous treatment effects of the different models, we compute the Kullback-Leibler divergence of the distribution of the out-of-bag treatment effects from our main model $\hat{\tau}_{\text{main}}(x)$ to the distribution of the out-of-bag treatment effects from the model tested $\hat{\tau}_{\text{tested}}(x)$. Define the Kullback-Leibler divergence as

$$D_{KL}(\hat{\tau}_{\text{main}}(x) \mid\mid \hat{\tau}_{\text{tested}}) = \int_{-\infty}^{\infty} \hat{\tau}_{\text{main}}(x) \ln \frac{\hat{\tau}_{\text{main}}(x)}{\hat{\tau}_{\text{tested}}(x)} dx \tag{A.2}$$

The Kullback-Leibler divergence is non-negative and the smaller it is, the more similar are the distributions [KL51].

	$(\hat{\tau}, \hat{\sigma}_{\hat{\tau}})$	(ATT, overlap-weighted ATE)	$D_{KL}(\hat{\tau}_{\text{main}}(x) \mid\mid \hat{\tau}_{\text{tested}})$	$(\hat{\alpha}, \hat{\beta})$-Calibration
Main Model				
	$(-0.044, 0.020)$	$(-0.054, -0.045)$	0	$(0.978^{**}, -0.865)$
Lag l				
$l = 5$	$(-0.038, 0.018)$	$(-0.049, -0.042)$	0.109	$(1.150^{**}, -1.040)$
$l = 8$	$(-0.037, 0.021)$	$(-0.043, -0.038)$	0.010	$(0.959^{*}, -0.382)$
$l = 10$	$(-0.037, 0.021)$	$(-0.043, -0.038)$	0.060	$(0.902, 0.092)$
$l = 14$	$(-0.039, 0.030)$	$(-0.069, -0.057)$	0.118	$(0.850, 0.684)$
Not Clustered				
`cluster=NULL`	$(-0.074, 0.007)$	$(-0.033, -0.073)$	0.216	$(0.782^{**}, 1.693^{***})$
Pre-selection				
$d = 6$	$(-0.042, 0.202)$	$(-0.054, -0.046)$		$(0.960^{**}, -0.690)$
Differenced				
Δx	$(-0.005, 0.046)$	$(-0.028, -0.027)$		$(1.130, -0.572)$

Table 6: This table lists the conducted robustness-checks. The second column contain the estimated ATE as well as the estimated standard error of the ATE. Further, ATT refers to the average treatment effect on the treated meaning ATT $= \mathbb{E}[Y_i^{(1)} - Y_i^{(0)} \mid W_i = 1]$ while the overlap-weighted ATE is an alternative calculation of the ATE that is more robust to small treatment propensities [LMZ18]. The fourth column contains the estimated Kullback-Leibler divergence while the last column contain the $(\hat{\alpha}, \hat{\beta})$-coefficients from the calibration regression described in section 5.4.2.

We test the robustness of our results along four general dimensions. Firstly, we examine the effect of the length of the lag l as described in section 5.3.1. Secondly, we have a look at the influence of using cluster-robust Causal Forests. We then inspect the influence of pre-selection of variables. This is achieved by running two Causal Forests where the first is used to assess variable importance and the second Causal Forest is run on the subset of variables with an above median variable importance. Lastly, we compare the results of our main approach to the approach where the difference-operator Δ is applied to all variables apart from the indicators.

We test lags $l \in \{5, 8, 10, 14\}$ which is in driven by different estimates of the incubation period as well as behavioral delays to COVID-19 policies [Che+21]. The results are stable across the choices of l which is reassuring. Running the estimation while not clustering on the cantons yields an average treatment effect of -0.074 and coupled with a very small standard error of 0.007. [AW19] observe a similar pattern. They hypothesize that it is a result of overfitting on

the canton-level effects as unclustered Causal Forests do not take within-canton correlations into account. This might drive the Causal Forests to overestimate the magnitude of the average treatment effect. Furthermore, we observe that the estimated treatment propensities $\hat{e}(X)$ are much closer 0 and 1 which is problematic. Thirdly, we consider pre-selection on the variables. The first Causal Forest selects $d = 6$ variables which are deployed to the second Causal Forest. The results remain nearly unchanged.

Overall, the results are very robust to alterations in the data such as the choice of lag l and pre-selection. The average treatment effect of the Causal Forest that is not cluster-robust is roughly twice as large as the average treatment effect from our main model which is most likely due to overfitting on the cantonal-level effects.