Image-based Treatment Effect Heterogeneity

Connor T. Jerzak CONNOR.JERZAK@AUSTIN.UTEXAS.EDU
Department of Government
The University of Texas at Austin
`ConnorJerzak.com`

Fredrik Johansson FREDRIK.JOHANSSON@CHALMERS.SE
Data Science and AI Division
Chalmers University of Technology
`fredjo.com`

Adel Daoud ADEL.DAOUD@LIU.SE
Institute for Analytical Sociology
Linköping University
`AdelDaoud.se`, `global-lab.ai`

Editors: Mihaela van der Schaar, Dominik Janzing and Cheng Zhang

Abstract

Randomized controlled trials (RCTs) are considered the gold standard for estimating the Average Treatment Effect (ATE) of interventions. One important use of RCTs is to study the causes of global poverty—a subject explicitly cited in the 2019 *Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel* awarded to Duflo, Banerjee, and Kremer "for their experimental approach to alleviating global poverty." Because the ATE is a population summary, researchers often want to better understand how the treatment effect varies across different populations by conditioning on tabular variables such as age and ethnicity that were measured during the RCT data collection. Although such variables carry substantive importance, they are often only observed only near the time of the experiment: exclusive use of such variables may fail to capture historical, geographical, or neighborhood-specific contributors to effect variation. In global poverty research, when the geographical location of the experiment units is approximately known, satellite imagery can provide a window into such historical and geographical factors important for understanding heterogeneity. However, there is no causal inference method that specifically enables applied researchers to analyze Conditional Average Treatment Effects (CATEs) from images. In this paper, we develop a deep probabilistic modeling framework that identifies clusters of images with similar treatment effect distributions, enabling researchers to analyze treatment effect variation by image. Our interpretable image CATE model also emphasizes an image sensitivity factor that quantifies the importance of image segments in contributing to the mean effect cluster prediction. We compare the proposed methods against alternatives in simulation; additionally, we show how the model works in an actual RCT, estimating the effects of an anti-poverty intervention in northern Uganda and obtaining a posterior predictive distribution over treatment effects for the rest of the country where no experimental data was collected. We make code for all modeling strategies available in an open-source software package and discuss their applicability in other domains (such as the biomedical sciences) where image data are also prevalent.

Keywords: Causal inference; Treatment effect heterogeneity; Earth observation; Image data; Probabilistic reasoning

© 2023 C.T. Jerzak, F. Johansson & A. Daoud.

1. Introduction

Field experiments in the social and health sciences help us understand the effects of interventions in the natural habitat where people live (Banerjee et al., 2011). Their primary goal is often to identify the overall effect of a treatment T_i on an outcome Y_i, marginalizing over units ($i \in \mathcal{I}$) in the sample population, and thereby, calculating the Average Treatment Effect (ATE). By collecting tabular characteristics, $\mathbf{X}_i$, such as age and ethnicity, investigators can unpack results by sub-populations, estimating the Conditional Average Treatment Effects (CATEs) (Künzel et al., 2019; Balgi et al., 2022). However, features in $\mathbf{X}_i$ are often measured only at baseline, just before the experiment is initiated. Thus, $\mathbf{X}_i$ rarely contain information on an experimental unit's historical characteristics, including its past neighborhood-level features and geographical context, features which may be useful in identifying sub-populations that react differently to the same treatment (Kino et al., 2021).

When experimental units i are associated with a particular location, satellite images, $\mathbf{M}_i$, can fill in this gap, providing important information about the otherwise missing historical, neighborhood, and geographical contexts (Burke et al., 2021; Daoud et al., 2021).

Indeed, in contrast to covariates measured during experiments, satellite imagery is collected passively from space for every geographic context on earth, and thus (except for clouds covering the line sight of a satellite) there is no systematic missingness in the data source. Moreover, these data have been collected for parts of the world since the CORONA intelligence satellites of the 1950s and for the entire world since the start of NASA and the US Geological Survey's joint Landsat mission in the 1970s. The Landsat data are publicly available with a revisiting time of on average 16 days. Therefore, by combining satellite imagery with experimental data, researchers can investigate not only the historical and geographical roots of effect variation, but they can also predict how an intervention will likely impact places outside the scope of the original study where no researcher-collected covariate data are available. That is, predictive distributions over treatment effects can be estimated not only for the experimental context but also investigated in places not originally conceived of during the experiment—where no tabular background covariates were measured. In this way, earth observation data have the potential to increase the applicability of ideas in casual transportability (Pearl and Bareinboim, 2022).

However, despite the potential offered by images for causal inference, as evident by the growing literature (Castro et al., 2020; Chalupka et al., 2016b,a, 2015; Schölkopf et al., 2021; Yi et al., 2020; Ding et al., 2021; Paciorek, 2010; Kaddour et al., 2021b; Louizos et al., 2017; Pawlowski et al., 2020; Lopez-Paz et al., 2017), it remains unclear how researchers should use these $\mathbf{M}_i$'s for CATE analysis. One key challenge is that $\mathbf{M}_i$ is high-dimensional and rarely annotated. A CATE analysis using $\mathbf{M}_i$ but relying on tabular methods, such as linear models or the generalized random forest Athey et al. (2019), would likely lead to poor interpretability or struggle to find heterogeneity within the high-dimensional image tensors.

For these raesons, there is a major research need to form a causal inference framework for CATE analysis in images. Equipped with such a method, applied researchers would not only be enabled to start incorporating satellite images in future RCTs, but also re-analyze data from past experiments that have already been performed, with the goal of deepening understanding using satellite image data, potentially uncovering missed yet significant CATE–informing policymaking.

In this article, we develop machine learning models which characterize image-based effect heterogeneity in the RCT setting. We introduce an interpretable CATE model that employs Bayesian convolutional neural network arms (CNNs) with categorical gates that allow us to directly model

mixtures of image clusters with similar effects. Our models estimate treatment effects for all units, conditional on treatment status, the images $\mathbf{M}_i$, and, if desired, accounting for available $\mathbf{X}_i$ by incorporating them into the cluster prediction or via orthogonalization. By residualizing, our model will identify what additional CATE that stems from $\mathbf{M}_i$, separately from $\mathbf{X}_i$, for enhanced heterogeneity analysis. The models construct image-type clusters that group units probabilistically based on their CATE similarity.

In the following sections, we develop our methods, show some of their properties analytically, and explore others in simulated experiments. We demonstrate the usefulness of our methods by replicating the results of an RCT study in Uganda—the Youth Opportunities Program (YOP) (Blattman et al., 2014). This study, conducted in 2008, was designed to help the poor break unemployment cycles by financially assisting their artisans and business activity. The government solicited young adults to participate in YOP, asking them to form teams and compose a business plan. After screening teams, the government randomly assigned some to receive one-time grants worth about $7,500, often more than members' joint annual income. Most of the applicants were young, rural farmers having low educational attainment ($\sim$ eighth grade), earning less than $1 per day, and working less than twelve hours per week. In many RCTs like YOP, the researchers collected a set of baseline covariates, but none of them explicitly capture historical neighborhood or geographical characteristics. Our replication uses satellite imagery collected independently of the original experiment and demonstrates the usefulness, and limitations, of using $\mathbf{M}_i$ for CATE, as a complement to the tabular version.

While our contribution focuses on the use of satellite images in global poverty research, our methods are designed such that they generalize to other RCT settings where complementary image data are available. In the last few decades, there has been a rapid increase in the availability of imaging technologies. Most notably, these technologies are readily available in biomedical fields in the form of X-ray, positron emission, MRI, and ultrasound modalities. These images data streams are likely useful not only for estimating ATE (Castro et al., 2020; Lopez-Paz et al., 2017; Chalupka et al., 2016b), but also for evaluating effect heterogeneity. More research will be needed to determine the usefulness of our modeling approach for such domains.

2. Background and Related Work

CATEs with Tabular Data Let $Y_i(t)$ denote the potential outcome (Rubin, 2005) of an intervention $t \in \{0, 1\}$ for a unit of study i. For example, $Y_i(1)$ may represent the poverty level in household i following an aid intervention, and $Y_i(0)$ is the level without intervention.

