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 30, 2022

Abstract

In this thesis, we examine the estimation of heterogeneous treatment effects from panel data using the Causal (Random) Forest. Firstly, we explain the algorithm in detail. Thereafter, a Monte Carlo simulation study is conducted that shows that the cluster-robust Causal Forest successfully estimates heterogeneous and average treatment effects from panel data. Furthermore, we find that the performance is robust in regards to the presence of some observed and unobserved confounding variables. Additionally, we put the Causal Forest to the test in an empirical application. We use the cantonal-level heterogeneity in COVID-19 containment policies in Switzerland over the summer of 2020 to infer the effect of introducing a stricter facial mask policy on the effective reproductive number. We find an average treatment effect that corresponds to lowering the effective reproductive number by 5% that is significant at the $\alpha = 5\%$ level.

Contents

List of Figures

List of Tables

List of Abbreviations

ATE	average treatment effect
ATT	average treatment effect of the treated
CATE	conditional average treatment effect
MSE	mean squared error
GRF	Generalized Random Forest

1 Introduction

A central objective in statistics is the estimation of causal treatment effects. For example, understanding the causal effect of a drug on some measured health outcome is crucial for public health in our modern societies. Yet, not all patients react identically to a given drug. Aiming to understand which patient characteristics influence the treatment effect creates the need to estimate heterogeneous treatment effects. To restrict the statistical machinery to experimental data as is mostly the case in medical studies however, would leave a great number of pressing questions unanswered. Questions concerning the effectiveness of public policies are a hot-topic especially during the COVID-19 pandemic. In an effort to answer these questions, a broad body of econometric literature focuses on the development of methods to estimate heterogeneous treatment effects from observational and experimental data and subsequently conduct statistical inference.

A field that has been particularly successful in answering questions by creating knowledge from data is Machine Learning. These methods aim to predict outcomes based on data that is provided as input. Generally, an algorithm is deployed on a set of training data with the objective of learning the specific estimand. Thereafter, predictions are made on a test set and the accuracy is evaluated. Through this procedure, a best estimator is chosen regarding the particular application. These approaches work extraordinarily well in the realm of forecasting. To make consequential decisions or to answer scientific questions with the help of Machine Learning however, has its limits. In order to adequately respond to the challenges enumerated, decision-makers or researchers often seek to identify the mechanism through which the observed outcome is produced. Methods that promise to enable such procedures are summarized under the field of Causal Inference which is very popular among econometricians. The marriage of the two fields of Machine Learning and Causal Inference has aroused a lot of attention in the last couple of years as it promises to combine the flexibility of Machine Learning with the methods to identify causal parameters from Causal Inference.

In the following, we name the most important advances of incorporating Machine Learning into Causal Inference in an effort to estimate heterogeneous

treatment effects. The sequence of work by Imai and Ratkovic (2013), Tian et al. (2014) and Weisberg and Pontes (2015) develops Lasso-methods to estimate heterogeneous treatment effects in a sparse high-dimensional setting. Estimating heterogeneous treatment effects through empirical loss-minimization is introduced by Beygelzimer and Langford (2009), Dudik et al. (2011) and Nie and Wager (2021). On the Bayesian side of statistics, there are methods using forest-based algorithms to estimate heterogeneous treatment effects from Green and Kern (2012), Hill (2011) and Hill and Su (2013). Lastly, there is the sequence of Athey and Imbens (2016), Wager and Athey (2018) and Athey et al. (2019) that employ the Random Forest as introduced by Breiman (2001) to estimate and conduct inference on heterogeneous treatment effects. In this thesis, the attention lies on the methods listed last as they provide a very powerful machinery for empirical applications as well as a very solid theoretical foundation that allows for inference on the estimated treatment effects. The authors name these algorithms Causal Tree and further Causal Forest which we follow for the remainder of the thesis due to it being shorter than Causal Random Forest.

In this thesis, we propose that the Causal Forest as described in Athey et al. (2019) can be used successfully to estimate heterogeneous treatment effects from panel data. Firstly, we motivate the Causal Forest by describing the algorithm in detail. We then provide an extensive simulation study to show that the algorithm that is developed for independent and identically distributed data is capable of handling the difficulties that go hand in hand with panel data. In doing so, we apply the research of estimating heterogeneous treatment effects through the lens of Machine Learning to one of the most important areas of econometrics. The estimation of treatment effects from panel data has been at the heart of econometrics with methods such as Differences in Differences (Bertrand et al. 2004), Synthetic Controls (Abadie and Gardeazabal 2003) or Fixed-Effects estimators (Bell et al. 2019) being part of every econometrician's education. In the simulation study, we show that the Causal Forest achieves great coverage rates for the heterogeneous treatment effects even in the presence of observed and unobserved confounding variables. This implies that the Causal Forest can be used as a flexible tool that allows the researcher to estimate heterogeneous treatment effects and average treatment

effects from panel data.

Further, we aim to put our claims to the test in an empirical application. As the world suffers from great uncertainty due to COVID-19, the task of scientific research to provide guidance to decision-makers becomes even more important. In an effort to provide such scientific research, we quantify the effect of imposing a strict facial mask policy in Switzerland over the period from August 21, 2020 to October 19, 2020. The identification is enabled by cantonal-level heterogeneity in the facial mask policies over the period of analysis. We estimate a considerable average treatment effect that corresponds to a 5%-reduction of the effective reproductive number R_e that is significant at the $\alpha = 5\%$ level. We do not find evidence for treatment effect heterogeneity induced by any observed variable nor do we estimate substantial differences in treatment effects across cantons and time.

The remainder of this thesis is structured as follows. Treatment effects are defined and the framework of potential outcomes is described in section 2. In section 3, we describe the Causal Forest algorithm and present the theoretical results that enable statistical inference. Further, we introduce panel data and heuristically motivate why the Causal Forest is suitable for such a setting. The following section 4 presents the simulation study which includes the data generating process and choice of treatment delay as well as discussing performance evaluation criteria and the results of the simulation. In section 5, we display the empirical application by explaining the data and discussing the results. Finally, we conclude in section 6.

2 Estimation of Treatment Effects & Potential Outcomes Framework

Firstly, we need to define what a treatment effect is to lay the foundation for estimating treatment effects. The potential outcomes framework from Neyman (1923) and Rubin (1974) allows for a proper definition of treatment effects and a mathematical tool to think about causality. We describe the potential outcomes framework in the first part of this section while we elaborate on the estimands of interest and specify the assumptions made in the second part of this section.

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 $W_i \in \{0,1\}$ (Wager and Athey 2018: 1230). Following Athey et al. (2019: 1148), 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 independent and identically distributed sample drawn from a large population (Athey and Imbens 2016: 7354).

Neyman (1923: 2) and Rubin (1974: 689) 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 $Y_i^{(0)}$ making causality inherently a problem of missing data (Ding and Li 2018: 216). 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 name $\tau(x)$ the heterogeneous treatment effect function. It is also called the conditional average treatment effect 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 $\mathbb{E}_X$ specifies the expectation over X.

As estimating $\tau(x)$ or τ from data (X_i, O_i) is generally not possible in an observational setting, Wager and Athey (2018: 1232) assume unconfoundedness, which is standard practice. Unconfoundedness translates into the 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 realizations from a randomized experiment (Wager and Athey 2018: 1232). The second assumption that enables the Causal Forest to work is overlap, meaning for some $\epsilon > 0$ and all $x \in [0,1]^d$, we have

$$\epsilon < \mathbb{P}\Big[W = 1 \mid X = x\Big] < 1 - \epsilon \tag{2.4}$$

This assures that there are enough treatment and control units near any test point x for large n. Note that we use the notation $\mathbb{P}\left[W = 1 \mid X = x\right] = e(x)$ throughout this thesis.

3 Causal Forest

We firstly describe the Causal Forest algorithm as developed in the sequence of work of Athey and Imbens (2016), Wager and Athey (2018) and Athey et al. (2019) in section 3.1 in an effort to motivate the methodology. In section 3.2, we present the results that enable statistical inference on the estimated heterogeneous treatment effects. Lastly, we discuss panel data and heuristically present why and how the Causal Forest can be adapted to suit treatment effect estimation from panel data in section 3.3.

3.1 Description of Causal Forest

With the aim to structure the problem of heterogeneous treatment effect estimation, Robinson (1988: 931) studies a class of semi-parametric problems

where we have a model of $\tau(x)$ given by

$$Y_i^{(w)} = f(X_i) + w\tau(X_i) + \epsilon_i(w), \;\; w \in \{0, 1\} \tag{3.1}$$

He restricts 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 Robins (2004) found a way to rewrite the heterogeneous treatment effect function $\tau(x)$ from Robinson (1988) as a loss minimizer. Under unconfoundedness as defined in section 2 and writing the conditional response surfaces 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 Robinson (1988: 936) 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 use this equation to formulate the following equation

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

This transformation is central as an oracle with access to $m^\star(x)$ and $e^\star(x)$ could estimate the heterogeneous treatment function $\tau^\star(x)$ via empirical loss minimization of equation 3.4[1]. This is the starting point for Machine Learning algorithms for Causal Inference such as the R-learner that estimates $\hat{e}(x)$ and $\hat{m}(x)$ separately in a first step and minimizes the empirical loss motivated by equation 3.4 via cross-fitting of $\hat{e}(x)$ and $\hat{m}(x)$ in a second step (Nie and Wager 2021: 303–306). The Causal Forest on the other hand takes a more indirect approach that is motivated by a miss-specification of the partial linear model

[1] The superscript * characterizes functions that solve equation 3.4

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

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

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 (Robinson 1988: 937). We then have

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

Note that $\hat{\tau}$ is the coefficient of $\{W_i - \hat{e}(X_i)\}$ in the simple regression with $\{Y_i - \hat{m}(X_i)\}$ as the response and no intercept. This property lies behind the idea of the Causal Forest. On a high level, the algorithm aims 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, the Causal Forest aims to use equation 3.6 for treatment effect estimation.

