



A structured comparison of causal machine learning methods to assess heterogeneous treatment effects in spatial data

Kevin Credit¹ · Matthew Lehnert²

Received: 30 August 2022 / Accepted: 27 April 2023
© The Author(s) 2023

Abstract

The development of the “causal” forest by Wager and Athey (J Am Stat Assoc 113(523): 1228–1242, 2018) represents a significant advance in the area of explanatory/causal machine learning. However, this approach has not yet been widely applied to geographically referenced data, which present some unique issues: the random split of the test and training sets in the typical causal forest design fractures the spatial fabric of geographic data. To help solve this issue, we use a simulated dataset with known properties for average treatment effects and conditional average treatment effects to compare the performance of CF models across different definitions of the test/train split. We also develop a new “spatial” T-learner that can be implemented using predictive methods like random forest to provide estimates of heterogeneous treatment effects across all units. Our results show that all of the machine learning models outperform traditional ordinary least squares regression at identifying the true average treatment effect, but are not significantly different from one another. We then apply the preferred causal forest model in the context of analysing the treatment effect of the construction of the Valley Metro light rail (tram) system on on-road CO₂ emissions per capita at the block group level in Maricopa County, Arizona, and find that the neighbourhoods most likely to benefit from treatment are those with higher pre-treatment proportions of transit and pedestrian commuting and lower proportions of auto commuting.

Keywords Causal forest · Heterogeneous treatment effects · Machine learning · Causal inference · Spatial · CO₂ emissions · Transit

JEL Classification C21 · C52 · C54 · C63

✉ Kevin Credit
kevin.credit@mu.ie

¹ National Centre for Geocomputation, Maynooth University, Maynooth, Co. Kildare, Ireland

² Sate.lytics, Perrysburg, Ohio, USA

1 Introduction

Initially, the use of machine learning techniques such as random forest and neural networks focused mostly on prediction tasks because of the “black box,” non-linear nature of the relationship between input variables and outputs. But given the high predictive performance of these models compared to traditional statistical methods (like linear regression) (Strittmatter 2019; Farbmacher et al. 2021; Hagenauer et al. 2019; Yoshida and Seya 2021; Credit 2022), there has been considerable interest in developing new approaches for using these models to answer explanatory—and even causal—research questions in the social sciences. The development of the “causal” forest by Wager and Athey (2018) represents a significant advance in this area. By applying the fundamental logic of the principles of causal inference—“the science” of Rubin (2005)—to the powerful random forest model format, the causal forest not only provides an estimate of the average treatment effect (ATE) across the entire dataset, but also unit-level conditional average treatment effects (CATE) that allow researchers to draw conclusions about the effectiveness of treatment across various subpopulations (Athey and Imbens 2016). Providing estimates of heterogeneous treatment effects (HTE) for *all* treated and untreated units are particularly valuable in the context of local urban policy decision-making and could allow planners to assess the viability of infrastructure investments or policies in candidate areas.

Several studies have used causal forests to analyse HTE for the effect of various interventions on outcomes such as student achievement, agricultural yields, crime, and corporate investment (Athey and Wager 2019; Deines et al. 2019; Hoffman and Mast 2019; Davisd and Heller 2020; Gulen et al. 2019). However, this approach has not yet been widely applied to geographically referenced data (Deines et al. 2019; Hoffman and Mast 2019), which present some unique issues, including the fact that the random split of the test and training sets in the typical causal forest design fractures the spatial fabric of geographic data (Nikparvar and Thrill 2021).

Whilst this randomised split makes sense in a controlled trial for independently selected individuals, geographic entities have intrinsic relationships with each other based on distance (Tobler 1970) that should be included as a part of the process of modelling treatment effects, particularly when spatially lagged characteristics are included in the model. At the same time, of course, concerns over bias from overfitting training data are well-understood and need to be addressed, especially in situations where the treatment assignment may not be randomly determined, as is the case in most urban planning applications (Hawkins 2004; Kuhn and Johnson 2013).

To help solve these issues, we use a simulated dataset with known properties for ATE and CATE to compare the performance of causal forest models across two different definitions of the test/train split. We also develop a “spatial” T-learner (STL) that can be implemented using predictive methods like random forest to provide estimates of HTE across all units (Künzel et al. 2019). Our results show that in cases where the treatment is not randomly assigned, all three

machine learning models *nearly* significantly outperform ordinary least squares regression (OLS) at identifying the true ATE, but are not significantly different from one another. This indicates that, in this instance, there are no significant differences in performance based on the nature of the train/test split. The spatial lag of X (SLX) and spatial Durbin specifications offer some small additional performance over the baseline (non-spatial) specification, with the Durbin causal forest model without train/test split displaying the best estimate of true ATE and second-lowest mean squared error (MSE) for CATE, although, again, the difference is not statistically significant.

We then apply this preferred model in the context of analysing the treatment effect of the construction of the Valley Metro light rail (tram) system on on-road CO₂ emissions per capita at the block group (BG) level in Maricopa County, Arizona. We estimate that building the Valley Metro system reduced on-road CO₂ emissions at the BG level by an average of 8.02% over the treatment period analysed (2009–2017), or about 1% per year. This is roughly in line with estimates of reduction in total Vehicle Miles Travelled (VMT) due to transit construction in other studies (e.g. 2.97% per year) (Ewing and Hamidi 2014). Interestingly, the characteristics of BGs that are most important to predicting treatment effect are all transportation-related: the proportion of auto commuters in neighbouring BGs, the proportion of cycling and walking commuters, the proportion of transit commuters in neighbouring BGs, and the proportion of cycling and walking commuters in neighbouring BGs. Block groups with higher levels of transit, walking, and cycling commuting tend to have larger estimated reductions in CO₂ per capita due to light rail construction (treatment effects), whilst BGs with higher levels of pre-treatment CO₂ auto commuting tend to have smaller treatment effects from light rail. These relationships are generally reflected in the resulting spatial pattern of estimated treatment effects, which suggests that the chosen corridor was reasonably well-suited for construction (in terms of impacts on CO₂). Areas in the more walkable city centres of Phoenix, Tempe, and Mesa are most likely to benefit from light rail construction, whilst areas with more auto-oriented development patterns (e.g. those near highways) are less likely to benefit in terms of on-road CO₂ reductions.

By employing a causal machine learning analysis framework designed specifically around spatial data, providing a structured comparison of various causal machine learning specifications, and developing the STL, this paper provides a useful contribution to the ongoing work on “explainable AI” approaches in general and causal machine learning methods more specifically. In addition, the empirical results provide insight on the positive impact of transit construction on automobile emissions at a fine-grained spatial scale, which can be used by urban planners and policy-makers to help inform future decision-making around transportation infrastructure investment.

2 Methods

2.1 Potential outcomes model

The general approach to identifying causal effects from empirical data is known as the “potential outcomes” or “Rubin Causal Model,” which provides a basic

formulation of the causal estimand (Rubin 1978; 2005). In the potential outcomes model, we have a number of individual units (i) which receive some treatment (T), and we are trying to find out whether that treatment has an effect on some outcome of interest in the post-intervention time period (Y). In a medical application of causal inference, this might be an analysis of the effect of some drug treatment (T) on the blood pressure (Y) of hospital patients (i), or in a geographic context, it might be an analysis of the effect of the construction of a new light rail line (T) on the on-road CO₂ emissions (Y) of Census block groups (i). In any case, in the potential outcomes model, conditional average treatment effects (CATE) are determined by taking the difference in Y between the treated condition $Y_i(1)$ (i.e. what would happen if a unit received the treatment) and the counterfactual condition $Y_i(0)$ (i.e. what would happen if a unit was never treated): $Y_i(1) - Y_i(0)$. Averaging these differences across an entire set of similar units provide the average treatment effect (ATE): $E[Y(1) - Y(0)] = E[Y(1)] - E[Y(0)]$. Unfortunately, it is of course not possible to observe both $Y(0)$ and $Y(1)$ empirically for the same unit (because in reality, some units are treated and others aren't), which is known as the “fundamental problem of causal inference” (Rubin 1978; 2005). Therefore, in practice, the TE can only be *estimated* by comparing the values of Y in the treated and untreated (or “control”) units, commonly by taking the average (or median) difference in Y between the treated and untreated observations.¹