We may define the unit-level treatment effect as $\tau_i = Y_i(1) - Y_i(0)$. When τ_i is greater than 0, the unit's outcome is greater under treatment than otherwise. The quantity τ_i cannot be exactly identified without strong assumptions (Pearl, 2009). Because a unit can only receive a single treatment at a given time, only one of the potential outcomes, $Y_i(1)$ or $Y_i(0)$, is observed, and thus, the counterfactual remains unobserved. Assuming consistency—that is, units comply with their treatment assignment—the observed outcome can be written as, $Y_i = Y_i(T_i) = Y_i(1)T_i + Y_i(0)(1 - T_i)$, where T_i denotes the (random) treatment status of i (Miguel and Robins, 2020). The ATE captures the population effect by averaging over all unit-level effects:

$$\textit{Average Treatment Effect (ATE):} \quad \mathbb{E}[\tau_i] = \mathbb{E}[Y_i(1) - Y_i(0)].$$

The ATE is useful as it marginalizes over the heterogeneity present in a population to form an overall assessment of an experiment. With treatment randomization, ATE can be estimated non-parametrically by the difference between treatment and control outcomes (Rubin, 2005). Despite the importance of aggregate quantities such as the ATE, it is useful to disaggregate average effects using based on contextual information. Such a disaggregation is critical for not only scientific understanding but also for policy learning (e.g., by personalizing treatments (Greenland et al., 2020; Balgi et al., 2022). This disaggregation can be a function of any type of general pre-treatment data variable, $\mathbf{G}_i$:

$$\textit{Conditional Average Treatment Effect (CATE):} \quad \tau(\mathbf{g}) = \mathbb{E}[Y_i(1) - Y_i(0) \mid \mathbf{G}_i = \mathbf{g}],$$

The literature has primarily focused on conditioning sets that contain tabular data, $\mathbf{X}_i$:

$$\textit{Tabular Conditional Average Treatment Effect:} \quad \tau(\mathbf{x}) = \mathbb{E}[Y_i(1) - Y_i(0) \mid \mathbf{X}_i = \mathbf{x}],$$

where $\mathbf{x} \in \mathbb{R}^D$ denotes the vector of D pre-treatment covariates (Athey and Imbens, 2016; Athey et al., 2019; Ding et al., 2016; Imai and Ratkovic, 2013; Shalit et al., 2017; Zhao et al., 2017; Luedtke and van der Laan, 2016; Nie and Wager, 2021). For example, the generalized random forest is one such machine-learning method (Athey et al., 2019), and it has proven useful in a variety of applied settings (Shiba et al., 2021; Daoud and Johansson, 2019). However, these methods are tailored for annotated tabular data and images are high-dimensional, often non-annotated. These non-annotated image features consist of image bands and pixels that may jointly induce effect heterogeneity. Thus, more research is required to improve the ability of investigators to understand CATE in the context of unstructured high-dimensional image data.

Causal Inference with Image Data While most causal-inference studies use tabular data, there is an increasing realization that image data provides a creative yet useful way to conduct causal inference (Castro et al., 2020; Ramachandra, 2019; Daoud and Dubhashi, 2020). To this end, there is a growing methodological literature investigating how images should be integrated to identify and estimate ATEs in the observational setting (Kallus, 2020; Kaddour et al., 2021a; Pawlowski et al., 2020; Jerzak et al., 2023). Yet these approaches tend to mainly treat images as proxies, for inclusion in the adjustment set, thereby securing causal identification; these approaches are not tailored for CATE analysis in images. Hence, little is known about how to use images for CATE.

Like tabular information in $\mathbf{X}_i$, images as a whole or through its segments may be associated with treatment effect heterogeneity. In the observational setting, images $\mathbf{M}_i$ could be part of both the conditioning and effect modification sets. In the RCT setting, $\mathbf{M}_i$ is not a confounder, since the treatment was randomized, but it may provide significant information for CATE (as discussed in §1). In both settings, one target estimand is the Image CATE,

$$\textit{Image CATE:} \quad \tau(\mathbf{m}) = \mathbb{E}[Y_i(1) - Y_i(0) \mid \mathbf{M}_i = \mathbf{m}],$$

where $\mathbf{m} \in \mathbb{R}^{W \times H \times D}$ is an image, obtained before treatment assignment, of width W, height H, and with D data channels. These data channels often contain reflectance information from various electromagnetic bands. In some applications involving earth observation data, the i subscript may correspond to a spatially defined neighborhood (or to a person living in such a place). In applications involving medical imaging, the i subscript may correspond to patients or tissue regions. In both, the Image CATE analysis will have a connection with Multiple Instance Learning (MIL) in the sense

that a single effect response (e.g., high or low) may be associated with multiple image segments (for a review of MIL, see Foulds and Frank (2010)).

Although image and tabular CATEs are both special cases of the general CATE, there are conceptual differences between them. First, images are unstructured, high-dimensional objects, and satellite images of each neighborhood are unique in RCT applications. This makes it difficult, if not impossible, to perform non-parametric inference. Second, since images are unstructured, it is unclear how to interpret the act of conditioning on an image. We need a conceptual language and modeling strategy for characterizing proximity between images in the space of conditional effects. Therefore, the remainder of our article will contribute to establishing this conceptual language for RCTs, leaving image CATE in the observational case for future study.

3. Modeling Causal Effect Heterogeneity in Images

We first introduce a baseline method of interpretable image CATE analysis which we will later contrast against the probabilistic Image-Type Effect Cluster Model, which will form the focus of our later application.

3.1. Comparative Baseline: Prediction Cluster CATE

In experimental settings, CATEs may be estimated readily by function approximation. Perhaps the simplest approach is to use a parameterized function, $f_{\hat{Y}_t}(\mathbf{m})$, to predict potential outcomes, $Y_i(t)$, of each intervention—a so-called *T-learner* (Künzel et al., 2019). The CATE may then be estimated as $\hat{\tau}(\mathbf{m}) = f_{\hat{Y}_1}(\mathbf{m}) - f_{\hat{Y}_0}(\mathbf{m})$. Shalit et al. (2017) found that learning a shared representation used to predict both treatment outcomes improved prediction quality and named this approach, *TARNet*.

Given a CATE model formed by such function approximation, we can aim to increase the interpretability of the model output by partitioning units by their predicted causal effect, creating a clustering of inputs *post hoc*. This post-hoc clustering will serve as a comparative baseline for the subsequent probabilitistic models. Let $f_{\text{Cluster}}(\hat{\tau}(\mathbf{m})) \in \{1, 2, ..., K\}$ denote a cluster labeling function determined by the output of $\hat{\tau}$, which partitions the space of effect sizes and, as a result, the space of images. C may be constructed by quantile binning of $\hat{\tau}$, or, as in our experiments, by k-means clustering. The CATE with respect to the post-hoc clustering labels is

$$\textit{Prediction Cluster CATE:} \quad \tau(c) = \mathbb{E}[Y_i(1) - Y_i(0) \mid f_{\text{Cluster}}(\hat{\tau}(\mathbf{M}_i)) = c],$$

In our experiments, we use the post-hoc clustering of the $\hat{\tau}(\mathbf{M}_i)$'s from TARNet as a point of comparison, as the approach is a standard one for high-dimensional causal estimation.

A drawback of a post-hoc approach is that it compounds approximations to arrive at a discrete representation of treatment effect heterogeneity. If two similar images yield very different predictions due to misestimation in either the model for $Y_i(0)$ or for $Y_i(1)$, they are likely to be placed in different clusters. We would prefer to cluster images in a way that smoothly best approximates the Image CATE in a single model. Moreover, it is difficult to quantify how the image affects the post-hoc clustering because it may be computationally prohibitive to propagate gradients through both the outcome and subsequent clustering model (a topic explored in §3.3).

To address these limitations, we therefore develop a probabilistic image CATE model that directly targets the prediction of the heterogeneity itself in a low-dimensional summary, with the goal of increasing the interpretability of the image heterogeneity dynamics.

3.2. Interpretable Models for Effect Clustering Based on Image Type

We now introduce a series of modeling strategies for directly targeting CATE clusters in images; to the best of our knowledge, this is the first work to use images to estimate CATE with a particular focus on interpretability. We will aim to fulfill the following criteria: (1) *model potential outcomes and treatment effects flexibly (e.g., allowing for non-linearities)* (2) *identify interpretable image clusters with similar in-cluster effect sizes, and different cross-cluster effects*, and (3) *allow for the modeling of uncertainty regarding the image clusters and treatment effects*.

Our target quantity of interest will be

$$\textit{Image-Type CATE:} \quad \tau(z) = \mathbb{E}[Y_i(1) - Y_i(0) \mid Z_i = z],$$

where $Z_i \in \{1, 2, ..., K\}$ denotes the effect cluster of image $\mathbf{M}_i$, where there are K total clusters. We will search for assignments Z_i of clusters to images that best explain treatment effect variation.[1]

Because the treatment effects are decomposed by image type, there are only K quantities needed to effectively summarize the heterogeneity attributable to images. Since human working memory can track around 5 distinct chunks at a time (Cowan, 2010), this low-dimensional probabilistic summary of the complex heterogeneity process stemming from images can in principle be communicated to human stakeholders—thereby facilitating future treatment-targeting decisions. With this substantive motivation in mind, we now discuss how we meet the modeling criteria for capturing treatment effect heterogeneity with images.