In the following, we describe how the Causal Forest concretely uses the proposed idea for the estimation of heterogeneous treatment effects. The algorithm seeks to build $b = 1, \ldots, B$ trees, for each of which a subsample of observations $\mathcal{S}_b \subseteq \{1, \ldots, n\}$ is drawn. Then, the B trees are grown via recursive partitioning, one on each subsample $\mathcal{S}_b$ (Athey and Wager 2019: 40). The partitioning-schema is explained in detail later in this section. Assuming the B trees are built, we define $L_b(x)$ as the set of training samples falling in the same leaf node as the test point x. We can construct weights $\alpha_i(x)$ that quantify how often the ith training sample falls into the same leaf node 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 estima-

tor of the constant treatment effect τ from equation 3.6. That allows for heterogeneous treatment effect estimation by using the importance-weighted counterpart of equation 3.6 which gives training samples that are similar to the test point x more weight. Concretely, 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 $\hat{m}^{(-i)}, \hat{e}^{(-i)}$ (Athey and Wager 2019: 41). To compute heterogeneous treatment effects from equation 3.8, we further need $\hat{m}(x)$ and $\hat{e}(x)$. This yields the final two-step approach of the Causal Forest where $\hat{m}(x)$ and $\hat{e}(x)$ are grown in a first step via separate standard 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 is how exactly the B trees are grown via recursive partitioning. It is obvious that the standard Random Forest 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 node. Athey and Imbens (2016) first introduce a new splitting rule for heterogeneous treatment effects that is further developed in Wager and Athey (2018). They propose two approaches to recursively split the covariate space where one approach is capable of handling treatment heterogeneity and one is suitable for situations with confounding variables present. Considering these drawbacks, Athey et al. (2019) develop what is now considered the best option to estimate treatment heterogeneity within the Causal Forest framework.

To get into the details of the splitting procedure proposed by Athey et al. (2019), we have to start with a general description of their Generalized Random Forest (GRF). 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[2]. The approach of Athey et al. (2019: 1152–1158) aims to estimate solutions to equation 3.9 by minimizing its empirical similarity-weighted counterpart. The 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 in order to estimate heterogeneous treatment effects via equation 3.8. We obtain the similarity weights $\alpha_i(x)$ via the splitting procedure described later in this section. In the framework of the Generalized Random Forest, they allow to estimate the targets of interest via

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

We 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 then proceeds greedily as proposed by Breiman (2001: 5) and divides P into two children C_1, C_2 using an axis-aligned split that maximizes the following criterion

$$\Delta(C_1, C_2) \coloneqq \frac{n_{C_1} n_{C_2}}{n_P^2} \Big(\hat{\theta}_{C_1}(\mathcal{S}) - \hat{\theta}_{C_2}(\mathcal{S})\Big)^2 \tag{3.12}$$

given $n_L = |\{i \in \mathcal{S} : X_i \in L\}|,\ L \in \{P, C_1, C_2\}$ (Athey et al. 2019: 1153–1156). This essentially corresponds to making the heterogeneity of in-sample $\hat{\theta}$ as

[2]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 (Athey et al. 2019: 1149).

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, Athey et al. (2019: 1156) 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, we can compute the forest-based similarity weights via aggregation of the tree-based weights as given in equation 3.7. Using these weights, the heterogeneous treatment effects can be estimated using equation 3.8 as described before.

To summarize: The Causal Forest takes advantage of a consistent estimator of the constant treatment effect proposed by Robinson (1988), which is given by equation 3.6. This is justified by partitioning the covariate space into subspaces for whom the assumption of constant treatment effects can be justified. The partition is created using the Generalized Random Forest that determines the splits using an approximate criterion of equation 3.12. Finally, the heterogeneous treatment effects are estimated via equation 3.8 while $\hat{e}(x)$ and $\hat{m}(x)$ are estimated in a first step.

3.2 Inference for Causal Forest

Being able to conduct statistical inference on the estimated treatment effects as described in section 3.1 is crucial. The ability to infer the significance of an estimate is critical for applied research. We hence briefly present the most important results.

Wager and Athey (2018) provide asymptotic theory for predictors resulting from averages over trees based on U-statistics (Hoeffding 1948). 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 constraints are more technical and can be found in Wager and Athey (2018: 1233–1234). Under relatively weak assumptions and writing $\mu(x) = \mathbb{E}[Y \mid X = x]$[3], they

[3]Note that we previously used $m(x)$ for the conditional expectation function $\mathbb{E}[Y \mid X = x]$ in section 3.1 but we accept the notation from Wager and Athey (2018: 1233–1234) for consistency.

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. Athey et al. (2019: 1158–1163) impose the same restrictions on their trees as Wager and Athey (2018). Further, they manage to restate the Generalized Random Forest as a pseudo-forest to make the estimates $\hat{\theta}(x)$ an average of estimates over different trees which enables them to use the asymptotic theory established by Wager and Athey (2018). In doing so, they show that equation 3.13 holds for $\hat{\theta}(x)$ estimated by the Generalized Random Forest as well.

Additionally, they develop a method to construct asymptotically valid Gaussian 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 Forest and Panel Data

As we are interested in the estimation of heterogeneous treatment effects from panel data, we shortly describe the nature of panel data and its emerging challenges. Further, we present the properties of the Causal Forest that make it suitable for such a setting. Note that we examine these claims carefully in the following simulation study in section 4.

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. The Causal Forest is however developed for independent and identically distributed data. Yet the algorithm exhibits properties that enable the integration of the correlation-structures induced within each entity and over time into the estimation of treatment effects.

Firstly, we have the possibility to adjust for within-entity correlations by using the cluster-robust Causal Forest, which is provided in the R-package `grf`. Athey and Wager (2019: 40) explain that the possibility of clustering the Causal Forest allows for non-parametric random effects modeling. 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. The Random Forest has been extensively used in panel data settings. For example, Gu et al. (2020) provide an extensive analysis showing that forest-based methods are well-suited for forecasting stock returns from panel data. Furthermore, the researcher can include variables that seek to capture the effect of time such as lagged variables and time-indicators.

Considering the enumerated properties of the Causal Forest, we are confident that it is capable of dealing with the discussed difficulties.

4 Monte Carlo Simulation

In order to understand the performance of the Causal Forest in the setting of panel data, we run an extensive simulation study. The specific aim of this section is to understand under which circumstances the Causal Forest works well and under which circumstances it works less reliably. To do this, we impose different variations on the data generating process. The data generating process is described in section 4.1 while we elaborate on the delay of the treatment effect in section 4.2. Further, we experiment with the cluster-robust Causal Forest and study its advantage over the standard Causal Forest. The evaluation criteria for the experiments can be found under section 4.3 while we present the results in section 5.3. The Causal Forest algorithm as described in section 3.1 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 $(X_{it}, Y_{it}, W_{it}), i = 1, \ldots, n$ and $t = 1, \ldots, T$ as observations of the same n entities over T time periods. Given this structure, we want to examine how well the Causal Forest handles 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 the Causal Forest can withstand observed and unobserved confounding variables. We discuss the data generating process in the following and explain how these difficulties are introduced into the data.

Define the following functions used in the data generating process:

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

The baseline data generating process is inspired by Wager and Athey (2018: 1238) which in turn draws from Robinson (1988: 942–943) 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 to use 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 Bernoulli distribution with treatment propensity $e(X)$ meaning $W \sim \text{Bernoulli}(e(X))$. The treatment propensity $e(X)$ is set according to Wager and Athey (2018: 1238). 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 as defined in section 2 holds, 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 which is demonstrated in section 5.3.1. 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 (Chatfield and Xing 1981: 35–38). The specification of the main effect $m(X) = X_1 + X_3$ is chosen to be simple and is inspired by Wager and Athey (2018: 1238) 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 Wager and Athey (2018: 1238). They use a slightly more complex treatment function $\tau(X)$ as they investigate the comparative performance of the Causal Forest against nearest-neighbor approaches and they thus want to understand which method reigns superior in learning a complex

estimand. Putting everything together, we get:

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

As discussed before, we want to test the performance of the Causal Forest 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 (Greenland et al. 1999: 30). We experiment with both types of confounding. Given access to the causal Bayesian network of the data generating process, we define confounding as follows (Pearl and Paz 2014: 76)

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 variable X_1. It follows from equation 4.1 that the treatment assignment and the potential outcomes $\{Y^{(1)}, Y^{(0)}\}$ are dependent. This structure is implemented in all our experiments as observed confounding is certainly present when estimating treatment effects from panel data and we thus want to understand how well the Causal Forest controls for these confounding variables.

The second issue concerning confounding when estimating treatment effects from observational panel data are unobserved confounding variables. Unobserved confounding is in principle the same as observed confounding but with

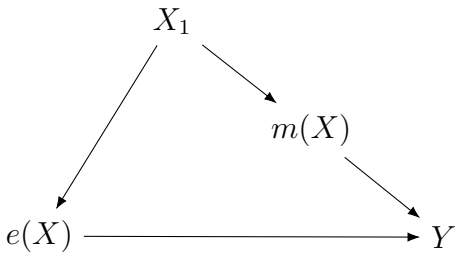

Figure 1: Graph of causal Bayesian network with confounding variable X_1. Note that there are two directed paths being $\{X_1 \to e(X) \to Y\}$ and $\{X_1 \to m(X) \to Y\}$ from X_1 to Y which means that $p(y \mid x_1) \neq p(y \mid do(x_1))$. If X_1 is observed, we call X_1 an observed confounder whereas we call X_1 an unobserved confounder if X_1 is not measured.

the confounding variable not being measured. Note that the presence of unmeasured confounding generally impedes identification and we do not claim that the Causal Forest bypasses that restriction (Veitch et al. 2019: 1). However, we are interested in gauging the robustness of the performance of the Causal Forest under unmeasured confounding.