However, identification of the estimated TE in this way rests on several assumptions, in particular “ignorability” or “conditional independence” and the Stable Unit Treatment Value Assumption (SUTVA). For ignorability, the simple idea is that the set of observations in the treatment and control groups must be as similar as possible in order to draw a valid causal inference about the effect of a treatment; or, put another way, the treatment assignment cannot be related in any way to the potential outcomes (Rubin 1978; 2005; Dorie et al. 2019; Matthay et al. 2020). In a true randomised experiment, the treatment assignment mechanism is random, thus allowing us to “ignore” it (i.e. the treatment and control groups are equally likely to have any given set of characteristics). This is of course impossible to recreate in most cases of urban-geographic or social science research where the treatment is (often very purposefully) not applied randomly.² The solution in these cases is to control for

¹ A common extension of the potential outcomes model, often used in economics, is the ‘difference in differences’ estimator or ‘gain score’ and evaluates the difference between the treatment and control observations in terms of their *difference* from pre- to post-intervention: $E[Y_{\text{POST}}(1) - Y_{\text{PRE}}(1)] - E[Y_{\text{POST}}(0) - Y_{\text{PRE}}(0)]$ (Card and Krueger 1994; Gelman and Hill 2006). The difference-in-differences estimator inherently incorporates the pre-intervention value of Y into the estimation of the causal effect (as the outcome is a difference ‘from’ that value); not knowing at the outset how this construction might behave in the causal forest model, we opted to test the more basic potential outcomes model with the pre-intervention value of Y as a covariate (that could be removed if necessary) in our simulation study. Given the vast and quickly growing body of literature on difference-in-differences, however, it would certainly be useful to extend the concepts explored in this paper to difference-in-differences in future work.

² In Fig. 2, we develop a directed acyclic graph (DAG) to describe the causal relationships in our simulated data design, in which the treatment is not randomly assigned. The magenta lines denote “biased” causal pathways where particular variables influence both the treatment assignment T and the potential outcomes inherent in the response variable, Y_{POST} . These must be controlled for in order to fulfil the “conditional” ignorability assumption.

the effect of confounding variables, i.e. pre-treatment characteristics of observations that affect both the treatment assignment and the potential outcomes, either by carefully matching treatment and control observations using a propensity score methodology (Rubin 2001) or by including the pre-treatment confounders in the estimation of the causal model, after which the model can be said to fulfil the assumption of “conditional” ignorability (Matthay et al. 2020; Gelman and Hill 2006).

For the SUTVA, the basic principles are that (1) each observation’s potential outcome is not dependent on other observations’ treatment status, and (2) that there are not multiple or “hidden” versions of treatment (Rubin 1980, 2005; VanderWeele and Hernán 2013). This is particularly difficult to fulfil in urban-geographic or other spatial contexts where global and local spatial spillovers and neighbourhood effects may be present (Baylis and Ham 2015; Kolak and Anselin 2020). For instance, extending the light rail example from above, for a given treated block group i , treatment in a neighbouring observation j may increase i ’s treatment effect since additional nearby stations multiply the usefulness of the rail line (and thus the reduction in on-road CO₂) due to there being more nearby destination options. Or perhaps the fact that j is treated means that complimentary transit-oriented development is established in j which attracts residents of i to that area, from which they take transit, replacing auto trips in a way that would not have happened had j not been treated. Whilst one common approach to mitigating the effects of the SUTVA is to aggregate the data to larger spatial units (so that i and j are no longer neighbours, but part of the same observation), this reduces the granularity of the data and what we can learn from it. For that reason, a spatially explicit modelling approach that takes into account these spatial effects has been suggested as a better alternative solution to the SUTVA (Baylis and Ham 2015; Huber and Steinmayr 2019; Kolak and Anselin 2020; Kosfeld et al. 2021; Butts 2021).

2.2 Spatial models

In fact, spatial effects can impact the basic causal inference framework in a variety of complex ways, including treatment assignment (ignorability), spillover effects (SUTVA), mismatched scales for spatial processes and outcomes, and more. Given this, Kolak and Anselin (2020, p.132) suggest taking a broadly spatial perspective that goes “beyond the uncritical implementation of spatial tools or methods...[to] consider the inherently spatially and temporally dynamic, interactive nature of the populations being studied, and, as such, inform the initial design of the model”. In this paper, we adapt the standard spatial model specifications for use in a machine learning context, with spatial lags of the dependent and independent variables directly entered into the model as individual variables (as in similar previous work) (Credit 2022). However, we’ll start here by describing the standard spatial econometric specifications in matrix notation for a two time-period analysis so that the reader has a point of comparison for the machine learning-adapted specifications

used in this paper³ (Gelman and Hill 2006; LeSage and Pace 2009; Kolak and Anselin 2020). We'll start with the baseline (non-spatial) OLS specification in Eq. (1):

$$Y_{\text{POST}_i} = \delta T_i + \beta X_i + \varepsilon_i, \quad (1)$$

where Y_i is the value of the outcome variable (Y) in the post-intervention time period (POST) for observation i , T is a binary treatment variable taking the value of “1” if an observation is treated and “0” if not, X is a vector of confounding covariates measured before the treatment, and ε is the error term. In the spatial lag or spatial autoregressive model (SAR) shown in Eq. (2), the spatial lag of the dependent variable (WY_{POST_i}) enters on the right-hand side of the equation, which controls for the value of post-intervention outcomes in neighbouring observations (i.e. spillover effects):

$$Y_{\text{POST}_i} = \rho WY_{\text{POST}_i} + \delta T_i + \beta X_i + \varepsilon_i, \quad (2)$$

In the spatial lag of X (SLX) specification shown in Eq. (3), spatial lags of the treatment variable (WT_i) and covariates (WX_i) are included. These lags control for the effect of neighbours' treatment status on a given observation's (i) outcome (the most straightforward SUTVA violation) as well as for the effect that *neighbouring* confounders might have on i , which could capture spillovers in the covariates or differing spatial extents between the determinants of a process and where it is being measured (i.e. a reduction in on-road CO₂ emissions in block group i may be related to the % of people commuting by auto in block group i as well as the % of people commuting by auto in all of the surrounding block groups as well):

$$Y_{\text{POST}_i} = \delta T_i + \zeta WT_i + \beta X_i + \lambda WX_i + \varepsilon_i, \quad (3)$$

Lastly, the spatial Durbin model simply combines the SAR and SLX models by including spatial lags of both the dependent and independent variables (including T), as described in Eq. (4):

$$Y_{\text{POST}_i} = \rho WY_{\text{POST}_i} + \delta T_i + \zeta WT_i + \beta X_i + \lambda WX_i + \varepsilon_i \quad (4)$$

These specifications form the basis for comparing the performance of different causal machine learning models in this paper under designed conditions of spatial dependence.

³ There are some important differences between linearly estimated spatial models (whose specifications are described in Eqs. (1)–(4)) and the machine learning versions used in this paper. First, the “forest” algorithms are nonlinear estimators, so they do not produce a typical regression coefficient (Krzywinski and Altman 2017). Second, because of this, causal forest and random forest do not use the typical procedures for estimating spatial econometric models, e.g. maximum likelihood estimation, Bayesian estimation, or general method of moments (GMM) (LeSage and Pace 2009; Anselin and Rey 2014). Thus, they do not produce estimates of “direct” and “indirect” spatial effects from the spatial lag model and cannot be used to directly estimate spatial error models.