Probabilistic Estimation of the Baseline Outcome We satisfy the first criterion by allowing the baseline potential outcome, $Y_i(0)$, to be estimated flexibly as a function of the image. In particular, we let the mean of the baseline potential outcome, conditional on the image, be parameterized by a Bayesian convolutional neural network,

$$\mathbb{E}[Y_i(0) \mid \mathbf{M}_i = \mathbf{m}] \;=\; \{\mu_{Y_i(0)} \mid \mathbf{M}_i = \mathbf{m}\} \;\sim\; \text{Bayesian CNN}(\mathbf{m}) \;, \tag{1}$$

where convolutional and dense parameters are not deterministic but instead defined according to a distribution. This model for the baseline mean is listed as (2) in Schema 1 and depicted as the arrow between $\mathbf{M}_i$ and $\mu_{Y_i(0)}$ in the probabilistic model depiction in Figure 1.

Effect Mixture Based on Image Type Having defined the baseline, we next turn our attention to the modeling component that targets the Image-Type CATE estimand. In this component, we first compute image type probabilities $\mathbf{P}_i$ using another Bayesian CNN. Given $P_{iz} = p_z$, the image takes on cluster type z with probability p_z. Intuitively, while the first CNN looks for image patterns indicative of $\mu_{Y_i(0)}$, the second looks for patterns associated with the type of treatment response. We develop two model variants for the type response characterization—one more interpretable and the other more flexible.

Variant 1. **Image-Type Effect Cluster Model** Here, conditional on the *image type* $Z = z$, mean treatment effects are drawn from a Normal with a mean $\mu_{\tau,z}$ and variance $\sigma^2_{\tau,z}$ indexed to that image type. We do not assume that there is a single treatment effect per image type, but instead that there is a specific *distribution* over treatment effects by image type. The cluster effect means and variances offer a complete summarization of this distribution. This model emphasizes interpretability: given the image type, treatment effects are characterized by a single, unique distribution.

1. A related quantity is targeted in the mixture-of-experts approach for CATEs in the conjoint setting using linear models with interactions (Goplerud et al., 2022).

Variant 2. **Image-Type Differential Effect Model** The first effect mixture model could be useful to investigators because conditional effects can be succinctly summarized with a small number of parameters, but in some circumstances, investigators may want a more flexible model for the heterogeneity structure. In that case, the distribution of treatment effects given the image type z is parameterized by a Bayesian CNN arm indexed to z:

$$\{\mu_{\tau_i} \mid \mathbf{M}_i = \mathbf{m},\ Z_i = z\} \sim \text{Bayesian CNN}_z(\mathbf{m})$$

Here, the image type acts as a stochastic gate that determines which image patterns will be used in predicting the mean treatment effect value given the image, $\mathbf{M}_i$.

Overall, the probabilistic generative modeling framework for image-based CATE is summarized visually in Figure 1 and as follows:

Schema 1: The Image Effect Cluster Model. See Figure 1 for visualization.

(1a) *Generating the Image Type*

$$\{\mathbf{P}_i \mid \mathbf{M}_i = \mathbf{m}\} \sim \text{Bayesian CNN}(\mathbf{m})$$
$$\{Z_i \mid \mathbf{P}_i = \mathbf{p}\} \sim \text{Categorical}(\mathbf{p})$$

$$\downarrow$$

(1b) *Generating the Treatment Effect Distributions*

i. Image-Type Effect Cluster Model: $\{\mu_{\tau_i} \mid Z_i = z\} \sim \mathcal{N}(\mu_{\tau,z},\ \sigma^2_{\tau,z})$

ii. Image-Type Differential Effect Model: $\{\mu_{\tau_i} \mid \mathbf{M}_i = \mathbf{m},\ Z_i = z\} \sim \text{Bayesian CNN}_z(\mathbf{m})$

$$\downarrow$$

(2) *Generating the Baseline Outcome, $Y_i(0)$*

$$\{\mu_{Y_i(0)} \mid \mathbf{M}_i = \mathbf{m}\} \sim \text{Bayesian CNN}(\mathbf{m})$$
$$\{Y_i(0) \mid \mu_i(0),\ \sigma^2_{0,z}\} \sim \mathcal{N}(\mu_{Y_i(0)},\ \sigma^2_{0,z})$$

$\rightarrow$ (3) *Generating the Outcome under Treatment, $Y_i(1)$*

$$\{Y_i(1) \mid \mu_{Y_i(0)}, \mu_{\tau_i},\ \sigma^2_{1,z}\} \sim \mathcal{N}(\mu_{Y_i(0)} + \mu_{\tau_i},\ \sigma^2_{1,z})$$

There are several advantages to these modeling strategies. There is improved interpretability from summarizing image-derived effect heterogeneity in K discrete clusters. Moreover, under the Image-Type Effect Cluster Model, we can efficiently summarize the cluster effects (see §A.1.2.1):

$$\tau(z) = \mathbb{E}[Y_i(1) - Y_i(0) \mid Z_i = z] = \mu_{\tau,z}, \qquad \text{Var}(Y_i(1) - Y_i(0) \mid Z_i = z) \ = \ \sigma^2_{0,z} + \sigma^2_{1,z} + \sigma^2_{\tau,z}.$$

Next, as the strategies are probabilistic, so we can explore uncertainty not only in the image cluster effects by also in the cluster assignment probabilities.[2] In addition, the cluster decomposition may facilitate scientific inquiry: an image type serves as a generalization tool for reasoning across images, facilitating theorizing about the causal mechanisms at play. Finally, we can readily compute the gradients of the expected cluster probabilities with respect to the image in order to identify *how* the image affects the typology, a matter explored in §3.3.

For both probabilistic model variants, estimation is performed via variational Bayesian methods[3] where we learn the joint distribution of the model parameters $\mathbf{\Theta}$ and the image clustering, $\mathbf{Z}$,

2. The approach outlined here can be readily adapted to outcomes having non-Normally distributed outcomes by selecting different observed data likelihoods.
3. We note in passing that an additional benefit of the approach proposed here, as opposed to post-hoc clustering, is that uncertainty of the variational clustering model can be further quantified under model misspecification using M-estimation theory (Westling and McCormick, 2019).

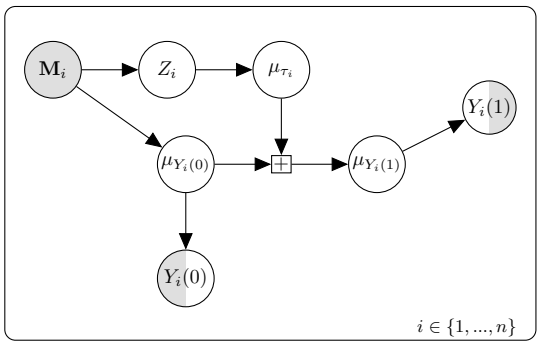

Figure 1: A stylized schematic depiction of the probabilistic treatment heterogeneity model for images. Gray circles denote observed random variables; white circles denote latent variables. Mixed gray/white nodes denote partially observed nodes (i.e., nodes observed for some, but not all, units). Square nodes denote deterministic transformations. Z_i denotes the image type generating a distribution over treatment effects. Arrows denote statistical (not causal) dependencies.

given the observed dataset, $\mathbf{D} = \{Y_i(T_i),\ \mathbf{M}_i,\ T_i\}_{i=1}^n$:

$$\textit{Target Posterior:} \quad p(\mathbf{Z},\ \mathbf{\Theta} \mid \mathbf{D}), \tag{2}$$

where $\mathbf{Z} = \{Z_i\}_{i=1}^n$. For details of how we model the uncertainties, see A.1.2. To estimate the posterior in (2), we maximize the Evidence Lower Bound (ELBO) (Ranganath et al., 2014),

$$\underset{q(\mathbf{Z},\ \mathbf{\Theta})}{\text{maximize}} \quad \mathbb{E}_{q(\mathbf{Z},\ \mathbf{\Theta})}\left[\log\left(p(\mathbf{Y}(\mathbf{T}) \mid \mathbf{Z},\ \mathbf{\Theta},\ \mathbf{M},\ \mathbf{T})\right)\right] \ - \ D_{\text{KL}}\left(q(\mathbf{Z},\ \mathbf{\Theta}) \mid\mid p(\mathbf{Z},\ \mathbf{\Theta})\right).$$

We solve the problem approximately using stochastic gradient descent with gradients passing through discrete sampling nodes using re-parameterization techniques (Parmas and Sugiyama, 2021). The choice of priors affects finite sample performance; when possible, we specify priors using observable marginal information (e.g., prior means for cluster effects are set to $\widehat{\mathbb{E}}[Y_i(1) - Y_i(0)]$).