4.2 Delay of Treatment Effects

To assume that the assignment of treatment has a contemporaneous effect on the response variable is unreasonable in the setting of panel data. In most applications, the treatment influences the response with some delay. In this section, we go through how we incorporate the delay when estimating treatment effects from panel data using the Causal Forest.

Note that the explanations below hold for any panel data set but we follow Chernozhukov et al. (2021: 30) who propose the following specification of the underlying data generating process for COVID-19 data:

$$\boldsymbol{Y}_{it+l} = \boldsymbol{\alpha}^\top \boldsymbol{B}_{it} + \boldsymbol{\pi}^\top \boldsymbol{P}_{it} + \boldsymbol{\mu}^\top \boldsymbol{I}_{it} + \boldsymbol{\delta}_y^\top \boldsymbol{W}_{it} + \boldsymbol{\epsilon}_{it}^y \tag{4.3}$$

where $\boldsymbol{Y}_{it}$ is the vector that contains the growth rate of COVID-19 infections and $\boldsymbol{B}_{it}, \boldsymbol{P}_{it}, \boldsymbol{I}_{it}, \boldsymbol{W}_{it}$ are the matrices that contain the behavioral, policy, further information and test-growth variables. The researcher has to pay attention to two pitfalls when applying such models being firstly the correct choice

of lag length l. There is a large body of research on sensible lag lengths l for COVID-19 data (Cheng et al. 2021: 2). Secondly, the model has to exhibit robustness in regards to modest changes in lag length l as the true lag length l will never be known. We follow these principles closely in the empirical application in section 5. For the simulation study however, we mainly test the scenario where the correct lag length l is chosen. For the model determined by equation 4.3, this implies that given access to data $\{\boldsymbol{Y}_{it}, \boldsymbol{B}_{it}, \boldsymbol{P}_{it}, \boldsymbol{I}_{it}, \boldsymbol{W}_{it}\}_{i,t=1}^{I,T}$ that are generated via equation 4.3, this would entail the application of the transformation $L^{-l}\boldsymbol{Y}_{it} = \boldsymbol{Y}_{it+l}$ where L is the lag-operator[4] (Chatfield and Xing 1981: 34–35). This is in our opinion the only sensible examination within the simulation study as the only channels through which information can pass from time t to time $t+l$ are the auto-correlated error term and the persistence of the treatment variable per entity i. We nevertheless report the results for the case where the choice of lag length l is off by $t = 7$ days[5].

In summary, we think that this issue is best approached by choosing a lag length l suggested by the literature concerning the specific empirical application and subsequently examining the robustness of the estimated results in regards to that choice.

4.3 Performance Evaluation

The performance of the Causal Forest is evaluated in terms of the mean squared error (MSE) and the Bias for estimating $\tau(X)$ and τ at a random test point x. Further, we compute the expected coverage of $\tau(X)$ and τ with a targeted coverage rate of 0.95 at a random test point x. We report both the MSE and Bias as it allows us to understand the composition of the MSE via the Bias-Variance decomposition[6] (James 2003: 117). Let S be the number of simulated data sets per configuration and n the number of observations. Further, let $\mathbb{1}_i = 1$ if the 95% Gaussian confidence interval for observation i contains the true parameter τ_i and $\mathbb{1}_i = 0$ otherwise. For each of the S data sets, we

[4] In general, the researcher wants to match the lag-structure encoded in the data generating process

[5] Concretely, we run the Causal Forest on $Y_{it+7} = m(X_{it}) + \tau(X_{it})W_{it} + \epsilon_{it}$ while the data generating process is given by $Y_{it} = m(X_{it}) + \tau(X_{it})W_{it} + \epsilon_{it}$

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

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 data sets 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. Note that we compute and report these measures for both the CATE as well as the ATE.

4.4 Results

In this section, we introduce the two main simulation designs and elaborate on the results. For all the setups, we make the following choices regarding the parameters of the data generating process: The dimension of the covariate 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 empirical application in section 5. The number of observations varies while we hold the number of entities $\mathcal{I} = |i \in \{1, \ldots, I\}|$ constant at $I = 10$ and change the number of time periods $\mathcal{T} = |t \in \{1, \ldots, T\}|$ with $T \in \{200, 1000\}$. The coefficient of the auto-correlated error term process $\{u_{it}\}$ is $\rho = 0.2$. We also implement $\{\alpha_i, \nu_{it}\} \sim \mathcal{N}(0, 1)$. Lastly, the number of simulations per configuration is $S = 100$. Note that we present the results for the correct choice of lag length l as explained in section 4.2 while the results for the incorrect lag length choice can be found in table 4.

In the following, we describe the two simulation designs and discuss the results. In section 4.4.1, we examine the importance of the cluster-robust Causal Forest while the influence of unobserved confounding variables is investigated in section 4.4.2.

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 auto-correlated error process $\{u_{it}\}$ and an entity fixed effect α_i that represents the effect of the entity i on the outcome Y_i. To gauge the ability of the Causal Forest to capture these entity fixed effects α_i, we employ the algorithm on the data described once while using the cluster-robust Causal Forest 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 across the standard and the cluster-robust Causal Forest. The cluster-robust Causal Forest however reigns supreme on the coverage rates achieving coverage rates that are on average 30% higher compared to the standard Causal Forest. 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 (Athey and Wager 2019: 40). We also note that the overall performance of the clustered Causal Forest 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 1 and table 2 in appendix A.1.

As clustering for entities improves the performance considerably, we work with the cluster-robust Causal Forest 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 the Causal Forest handles data with observed confounding present adequately as seen from Design 1, we want to test the performance when there is an unobserved confounding variable included in the data generating process.

We generate the data as described in equation 4.1 while we deviate in regard to two aspects. Firstly, we make the treatment propensity function $e(X)$ 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

$$e(X_{it}) = \frac{1}{4}(1 + \beta_{2,4}(\bar{X}_{12it}))\kappa, \ \kappa = 0.05 \tag{4.4}$$

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.

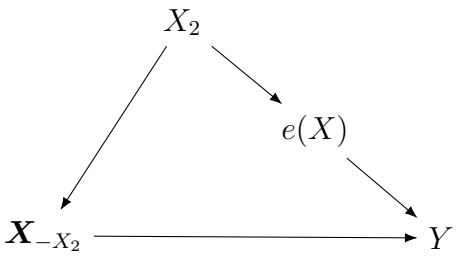

Figure 2: Graph of causal Bayesian network with unobserved X_2. X_2 is a confounding variable in this structure as X_2 influences $\boldsymbol{X}_{-X_2}$ through $\tilde{X}_1$ and Y through $e(X)$.

We now present the results for Design 2 in more detail as it resembles the estimation of treatment effects from panel data 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 magnitude of the MSE and Bias to the treatment effect function $\tau(X)$[7]. The performance regarding MSE and Bias is good which can be seen in table 3. We observe that MSE and Bias are smaller for $T = 1000$ compared to $T = 200$ when holding d constant. The coverage rate falls short of the targeted coverage rate of 0.95 by a small margin. 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 the data contains entity

[7]Remember that $\tilde{X}_1 = X_1 + 0.5X_2$ 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

fixed effects, an auto-correlated error term process, unobserved and observed confounding variables and the Causal Forest estimates a complex treatment function $\tau(X)$, 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 the Causal Forest as described in section 5.3.2. To achieve this, we use data generated from Design 2 while choosing $d = 24$, $T = 200$ and $I = 10$. We chose this combination as it most closely represents the data from the empirical application in section 5. We generate $S = 100$ data sets. The calibration regression explained in section 5.3.2 is estimated on all $S = 100$ runs. In figure 3, we see that $\hat{\alpha}_s$ are centered around 1.00 with 95% of the mass being between 0.89 and 1.17. All the p-values of the one-sided hypothesis tests of $\mathrm{H}_0 : \hat{\alpha}_s \leq 0$ lie within $[0, 0.01]$, which implies that the ATE is well estimated. The heterogeneous treatment effects on the other side appear to be calibrated worse with the $\hat{\beta}_s$ being far from 1. We note that $\tau(X)$ is flat for all $\tilde{X}_1 \in [1, 1.5]$ which is a possible explanation for the bad results in the calibration regressions. To investigate this issue, we run the proposed procedure with $\tilde{X}_1 \in [0, 1]$ such that the heterogeneous treatment effects are not largely constant. In doing so, we observe the same pattern as before. According to the calibration regressions, the Causal Forest accurately estimates the average treatment effect but fails 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 Forest's ability to estimate heterogeneous treatment effects from from panel data. This conclusion can be drawn as we know that the heterogeneous treatment effects are well estimated using the criteria set forth in section 4.3. This is of course problematic as the applied researcher heavily depends on the calibration regression to evaluate the performance of his Causal Forest.

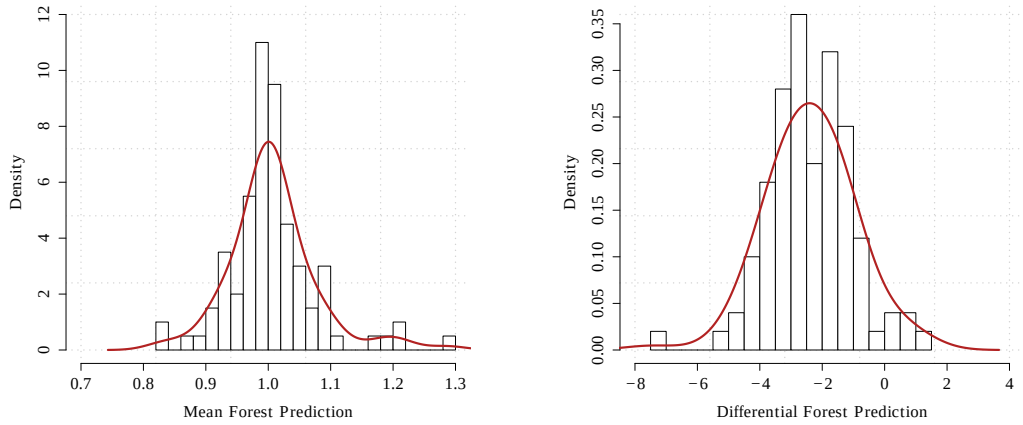

Figure 3: Histogram of $\hat{\alpha}_s$ over the $S = 100$ runs on the left and histogram of $\hat{\beta}_s$ over the $S = 100$ runs on the right. The $\hat{\alpha}_s$ are centered around 1.00 with $\hat{\sigma}_{\hat{\alpha}_s} = 0.07$ whereas the $\hat{\beta}_s$ are centered around -2.45 with $\hat{\sigma}_{\hat{\beta}_s} = 1.33$

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

The great uncertainty caused by the COVID-19 pandemic poses large difficulties for all societies around the world. In close collaboration with science, politicians try to balance the containment of the spreading of COVID-19 and to return back to life as it was before COVID-19. In doing so, they have to make decisions based on scarce information. To quantify the effects 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. For this reason, we dedicate the empirical application of this thesis to the evaluation of COVID-19 related policies.

