Fit is all you need: Counterfactual Image Generation for Spatial Intervention Evaluation

Daniel Lasso-Jaramillo

Abstract

This paper proposes a Deep Learning method for estimating the effect of a spatial intervention, measured by satellite images, by generating a counterfactual image that accounts for possible SUTVA violations. I train a Convolutional Neural Network (CNN) with control units' pre-intervention data to learn a latent representation of their post-intervention states. This representation is further applied to the treated unit's pre-intervention data to generate a counterfactual raster. CNNs are structured to preserve the spatial structure of the raster and generalize the evolution of outcomes on a higher-level unit than a pixel, which allows for a context-driven estimation that addresses exposure to treatment common SUTVA violations on spatial interventions. Also, it allows a flexible estimation of the Average Treatment Effect (ATE) that avoids misspecification and can rule out the effects of common confounders on the estimation. Also, I show that the method is able to generalize the confounders evolution trough time, allowing to account for them I present novel insights that articulate the causal machine learning literature, the computer vision literature, and the spatial impact evaluation literature. Results suggest that the method can generate counterfactual satellite images. Also, the proposed method outperforms Synthetic Control in 95% of the simulations based on real data presented.

Keywords: Causal Machine Learning, Image Generation, Remote Sensing, Spatial intervention, Convolutional Neural Networks.

1. Introduction

How can we evaluate the impact of a spatial intervention on an outcome of interest using satellite images? How could we use image generation and computer vision to accomplish this task? Literature related to spatial interventions approaches this issue on a similar basis. First, they extract tabular data from spatial sources such as maps or rasters, and then apply causal inference techniques such as Synthetic Control (SC), Regression Discontinuity (RD), or Differences in Differences (DD) to identify causal effects across the interested outcomes (Robert, M., 2021; Ruining, 2021; Fick, S., 2021).

Yet, this two-stage approach encounters several challenges. Firstly, the assumption of identical independent distribution (i.i.d.) is frequently violated in spatial interventions, rendering statistical inference invalid unless corrected (Abadie et al., 2023). Secondly, owing to spatial correlation or spill-overs, defining exposure to treatment becomes inherently ambiguous, thus violating the Stable Unit Treatment Value Assumption (SUTVA) (Keele, L., 2015). Also, interference in the spatial treatment multiple treatment locations affect an individual pixel (Pollmann, 2022). Thirdly, estimation is constrained to a rigid level unit (i.e. pixel), precluding contextual estimation that may arise after understanding how features or entities are affected by treatment as a whole (i.e. rivers, forests, communities, etc.). Also, specification and/or omitted variable bias may arise if a confounder is not correctly controlled for (Callaway, B. et al., 2021).

I propose a methodology, Deep Network Synthetic Control (DN-SC), based on Synthetic Control Methods (SCM) and Deep Learning (DL) to generate a counterfactual raster that allows to evaluate the effect on certain dimensions of spatial interventions. I train a Deep Network (DN) that uses control units' pre-intervention raster data to learn a latent representation of their post-intervention states. This representation is further applied to the treated unit's pre-intervention data to generate a counterfactual raster. The causal effect is estimated by contrasting the predicted raster to the observed one. Similarly, I develop a framework that exploits data augmentation techniques common in computer vision as an extension for traditional placebo inference over the causal effect estimated on an image.

Finally, I present a SHAP values approach that helps to explain the generated counterfactual image.

Results suggest that it is possible to provide a counterfactual context-coherent raster image that considers the evolution of spatially related features over time and creates a proper counterfactual when SUTVA is violated. Moreover, Monte Carlo simulations suggest that this methodology outperforms traditional Synthetic Control estimation methods, providing more accurate point estimates for 95% of the simulations carried. Finally enhances a transparent and visually accurate discussion over identification validity.

The proposed method contributes to the literature through the following approaches: i) It effectively deals with violations of SUTVA by leveraging Convolutional Neural Network (CNN) neurons, which specialize in the evolution of features or entities beyond the level of unit (pixels), allowing for a data-driven definition of interference across units (pixels). This capability is acquired during the training phase, where the CNN learns the evolution through time of similar features among control units (i.e. rivers, forests, communities, etc.). This allows for a joint estimation of counterfactual predictions within the generated raster. ii) The flexibility of the CNN lets the algorithm learn the evolution of confounders on the matrix, allowing for accounting for this effect on the counterfactual estimation.