2.3 Heterogeneous treatment effects (HTE)

Heterogeneous treatment effects (HTE), i.e. variations in the extent to which a given treatment impacts different types of observations, are important to understand in nearly all applications of causal inference. However, given the fundamental problem of causal inference, the true CATE are not readily observable; only aggregate measures of the treatment effect can be observed from the data. Researchers initially approached this problem via “sub-group” analysis, which essentially entails splitting the sample into relevant subsets (e.g. women vs. men) and finding the difference in ATE between the two, and/or inserting an interaction term into a causal regression between the characteristics of interest and the T variable (Parker and Naylor 2000; Lagakos 2006; Imai and Ratkovic 2013). However, several issues have been noted with this approach, including the fact that subsetting the data reduces the sample size in each sub-group, increasing the number of sub-groups increases the chance of a false positive, and this approach still does not provide estimates of CATE. To address these shortcomings, in recent years, many semi- and nonparametric machine learning methods for estimating HTE have been developed, including Bayesian regression trees (Hill 2011), Support Vector Machines (SVM) with LASSO (Imai and Ratkovic 2013), neural networks (Shalit et al. 2017) and the causal forest (Wager and Athey 2019). Given the relative novelty of the causal forest and the demonstrated high predictive power of the (similar) random forest (Strittmatter 2019; Farbmacher et al. 2021; Credit 2022), in this paper, we are interested primarily in comparing the performance of the causal forest alongside a new, spatially informed, T-learner model for estimating HTE (using random forest). We describe both in more detail below.

2.3.1 Causal forest

The causal forest is a type of “generalised” random forest that produces predicted values of the *unit-level* conditional average treatment effects rather than predicted values of the outcome variable, as in the traditional random forest (Athey et al. 2019). Statistically, the generalised random forest is broadly similar to the random forest in that it is an ensemble of decision trees that are grown on random subsamples of the training data. It uses an “honest” estimation strategy in which a given tree’s (random) subsample is split again into two subsamples: one used to estimate the tree’s partitions, and the other to “populate” the tree based on these partitions. For building the tree, a random subset of candidate covariates is selected; for each covariate, the algorithm takes each possible value and considers making a split based on which value will *maximise heterogeneity* in the resulting “child” nodes, which is importantly different than the random forest’s goal of *minimising prediction error*. For the causal forest, this involves maximising the difference in estimated TE between the children according to a linear approximation of the mean difference gradient (Wager and Athey 2018; Knaus et al. 2018; Athey et al. 2019; Tibshirani et al. 2022a). For predicting values, the generalised random forest can be conceptualised as an adaptive-kernel nearest-neighbour method, where “proximity” is measured in terms of the characteristics of the training observations that fall into the same leaf

as a given test observation. In this way, the test set is “dropped” into the built tree; each test observation lands in a particular leaf based on its characteristics, and a list of similar training observations (leaf-mates) is generated. A neighbourhood weight for each training observation is then calculated (across all trees in the forest) based on the number of times it falls into the same leaf as a given test observation. The predicted TE for each test data point, then, is the neighbourhood-weighted average difference of the outcome variable (Y) between treated and untreated observations (Wager and Athey 2018; Knaus et al. 2018; Athey et al. 2019; Tibshirani et al. 2022a). The causal forest as estimated in the *grf* package in R also includes some items pertaining specifically to the estimation of causal effects, including orthogonalization and balanced splits (between observations in the treatment and control groups) (see Tibshirani et al. 2022a for more detail).

2.3.2 Spatial T-learner (STL)

This paper also explores an alternative “forest”-based approach to estimating the CATE, which we have called the “spatial” T-learner (STL) after the two-stage (T) metalearner described in Künzel et al. (2019). The basic concept for the spatial T-learner is the same as the T-learner but with explicitly spatial components in terms of both a) the delineation of the treatment and control areas and b) the inclusion of spatially lagged variables in the predictive models at both states. The STL first involves *predicting* outcomes (Y_{POST}) using only data from the control group (training set) in the pre-intervention period. In this paper, we have estimated all steps of the STL using random forest due to its strong predictive accuracy, but other models could be used. Equation (5) shows the spatial Durbin specification (without regression coefficients) for first stage of the STL model:

$$Y_{\text{POST}_i} = WY_{\text{POST}_i} + WT_i + X_i + WX_i, \quad (5)$$

where $T_i=0$. Dropping the values of the “test” set—in this case, only the treatment area—into this tree (using the ‘predict’ function in R) produces an estimate of the *counterfactual* condition in the treated area, i.e. an estimate of Y_{POST} if the treatment had never occurred in the treated observations. This is a reasonable assumption as long as the number and quality of covariates is enough to produce a good prediction and the control and treatment groups are *relatively* similar (e.g. within the same city or subpopulation). In this case, we can say that \hat{Y} from Eq. (5) is an estimate of the true $Y(0)$, i.e. $\hat{Y}(0)$.

Now we can obtain estimates of the CATE (TE_i) directly by subtracting $\hat{Y}_i(0)$ from observed values for $Y_i(1)$ in the treatment area, Y_{POST_i} (where $T_i=1$). To find the CATE for *untreated* observations, a secondary random forest model is trained with TE_i as the dependent variable, as shown in Eq. (6):

$$TE_i = WY_{\text{POST}_i} + WT_i + X_i + WX_i, \quad (6)$$

where $T_i=1$. Dropping the values of the “training” set, i.e. the control group, into this tree produces predictions for the TE in the untreated observations (based on their other characteristics). From here, all summary statistics of treatment effect can

be calculated, including the ATE, which in this case is simply the average of TE_i across the entire study area.

3 Simulated data design

It is difficult to structure a true comparison of the performance of different causal machine learning methods at identifying average and unit-level TE without knowing the values of these effects beforehand, which necessitates the use of some simulated data with known properties. In this case, we created a simulated dataset that attempted to mimic the conditions present in many urban-geographic applications, including spatial effects and relationships. Our idea here is to design a system where the treatment assignment, pre-intervention values of Y , and post-intervention values of Y are all related (to various extents) to the spatial configuration of the data. This is motivated by a thought experiment—similar to the one we investigate in this paper’s application—in which city centre locations attract populations with certain characteristics, e.g. educational attainment and propensity to cycle or walk to work (U_1 and U_2), that generate lower pre-intervention levels of on-road CO₂ emissions (Y_{PRE}). There are also unobserved variables (UO) and variables whose characteristics are randomly distributed with respect to spatial location (X_1 and X_2).

To do this, we first create a rectangular study area of 500 equidistant points, 35 wide \times 25 long. Next, we identify a smaller rectangular “centre” area in the middle of the study area, as shown in Fig. 1a, and create two confounding factors, U_1 and U_2 , correlated with the *Centre* dummy variable at 0.7 and 0.5, respectively, using the R package *faux* version 1.1.0. We then add a “pre-intervention” value of Y , Y_{PRE} , correlated with the *Centre* dummy at -0.7 and an “unobserved” variable (UO) that would not be added to any of the models but is also correlated with *Centre* at 0.8. This is intended to reflect many real-world situations where numerous factors are correlated with specific geographic locations but are not available to model. Finally, we add two uncorrelated random variables for additional noise, X_1 and X_2 . Full characteristics of the variables can be found in Appendix 2.

We can imagine that true counterfactual values of Y are in some sense a product of all of these factors, so we create a new variable, $Y(0)_f$, that is correlated with the entire existing bundle of variables at 0.3, and add that to Y_{PRE} to obtain the true counterfactual value of Y in all locations, $Y(0)$. Importantly, this is a product of each observation’s characteristics (some of which in turn are correlated with the central location), but not related to the treatment in any way. Next, we model a (very realistic) scenario in which treatment assignment is correlated with existing characteristics (U_1 , U_2 , and UO^4) of observations (0.7). Continuing our example, this assignment mechanism simulates a condition in which the characteristics of populations drawn to the centre not only influence the pre-treatment emissions and counterfactual emissions, but also make it more likely (through activism, planning analysis, etc.) to have a new transit system built in their neighbourhoods. Thus, in this way,

⁴ In this case, Y_{PRE} was excluded since it is negatively correlated with the other variables, which prevented high correlations from being possible. X_1 and X_2 were excluded for similar reasons.