5.1 Introduction

After the country-wide lockdown that Switzerland started to exit beginning on April, 27, 2020, each of the 26 cantons that make up Switzerland was given partial political autonomy concerning the introduction of COVID-19 containment measures on June 19, 2020[8]. The partial autonomy created substantial differences concerning the implementation of COVID-19 related policies across cantons. Concretely, 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 compared to the lower bound determined by the federal government. Over the course of the next two months, 10 out of the 26 cantons implemented the stricter facial mask policy. On October 19, 2020, the federal government established multiple country-wide policies, including a tighter facial mask policy as COVID-19 cases increased drastically. We use these cantonal-level differences over the specified period of analysis to infer the effect of the establishment of the stricter facial mask policy on the containment of the spreading of COVID-

[8]The events are described here: https://www.bag.admin.ch/bag/de/home/das-bag/aktuell/medienmitteilungen.msg-id-79522.html

19 in Switzerland. In Switzerland, the stricter facial mask policy corresponds to mandatory mask wearing in all public or shared spaces where social distancing is not possible. See figure 5 for more details.

The contribution of this section lies firstly in the provision of an estimate of the effect of a stricter facial mask policy on reducing the dissemination of COVID-19. We measure the spreading of COVID-19 through the effective reproductive number denoted by R_e. To our knowledge, this is the first estimate of the effectiveness of facial mask policies in Switzerland. Secondly, we apply the Causal Forest in a scenario where methods such as Differences in Differences (Bertrand et al. 2004), Synthetic Controls (Abadie and Gardeazabal 2003) or Fixed-Effects estimators (Bell et al. 2019) are usually employed. We show that the Causal Forest provides a viable alternative that produces convincing estimates.

The results show that the introduction of making facial masks mandatory in all public or shared spaces where social distancing is not possible has a significant estimated average treatment effect of ATE $= -0.044$ on $\ln R_e$. This corresponds to lowering the effective reproductive number by 5% which is explained in more detail in section 5.3.2. We do not find evidence for treatment effect heterogeneity induced by any observed variable nor do the treatment effects vary substantially across cantons and time.

In the next section 5.2, we present the data and in section 5.3, the results are discussed while we elaborate on the estimated average treatment effects and the estimated heterogeneous treatment effects separately.

5.2 Data

The identification of heterogeneous treatment effects relies heavily on the Causal forest's ability to create a sensible partition of the covariate space $\mathcal{X}$. Therefore, we construct a rich data set that contains different types of information. The data can be classified into four categories being epidemiological data in section 5.2.1, baseline cantonal data in section 5.2.2, behaviorial data in section 5.2.3 and policy data in section 5.2.4. We present all types of data and document where they can be accessed in table 5. We further include a complete composition of the variables in table 6. Note that most of our data is at the cantonal-level and at a daily frequency to suit the analysis.

5.2.1 Epidemiological Data

To accurately represent the epidemiological situation, we construct a data set that contains information on COVID-19 cases, capacity and utilization of the health case system as well as R_e and related measures.

The logarithm of the effective reproductive number R_e 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 (Ashcroft et al. 2020: 2). Pleninger et al. (2021: 44) show that the γ-day growth rate of new infections can be approximated by $\ln \mathrm{R}_{e,t}$ assuming that the transmission takes place γ days after infection. Hence, it is very natural to use it as the response variable.

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

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

Using $l = 7$ days lies in accordance with estimates of the incubation period and further with the delay of political measures on behavior changes (Cheng et al. 2021: 2). We additionally work with $l \in \{5, 7, 8, 10, 14\}$ and show that the results are robust to the choice of l in appendix A.2.3.

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

5.2.2 Baseline Cantonal Data

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

These variables are important as evidence shows that infection growth is

strongly linked to measures of residential density (Wheaton and Kinsella Thompson 2020: 5–8). Further, Zoran et al. (2020: 5–7) suggest that weather conditions are closely linked to infection growth of COVID-19. They find in particular that dry air supports the transmission of COVID-19[9]. To seize these relationships, we assemble weather data from 14 weather stations. We solely use weather stations that represent the canton's weather accurately which means that we exclude stations situated on mountains. Next, we match each of the 26 cantons to the nearest weather station to get an accurate characterization of each canton's weather.

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

Another important factor in the composition of R_e is geographical proximity to an area of large growth rates of COVID-19 cases. Kapitsinis (2020: 1038–1040) find that there are distinct spatial patterns in the dynamics of the infection growth of COVID-19 which means that geographical proximity of cantons has to be modeled. As we need a cantonal indicator to run the cluster-robust Causal Forest, we seek to build an indicator with 26 distinct values that represent the spatial proximity of the cantons. To do this, we employ classical MDS (Kruskal 1964). First, the euclidian distance matrix $\boldsymbol{D} = \{d_{ij}\}_{i,j=1}^{26}$ between all the canton's main cities is computed[10]. We then apply classical MDS on $\boldsymbol{D}$ to project $\boldsymbol{D}$ from $\mathbb{R}^2 \rightarrow \mathbb{R}$ while preserving the euclidian distances as best as possible. Lastly, we sort the projected distances and replace them with the sequence of natural numbers from 1 to 26 preserving the previous sorting.

5.2.3 Behavioral Data

We collect data on household spending as well as mobility data to quantify the population's behavior during the period of analysis.

To obtain a measurement of household spending, we use credit card and debit card transactions as well as bank transfers from mobile phones. We

[9]Their findings are supported by Zhu et al. (2020) and Fattorini and Regoli (2020)

[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 (Rapp 1991: 24)

include transfers where the origin of the cardholder can either be domestic or foreign. As we work with cantonal-level data and absolute numbers are not of particular interest, we compute the daily growth rates of the number of transactions and the amount spent in CHF. Pleninger et al. (2021: 19) argue that consumption can be used as a proxy for the level of social distancing. We therefore do not include e-commerce transfers.

Changes in behavior can also be identified through changes in personal mobility. We compute the daily growth rate of the median distance traveled per person and day in kilometers.

5.2.4 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.

The approach that is implemented by the KOF closely follows the Oxford Stringency index (Pleninger et al. 2021: 11–15; Hale et al. 2021: 533–542). 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 government in regards to COVID-19 policies. The Oxford Stringency index is computed at the country-level. The KOF Stringency-Plus index (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 regulations concerning facial masks, school closings, workplace closings, 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 campaigns (Pleninger et al. 2021: 13). Compared to the Oxford Stringency index, they add the 10^{th} indicator on facial masks and make a slight modification to the coding of workplace closing[11]. We exclude

[11] Indicator-coding: https://github.com/OxCGRT/covid-policy-tracker/blob/master

indicators from the analysis that neither vary across cantons over the period of analysis nor change at the national-level during that period. The three remaining indicators are cancellation of public events, restrictions on gatherings as well as facial masks. See figure 5 for a graphical overview of the facial mask policies over the period of analysis.

5.3 Results

All our models are estimated using the Causal Forest from the R package `grf` while clustering at the cantonal-level. We use the design matrix $\boldsymbol{X}$ with $\dim(\boldsymbol{X}) = 1560 \times 24$ described in table 6. Given that there are 26 cantons, we end up with data for $T = 60$ days. We dispense with preselection of variables by firstly running a Causal Forest to determine feature importance as described by Athey and Wager (2019: 42) as it unnecessarily reduces the number of features while not improving performance in our case[12].

In the following, we firstly discuss if the baseline assumption of overlap is fulfilled in section 5.3.1. We then evaluate the estimated average treatment effect while we further report the average treatment effect of the treated (ATT) and the overlap-weighted ATE in the appendix A.2.3. In section 5.3.3, we search for treatment effect heterogeneity and examine if an observed variable is predictive of the heterogeneity.

5.3.1 Assumptions

The identifying assumption of overlap as discussed in section 2 can be graphically evaluated. Remember 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 the Causal Forest as discussed in section 3.1 means that we have access to an estimate of $e(X)$. By looking at the histogram of $\hat{e}(X)$ in figure 7, 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

[12] We show in the appendix A.2.3 that our results are robust to preselection

is expected for example from data that come from a randomized control trial. Given $\min(\hat{e}(X)) = 0.045$ and $\max(\hat{e}(X)) = 0.880$ however, we can conclude that the assumption of overlap is met.

5.3.2 Average Treatment Effect

To understand the average effect of introducing the stricter facial mask policy as explained in section 5.1 in Switzerland on the effective reproductive number is of central interest. We hence first abstract from the ability to estimate heterogeneous treatment effects and estimate the average treatment effect denoted as τ. Athey and Wager (2019: 42) define the aggregation of the estimated ATE $\hat{\tau}$ for the cluster-robust Causal Forest 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 (Glynn and Quinn 2009: 39). $\hat{\tau}_j$ is the average treatment effect for canton j. 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 the Causal Forest 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[13]. 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 spreading of COVID-19 in the period of analysis.