DN-SC method relates to the SCM as it fits in a category-bounding predictive technique used to create a synthetic counterfactual of a treated unit. Abadie (2003) introduced a method where the treated unit is represented as a linear weighted sum of untreated units, minimizing mean squared prediction error (MSPE). Other approaches have explored the possibility of using Machine Learning to improve the synthetic control. For example, by using elastic net to deal with multidimensional data (Doudchenko & Imbens, 2016; Bône A, et al., 2022). This paper differs from this literature as it is not focusing on creating a combination of untreated units, but to learn a latent representation of the effect of confounders on outcomes over time.

It also relates to the Causal Machine Learning literature, which combines causal inference with novel machine learning techniques (Athey, 2019; Huber M. 2021). Especially relevant, Athey et al. (2021) proposes an application of the Machine Learning method called Matrix

Completion for causal panel data models (Candes and Recht (2009); Candes and Plan (2010); Mazumder et al. (2010)). The idea is to impute the "missing" potential control outcomes of a time-series dependent matrix for the untreated unit. However, DN-SC goes beyond this literature, as its main objective is not to complete a matrix, but to entirely generate it.

DN-SC is also pertinent to Neural Networks, which constitute nonlinear functions enabling the prediction of covariate responses in a flexible manner (Hastie, T., 2013). These networks have found application in diverse domains including image classification and generation, forecasting, and medical research. Notably, they have been employed to generate contrast media states in Magnetic Resonance Imaging (MRI) scans of individuals afflicted with brain cancer, leveraging pre-contrast images (Bône A, et al., 2022). Additionally, Neural Networks have contributed to the field of future generation and image prediction, a subdiscipline of computer science focusing on forecasting subsequent frames in a video sequence based on previous pixel data (Vondrick, 2017; Oprea, 2022; Gao, Z., 2022). Both applications provide a structured approach for predicting the future status of raster images.

Section 1 in this paper presents the introduction, section 2 presents the potential outcomes framework and addresses the SUTVA violation problem. Section 3 presents the training and architecture of the Convolutional Neural Network and section 4 provides a simulation framework to compare the proposed method fit when compared to the traditional causal inference approaches.

2. Spatial Problem Setup

Consider the scenario where there are N_0 untreated units and N_1 "treated"[1] units, for a total of N individual images. These units are observed over T time periods, where t0 denotes the period when spatial treatment occurs. Let $\boldsymbol{Y_{i,t}}$be a matrix of size J x K, where $y_{i,t}^{j,k}$ represents the value in the jth row and kth column of the matrix $\boldsymbol{Y_{i,t}}$ with $j, k \in (J, K)$. Therefore, the observed value of the outcome for unit *i* can be represented as follow:

[1] In this context I assume that untreated units are objectively not exposed to the treatment, where for the "treated" matrix the exposure of the individual pixels is ambiguous.

$$Y_{it} = \begin{pmatrix} y_{i,t}^{1,1} & \cdots & y_{i,t}^{1,K} \\ \vdots & \ddots & \vdots \\ y_{i,t}^{J,1} & \cdots & y_{i,t}^{J,K} \end{pmatrix} \quad (1)$$

Consider the potential outcome for a pixel $y_{i,t}^{j,k}(s_{jk})$ a function of a treatment matrix s_{jk} that establishes the relationship between the matrix's pixels and $y_{i,t}^{j,k}$. Under this framework, for every matrix Y_{it} in the control group, $\forall t$ the observed potential outcome will be defined as:

$$Y_{it}(0) = \begin{pmatrix} y_{i,t}^{1,1}(0) & \cdots & y_{i,t}^{1,K}(0) \\ \vdots & \ddots & \vdots \\ y_{i,t}^{J,1}(0) & \cdots & y_{i,t}^{J,K}(0) \end{pmatrix} \quad (2)$$

with 0 a special case of the assignment to treatment matrix s_{jk} that implies that none of the pixels were treated. On the other hand, for the treated unit pre-intervention period ($t < t_0$) the observed potential outcome is described as in equation (2), but posterior to the intervention is defined as:

$$Y_{it}(s_{jk}) = \begin{pmatrix} y_{i,t}^{1,1}(s_{11}) & \cdots & y_{i,t}^{1,K}(s_{1k}) \\ \vdots & \ddots & \vdots \\ y_{i,t}^{J,1}(s_{j1}) & \cdots & y_{i,t}^{J,K}(s_{jk}) \end{pmatrix} \quad (3)$$