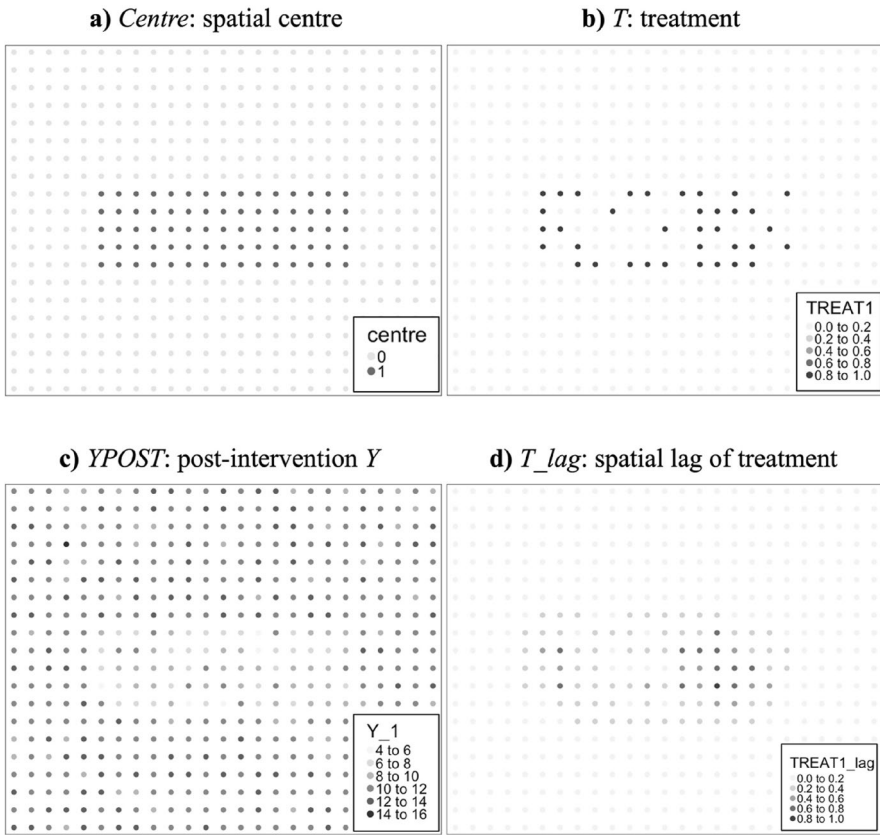


Fig. 1 a–d Spatial distributions of a selection of simulated variables

U_1 , U_2 , and U_0 are true confounders that must be controlled for in order to meet the assumption of ignorability (Matthay et al. 2020; Gelman and Hill 2006). The resulting treated observations (T) are shown in Fig. 1b.

Since we want to create a scenario where the treatment has some impact, we specify a TE variable with a mean of -2 that is highly correlated with the treatment (-0.8). Adding that to $Y(0)$ yields $Y(1)$, the true treated value of Y in every observation. Since we have created the data, in this case, we *can* overcome the fundamental problem of causal inference, but in order to mimic real-world data, we create a final Y_{POST} variable that takes the value of $Y(0)$ for untreated observations and $Y(1)$ for treated observations. Figure 2a shows a directed acyclic graph⁵ (DAG) of the causal relationships that we have generated through this simulated data—because the *Centre* influences both the treatment assignment T and the values of U_1 , U_2 , and Y_{PRE} , which all subsequently influence Y_{POST} (through $Y(0)$), U_1 , U_2 , and Y_{PRE} are true confounders producing bias that must be adjusted for in order to identify the treatment effect. Figure 2b shows the

⁵ Graphs were created using DAGitty v.3.0 in a Chrome browser.

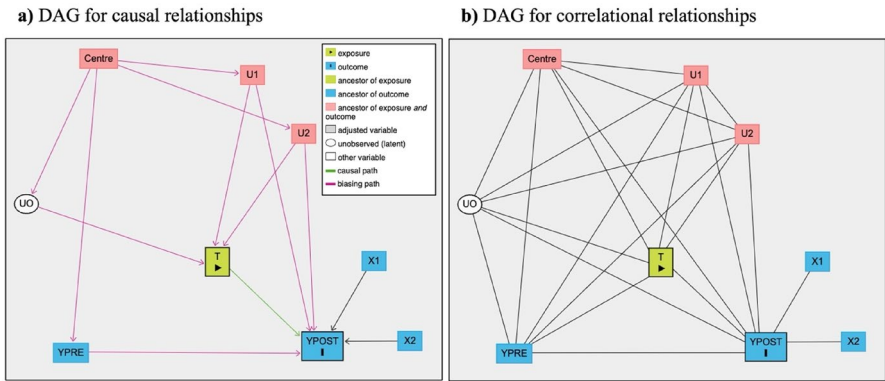


Fig. 2 a, b Directed acyclic graphs (DAG) showing causal and correlational relationships for simulated variables

Table 1 Average values for average treatment effect (ATE) and mean squared error (MSE) for simulated data across 16 different combinations of model choice and specification

Model	Baseline		SAR		SLX		Durbin	
	ATE	MSE	ATE	MSE	ATE	MSE	ATE	MSE
A Ordinary least squares (OLS)	-2.050		-2.110		-2.123		-2.128	
B Causal forest: No split	-1.992	0.026	-2.007	0.027	-2.010	0.027	-2.000	0.027
C Causal forest: Random 70/30 test/train split	-1.965	0.034	-1.967	0.033	-1.974	0.035	-1.963	0.035
D Spatial T-learner: Random forest	-2.036	0.108	-1.965	0.137	-1.983	0.146	-2.003	0.129

Values nearing statistically significant deviation from true values (based on 999 iterations each) are shown in orange. Boxes highlight the best estimates of ATE and MSE across all model-specification combinations

correlational relationships embodied by this data structure, which are of course quite tangled, as is the case in many real-world urban-geographic applications of causal inference.

Due to the spatial nature of these correlations, the influence of the *Centre* on *UO* can indirectly be controlled for by adding spatial lags of the observed variables to the model, which we do using an eight nearest-neighbour spatial weights matrix. The spatial pattern for Y_{POST} and the spatial lag of *T* are shown in Fig. 1c, d, respectively.

4 Structured model comparison using simulated data

The primary goal of this paper is to compare the performance of causal machine learning methods (Models B–D in Table 1) with traditional OLS (Model A) across four different spatial specifications: a baseline (non-spatial) model based in Eq. (1),

a spatial lag (SAR) model based in Eq. (2), a spatial lag of X (SLX) model based in Eq. (3), and a spatial Durbin model based in Eq. (4).⁶ In addition, we are interested in assessing the risk of overfitting when using the entirety of the data for training in the causal forest (Model B)—and thus preserving the underlying spatial configuration of the data as compared to the usual 70% train/test split (Model C)—and in testing the viability of the “predicted counterfactual” model (Model D).

To understand the distributional properties of this data—and be able to make comparisons based on statistical significance—we ran 999 simulations on each estimation method. This approach allows us to test the statistical significance of each estimation technique against each other. We assessed the performance of each respective estimation technique and specification through the construction of violin plots. These plots show the distribution of the given metrics of interest along with their mean and ninety-five percent confidence interval.

It is also important to note that when estimating with decision tree-based algorithms, various hyperparameters can be tuned. For the sake of simplicity, we do not tune the hyperparameters of each run for causal forest and random forest. In practice, this would make the simulations too computationally expensive as there is an infinite combination of hyperparameters that could be used. Instead, we use the default hyperparameters that are set within the given libraries we used for estimation. These default values can be found in the documentation for causal forest and random forest in the *grf* package and the *Random Forest* package in R.