3.3. Determination of Salience Regions in the Posterior Mean Probabilities

One benefit of the Bayesian image-type heterogeneity model is that we can examine the model in order to assess *how* image information translates into the predicted effect cluster. In particular, we can take the derivative of the posterior mean cluster probabilities with respect to pixel (w, h):

$$s_{whk}^{\text{Direction}} = \sum_{c=1}^{C} \frac{\partial\, \mathbb{E}\left[\Pr(Z_i = k \mid \mathbf{M}_i = \mathbf{m})\right]}{\partial\, m_{whc}},$$

where $c \in \{1, ..., C\}$ denotes the channel (band) dimension. The quantity, s_{whk} is a scalar summary of how changing pixel (w, h) would induce changes in the predicted effect cluster probability. Because the modeling strategy is probabilistic, the salience must average over the randomness in the predicted cluster probabilities, hence the expectation on the inside of the derivative. This expectation is approximated via Monte Carlo. Positive values of this quantity would indicate that increasing the pixel intensity at (w, h) would increase the probability of cluster k; with the same logic, negative values indicate that increasing pixel intensity at (w, h) would decrease the probability of cluster k.

Because directional salience may not be interpretable when increasing the pixel intensity is not itself interpretable, we can also consider an approach based on magnitudes, bypassing the potentially

difficult interpretation of pixel intensities in different bands:

$$s_{whk}^{\text{Magnitude}} = \sqrt{\sum_{c=1}^{C} \left(\frac{\partial \, \mathbb{E}\left[\Pr(Z_i = k \mid \mathbf{M}_i = \mathbf{m})\right]}{\partial \, m_{whc}} \right)^2}.$$

Large values of $s_{whk}^{\text{Magnitude}}$ reveal locations in the image that, if changed, would lead to large changes in cluster k probabilities. This measure is agnostic about whether those changes would be associated with increases or decreases in those expected probabilities. This salience information is difficult to compute for post-hoc clustering methods, as the clustering of the $\hat{\tau}_i$'s needs to solve a second optimization problem (as in k-means) through which gradients with respect to the initial outcome model(s) may not be efficiently traced. Moreover, in contrast to the post-hoc approach of §3.1, the salience measure here incorporates uncertainty in the cluster prediction itself (since the salience averages over randomness in the cluster probabilities).

3.4. Policy Action Using the Image-based Heterogeneity Model

A major motivation for considering satellite-image-based CATEs is that we can readily generate predictive distributions over treatment effects for contexts outside the experimental setting and where no tabular covariates were measured by researchers—a possibility that may meaningfully expand the reach of causal analyses. In this context, for a new out-of-sample point not from the original dataset, $i^{\text{Out}} \in \mathcal{I}(\text{Out})$, we form a predictive distribution over image treatment effects using

$$p(\tau_{i^{\text{Out}}} \mid \mathbf{M}_{i^{\text{Out}}}, \, \mathbf{D}) = \sum_{z=1}^{K} \int p(\tau_i \mid Z_{i^{\text{Out}}} = z; \, \boldsymbol{\Theta} = \boldsymbol{\theta}) \cdot p(Z_{i^{\text{Out}}} = z \mid \mathbf{M}_{i^{\text{Out}}}; \, \boldsymbol{\Theta} = \boldsymbol{\theta}) \cdot p(\boldsymbol{\Theta} = \boldsymbol{\theta} \mid \mathbf{D}) \, \mathrm{d}\boldsymbol{\theta}$$

We can use the predictive distribution over treatment effects to improve treatment targeting for out-of-sample individuals. With a fixed treatment budget of size n_1^O for the new dataset of size $n^O = |\mathcal{I}(\text{Out})|$, this policy can be written as $\boldsymbol{\Pi}(\{\mathbf{M}_{i^{\text{Out}}}\}_{\mathcal{I}(\text{Out})}) \rightarrow \{0,1\}^{n^O}$. There are many approaches to this problem, and we refer readers to the relevant literature (Hitsch and Misra, 2018).[4]

3.5. Multi-modal Learning with Image and Tabular Heterogeneity

Tabular information can be readily incorporated into this modeling pipeline. For example, tabular covariates can be appended to the input of the dense part of the cluster type and baseline outcome models. The resulting treatment effect heterogeneity clusters are therefore conditional on both individual-level and also image-context-level information. This image and tabular data integration can be useful when investigators are focused on understanding the holistic heterogeneity dynamic in an experimental context, integrating both individual- and neighborhood-level information. Such a combined approach, an example of multi-modal learning (Ullah et al., 2022), can also be potentially useful in the medical domain, where image and high-dimensional medical records text can form the basis for improved patient response modeling.

4. We also note that, for the predicted treatment effects to be reliable, there should be minimal distribution shift between in- and out-of-sample points. In practice, this means that the experimental areas should ideally be selected randomly from within the geographic unit of interest, so that extrapolations into data-sparse regions are minimal.

3.6. Distinguishing Image from Tabular Heterogeneity

When tabular covariates were not measured or when we aim to generate predictions for observations having no measured tabular covariates, it may be useful to perform image-type effect clustering directly using images alone. When other tabular covariates are measured for the experimental sample, researchers may seek to understand the heterogeneity stemming from image information that is unique when compared with tabular covariates.

For example, information about economic class is embedded in earth observation images, since the urban poor are often concentrated in dense city centers while the affluent often live in green-space-rich areas outside cities (Venter et al., 2020). We may wonder about the remaining heterogeneity after accounting for tabular variables such as income. In this context, we can perform the image clustering on the orthogonalized outcomes: $Y_i(t)^{\perp} = Y_i(t) - \widehat{\mathbb{E}}[Y_i(t)|\mathbf{X}_i]$. The resulting clusters can then be interpreted as image effect types after accounting for the additive heterogeneity from measured tabular data. This image-specific decomposition can be helpful when geographic or neighborhood information is the focus of study.

4. Treatment Effect Cluster Recovery in Simulation

We now explore the dynamics of the proposed methods in simulation, where true treatment effects are known. We generate image-based treatment effect heterogeneity using,

$$H_i = \text{GN}(\max(f_l(\mathbf{M}_i))), \qquad H_i^+ = \underbrace{|\min\{H_i\}_{i=1}^n|}_{\text{Ensures } \tau_i > 0} + \underbrace{\text{sign}(H_i) \cdot |H_i|^{1/\gamma}}_{\text{Generates bimodality as } \gamma \to \infty} \tag{3}$$

where $f_l(\cdot)$ denotes the application of a $l \times l$ filter to the image, $\max(\cdot)$ denotes the global maximum operation across the image, and $\text{GN}(\cdot)$ denotes a global normalization function scaling the H_i values to have mean 0 and variance 1 across the image pool. The specific transformation generating H_i^+ is selected to ensure all the treatment effects are in the same direction (i.e., all positive) and to generate heterogeneity in the effect distribution, with greater bimodality in the treatment response as $\gamma \to \infty$. We let $\gamma = 2$. We define the treatment and outcome:

$$T_i \sim \text{Binomial}(0.5), \quad Y_i = T_i H_i^+ + \epsilon_i,$$

with $\epsilon_i \sim \mathcal{N}(0,\ \nu \cdot \text{Var}(H_i^+))$. The value of ν controls the extent to which the image is predictive of the outcomes, where smaller values indicate a stronger image heterogeneity signal. To explore the effect of the signal-to-noise ratio, we vary $\nu \in \{0.01, 0.1, 1\}$.

The filter used in the convolution function of Equation 3 is visualized in Figure A.1, along with high and low responders from the set of images used in the simulation that we take from Landsat mosaics of sub-Saharan Africa. We have some degree of model misspecification here, as the way the data are generated is distinct from the estimation models; given this, we will examine the degree to which the various models will recover key properties of the image-based causal system.