Due to the fundamental problem of causal inference only the treated potential outcomes are observed for the treated group. If both potential outcomes $Y_{it}(s_{jk})$ and $Y_{it}(0)$ then, identification of the causal parameter $E(\tau_{it}) = E(Y_{it}(s_{jk}) - Y_{it}(0))$ corresponds to the Average Treatment Effect (ATE) of the spatial intervention.

Assumption 1:

For DN-SC to be able to infer about the confounders effect through time the assumption of common shocks implies that if an unobserved confounder z_i exists, then it should occur on the treated and untreated units. This can also be seen as a positive assumption common in the Causal Machine Learning literature (Athey, 2019; Poulous, Z., 2019)

Spatial Disturbance Distance

I begin by assuming that the closer a pixel is to others, the more it is affected by them. Conversely, if a pixel is far from another, there will not be any correlation between them. This implies that the treatment allocation matrix s_{jk} is a sparse matrix, taking values of 1 for the neighbors related to the pixel $y_{i,t}^{j,k}$ and 0 for those distanced. Moreover, I will generalize that for a pixel $j, k \in (J, K)$ the neighbors of degree z can be represented as $N_z{}^{j,k} = \{j', k' \in \omega : d([j, j'], [k, k']) \leq z\} \;\; \forall j, k \; \neq j', k'$ where $d([j, j'], [k, k'])$ explains the distance between pixels in the matrix.

Assumption 2:

I will assume that the spatial disturbance $d([j, j'], [k, k'])$ follows the *Chebyshev* distance, so the neighbors for a pixel can be expressed as follows:

$$N_z{}^{j,k} = \left\{j^{'}, k^{'} \in \omega : \; \max\{|j - j'|, |k - k'|\}, \leq z\right\} \;\; \forall j, k \; \neq j', k' \qquad (4)$$

Intuitively, for the first degree this metric will contemplate as neighbors the pixels directly connected to the original one.

Assumption 3:

I will assume that s_{jk} is a sparce matrix that for each pixel it takes the value of 1 for the neighbors $N_z{}^{j,k}$ and 0 for the rest of the pixels. Intuitively this matrix expresses the connection between pixels in an individual raster. I will also assume that s_{jk} remains constant for every time (*t*) and for every matrix (*i*).

The Stable Unit Treatment Value Assumption

In most applications, the ATE can be identified under the basic assumption of causal effect stability that requires that the potential outcomes of individuals be unaffected by potential changes in the treatment exposures of other individuals. Simply put, SUTVA implies that if pixel i gets treated in period t, it should have no effect on any other pixel's outcome at any given time, nor should it affect that individual's outcome in other time periods (Imbens and

Rubin, 2015). In terms of the setup, this implies that s_{jk} should be a matrix with only one positive value in the position (j,k).

In spatial settings, this SUTVA assumption can be extremely difficult to satisfy. In the following section, I describe a common setting in spatial interventions where SUTVA is violated and develop a framework with CNNs to generate possible alternative counterfactuals that allows the flexibilization of the assumption.

3. Convolutional Neural Networks

I use a convolutional neural network (CNN) to generate possible alternative counterfactuals. Firstly, I explain convolution filters and explain why it addresses the SUTVA violations. Secondly, I explain CNN´s and the proposed method DN-SC in the matrix generation problem. Finally I explain the architecture of the network, the forward pass and the back propagation to determine the optimal parameters of the estimation.