For each model-specification combination, two types of assessments of accuracy are made: ATE and MSE. ATE is calculated differently for each model type: for OLS, we take the regression coefficient on the T variable. For the two causal forests, we use the ‘average_treatment_effect’ function from the *grf* package in R (version 2.0.2) on the “all” target sample using the augmented inverse-propensity weighting method from Robins and Rotnitzky (1995). For the STL, we use the predictive method described in Sect. 2.3.2 above. Since the three machine learning models produce heterogeneous estimates of CATE—which we have also generated in the simulated dataset—we can also find the accuracy of those individual values by calculating the MSE between known CATE and predicted CATE for each observation. The average value for ATE and MSE after 3,966 iterations of each model-specification combination can be found in Table 1. Violin plots showing the full distribution of results for ATE and MSE for the preferred Durbin model (discussed in more detail below) are also shown in Fig. 3a, b (respectively). The Durbin plots are broadly representative of the results of each of the specifications.⁷

Four primary results stand out. First, the three machine learning models *nearly* significantly outperform OLS in terms of identifying the true ATE of -2 for the SAR, SLX, and Durbin models; as Fig. 3a shows, the true estimate falls just within the upper 95% confidence interval for the OLS estimate, whilst each of the other

⁶ Complete specifications of these models using the simulated data variables can be found in Appendix 4. It is important to note that the OLS models in this paper are *not* estimated using standard spatial econometric techniques, but—to maintain consistency with the machine learning models—simply employ the same spatially lagged variables.

⁷ Violin plots for all specifications tested in the paper available by request.

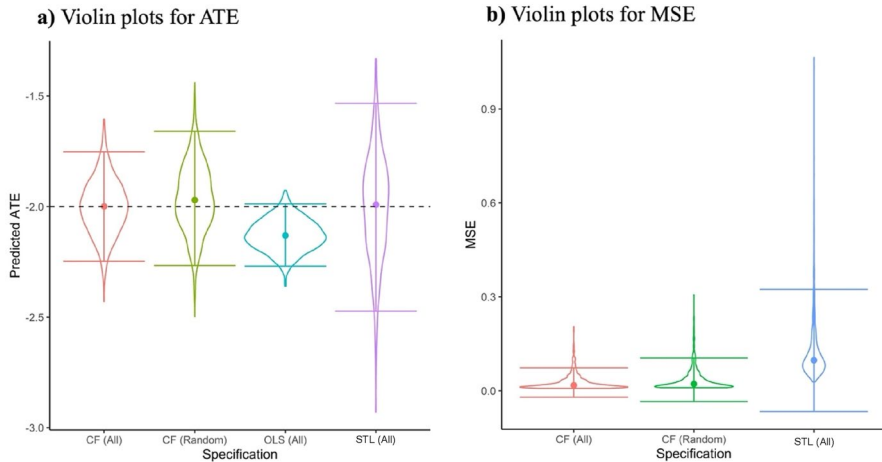


Fig. 3 a, b Violin plots showing distributional results (based on 999 iterations each) for average treatment effect (ATE) and mean squared error (MSE) for the causal forest (CF), ordinary least squares (OLS), and spatial T-learner (STL) models using the spatial Durbin specification

three averages very nearly pinpoint the true value. However, the three models’ estimates are not significantly different from one another, although the STL model’s variance is somewhat larger than the two causal forest models. Second, these results also hold for MSE of the CATE, except that in this case, as Fig. 3b shows, the STL models are slightly worse than the causal forests, nearing (but not quite attaining) statistical significance.

Third, for the machine learning models, the SAR specification tends to marginally outperform the baseline across the board, which makes sense given the spatial construction of the dataset. The SLX or Durbin models (or both) also offer some small additional performance over the SAR in most cases. For example, the Durbin performs better than the SAR for Models B and D (in terms of both ATE and MSE), whilst the SLX performs better than the SAR for Model 3 in terms of ATE only. However, it is important to note that none of these differences are statistically significant. Finally, along those lines, the Durbin specification for Model B displays the most accurate average estimate of ATE and the second-lowest average MSE, although the difference is not statistically significant. This is chosen as the preferred model for this reason.

4.1 Application: assessing the impact of light rail construction on on-road CO₂ emissions in phoenix

This paper’s application is motivated primarily by a growing interest in the literature on the impact of transportation patterns, infrastructure, and urban form on reducing greenhouse gas (GHG) emissions that are harmful to the global climate (IPCC 2018). Existing empirical research has primarily focused either on the impacts of transit construction on reductions in Vehicle Miles Travelled (VMT) (Newman and

Kenworthy 1999; Holtzclaw 2000; Bailey et al. 2008; Ewing and Hamidi 2014; Boarnet et al. 2020) or the impact of more aggregate features of urban spatial structure (e.g. compactness, density) on reductions in greenhouse gas (GHG) emissions (Glaeser and Kahn 2010; Jones and Kammen 2013; Lee and Lee 2014; Gately et al. 2015; Gudipudi et al. 2016; Mitchell et al. 2017). However, relatively little existing work looks at the direct empirical impact of transit construction on CO₂ emissions using methods of causal inference or at fine-grained spatial scales of analysis.

To do this, we employ the Database of Road Transportation Emissions (DARTE), which contains block group-level estimates of on-road CO₂ emissions for the entire continental US for every year from 1980 to 2017 (Gudipudi et al. 2016). Whilst this dataset obviously has immense utility for several potential applications of spatial and causal inference, in this paper, we are interested in presenting a relatively straightforward analysis. Thus, we have focused on one region in particular, Phoenix, AZ, which finished construction of its Valley Metro light rail system in 2009. Phoenix offers an ideal case from the perspective of the DARTE time series, with sufficient data on the outcome variable both pre- and post-construction. Also, by using Phoenix, we are able to directly import study design considerations from previous quasi-experimental work in the region looking at the impact of transit construction on new business creation (Credit 2018). Phoenix is also an interesting test case from a scientific standpoint—it is one of the largest and fastest-growing regions in the US (US Census Bureau 2022), and at the same time has one of the most auto-oriented, “sprawling” urban development patterns in the country. Given this context, finding a measurable impact of transit construction on reducing on-road CO₂ emissions in Phoenix (even if small) would certainly be interesting for urban planners, researchers, and policy-makers.

4.2 Treatment and control groups, variables, and model specification

To test the impact of light rail construction on CO₂ emissions at the block group scale in Phoenix, we first must delineate suitable “treatment” and “control” areas. Whilst distances from 1/4 mile to one mile have been put forward in previous research as boundaries of economic development impact around transit stations (Calthorpe 1993; Zhao et al. 2003; Guerra et al. 2011; Mohammad et al. 2013), for this application, we have chosen the 1 mile buffer used in a previous analysis of Phoenix’s light rail, which roughly corresponds to a 20-min walk (Credit 2018). For the control area, again following an approach offered in Credit (2018), we have chosen a 2.5-mile buffer around major highway interchanges in Phoenix (excluding areas already within the treatment area), which roughly corresponds to a 10-min drive at 15 miles per hour. This choice of control area is motivated primarily by theoretical considerations, i.e. in Phoenix, areas near highway interchanges are mostly likely to represent a counterfactual condition to areas around transit stations; both are influenced by the development and use of transportation infrastructure, but