Cluster Recovery Measure We compare the estimated effect clustering with an oracle baseline from the true, in practice unknown, τ_i's. That is, we first compute the oracle k-means clustering:

$$\tau(z)_{\text{Oracle}} = z^{\text{th}} \text{ center from the oracle } k\text{-means applied to the true (in practice unknown) } \tau_i\text{'s}$$

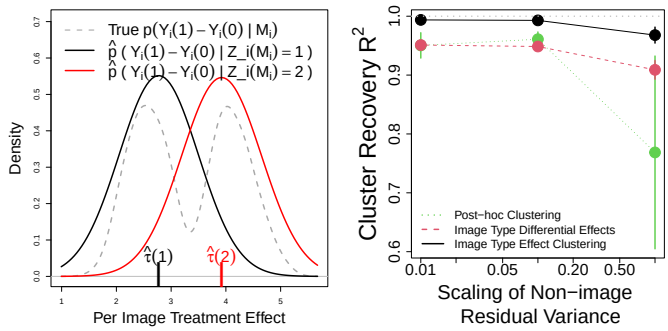

Figure 2: *Left:* Capturing the treatment effect heterogeneity with our Image-Type Effect Model. *Right:* Comparing models as we vary the signal-to-noise ratio.

The clustering quality measure then compares the oracle with estimated cluster means:

$$\text{Cluster Recovery } R^2 \;=\; 1 - \frac{\sum_{z=1}^{K} \min_{z'} (\widehat{\tau}(z) - \tau(z')_{\text{Oracle}})^2}{\sum_{z''=1}^{K} (\tau(z'')_{\text{Oracle}} - \overline{\tau}_{\text{Oracle}})^2},$$

where $\overline{\tau}_{\text{Oracle}}$ denotes the mean across the oracle cluster centers. This measure is equivalent to the R^2 in predicting the oracle cluster means from the estimated ones, where the ordering of the clusters has been arranged so that each oracle center is compared to its nearest estimated counterpart.

Simulation Results We see in the left panel of Figure 2 one representative posterior distribution over $Y_i(1) - Y_i(0)$ given estimated cluster information. We find that the estimated clusters capture the bimodality present in the true distribution of τ_i's. The right panel shows how the cluster recovery measure for the Image-Type Differential Effects Model and TARNet post-hoc clustering are similar in the low residual variance setting. In the high residual variance setting, the TARNet clustering struggles somewhat in recovering the oracle cluster centers. In contrast, the parsimonious Image-Type Effect Clustering Model performs best at recovering the clustering of the treatment effects across the noise range. This is encouraging for the application of the Image-Type Effect Cluster Model in practice, as presumably, the signal-to-noise ratio for real tasks involving earth observation images is relatively high.

5. Application to an Anti-Poverty Experiment in Uganda

Data In our application, we explore the effects of the anti-poverty experiment performed in Uganda and described in §1. The treatment variable is the random assignment of small teams to receive grants for business ventures. The outcome variable is an aggregate summary of skilled labor (see A.1.4.1) measured at the end of the experiment (two years after treatment assignment). De-identified outcome and treatment data were given voluntarily by subjects and are available under CC0 1.0 license. Longitude/latitude information about respondents' villages is found using OpenStreetMap.

Pre-treatment image data are taken from Landsat. We use the Orthorectified ETM+ pan-sharpened data product, processed to contain minimal cloud cover. Reflectance is measured in the green, near-infrared, and short-wave infrared bands. These bands are useful in capturing information about peak vegetation, water content, and thermal dynamics, in addition to structural land features.

Empirical Results Due to space constraints, we focus on results from the Image-Type Effect Cluster Model (details in §A.1.6). We set the cluster number to 2 after finding that cluster probabilities become highly correlated with additional clusters. The top three rows in the left panel of Figure 3 show results for the highest images having the highest posterior mean probabilities for cluster

1. For each image, this figure visualizes the salience measures defined in §3.3. The bottom portion shows results for the highest posterior mean cluster 2 images. The effect for cluster 1 is substantially different than for cluster 2. Visually, we see that smaller effects exist for places with harsher terrain and less developed transportation networks, hampering economic growth. These low responders are found in the harsh mountainous northern part of Uganda. This is logical, as skilled labor tends to thrive in areas that are connected via transportation networks (Ashraf and Galor, 2011).

We show in the right panel of Figure 3 how the results of the experiment may be generalized to the entire country of Uganda, assuming no systematic bias in places chosen conditional on the image information. In particular, we show the posterior predictive mean cluster 2 probability for the entire country. This kind of analysis can provide policymakers with potentially useful information for how to improve the targeting of treatments in the future across larger geographic contexts.

The Appendix contains supplementary analyses. For example, we show in Figure A.4 the correlation between the estimated Image CATEs and Tabular CATEs using various conditioning sets, as well as, in Table A.1, between the cluster probabilities and other individual-level covariates. We show in Figure A.3 the images having the greatest uncertainty in the cluster probabilities (estimated by the posterior standard deviation). In §A.1.5, we orthogonalize the potential outcomes using tabular information, and the results remain similar: the correlation between raw and non-orthogonalized cluster probabilities is 0.85. Because our results remain similar after orthogonalization, the satellite images seem to supply independent and thought-provoking information about effect heterogeneity.

6. Discussion and Conclusion

Scientists and policymakers use RCTs to estimate population-wide effects (ATE) and sub-population effects (CATE), using tabular data collected at baseline, often near-time to when the RCT is launched. However, these near-time variables tend to miss important historical or neighborhood-level features. While such features are often unavailable or expensive to collect, satellite images are a data stream that captures such characteristics in an unstructured form. As no CATE method exists explicitly for image analysis, this paper presents principles and modeling strategies for analyzing image-based CATE using probabilistic image-type models. After deriving some model properties, we perform approximate inference using variational methods. Dynamics are explored via simulation; an anti-poverty field experiment from Uganda is analyzed, where we seem to find interesting heterogeneity.

Our approach has limitations, which serve to motivate future research. First, our models estimate heterogeneity clusters at the image level, but not explicitly for smaller segments of an image. Having such within-image heterogeneity segmentation would further improve understanding of what in the image is generating heterogeneity. Second, our methods estimate heterogeneity with respect to a fixed baseline (i.e., the control intervention). While the choice of baseline is clear in most settings, in unclear cases, investigators may need to explore different baselines and compare results. Third, our model is tailored for RCTs (i.e., assuming unconfoundedness); more research is required to adapt it for observational settings. Using experimental data, effect estimates are confounding-free by design; heterogeneity can be studied independently of identification. Observational data are more plentiful but require adjustment (Rosenbaum et al., 2010). We have also viewed images from solely the perspective of surrogate effect modification (in the language of Miguel and Robins (2020)); the use of images as causal modifiers or mediators is left for future study.

Finally, our focus on *CATE for images* opens exciting possibilities. It not only encourages others to start incorporating satellite images in their planned experiments but also reanalyze past experi-