Convolution filters

A convolution filter relies on an operation called a convolution, which corresponds to repeatedly multiplying matrix elements and then adding the results (Hastie, T., 2013). To understand how a convolution filter works and solves the SUTVA inquiries presented, consider a 3x2 image and a 2x2 convolution filter such as:

$$\boldsymbol{Image_{3x2}} = \begin{pmatrix} y_{i,t}^{1,1} & y_{i,t}^{1,2} \\ y_{i,t}^{2,1} & y_{i,t}^{2,2} \\ y_{i,t}^{3,1} & y_{i,t}^{3,2} \end{pmatrix} \qquad \boldsymbol{Convolution\ filter_{2x2}} = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix}$$

The convolution operation is then a multiplication and posterior sum of the results of the convolution filter and all the 2x2 submatrix of the image[2]. After the convolution, the result is a new lower rank matrix that preserve the information and interactions between pixels:

[2] There are different convolution operations such as full convolution. It involves applying the entire kernel to the input image, resulting in a convolved image size that matches the input size.

$$\boldsymbol{convolved\ image_{2x1}} = \begin{pmatrix} \alpha * y_{i,t}^{1,1} + \beta y_{i,t}^{1,2} + \gamma y_{i,t}^{2,1} + \delta y_{i,t}^{2,2} \\ \alpha * y_{i,t}^{2,1} + \beta y_{i,t}^{2,2} + \gamma y_{i,t}^{3,1} + \delta y_{i,t}^{3,2} \end{pmatrix} \quad (\mathbf{5})$$

I will argue that under *assumptions 2* and 3, by construction, the convolved image expresses the representation of the matrix s_{jk} for each individual pixel. As presented, the new matrix contains information over the interference of the neighbors on the outcomes, which further will be used on the CNN to learn a latent representation of the evolution of features through time and will help to generate a proper counterfactual. Also, I argue that there is a relation between the z degree of the neighbor function ($N_z{}^{j,k}$) and the dimensions of the convolution filter.

Convolutional neural networks

CNNs are a subdivision of neural networks that efficiently processes a grid-like topology data. In general, multiple filters are convolved with the input to extract a representation of features in the topology. The optimal filters are learned during the training process, allowing for a data driven approximation of the meaningful patterns of the image (Goodfellow, I., 2016).

The proposed method consists of constructing a synthetic counterfactual untreated unit $\widehat{Y_{it}}(0)$ for the treated matrix. For doing so, in the training stage, for the untreated unit I will estimate using CNN a general function that maps the pre-treatment image to its post-treatment status conditional on all the previously observed images.

$$Y_{i,t+1} = \hat{f}(Y_{i,t} | \mathrm{D_{i,t}} = 0, Y_{i,t-1}, \ldots, Y_{i,t-T})$$

Then, the learned representation is applied to the treated unit, so the counterfactual prediction is obtained $\widehat{Y_{it}}(0)$. By doing this the generalized function $\hat{f}()$ learns the evolution of the confounders over entities. Also, learns a representation on how spatially correlated pixels are affected among them, internalizing the way potential outcomes correlate.

DN-SC Architecture, Forward Pass and Back Propagation

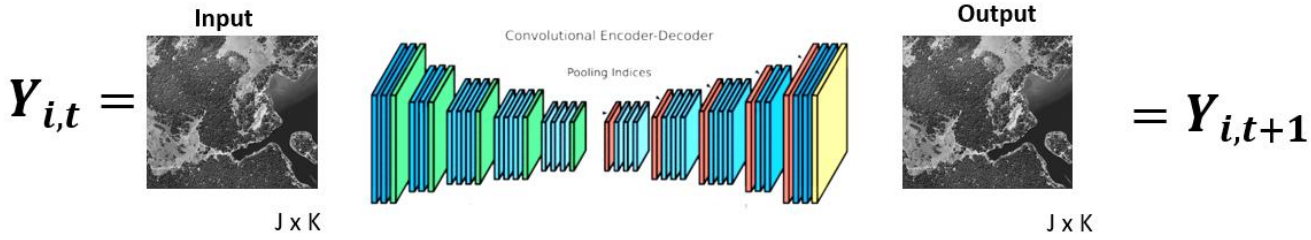
Figure 1 provides an example of the proposed network architecture. In the forward pass stage a distribution of N_0 matrices of sizes (J,K) is transformed throughout multiple convolutions of kernels of size nxm each denoted by $\boldsymbol{Ke_{nxm}}$ and is added a bias matrix or constant, denoted by $\boldsymbol{B_{nxm}}$ until a lower rank latent representation of the matrix is achieved. After that, this representation is passed to a neural network which estimates the new matrix as follows:

$$Y_{i,t} = B_i + \sum_{m=1}^{n} Y_{i,t-1} * K_{i,m}$$

where * represents the convolution operation presented.