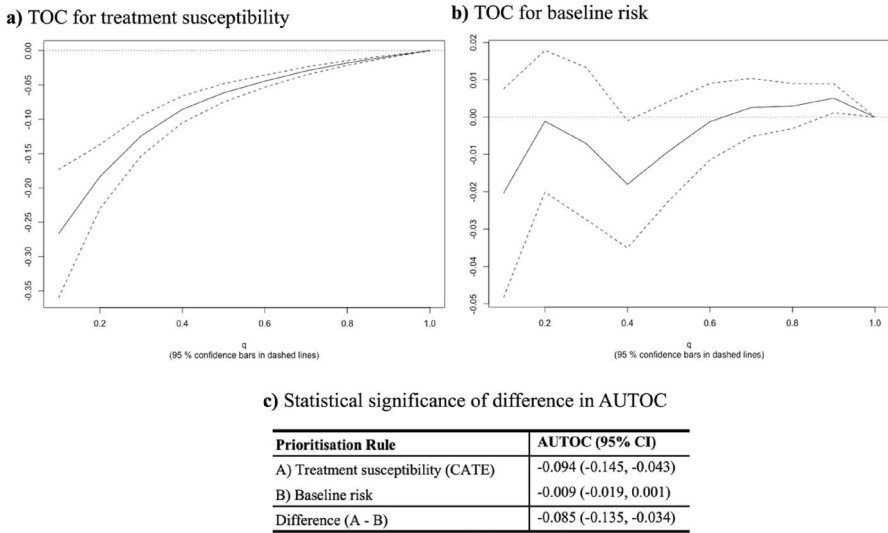


Fig. 4 a–c Rank-weighted average treatment effects (RATE) and targeting operator characteristic (TOC) curves for **a** priority based on treatment susceptibility (conditional average treatment effects), **b** priority based on baseline risk, and **c** a comparison of the statistical significance of the area under the TOC curve (AUTOOC) for both rules. Negative values indicate larger benefit for *reducing* on-road CO₂ per capita. For more details, see Yadlowsky et al. (2021) and Tibshirani et al. (2022b)

the highway (automobile) infrastructure serves the “business as usual” mode.⁸ Both buffers are calculated using street network distance, and the resulting treatment and control areas are determined by the block groups which intersect each respective buffer. Figure 4 shows the full set of block groups chosen for analysis, with the treatment buffer overlaid on the “treated” block groups. The remaining block groups belong to the control group.

Following the results of Sect. 4 above, we have chosen to use a causal forest spatial Durbin model specification with no train/test split to analyse the causal effect of transit construction on CO₂ emissions. Given the lack of useful covariates in non-decennial Census years through the introduction of the American Community Survey (ACS) in 2005, we employ a two time-period strategy with 2000 as the “pre-intervention” year and 2017 as the “post-intervention” year. To deal with the fact that the spatial boundaries of block groups changed between 2000 and 2010, counts of demographic variables at the centroids of 2000 block groups are summed within the 2010 block group boundaries (which are used as a stable unit of reference for all years of DARTE data) and then re-divided by new totals (or area, in the case of population density). Block length was calculated by block group in 2000 and averaged within the 2010 block group boundaries. Other variables were available only at the Census tract scale in 2000, so these values were applied evenly to the 2017 block group centroids within 2000 Census tract boundaries.

⁸ For additional detail on the specific methodology and justification for this choice of treatment and control areas, please see Credit (2018, pp. 2844–2845).

Appendix 3 provides a description for the covariates used in the analysis, each of which falls generally into one of three broad, theoretically relevant categories for possible confounders: socio-demographics, transportation, and built environment (based on the so-called “D” variables) (Ewing and Cervero 2010). On-road CO₂ per capita (i.e. emissions divided by population) was chosen as the outcome variable since emissions are primarily generated by people driving vehicles, not by the areal size of a block group (as in a density). The natural log of this variable was taken due to right-skewness. Spatial lags of each variable were calculated using a six nearest-neighbour spatial weights matrix, according to the Durbin specification laid out in Appendix 4.

5 Results

To better understand the impact of light rail construction on CO₂ emissions and the spatial nature of heterogeneities in this impact, we are primarily interested in three items: (1) the estimated ATE from the causal forest model, (2) the heterogeneous relationships between the CATE and the most important variables, and (3) the general spatial patterns of the CATE. To calculate the ATE, as above, we use the ‘average_treatment_effect’ function on the entire target sample and find a value of -0.084 , with an estimated standard error of 0.034 . Since the outcome variable is logged, we take the exponentiated value of this variable -1 to find the aggregate % decrease in on-road CO₂ per capital attributable to light rail over the impact period (2009–2017), which is 8.02%.

In addition, one of the most useful features of models that produce HTE is that we can learn something about the “susceptibility” of certain observations to treatment. This can be done formally by constructing a rank-weighted average treatment effect (RATE) measure based on a given treatment prioritisation rule and evaluating its significance using a targeting operator characteristic (TOC) curve (Yadlowsky et al. 2021). Figure 4 shows the TOC graphs and results for two prioritisation rules: (1) priority based on treatment susceptibility, i.e. conditional average treatment effect (CATE), and (2) priority based on baseline risk, i.e. composition of features that predict high levels of 2017 on-road CO₂ per capita. In this case, for the first rule, calculating the area under the TOC curve (AUTOC) suggests that HTE are significantly present and that this is a useful prioritisation rule (Tibshirani et al. 2022b).

To get a sense for the relationships driving the CATE, we can examine the importance (proportion of times each variable was split on at each depth in the causal forest) of each of the covariates used in the model, shown in Appendix 5 ranked from highest to lowest. Interestingly, the four most important variables all involve transportation characteristics of block groups, and nine of the top eleven are either transportation or built environment-related.

To understand the direction of these relationships—and whether they vary between treated and untreated observations—we need to look at the correlations between the predicted TE and these variables. Figure 5 shows the linear association between predicted CATE and all six of the transportation variables, as well as two of the most important built environment variables, population density and average

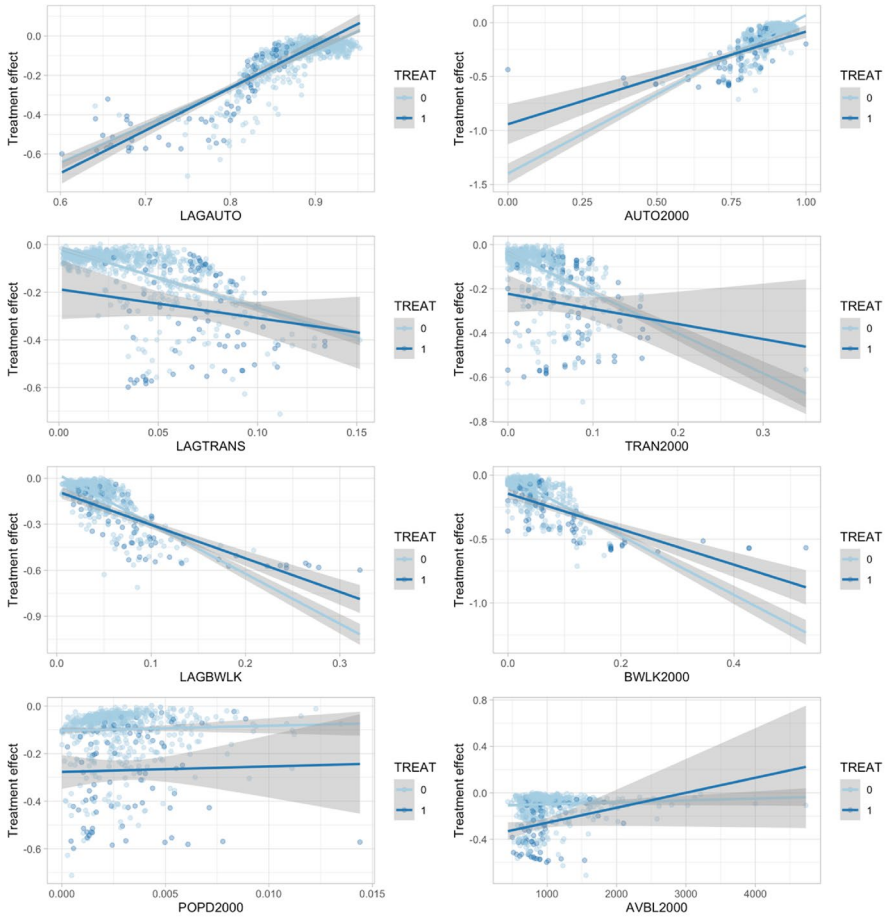


Fig. 5 Scatterplots showing linear relationships between selected covariates and predicted conditional average treatment effects (CATE) from causal forest model with no test/train split using spatial Durbin specification

block length.⁹ In general, we can see that the spatially lagged variables here tend to follow the associations in the raw pre-treatment variables: higher proportions of auto commuting, longer average block length, and (surprisingly) higher population densities are all associated with smaller treatment effects (smaller reductions in on-road CO₂). On the other hand, higher proportions of transit, cycling, and walking

⁹ Appendix 6 shows the full table of regression results between the doubly robust method for estimating the CATE from Cui et al. (2023) and each of the covariates of interest (with coefficient values and estimates of standard error) from the ‘best_linear_projection’ function in the *grf* package. Whilst we prefer the graphical representation of results in Fig. 5 because it allows the reader to see the context of each individual relationship – and differences in the relationship between treated and untreated units, which may be obscured in a single regression coefficient – without the effects of multicollinearity, we understand that the regression table may be of interest to some readers.