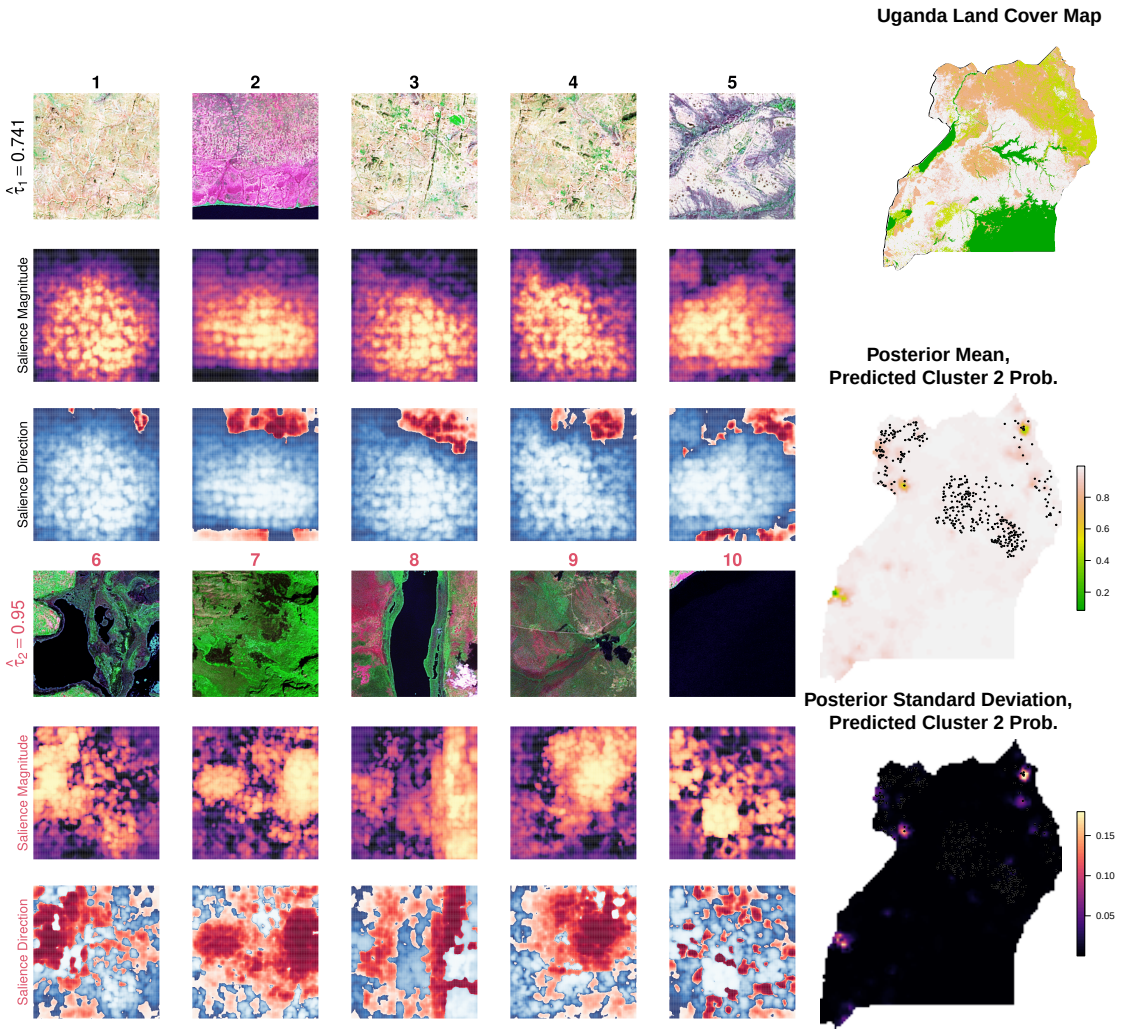

Figure 3: *Left, top 3 rows:* High probability cluster 1 images. *Left, bottom 3 rows:* High probability cluster 2 images. "Salience Magnitude" and "Direction" are defined in §3.3. *Right, top:* A land use map of Uganda. *Right, center:* Posterior predictive mean cluster 2 probabilities for the entire country. Black circles represent observed sample points. *Right, bottom*. Posterior standard deviation of the cluster 2 probabilities.

ments, potentially unraveling previously undetected yet significant sources of effect heterogeneity. As demonstrated in our analysis of the Ugandan anti-poverty experiment, our method identifies heterogeneity not initially detected by incorporating informative satellite data. Thus, our image-based methods have the potential to contribute to policy by complementing traditional RCT heterogeneity analysis based on tabular $\mathbf{X}_i$—and to analyses in other fields such as agriculture, disaster relief, climate science, and medicine where image data are also prevalent. □

References

Quamrul Ashraf and Oded Galor. Cultural diversity, geographical isolation, and the origin of the wealth of nations. Technical report, National Bureau of Economic Research, 2011.

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

Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019.

Sourabh Balgi, Jose M. Pena, and Adel Daoud. Personalized Public Policy Analysis in Social Sciences using Causal-Graphical Normalizing Flows. *Association for the Advancement of Artificial Intelligence: AI for Social Impact track*, February 2022. URL http://arxiv.org/abs/2202.03281. arXiv: 2202.03281.

Abhijit Banerjee, Abhijit V Banerjee, and Esther Duflo. *Poor economics: A radical rethinking of the way to fight global poverty*. Public Affairs, 2011.

Christopher Blattman, Nathan Fiala, and Sebastian Martinez. Generating skilled self-employment in developing countries: Experimental evidence from uganda. *The Quarterly Journal of Economics*, 129(2):697–752, 2014.

Marshall Burke, Anne Driscoll, David B. Lobell, and Stefano Ermon. Using satellite imagery to understand and promote sustainable development. *Science*, 371(6535):eabe8628, March 2021. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.abe8628. URL https://www.sciencemag.org/lookup/doi/10.1126/science.abe8628.

Daniel C. Castro, Ian Walker, Ben Glocker, Ian Walker, and Ben Glocker. Causality matters in medical imaging. *Nature Communications*, 11(1):3673, July 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-17478-w. URL https://www.nature.com/articles/s41467-020-17478-w.

Krzysztof Chalupka, Pietro Perona, and Frederick Eberhardt. Visual Causal Feature Learning. Technical Report arXiv:1412.2309, arXiv, June 2015. URL http://arxiv.org/abs/1412.2309. arXiv:1412.2309 [cs, stat] type: article.

Krzysztof Chalupka, Tobias Bischoff, Pietro Perona, and Frederick Eberhardt. Unsupervised Discovery of El Nino Using Causal Feature Learning on Microlevel Climate Data. Technical Report arXiv:1605.09370, arXiv, May 2016a. URL http://arxiv.org/abs/1605.09370.

Krzysztof Chalupka, Frederick Eberhardt, and Pietro Perona. Multi-Level Cause-Effect Systems. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pages 361–369. PMLR, May 2016b. URL https://proceedings.mlr.press/v51/chalupka16.html. ISSN: 1938-7228.

Nelson Cowan. The magical mystery four: How is working memory capacity limited, and why? *Current directions in psychological science*, 19(1):51–57, 2010.

Adel Daoud and Devdatt Dubhashi. Statistical modeling: the three cultures. *arXiv:2012.04570 [cs, stat]*, December 2020. URL http://arxiv.org/abs/2012.04570. arXiv: 2012.04570.

Adel Daoud and Fredrik Johansson. Estimating treatment heterogeneity of international monetary fund programs on child poverty with generalized random forest. 2019.

Adel Daoud, Felipe Jordan, Makkunda Sharma, Fredrik Johansson, Devdatt Dubhashi, Sourabh Paul, and Subhashis Banerjee. Using satellites and artificial intelligence to measure health and material-living standards in India. Technical Report arXiv:2202.00109, arXiv, December 2021. URL http://arxiv.org/abs/2202.00109. arXiv:2202.00109 [cs, econ, q-fin] type: article.

Mingyu Ding, Zhenfang Chen, Tao Du, Ping Luo, Josh Tenenbaum, and Chuang Gan. Dynamic Visual Reasoning by Learning Differentiable Physics Models from Video and Language. In *Advances in Neural Information Processing Systems*, volume 34, pages 887–899. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/07845cd9aefa6cde3f8926d25138a3a2-Abstract.html.

Peng Ding, Avi Feller, and Luke Miratrix. Randomization inference for treatment effect variation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(3):655–671, 2016.

James Foulds and Eibe Frank. A review of multi-instance learning assumptions. *The knowledge engineering review*, 25(1):1–25, 2010.

Max Goplerud, Kosuke Imai, and Nicole E Pashley. Estimating heterogeneous causal effects of high-dimensional treatments: Application to conjoint analysis. *arXiv preprint arXiv:2201.01357*, 2022.

Sander Greenland, Michael P Fay, Erica H Brittain, Joanna H Shih, Dean A Follmann, Erin E Gabriel, and James M Robins. On causal inferences for personalized medicine: How hidden causal assumptions led to erroneous causal claims about the d-value. *The American Statistician*, 74(3):243–248, 2020.

Günter J Hitsch and Sanjog Misra. Heterogeneous treatment effects and optimal targeting policy evaluation. *Available at SSRN 3111957*, 2018.

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

Connor T Jerzak, Fredrik Johansson, and Adel Daoud. Integrating earth observation data into causal inference: Challenges and opportunities. *arXiv preprint arXiv:2301.12985*, 2023.

Jean Kaddour, Yuchen Zhu, Qi Liu, Matt J Kusner, and Ricardo Silva. Causal effect inference for structured treatments. *Advances in Neural Information Processing Systems*, 34, 2021a.

Jean Kaddour, Yuchen Zhu, Qi Liu, Matt J. Kusner, and Ricardo Silva. Causal Effect Inference for Structured Treatments. *Advances in Neural Information Processing Systems*, 34, 2021b.

Nathan Kallus. Deepmatch: Balancing deep covariate representations for causal inference using adversarial training. In *International Conference on Machine Learning*, pages 5067–5077. PMLR, 2020.

Shiho Kino, Yu-Tien Hsu, Koichiro Shiba, Yung-Shin Chien, Carol Mita, Ichiro Kawachi, and Adel Daoud. A scoping review on the use of machine learning in research on social determinants of health: Trends and research prospects. *SSM - Population Health*, 15:100836, September 2021. ISSN 2352-8273. doi: 10.1016/j.ssmph.2021.100836. URL https://www.sciencedirect.com/science/article/pii/S2352827321001117.