[13]The analysis for the ATT and overlap-weighted ATE can be found in the appendix A.2.3

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 $\mathrm{R}_0 > 1$ correspond to exponential spreading of infectious diseases while values of $\mathrm{R}_0 < 1$ are associated a decline in the spreading of the disease (Purkayastha et al. 2021: 5). 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 6. Going with the prediction of 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 Chernozhukov et al. (2018: 5–6). 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 the Causal Forest, this translates 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 $\bar{\tau} = n^{-1} \sum_{i=1}^{n} \tau^{(-i)}(X_i)$. 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 (Athey and Wager 2019: 43). We obtain a coefficient of $\hat{\alpha} = 0.978$ with a standard error of 0.638. We thus can reject the null hypothesis $H_0 : \hat{\alpha} \leq 0$ in the one sided test at the 10% confidence level.

This proposes that the average prediction is correct, which is of course very promising. Further, Athey et al. (2017: 278) derive an expression for the Bias of the estimated average treatment effect given by

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

where $b(x)$ 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 $b(x) = (e(x) - p)p(\mu_0(x) - \mu_0) + (1-p)(\mu_1(x) - \mu_1)$. In figure 8, 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.

5.3.3 Treatment Heterogeneity

As the Causal Forest is capable of estimating heterogeneous treatment effects, we want to examine if there is heterogeneity in the estimated treatment effects and subsequently how well the heterogeneity is estimated.

Looking at figure 4 might lead to the belief 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 $\hat{\tau}$ (Athey and Wager 2019: 43).

To understand whether a variable of our design matrix $\boldsymbol{X}$ is predictive of the estimated heterogeneous treatment effects helps to grasp through which channels the introduction of the stricter facial mask policy affects the spreading of COVID-19. To test this, we compute the best linear projection of the estimated heterogeneous treatment effects onto $\boldsymbol{X}$. Writing $\hat{\boldsymbol{\tau}}$ for the vector that contains the estimated heterogeneous treatment effects with $\dim(\hat{\boldsymbol{\tau}}) = 1560$, we estimate the linear model $\hat{\boldsymbol{\tau}} = \boldsymbol{\beta}_0 + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ using heteroskedastic-robust

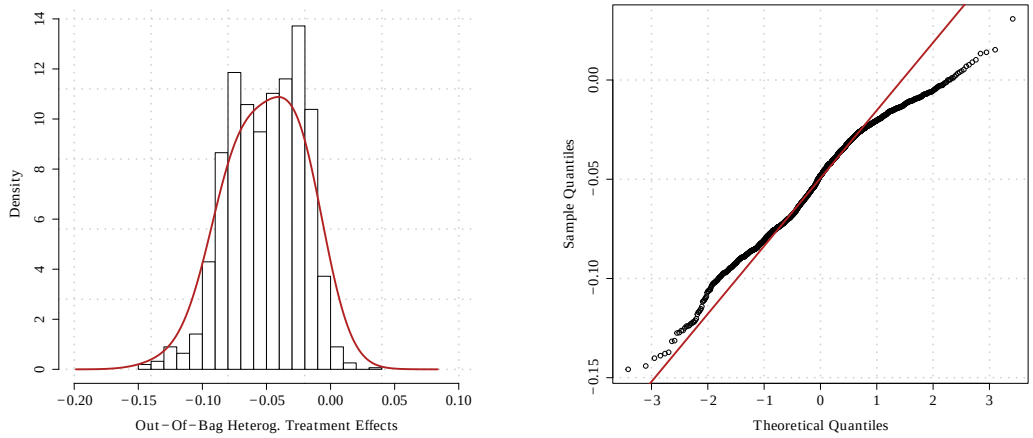

Figure 4: Histogram of out-of-bag heterogeneous treatment effects from the Causal forest on the left. QQ-plot against normal distribution of the out-of-bag heterogeneous treatment effects displaying some light tails on the right

(HC3) standard errors. Apart from the coefficient on government response index (`gov_resp`), we cannot reject the null-hypothesis' of $H_0 : \hat{\beta}_j = 0,\ j = 1, \ldots, 24$ 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.

With the conclusion that none of the variables seem to have predictive power on the estimated heterogeneous treatment effects, we seek to investigate cantonal-level heterogeneity. To achieve this, we compute the best linear projection of the estimated cantonal average treatment effects $\hat{\tau}_j$ from equation 5.5 on $\boldsymbol{X}_{Cantonal} \subset \boldsymbol{X}$. The subset $\boldsymbol{X}_{Cantonal}$ 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 conclude that we do not find evidence for treatment heterogeneity. See table 7 for more details.

Judging the accuracy of the estimates of treatment heterogeneity is enabled

by the calibration procedure described in section 5.3.2. Estimating equation 5.6 yields a coefficient of $\hat{\beta} = -0.670$ with a standard error of 0.974. 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 $\tau(X_i)$ (Athey and Wager 2019: 43). To distinguish 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 the Causal Forest. With this scenario in mind, we run calibration regressions on estimates from the Causal Forest on simulated data where treatment heterogeneity is present in section 4.4.2. Remember that we find that the calibration regressions underestimate the ability of the Causal Forest to accurately estimate heterogeneous treatment effects. This entails that the heterogeneous treatment effects might be estimated more accurately than suggested by $\hat{\beta} = -0.670$.

To summarize, we do not find a strong association of any observed variable with the estimated heterogeneous treatment effects. It is further not possible to judge adequately if the CATE is estimated precisely using the calibration regression procedure. Hence, the estimated average treatment effect is more credible in our opinion.

6 Conclusion

This thesis investigates using the Causal Forest to estimate treatment effects from panel data. We conduct a first of a kind Monte Carlo simulation study for the Causal Forest in the setting of panel data. We demonstrate the importance of the cluster-robust Causal Forest as the ability of clustering on the entity-level improves the coverage rate on average by 30%. Furthermore, we show that the cluster-robust Causal Forest adequately deals with the difficulties encompassed in panel data. This not only includes within-entity correlations, correlation-structures induced over time but also observed and unobserved confounding variables. For the main simulation design, the Causal Forest reaches an equally-weighted average coverage rate of 89% for the CATE and 88% for the ATE. Additionally, we propose an application-driven approach for the choice of lag length l where l should be chosen according to specific a-priori knowledge of the subject examined and subsequently checked for robustness.

Further, we apply the Causal Forest to policy evaluation. Using cantonal-level and daily-frequency data from Switzerland, we find that the introduction of the stricter mask policy reduced the effective reproductive number R_e on average by 5% which is statistically significant at the $\alpha = 5\%$ level. We do not find evidence for treatment heterogeneity. This work provides a first of a kind estimate in regards to the effectiveness of facial mask policies on the containment of the spreading COVID-19 for Switzerland.

There are natural continuations of this thesis regarding both the simulation study and the empirical application. We would love to conduct a comparative simulation study where we examine the performances of the Causal Forest, Difference in Difference estimators, Synthetic Control estimators and Fixed-Effects estimators. This allows for a better classification of the results obtained in this simulation study. Another interesting area of further research is the investigation of the ability of the calibration regressions to accurately represent the Causal Forest's ability to estimate the treatment effects. In this simulation, they underestimate the Causal Forest's ability to estimate CATE which is problematic as the applied researcher depends largely on these regressions to evaluate the performance of his Causal Forest. Lastly for the simulation study, we would like to examine how violations of the assumption of overlap translate

into the performance of the Causal Forest as this is something that can occur when working with panel data.

For the analysis of COVID-19 policies, we would love to examine the effects of restrictions on gatherings as well as cancellation of public events on slowing down the spreading of COVID-19 in Switzerland. These policies exhibit large differences across the cantons in the period of analysis described in this thesis (and are thus controlled for in this thesis). The estimation of these effects however calls for an extension of the Causal Forest called the Multi-Armed Causal Forest that allows for non-binary treatments which is the case for these policies.

References

Abadie, A. and J. Gardeazabal (2003). The economic costs of conflict: A case study of the Basque country. *American Economic Review 93*(1), 113–132.

Ashcroft, P., J. S. Huisman, S. Lehtinen, J. A. Bouman, C. L. Althaus, R. R. Regoes, and S. Bonhoeffer (2020). COVID-19 infectivity profile correction. *Swiss Medical Weekly 150*(32), Article 20336.

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

Athey, S., G. Imbens, T. Pham, and S. Wager (2017). Estimating average treatment effects: Supplementary analyses and remaining challenges. *American Economic Review 107*(5), 278–281.

Athey, S., J. Tibshirani, and S. Wager (2019). Generalized random forests. *Annals of Statistics 47*(2), 1148–1178.

Athey, S. and S. Wager (2019). Estimating Treatment Effects with Causal Forests: An Application. *Observational Studies 5*(2), 37–51.

Bell, A., M. Fairbrother, and K. Jones (2019). Fixed and random effects models: making an informed choice. *Quality and Quantity 53*(2), 1051–1074.

Bertrand, M., E. Duflo, and S. Mullainathan (2004). How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics 119*(1), 249–275.

Beygelzimer, A. and J. Langford (2009). The offset tree for learning with partial labels. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 15*(1), 129–138.

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

Chatfield, C. and H. Xing (1981). *The Analysis of Time Series: An Introduction* (1 ed.). London: Chapman and Hall.

Chen, S., Q. Chen, W. Yang, L. Xue, Y. Liu, J. Yang, C. Wang, and T. Bärnighausen (2020). Buying Time for an Effective Epidemic Response: The Impact of a Public Holiday for Outbreak Control on COVID-19 Epidemic Spread. *Engineering 6*(10), 1108–1114.

Cheng, C., D. D. Zhang, D. Dang, J. Geng, P. Zhu, M. Yuan, R. Liang, H. Yang, Y. Jin, J. Xie, S. Chen, and G. Duan (2021). The incubation period of COVID-19: a global meta-analysis of 53 studies and a Chinese observation study of 11 545 patients. *Infectious Diseases of Poverty 10*(1), Article 119.