After that, the second stage consists of deconvolutional layer that makes the inverse process of convolutional layer and allows to expand the shape of the latent representation to the original one.

Figure 1: Example of an encoder-decoder architecture for image prediction



$$\boldsymbol{Y_{i,t}} = \qquad = \boldsymbol{Y_{i,t+1}}$$

Source: Badrinarayanan et al., 2017

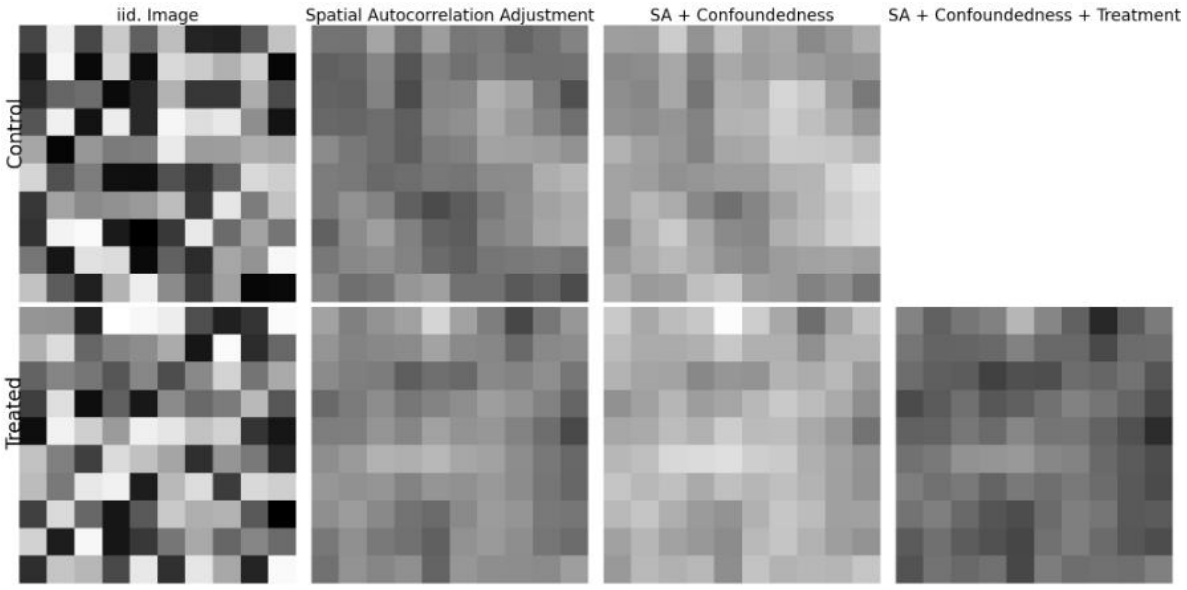
After reconstructing a predicted image, an error measurement can be computed. Literature uses Mean Squared Error as a loss function for image prediction. This can be interpreted as the sum of the subtraction of the predicted and observed matrix $E = (\sum_{i=1}^{N_1} \widehat{Y_{i,t}} - Y_{i,t})^2$.

Backward propagation computes the directional derivates of E with respect to the kernel and biases following the equations $\frac{\partial E}{\partial K_{i,j}} = Y_{i,t-1} * \frac{\partial E}{\partial Y_{i,t}}$ and $\frac{\partial E}{\partial B_i} = \frac{\partial E}{\partial Y_{i,t}}$. For the next propagation (epoch) the weights $K_{i,j}$ and B_i are updated by moving towards the directional derivative, throughout a process called gradient descent with a defined learning rate. Finally, the learned function $\hat{f}()$ is then applied to the N_1 treated matrix.