Ranganath Krishnan, Mahesh Subedar, and Omesh Tickoo. Specifying weight priors in bayesian deep neural networks with empirical bayes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4477–4484, 2020.

Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165, 2019.

David Lopez-Paz, Robert Nishihara, Soumith Chintala, Bernhard Scholkopf, and Leon Bottou. Discovering Causal Signals in Images. pages 6979–6987, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Lopez-Paz_Discovering_Causal_Signals_CVPR_2017_paper.html.

Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal Effect Inference with Deep Latent-Variable Models. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/hash/94b5bde6de888ddf9cde6748ad2523d1-Abstract.html.

Alexander R Luedtke and Mark J van der Laan. Super-learning of an optimal dynamic treatment rule. *The international journal of biostatistics*, 12(1):305–332, 2016.

Hernán Miguel and Jamie Robins. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC, 2020.

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

Christopher J Paciorek. The importance of scale for spatial-confounding bias and precision of spatial regression estimators. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1):107, 2010.

Paavo Parmas and Masashi Sugiyama. A unified view of likelihood ratio and reparameterization gradients. In *International Conference on Artificial Intelligence and Statistics*, pages 4078–4086. PMLR, 2021.

Nick Pawlowski, Daniel C. Castro, and Ben Glocker. Deep Structural Causal Models for Tractable Counterfactual Inference. *arXiv:2006.06485 [cs, stat]*, October 2020. URL http://arxiv.org/abs/2006.06485. arXiv: 2006.06485.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Judea Pearl and Elias Bareinboim. External validity: From do-calculus to transportability across populations. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pages 451–482. 2022.

Vikas Ramachandra. Causal inference for climate change events from satellite image time series using computer vision and deep learning. *arXiv preprint arXiv:1910.11492*, 2019.

Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In *Artificial intelligence and statistics*, pages 814–822. PMLR, 2014.

Paul R Rosenbaum, PR Rosenbaum, and Briskman. *Design of observational studies*, volume 10. Springer, 2010.

Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Towards Causal Representation Learning. *arXiv:2102.11107 [cs]*, February 2021. URL http://arxiv.org/abs/2102.11107. arXiv: 2102.11107.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pages 3076–3085. PMLR, 2017.

Koichiro Shiba, Adel Daoud, Hiroyuki Hikichi, Aki Yazawa, Jun Aida, Katsunori Kondo, and Ichiro Kawachi. Heterogeneity in cognitive disability after a major disaster: A natural experiment study. *Science advances*, 7(40):eabj2610, 2021.

Ubaid Ullah, Jeong-Sik Lee, Chang-Hyeon An, Hyeonjin Lee, Su-Yeong Park, Rock-Hyun Baek, and Hyun-Chul Choi. A review of multi-modal learning from the text-guided visual processing viewpoint. *Sensors*, 22(18):6816, 2022.

Zander S Venter, Charlie M Shackleton, Francini Van Staden, Odirilwe Selomane, and Vanessa A Masterson. Green apartheid: Urban green infrastructure remains unequally distributed across income and race geographies in south africa. *Landscape and Urban Planning*, 203:103889, 2020.

Yeming Wen, Paul Vicol, Jimmy Ba, Dustin Tran, and Roger Grosse. Flipout: Efficient pseudo-independent weight perturbations on mini-batches. *arXiv preprint arXiv:1803.04386*, 2018.

T Westling and TH McCormick. Beyond prediction: A framework for inference with variational approximations in mixture models. *Journal of Computational and Graphical Statistics*, 28(4): 778–789, 2019.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. CLEVRER: CoLlision Events for Video REpresentation and Reasoning. Technical Report arXiv:1910.01442, arXiv, March 2020. URL http://arxiv.org/abs/1910.01442. arXiv:1910.01442 [cs] type: article.

Qingyuan Zhao, Dylan S Small, and Ashkan Ertefaie. Selective inference for effect modification via the lasso. *arXiv preprint arXiv:1705.08020*, 2017.

Appendix

A.1.1. Open-source Software & Reproducability

We make the modeling strategies introduced in this paper accessible in an open-source software package available at `github.com/cjerzak/causalimages-software`. For an up-to-date tutorial regarding package use, see `github.com/cjerzak/causalimages-software#readme`. Replication data for the experiment analyzed in the application are contained in this `GitHub` repository as well (we include both the experimental data from the original investigators and the geo-referenced satellite images).

A.1.2. Supplementary Information for the Image-Type Probabilistic Models

A.1.2.1. DERIVING THE CONDITIONAL DISTRIBUTION, $\{\tau_i = Y_i(1) - Y_i(0) | Z_i = z\}$

Using the model outlined in the main text, conditioning on τ_i, and exploiting Normality,

$$\{Y_i(1) - Y_i(0) \mid Z_i = z, \mu_{\tau_i}\} \sim \mathcal{N}(\mu_{\tau_i},\ \sigma^2_{0,z} + \sigma^2_{1,z})$$

Integrating out μ_{τ_i}:

$$\begin{aligned}
p(Y_i(1) - Y_i(0) = \tau_i \mid Z_i = z) &= \int_{-\infty}^{\infty} p(Y_i(1) - Y_i(0) = \tau_i \mid Z_i = z,\ \mu_{\tau_i})\, p(\mu_{\tau_i} \mid Z_i = z)\, \mathrm{d}\mu_{\tau_i} \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi(\sigma^2_{0,z} + \sigma^2_{1,z})}} \exp\left\{\frac{-(\tau_i - \mu_{\tau_i})^2}{2(\sigma^2_{0,z} + \sigma^2_{1,z})}\right\} \\
&\qquad \times \frac{1}{\sqrt{2\pi\sigma^2_{\tau,z}}} \exp\left\{\frac{-(\mu_{\tau_i} - \mu_{\tau,z})^2}{2\sigma^2_{\tau,z}}\right\} \mathrm{d}\mu_{\tau_i} \\
&= \frac{1}{\sqrt{2\pi([\sigma^2_{0,z} + \sigma^2_{1,z}] + \sigma^2_{\tau,z})}} \exp\left\{\frac{-(\tau_i - \mu_{\tau,z})^2}{2([\sigma^2_{0,z} + \sigma^2_{1,z}] + \sigma^2_{\tau,z})}\right\}.
\end{aligned}$$

Therefore,

$$\{\tau_i = Y_i(1) - Y_i(0) \mid Z_i = z\} \ \sim\ \mathcal{N}\left(\mu_{\tau,z},\ \sigma^2_{0,z} + \sigma^2_{1,z} + \sigma^2_{\tau,z}\right).$$

A.1.3. Simulation Details

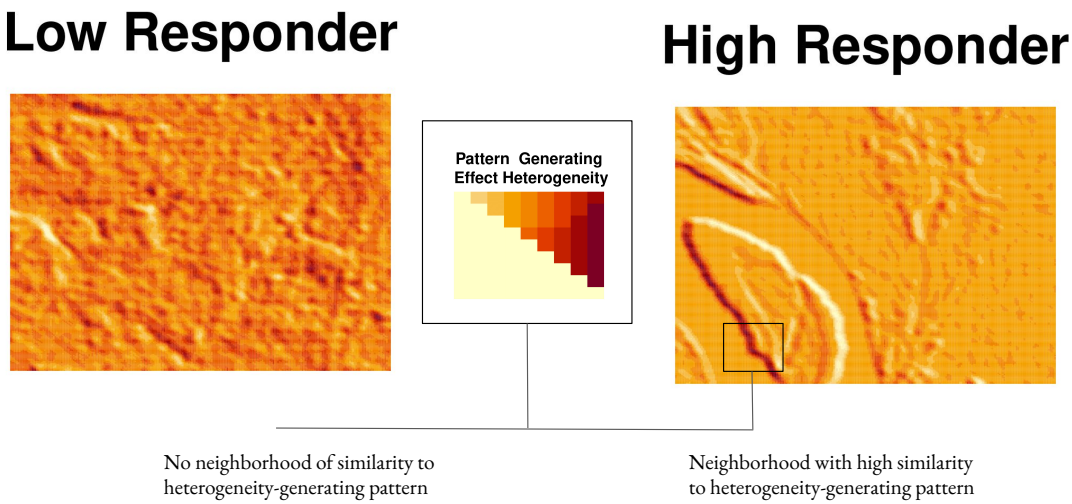

Figure A.1: Simulation design illustration. *Center:* The image pattern used in generating the heterogeneity response in the simulation design of §4. *Left:* An image having no regions of strong similarity to the heterogeneity-generating pattern (leading to a low treatment effect). *Right:* An image with many regions of strong similarity to the heterogeneity-generating pattern (leading to a high treatment effect).