Chernozhukov, V., M. Demirer, E. Duflo, and I. Fernández-Val (2018, 6). Generic Machine Learning Inference on Heterogeneous Treatment Effects in Randomized Experiments, with an Application to Immunization in India. *NBER Working Paper Series*, Article 24678.

Chernozhukov, V., H. Kasahara, and P. Schrimpf (2021). Causal impact of masks, policies, behavior on early covid-19 pandemic in the U.S. *Journal of Econometrics 220*(1), 23–62.

Crimmins, E. M. (2020). Age-Related Vulnerability to Coronavirus Disease 2019 (COVID-19): Biological, Contextual, and Policy-Related Factors. *Public Policy & Aging Report 30*(4), 142–146.

Ding, P. and F. Li (2018). Causal inference: A missing data perspective. *Statistical Science 33*(2), 214–237.

Dudik, M., J. Langford, and H. Li (2011). Doubly robust policy evaluation and learning. *Proceedings of the 28th International Conference on Machine Learning, ICML 2011 28*(1), 1097–1104.

Eckert, F. and H. Mikosch (2020). Mobility and sales activity during the Corona crisis: daily indicators for Switzerland. *Swiss Journal of Economics and Statistics 156*(1), Article 9.

Fattorini, D. and F. Regoli (2020). Role of the chronic air pollution levels in the Covid-19 outbreak risk in Italy. *Environmental Pollution 264*(1), Article 114732.

Glynn, A. N. and K. M. Quinn (2009). An introduction to the augmented inverse propensity weighted estimator. *Political Analysis 18*(1), 36–56.

Green, D. P. and H. L. Kern (2012). Modeling heterogeneous treatment effects in survey experiments with bayesian additive regression trees. *Public Opinion Quarterly 76*(3), 491–511.

Greenland, S., J. M. Robins, and J. Pearl (1999). Confounding and collapsibility in causal inference. *Statistical Science 14*(1), 29–46.

Gu, S., B. Kelly, and D. Xiu (2020). Empirical Asset Pricing via Machine Learning. *Review of Financial Studies 33*(5), 2223–2273.

Guidotti, E. and D. Ardia (2020). COVID-19 Data Hub. *Journal of Open Source Software 5*(51), Article 2376.

Hale, T., N. Angrist, R. Goldszmidt, B. Kira, A. Petherick, T. Phillips, S. Webster, E. Cameron-Blake, L. Hallas, S. Majumdar, and H. Tatlow (2021). A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker). *Nature Human Behaviour 5*(4), 529–538.

Hill, J. and Y. S. Su (2013). Assessing lack of common support in causal inference using bayesian nonparametrics: Implications for evaluating the effect of breastfeeding on children's cognitive outcomes. *Annals of Applied Statistics 7*(3), 1386–1420.

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics 20*(1), 217–240.

Hoeffding, W. (1948). A Class of Statistics with Asymptotically Normal Distribution. *The Annals of Mathematical Statistics 19*(3), 293–325.

Huisman, J. S., J. Scire, D. C. Angst, R. A. Neher, S. Bonhoeffer, and T. Stadler (2020, 1). Estimation and worldwide monitoring of the effective reproductive number of SARS-CoV-2. *MedRxiv Working Paper Series*, Article 20239368.

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

James, G. M. (2003). Variance and bias for general loss functions. *Machine Learning 51*(2), 115–135.

Kapitsinis, N. (2020). The underlying factors of the COVID-19 spatially uneven spread. Initial evidence from regions in nine EU countries. *Regional Science Policy and Practice 12*(6), 1027–1045.

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika 29*(1), 1–27.

Kullback, S. and R. A. Leibler (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics 22*(1), 79–86.

Li, F., K. L. Morgan, and A. M. Zaslavsky (2018). Balancing Covariates via Propensity Score Weighting. *Journal of the American Statistical Association 113*(521), 390–400.

Neyman, J. (1923). *Sur les applications de la théorie des probabilitès aux experiences agricoles: Essai des principes*. Ph. D. thesis, University of Warsaw.

Nie, X. and S. Wager (2021). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika 108*(2), 299–319.

Pearl, J. and A. Paz (2014). Confounding Equivalence in Causal Inference. *Journal of Causal Inference 2*(1), 75–93.

Pleninger, R., S. Streicher, and J.-E. Sturm (2021, 6). Do COVID-19 Containment Measures Work? Evidence from Switzerland. *KOF Working Paper Series*, Article 494.

Purkayastha, S., R. Bhattacharyya, R. Bhaduri, R. Kundu, X. Gu, M. Salvatore, D. Ray, S. Mishra, and B. Mukherjee (2021). A comparison of five epidemiological models for transmission of SARS-CoV-2 in India. *BMC Infectious Diseases 21*(1), Article 533.

Rapp, R. (1991). *Geometric Geodesy* (1 ed.). Ohio: Ohio State University, Department of Geodetic Science and Surveying.

Robins, J. M. (2004). Optimal Structural Nested Models for Optimal Sequential Decisions. *Proceedings of the Second Seattle Symposium in Biostatistics 179*(1), 189–326.

Robinson, P. M. (1988). Root-N-Consistent Semiparametric Regression. *Econometrica 56*(4), 931–954.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology 66*(5), 688–701.

Thompson, R. and E. Brooks-Pollock (2019). Detection, forecasting and control of infectious disease epidemics: modelling outbreaks in humans, animals and plants. *Royal Society Philosophical Transactions B 374*(1775), Article 20180257.

Tian, L., A. A. Alizadeh, A. J. Gentles, and R. Tibshirani (2014). A Simple Method for Estimating Interactions Between a Treatment and a Large Number of Covariates. *Journal of the American Statistical Association 109*(508), 1517–1532.

Veitch, V., Y. Wang, and D. M. Blei (2019). Using embeddings to correct for unobserved confounding in networks. *Advances in Neural Information Processing Systems 33*(1), Article 1237.

Wager, S. and S. Athey (2018). Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association 113*(523), 1228–1242.

Weisberg, H. I. and V. P. Pontes (2015, 6). Post hoc subgroups in clinical trials: Anathema or analytics? *Clinical Trials 12*(4), 357–364.

Wheaton, W. C. and A. Kinsella Thompson (2020). The Geography of COVID-19 growth in the US: Counties and Metropolitan Areas. *SSRN Electronic Journal*, Article 3570540.

Zhu, Y., J. Xie, F. Huang, and L. Cao (2020). Association between short-term exposure to air pollution and COVID-19 infection: Evidence from China. *Science of the Total Environment 727*(1), Article 138704.

Zoran, M. A., R. S. Savastru, D. M. Savastru, and M. N. Tautan (2020). Assessing the relationship between surface levels of PM2.5 and PM10 particulate matter impact on COVID-19 in Milan, Italy. *Science of the Total Environment 738*(1), Article 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 the cluster-robust and the standard Causal Forest. The data is generated according to Design 1 with $d \in \{3, 4, 5, 6, 7, 8, 24\}$ and $T \in \{200, 1000\}$ and $I = 10$ entities. The standard Causal Forest outperforms the cluster-robust Causal Forest slightly in terms of MSE and Bias^2. However, it clearly falls short in terms of coverage as it fails 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 the cluster-robust and the standard Causal Forest. The data is generated according to Design 1 with $d \in \{3, 4, 5, 6, 7, 8, 24\}$ and $T \in \{200, 1000\}$ and $I = 10$ entities. As in table 1, the standard Causal Forest outperforms the cluster-robust Causal Forest slightly in terms of MSE and Bias^2 but clearly falls short in terms of coverage as it fails 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 1 which suggests that the standard Causal Forest considerably underestimates 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 $I = 10$ 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 it is generally very hard to test an estimator's 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.

True Unobserved Confounding: Results for Incorrect Lag Choice							
		MSE	**Bias**^{2}	**Coverage**	**MSE**	**Bias**^{2}	**Coverage**
d	T	CATE	CATE	CATE	ATE	ATE	ATE
3	200	0.240	0.150	0.812	0.239	0.149	0.800
4	200	0.279	0.193	0.772	0.280	0.193	0.760
5	200	0.335	0.226	0.716	0.337	0.226	0.690
6	200	0.229	0.159	0.777	0.229	0.159	0.760
7	200	0.287	0.195	0.728	0.287	0.195	0.700
8	200	0.232	0.136	0.819	0.232	0.136	0.820
24	200	0.243	0.161	0.765	0.242	0.161	0.710
3	1000	0.291	0.182	0.774	0.286	0.180	0.670
4	1000	0.324	0.212	0.714	0.322	0.209	0.670
5	1000	0.207	0.138	0.771	0.207	0.138	0.760
6	1000	0.268	0.160	0.754	0.270	0.162	0.760
7	1000	0.248	0.156	0.766	0.246	0.156	0.780
8	1000	0.253	0.168	0.707	0.253	0.169	0.710
24	1000	0.210	0.147	0.779	0.211	0.147	0.750

Table 4: Results in terms of the CATE and ATE for the Design 2 while applying the transformation L^{-l} to the response variable Y_{it} to create Y_{it+l} using $l = 7$. We simulate data with $d \in \{3, 4, 5, 6, 7, 8, 24\}$ and $T \in \{200, 1000\}$ with $I = 10$ entities. The results are much worse compared to table 3. The equally-weighted coverage rate for the CATE drops from 89% to 76%. The equally-weighted coverage rate for the ATE drops from 88% to 74%. This means that the Gaussian confidence intervals with confidence level $\alpha = 5\%$ do not contain the true parameter in roughly a fourth of the cases.