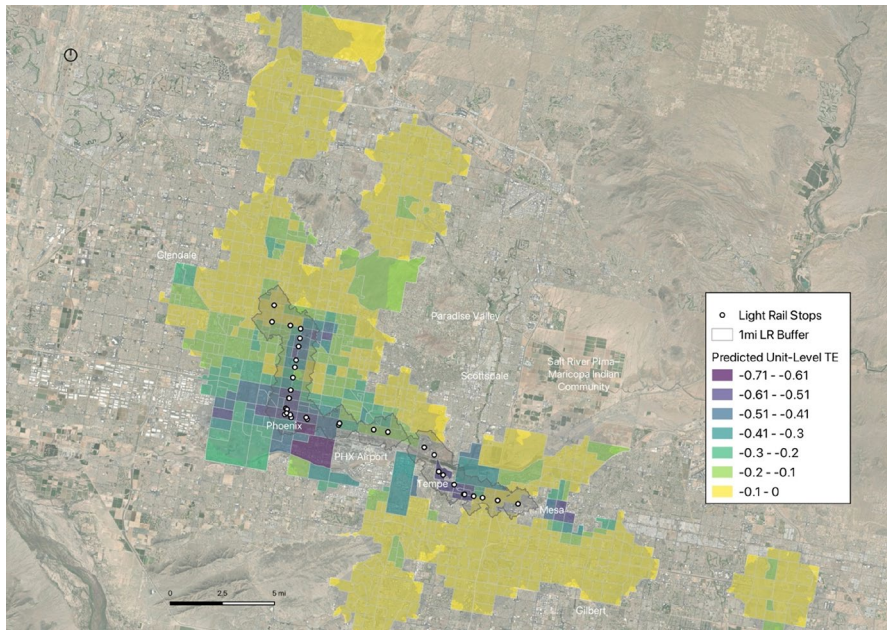


Fig. 6 Map of predicted conditional average treatment effects (CATE) from causal forest model with no test/train split using spatial Durbin specification

commuting are associated with larger treatment effects (larger reductions in on-road CO₂). In general, this confirms our fundamental notions of transit ridership, i.e. that neighbourhoods with higher proportions of non-auto travel are likely to take advantage of new light rail system construction (Werner et al. 2017), but it also reveals important empirical information for planners and policy-makers interested in lessons for future transit line planning.

The spatial pattern of CATE, shown in Fig. 6, generally confirms the relationships shown in Fig. 5 in a spatial context. More walkable areas in the city centres of Phoenix (along N Central Ave.) and Tempe (Mill Ave. and Arizona State University) that are within one mile of stations appear to be most susceptible to the benefits of light rail construction, as we would expect given the relationships in Fig. 5. In general, these results suggest that the planned alignment for the light rail was well-suited to benefit from transit construction. Given the significant results of the CATE RATE, prioritising treatment based on CATE (as estimated from the causal forest Durbin model) appears to be an effective approach for maximising the benefits of light rail construction. These results also speak to areas outside of the initial alignment that could be targeted for future extension of the line. In fact, two of the areas outside of the original “treatment” area with the highest CATE did see small extensions of the system within the time period of the analysis: the PHX Sky Train people mover opened in 2013, providing direct service to the airport from the 44th Street/Washington, which shows up on the map just to the west of Phoenix Sky Harbour Airport. In addition, the Central Mesa Extension to downtown Mesa was completed

in 2015; this area likewise shows up with large CATE or “susceptibility” to the transit “treatment”. Similarly, more auto-oriented, suburban portions of Glendale, south Tempe, and Mesa appear to be least susceptible to treatment, as might be expected given their land use/development patterns.

In general, the spatial Durbin causal forest model appears to be well-suited to the task of identifying CATE in application to an analysis of the effect of light rail construction on reducing on-road CO₂ emissions. Whilst the relationships between pre-treatment characteristics and predicted TE that we found are generally as expected, they provide useful confirmation of existing urban planning theory, namely that neighbourhoods with higher levels of non-auto commuting and more walkable built environments are best-suited for transit construction (in terms of reducing on-road emissions through high levels of ridership). This application also demonstrates how this kind of causal machine learning model could be used in a wide variety of urban planning contexts to identify and evaluate the spatial location of policy interventions.

6 Conclusions

In this paper, we have compared the performance of three causal machine learning techniques—including a newly developed “spatial” T-learner model—using simulated spatial data to better understand whether or not these methods outperform traditional OLS, the benefits and drawbacks of randomly splitting the test and training sets, and the role of “spatial” specifications in identifying treatment effects. We also applied the preferred model from the simulation analysis to real data, analysing the causal impact of light rail construction on reducing on-road CO₂ emissions in Phoenix, Arizona.

Overall, we found that (1) the three causal machine learning models consistently outperform traditional OLS at identifying ATE in this context, but (2) are not significantly different from one another, which suggests that neither splitting the test/training set nor training on a random sample of spatial data change the underlying results much. (3) Ultimately, the causal forest with no test/train split using the spatial Durbin specification, which includes lags of the dependent and independent variables, performed marginally better (on average) than all other models at identifying the true ATE and was second-best at reducing MSE of the CATE, though these differences were not statistically significant. We then applied this model to the question of assessing the causal impact of light rail (tram) construction on reducing on-road CO₂ emissions per capita and found that the (4) estimated average reduction in on-road CO₂ per capita due to light rail construction over the entire impact period is 8.02%, or about 1% per year, which is on the low end of previous estimates of VMT reduction due to transit construction (Ewing and Hamidi 2014). Finally, (5) analysis of the HTE from the model found that higher proportions of transit, cycling, and walking commuting in the pre-treatment period are associated with larger reductions in on-road CO₂.

There are of course several limitations to our current approach that should be explored and expanded on in future work. Whilst our strategy for simulating data

represents a useful first step in structurally diagramming how causal relationships work in complex urban environments, a comprehensive comparison of model performance in a wider range of different scenarios, accounting for, e.g. multiple types of bias, larger datasets, additional covariates, multiple spatial “poles”, and global/local spatial spillovers, would be a useful contribution to the burgeoning literature on spatial and causal machine learning. It would also be interesting to extend the model specifications tested here to the very large difference-in-differences literature; whilst a similar approach to this papers’ using “gain” models could be implemented, recent advances to the difference-in-differences methodology regarding temporal dynamics, staggered adoption, and treatment intensity would be interesting to adapt to the causal forest and STL specifications laid out here (Roth et al. 2022). Empirical applications of the causal forest “Durbin” (or other spatial specifications used here) in a wider variety of urban-geographic contexts would also be highly useful.

Overall, we feel this paper has important implications for both future research and policy. From a scientific standpoint, this paper represents an early attempt to bridge the machine learning, causal inference, and spatial econometrics literature. Since work on both (1) spatially informed methods for causal inference (Baylis and Ham 2015; Pollmann 2020; Kolak and Anselin 2020) and (2) causal machine learning methods (Wager and Athey 2018; Athey and Wager 2019; Deines et al. 2019; Hoffman and Mast 2019; Davis and Heller 2020; Gulen et al. 2019) are still relatively new, there is a real need to combine insights from these two subfields in a systematic way and in particular to create structured approaches for dealing with spatial data in causal machine learning models. The framework that we have developed in this paper for thinking about—and simulating—the overlapping correlations in urban spatial data and processes is hopefully a useful contribution in this direction and may be helpful to future researchers interested in causal inference for spatial data.