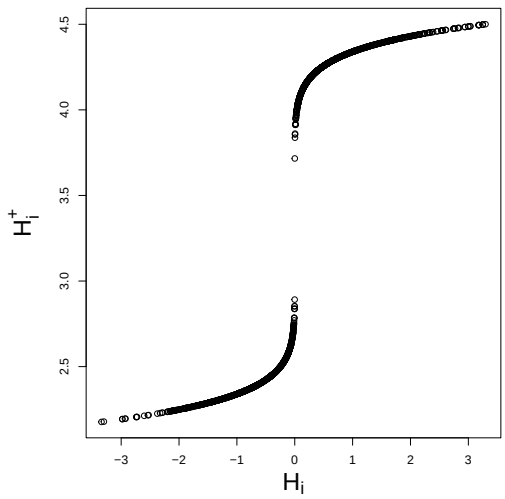

Figure A.2: Illustration of the non-linear transformation used in the simulation for generating H_i^+ from H_i.

A.1.4. Supplementary Analyses for the Application

A.1.4.1. ADDITIONAL DATA DESCRIPTION

We obtain satellite data for the neighborhood around each experimental unit in the following way. First, the place name of residence associated with each unit was geo-referenced using OpenStreetMap and, of this geo-referencing failed, the Google Geocoding API. When this second geo-referencing attempt failed, we use the geometric center for the layer associated with the geographic unit as our focal point for the given unit. Satellite information was then obtained for a cube around focal points with side lengths of 5000 meters. For the skilled work outcome, we take the scaled sum of the log hours worked by experimental units in the last 7 days in skilled or highly skilled trades.

Despite our best efforts, there is still room for error in this geo-coding process. We expect that such errors would introduce random noise into the analysis, drowning out potential signal and introducing attenuation bias of conditional effects towards 0.

A.1.4.2. ADDITIONAL APPLICATION ANALYSES

Here, we include additional analyses associated with the main application. In particular, in Figure A.4, we see the correlation between the estimated Image CATEs and Tabular CATEs using various conditioning sets. These correlations are further broken down by tabular covariate in Table A.1. We show in Figure A.3 the images having the greatest uncertainty in the cluster probabilities (estimated by the posterior standard deviation).

Table A.1: Correlation of estimated image cluster 1 probabilities with key tabular covariates.

	Correlation
Urban	0.18
Longitude	-0.01
Latitude	0.27
Female indicator	0.04
Human capital score	-0.11

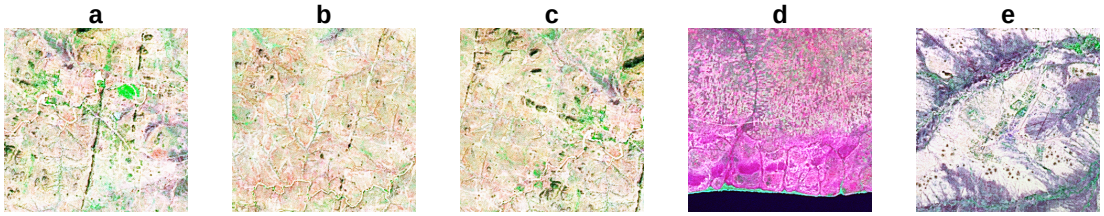

Figure A.3: Images with greatest uncertainty in cluster probabilities from the main empirical analysis.

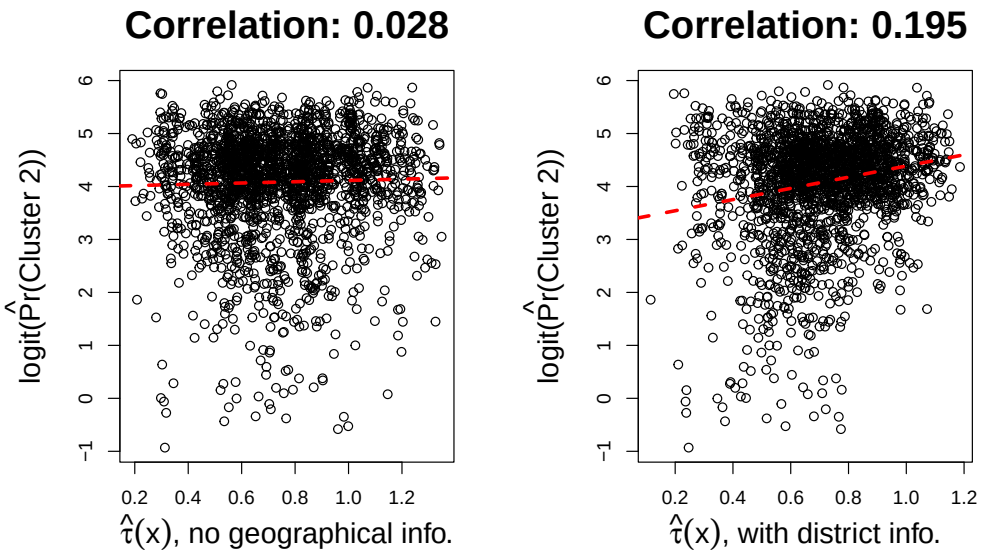

Figure A.4: *Left*. Correlation between estimated treatment effects using a causal forest with individual-level tabular covariates and the posterior mean cluster 2 probabilities from the image heterogeneity model. Individual-level covariates include gender, education, parental education, and indicators for whether a unit's mother or father were alive at the start of the experiment. *Right*. Correlation between estimated treatment effects using a causal forest with individual-level tabular covariates along with district-level indicators and the posterior mean cluster 2 probabilities. As expected, the correlation increases, but there is still considerable information present in the estimated clusters not reducible to district indicators alone.

A.1.5. Empirical Analysis with Orthogonalized Potential Outcomes

To understand the degree to which the neighborhood-level satellite information is a proxy for tabular covariate information, we perform an image CATE analysis in the space of orthogonalized potential outcomes. We orthogonalize potential outcomes by fitting a model for the observed potential outcomes using tabular covariates and residualizing. The potential outcome model here used is a linear regression model predicting each observed outcome using main treatment effects and interactions between treatment and gender, treatment and baseline human capital, and treatment and baseline business capital (as well as the main effect terms for the associated interaction). We use this functional form because it is similar to a model used in the original experimental analysis. We find an absolute correlation of 0.85 between the cluster probabilities using the orthogonalized and raw outcomes.

A.1.6. Model Implementation Details

In the implementation of our models using Bayesian CNN arms, we leave the number of hidden layers, the filter size, the stride length and other quantities as hyper-parameters that can be set by investigators. Future work should explore the implications of these choices on the practical considerations of probabilistic causal image analysis. That said, there are some general principles

that are evident from prior research, such as the idea that, with more data, the number of hidden layers can be increased.

We also here add information about the choice of priors in the Bayesian model. The unconstrained components of the uncertainties are drawn from Gaussians with mean and variance scaled indexed to z; the non-negativity of the variance is enforced through the softplus transformation (where $\text{softplus}(x) = \log(1 + \exp(x))$). Neural network parameters receive priors using the Empirical Bayes' approach described in Krishnan et al. (2020). We have found the use of Empirical Bayes to be important practice: given the highly non-linear parametric functions applied here, seemingly non-informative priors in the parameter space (e.g., $\mathcal{N}(0, 10)$) can be highly informative in the space of induced transformations. The prior for the treatment effect mixture components are centered around the non-parametric difference-in-means estimator for the ATE.

In our application, we use four convolutional layers (filter dimension 5×5), separated by max-pooling layers (2×2). Each convolutional layer applies 32 filters. Bottleneck projection layers are used after each convolutional layer, projecting the 32 dimensions down to 3 to keep the number of parameters reasonably low given the small sample size available in the application. Batch normalization layers are used across the feature dimension after each non-linearity (batch normalization momentum across each update step is $= 0.90$). The swish activation is used. We apply the Gumbel-Softmax to approximate the random categorical sampling with the temperature parameter set to 0.5. We use the flipout estimator for the neural parameter sampling (Wen et al., 2018). Five Monte Carlo iterations are used in each variational inference training step. With this model structure, each batch sample of 20 units takes about one second on a single Apple M1 GPU using Metal-optimized tensorflow 2.11. The full simulation suite takes about 12 hours on local hardware.