A.2 Empirical Application

A.2.1 Tables

Nr.	Data set	URL	Literature
1	COVID19Re_geoRegion	R-Values	✗
2	COVID19Cases_geoRegion	Cases	✗
3	Demographics	Demographics	✗
4	Population	Population	✗
5	nbcn_daily_Kanton	Weather	✗
6	Holidays	Holidays	✗
7	ACQ_POS_Canton	Consumption	✗
8	MobilityKOF	Mobility	(19)
9	Additional Data Switzerland	Policy-Indices	(25)
10	KOFMeasures	COVID-19 Policies	(40)

Table 5: This table lists the 10 data sets used to construct the master data set. The data is separated along their types as described in section 5.2. The data can be downloaded using the specified link under URL. If the authors name an academic paper that has to be cited when using the data, the paper can be found under Literature. The Nr. column is used in table 6 to indicate from which data set a variable stems.

Nr.	Acronym	Description	Frequency & Unit	Literature
1	`r_median`	median R_e	Daily & Cantonal	(40)
2	`deaths`	cum. deaths	Daily & Cantonal	(46)
2	`recovered`	cum. recovered	Daily & Cantonal	(46)
2	`tests`	cum. tests	Daily & Cantonal	(46)
2	`hosp`	curr. hospitalized	Daily & Cantonal	(46)
4	`perc_age`	age ≥ 80 years in %	Constant & Cantonal	(16)
3	`abs_age`	age ≥ 65 years	Constant & Cantonal	(16)
3	`area`	area in ha	Constant & Cantonal	(51)
3	`density`	people per km^2	Constant & Cantonal	(51)
4	`pop`	population	Constant & Cantonal	(51)
3	`beds`	hospital beds per capita	Constant & Cantonal	(46)
6	`holidays`	official school holidays	Daily & Cantonal	(12)
5	`sunshine`	sunshine in minutes	Daily & Cantonal	(53)
5	`temp`	mean air-temperature in C°	Daily & Cantonal	(53)
5	`humidity`	relative humidity in %	Daily & Cantonal	(53)
7	`amount_spent`	growth CHF spent debit, credit and mobiles	Daily & Cantonal	(40)
7	`transactions`	growth transactions debit, credit and mobiles	Daily & Cantonal	(40)
8	`mobility`	median distance in km	Daily & National	(40)
9	`eco_supp`	economic support index	Daily & National	(15)
9	`gov_resp`	government response index	Daily & National	(15)
10	`canc_events`	cancellation of events indicator	Daily & Cantonal	(40)
10	`rest_gatherings`	restrictions on gatherings indicator	Daily & Cantonal	(40)
10	`facial_covering`	facial mask policy indicator	Daily & Cantonal	(40)
10	`kof_stringency`	KOF stringency index	Daily & Cantonal	(40)
✗	`week`	weekly indicator	Weekly & National	(24)
✗	`canton`	cantonal indicator	Constant & Cantonal	(33)

Table 6: 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 5.2 which is represented by the horizontal lines. The column Nr. describes from which data set described in table 5 the respective variable stems from. Note that cum. stands for cumulative. Check out the respective links referenced in table 5 for a detailed description of the policy indicators.

Best Linear Projections of Treatment Effects				
	CATE	**Cantonal ATE**		**CATE**
intercept	2.995	-0.006	temp	0.003
	(5.878)	(0.081)		(0.014)
abs_age	2.262	-0.132	humidity	0.001
	(14.887)	(0.4518)		(0.004)
area	0.000	0.000	transactions	0.022
	(0.000)	(0.000)		(0.021)
density	0.000	0.000	mobility	-0.010
	(0.000)	(0.000)		(0.046)
pop	0.000	0.000	eco_supp	0.000
	(0.000)	(0.000)		(0.003)
holidays	0.105		gov_resp	0.025*
	(0.119)			(0.010)
deaths	-0.001		canc_events	0.435
	(0.005)			(2.608)
recovered	0.000		rest_gatherings	0.198
	(0.001)			(0.450)
tests	0.000		kof_stringency	-0.082
	(0.000)			(0.121)
hosp	0.004		week	-0.006
	(0.013)			(0.066)
sunshine	0.000		canton	-0.009
	(0.000)			(0.060)

Table 7: Table with the results of the best linear projection approach explained in section 5.3.3. Note that the table contains the results for the heterogeneous treatment effects under CATE and the cantonal average treatment effects under Cantonal ATE. The numbers in the brackets are the estimated standard errors and the superscript * denotes statistical significance at the level $\alpha = 1\%$. We do not observe significant relationships apart from `gov_resp`. The estimated coefficient on `gov_resp` is very small and thus not of particular interest. Further, we exclude `beds`, `perc_age` and `transactions` for this model due to problems with multicollinearity. Note that we estimate a variety of models using different subsets of $\boldsymbol{X}$ but none of the approaches suggest that the observed variables have predictive power on the estimated treatment effects. Lastly, the cantonal ATE are only projected on variables that are constant at the cantonal-level as explained in section 5.3.3.

A.2.2 Figures

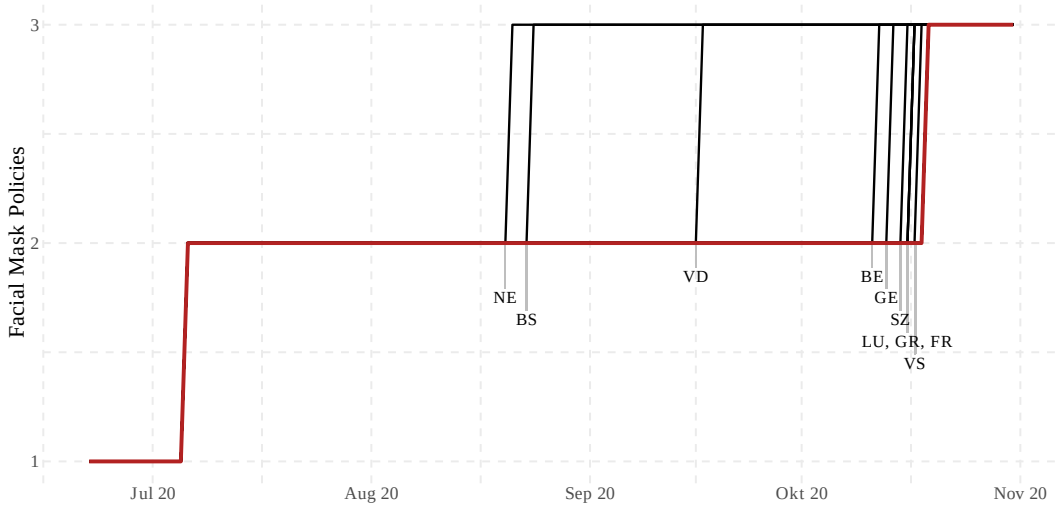

Figure 5: The red line represents the lower bound concerning facial mask policies put in place at the national level. 10 out of the 26 cantons enforced a stricter mask policy than nationally required which is indicated by the lines diverging from the red line. From the $26 \times T = 1560$ data points, we observe $195 = 12.5\%$ where a stricter policy is in place. 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. This is confirmed under: https://www.admin.ch/gov/de/start/dokumentation/medienmitteilungen.msg-id-79711.html. 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 which is confirmed under: https://www.bag.admin.ch/bag/de/home/das-bag/aktuell/news/news-18-10-2020.html. 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 corresponding to a value of 3 and untreated otherwise.

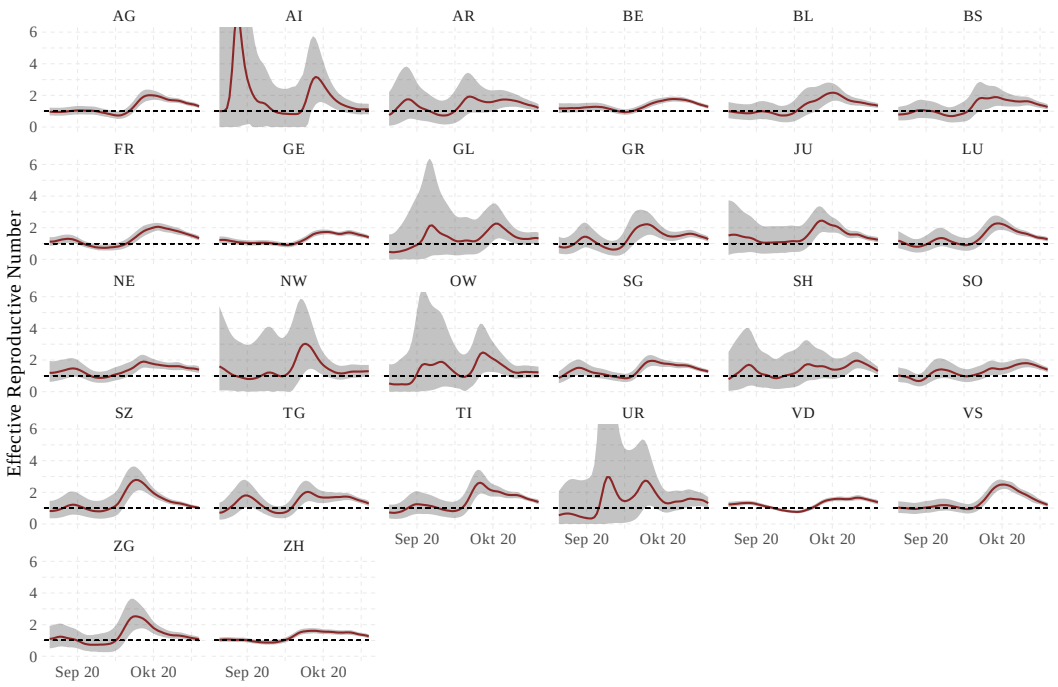

Figure 6: The red line represents the point estimate of effective reproductive number R_e per canton over the period of analysis. The grey band represents the confidence interval of the effective reproductive number as computed in Huisman et al. (2020). The confidence intervals are tight apart from mainly very small cantons where very little infections occurred making inference very difficult. We observe large differences across the cantons which is promising for our approach.