4. Simulation

To show the performance of the proposed method against traditional causal inference approaches this section simulates raster images for two time periods (pre-intervention and post-intervention) on a control group consisting of 99 observations and 1 treated image. First, an independently and identical distributed image is created using a random uniform distribution with values from 0 to 256. After that, another image with spatial autocorrelation is created by replacing the pixel value $y_{i,t}^{j,k}$ for the mean value of the first-degree neighbors. Moreover, the post-treatment matrixes are created by adding the effect of an unobserved variable (confounder) distributed N(40, 2) to both treated and untreated matrixes. This will prove that DN-SC is capable of learning a latent representation of the evolution of confounders for the untreated units, that should be leveraged to the treated unit. Finally, in the post-treatment matrix of the treated unit, an homogenous treatment effect is included for every pixel. Figure 2 presents a visual explanation of the simulation process.

Figure 2: example of the simulation of the DGP of the images.



References

Abadie, A., Athey, S., Imbens, G., Wooldridge, J., When Should You Adjust Standard Errors for Clustering?, *The Quarterly Journal of Economics*, Volume 138, Issue 1, February 2023, Pages 1–35, https://doi.org/10.1093/qje/qjac038

Athey S, Imbens G. 2019. Machine learning methods that economists should know about. Annual Review of Economics 11:685-725.

Susan Athey, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens & Khashayar Khosravi (2021) Matrix Completion Methods for Causal Panel Data Models, Journal of the American Statistical Association, 116:536, 1716-1730, DOI: 10.1080/01621459.2021.1891924

Callaway, B., & Sant'Anna, P. H. C. (2020). Difference-in-Differences with multiple time periods. Journal of Econometrics. doi:10.1016/j.jeconom.2020.12

Huber M. 2021. Causal analysis: impact evaluation and causal machine learning with applications in R. https://drive.switch.ch/index.php/s/tNhKQmkGB48bjfz

Ruining, J., Shuai S., Lili, Y., 2021. "High-speed rail and CO2 emissions in urban China: A spatial difference-in-differences approach," Energy Economics, Elsevier, vol. 99(C).

Fick, S. E., Nauman, T. W., Brungard, C. C., & Duniway, M. C. (2021). Evaluating natural experiments in ecology: using synthetic controls in assessments of remotely sensed land treatments. *Ecological Applications*, *31*(3), 1–16. https://www.jstor.org/stable/27029215

Gonzalez, Robert M. 2021. "Cell Phone Access and Election Fraud: Evidence from a Spatial Regression Discontinuity Design in Afghanistan." *American Economic Journal: Applied Economics*, 13 (2): 1-51.DOI: 10.1257/app.20190443

Abadie, A., and Gardeazabal, J., 2003, The Economic Costs of Conflict: A Case Study of the Basque Country. American Economic Review 93 (1): 113–32.

Brodersen, K., Koehler, J., Remy, N., Scott, S., Gallusser, F., Inferring causal impact using Bayesian structural time-series models, Annals of Applied Statistics, vol. 9 (2015), pp. 247-274

Doudchenko, N., Imbens, G., 2016, Balancing, Regression, Difference-In-Differences and Synthetic Control Methods: A Synthesis. Cambridge, MA. National Bureau of Economic Research.

A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta and A. A. Bharath, 2018, "Generative Adversarial Networks: An Overview," in *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 53-65, Jan., doi: 10.1109/MSP.2017.2765202.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (1st ed.) [PDF]. Springer.

Bône A, et al., 2022, From Dose Reduction to Contrast Maximization: Can Deep Learning Amplify the Impact of Contrast Media on Brain Magnetic Resonance Image Quality? A Reader Study. Invest Radiol. 2022 Aug 1;57(8):527-535. doi: 10.1097/RLI.0000000000000867.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Pollmann, M., 2022, Causal Inference for Spatial Treatments (Working Paper).

Keele, Luke. 2015. “The Statistics of Causal Inference: A View from Political Methodology.” Polit. Anal. 23 (3): 313–35.

Imbens, Guido W, and Donald B Rubin. 2015. Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. Cambridge University Press.