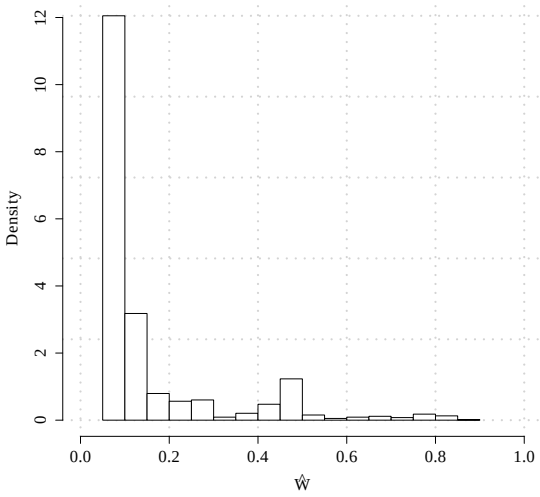

Figure 7: Histogram of $\hat{e}(X)$ from the Causal Forest. The estimated propensity scores are not centered around 0.5 as is expected for example from data coming from a randomized control trial. The propensity scores are however sufficiently bounded away from 0 and 1 with $\min(\hat{e}(X)) = 0.045$ and $\max(\hat{e}(X)) = 0.880$. These estimated propensities are nevertheless the weakest link in the described procedure. However, the estimated overlap-weighted ATE is very similar to the standard estimated ATE which can be seen in appendix A.2.3 which suggests that the estimated propensities do not pose a problem for the Causal Forest in this application.

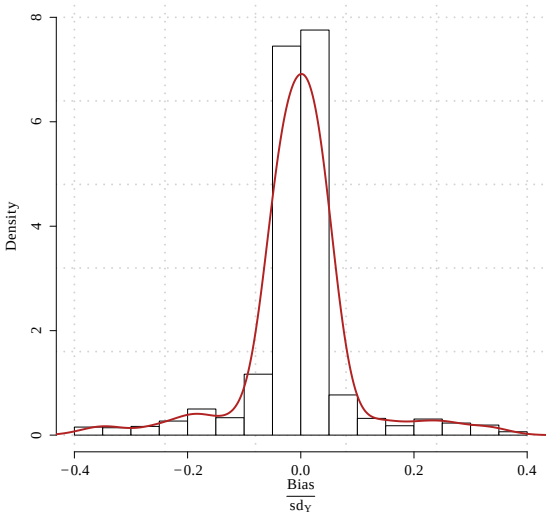

Figure 8: Histogram of $\hat{b}(X)$ from the Causal Forest. The contributions are centered around 0 while 95% of the mass lies within -0.05 and 0.05. This implies that the Bias of estimating the ATE is small.

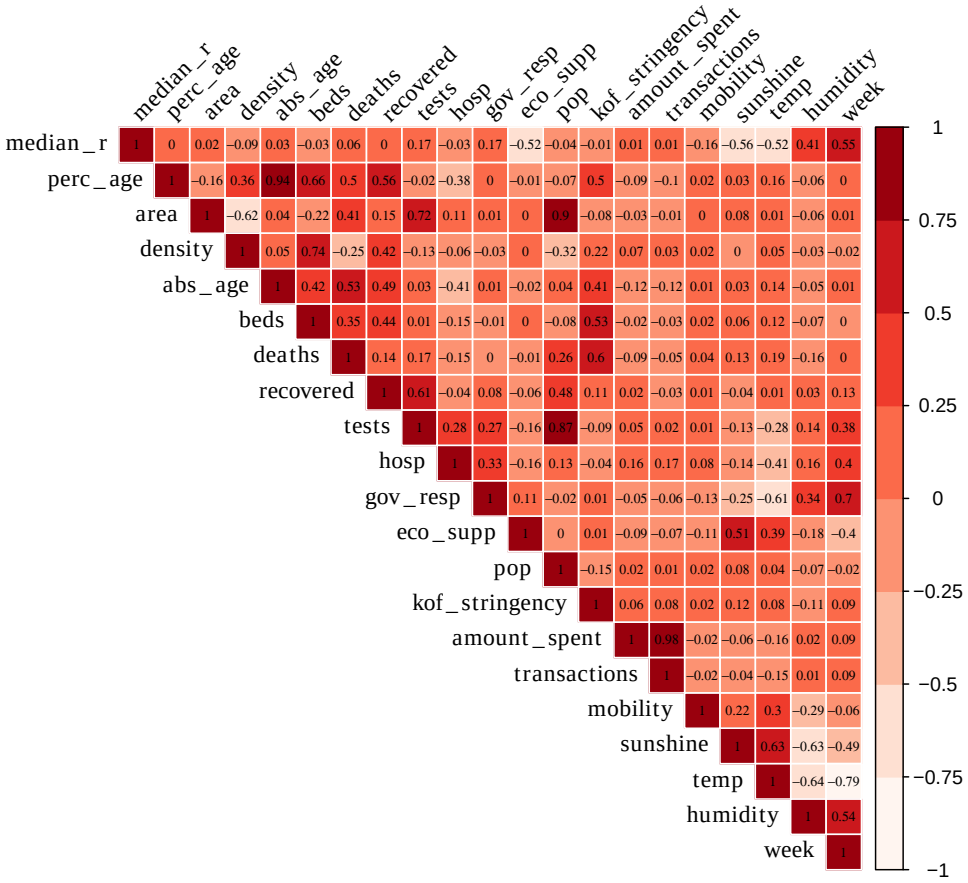

Figure 9: Correlation-matrix of all continuous variables including the indicator week as low values of week can be interpreted as an early time period and high values as a later time period with respect to the period of analysis. Particularly interesting are the high correlations (in magnitude) of the weather variables with the response variable `median_r`.

A.2.3 Robustness-Checks

We present various robustness-checks related to our main results from sections 5.3.2 and 5.3.3. We test the robustness of our results along three general dimensions. Firstly, we examine the effect of the length of the lag l as described in section 5.2.1. Secondly, we have a look at the influence of using the cluster-robust Causal Forest. Lastly, we inspect the influence of preselecting the variables. This is achieved by running two Causal Forests where the first is used to assess variable importance while the second Causal Forest runs on the subset of variables with an above median variable importance. Furthermore, to compare the estimated heterogeneous treatment effects from the different models, we compute the Kullback-Leibler divergence of the distribution of the out-of-bag heterogeneous treatment effects from our main model $\hat{\tau}_{\text{main}}(x)$ to the distribution of the out-of-bag heterogeneous treatment effects from the model tested $\hat{\tau}_{\text{tested}}(x)$. Define the Kullback-Leibler divergence as

$$D_{KL}(\hat{\tau}_{\text{main}}(x) \mid\mid \hat{\tau}_{\text{tested}}) = \int_{-\infty}^{\infty} \hat{\tau}_{\text{main}}(x) \ln \frac{\hat{\tau}_{\text{main}}(x)}{\hat{\tau}_{\text{tested}}(x)} dx \tag{A.1}$$

The Kullback-Leibler divergence is non-negative and the smaller it is, the more similar are the distributions (Kullback and Leibler 1951).

	$(\hat{\tau}, \hat{\sigma}_{\hat{\tau}})$	(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	$(0.978^{\star\star}, -0.865)$
Lag l				
$l = 5$	$(-0.038, 0.018)$	$(-0.049, -0.042)$	0.109	$(1.150^{\star\star}, -1.040)$
$l = 8$	$(-0.039, 0.021)$	$(-0.040, -0.038)$	0.010	$(0.959^{\star}, -0.382)$
$l = 10$	$(-0.037, 0.021)$	$(-0.043, -0.035)$	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`	$(-0.074, 0.007)$	$(-0.033, -0.073)$	1.260	$(0.782^{\star\star}, 1.693^{\star\star})$
Preselection				
$d = 6$	$(-0.042, 0.202)$	$(-0.054, -0.046)$	0.015	$(0.960^{\star\star}, -0.690)$

Table 8: This table lists the conducted robustness-checks. The second column contains 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 $\text{ATT} = \mathbb{E}[Y_i^{(1)} - Y_i^{(0)} \mid W_i = 1]$ while the overlap-weighted ATE is an alternative calculation of the ATE that is more robust to extreme treatment propensities (Li et al. 2018: 393). The fourth column contains the estimated Kullback-Leibler divergence while the last column contains the $(\hat{\alpha}, \hat{\beta})$-coefficients from the calibration regression described in section 5.3.2. Note that the superscript * denotes statistical significance at the $\alpha = 1\%$ and so forth.

We test lags $l \in \{5, 8, 10, 14\}$ which is driven by different estimates of the incubation period as well as behavioral delays to COVID-19 policies (Cheng et al. 2021: 2). The results are stable across the choice of l which is reassuring. Running the estimation while not clustering on the cantons yields an average treatment effect of -0.074 coupled with a very small standard error of 0.007. Athey and Wager (2019: 47) observe a similar pattern. They hypothesize that it is a result of overfitting on the canton-level effects as the unclustered Causal Forest does not take within-canton correlations into account. This might drive the Causal Forest to overestimate the magnitude of the average treatment effect. Furthermore, we observe that the estimated treatment propensities $\hat{e}(X)$ are much closer to 0 and 1 which is problematic. Thirdly, we consider preselection on the variables. The first Causal Forest selects $d = 6$ variables on which the second Causal Forest is deployed. The results remain nearly unchanged, which is encouraging. The estimated Kullback-Leibler divergences are very small apart from the unclustered Causal Forest, which indicates that

the different models estimate similar heterogeneous treatment effects.

Overall, the results are very robust to alterations in the data such as the choice of lag l and preselection. The estimated average treatment effect of the Causal Forest that does use the natural clusters is roughly twice as large than the average treatment effect from our main model which is most likely due to overfitting on the cantonal-level effects.

Statutory Declaration

I hereby declare that the thesis with title

Estimation and Inference of Heterogeneous Treatment Effects with Causal Random Forest

has been composed by myself autonomously and that no means other than those declared were used. In every single case, I have marked parts that were taken out of published or unpublished work, either verbatim or in a paraphrased manner, as such through a quotation.

This thesis has not been handed in or published before in the same or similar form.

Zurich, January 30, 2022