From a policy perspective, the application of the causal forest Durbin model to analysing the impact of transit construction on CO₂ emissions provides some useful information for urban planners and policy-makers looking to undertake new transit system planning in the future. Even more broadly, this paper’s use of methods for a question related to urban policy analysis hopefully provides a useful contribution for analytically minded practitioners. In general, models that produce (spatial) HTE can be powerful tools for determining and evaluating the location of existing and future policy interventions.

However, we would like to stress here that applications of these methods to urban planning and policy-making must be carefully designed to avoid algorithmic bias. Whilst we have attempted in this paper to *begin* to disentangle a small piece of the web of overlapping (spatial) correlations at play in many common urban-geographic analyses, many of these factors are correlated with historical disadvantage, structural racism, and ongoing politics of oppression (Braveman et al. 2022). These models cannot be applied in an urban context without a conscious knowledge of that history and an understanding of the structural forces that might influence the results. Interestingly, in our application, the causal forest Durbin model did a relatively “good” job of picking up on characteristics directly related to transportation patterns rather than correlated socioeconomic or demographic features, which is encouraging; but models (and their outputs)

will ultimately reflect the biases and errors of the data, analysts, and societies that create them, and future users of these causal machine learning approaches have a grave responsibility to carefully interrogate their data, methodologies, assumptions, and biases when applying them to policy-making and planning.

7 Appendix

See Tables 2, 3, 4, 5, 6

Table 2 Full characteristics of the simulated variables

Name	Description	Avg	Std. Dev	Correlated with	<i>r</i>
Centre	Spatial centre: $x=[6, 20], y=[8, 12]$				
U1	Confounder 1	1	0.5	Centre	0.7
U2	Confounder 2	2	1	Centre	0.5
UO	Unobserved	4	2	Centre	0.8
Tf	Treatment factor	0	1	[U1, U2, UO]	0.7
T	Treatment: $Tf > 1.75$				
TE	Unit-level treatment effects	-2	0.1	<i>T</i>	-0.8
X1	Random variable 1	10	5		
X2	Random variable 2	5	1		
Y_{PRE}	Pre-intervention <i>Y</i>	9	1	Centre	-0.7
$Y(0)f$	Counterfactual <i>Y</i> factor	2	0.5	[U1, U2, UO, X1, X2, Y_{PRE}]	0.3
$Y(0)$	Counterfactual <i>Y</i> : $Y_{PRE} + Y(0)f$				
$Y(1)$	Treatment effected <i>Y</i> : $Y(0) + TE$				
<i>Y</i>	Post-intervention <i>Y</i> : $(T=0 \rightarrow Y(0)) \wedge (T=1 \rightarrow Y(1))$				
T_lag	Spatial lag of <i>T</i> (8NN)				
U1_lag	Spatial lag of U1 (8NN)				
U2_lag	Spatial lag of U2 (8NN)				
X1_lag	Spatial lag of X1 (8NN)				
X2_lag	Spatial lag of X2 (8NN)				
Y_{POST_lag}	Spatial lag of Y_{POST} (8NN)				
T_lag	Spatial lag of <i>T</i> (8NN)				

Table 3 Variables, data sources, and descriptions used in the Phoenix application

Type	Name	Description	Source	YearS
Outcome variable	LCCO2C17	Natural log of 2017 on-road CO ₂ emissions (kg) per capita; post-treatment outcome	DARTE	2017
	LGO2C00	Natural log of 2000 on-road CO ₂ emissions (kg) per capita; pre-treatment outcome	DARTE	2000
Socio-demographics	WHTP2000	White non-Hispanic population %	US Census	2000
	BLKP2000	Black non-Hispanic population %	US Census	2000
	HSP2000	Hispanic population %	US Census	2000
	U18P2000	Population % aged 0–17	US Census	2000
	O65P2000	Population % aged 65 +	US Census	2000
	PINC2000	Per capita income	US Census	2000
	AUTO2000	Commuting by automobile %	US Census	2000
Transportation	TRAN2000	Commuting by public transportation %	US Census	2000
	BWLK2000	Commuting by cycling and walking %	US Census	2000
	TT302000	% Workers commuting 30+ minutes on average	US Census	2000
Built environment	POPD2000	Population density	US Census	2000
	AGED2000	Building age diversity*	US Census	2000
	AVBL2000	Average Census block length (within block group)	US Census	2000
	TREAT	Location within 1 mile (street network distance) of station	US Census	2000
Outcome variable	LAGCO2C17	Spatial lag (6NN) of LCCO2C17	Author	2009
Socio-demographics	LAGWHTP	Spatial lag (6NN) of WHTP2000	Author	2017
	LAGBLKP	Spatial lag (6NN) of BLKP2000	Author	2000
	LAGHISPP	Spatial lag (6NN) of HSP2000	Author	2000
	LAGU18P	Spatial lag (6NN) of U18P2000	Author	2000
	LAGO65P	Spatial lag (6NN) of O65P2000	Author	2000
	LAGPINC	Spatial lag (6NN) of PINC2000	Author	2000

Table 3 (continued)

Type	Name	Description	Source	YearS
Transportation	LAGAUTO	Spatial lag (6NN) of AUTO2000	Author	2000
	LAGTRANS	Spatial lag (6NN) of TRAN2000	Author	2000
	LAGBWLK	Spatial lag (6NN) of BWLK2000	Author	2000
Built environment	LAGTT30	Spatial lag (6NN) of TT302000	Author	2000
	LAGPOPD	Spatial lag (6NN) of POPD2000	Author	2000
	LAGAGED	Spatial lag (6NN) of AGED2000	Author	2000
	LAGAVBL	Spatial lag (6NN) of AVBL2000	Author	2000
Treatment	LAGTREAT	Spatial lag (6NN) of TREAT	Author	2009

*Calculated using a Herfindahl Index (HI) based on the proportion of buildings built in each decade: pre-1939, 1940s, 1950s, 1960s, 1970s, 1980s, 1990s

Table 4 Detailed specifications for model comparisons performed in Sect. 4

Model	Specification
Baseline	$Y_{POST} = T + U1 + U2 + X1 + X2 + Y_{PRE}$
SAR	$Y_{POST} = T + U1 + U2 + X1 + X2 + Y_{PRE} + Y_{POST_lag}$
SLX	$Y_{POST} = T + U1 + U2 + X1 + X2 + Y_{PRE} + U1_lag + U2_lag + X1_lag + X2_lag + T_lag$
Durbin	$Y_{POST} = T + U1 + U2 + X1 + X2 + Y_{PRE} + U1_lag + U2_lag + X1_lag + X2_lag + T_lag + Y_{POST_lag}$

“_lag” suffix denotes a spatially lagged variable using an 8NN spatial weights matrix. “ Y_{PRE_lag} ” was tested but not included in the SLX or Durbin specifications due to its high correlation with Y_{PRE} . In the end the Durbin model without Y_{PRE_lag} performed best out of any model tested, so, to be consistent, the versions of the SLX and Durbin models without Y_{PRE_lag} were chosen for the paper. Full model results available by request

Table 5 Percentage of times each feature was split on at each depth in the Durbin causal forest (top 12 values shown)

Name	Description	Importance (%)
LAGAUTO	Spatial lag of commuting by automobile %	15.77
BWLK2000	Commuting by cycling and walking %	11.25
LAGTRANS	Spatial lag of commuting by public transportation %	10.46
LAGBWLK	Spatial lag of commuting by cycling and walking %	7.97
LAGBLKP	Spatial lag of Black non-Hispanic population %	5.67
TRAN2000	Commuting by public transportation %	4.95
AUTO2000	Commuting by automobile %	4.36
BLKP2000	Black non-Hispanic population %	3.98
LAGAVBL	Spatial lag of average Census block length (within block group)	3.33
AVBL2000	Average Census block length (within block group)	2.97
POPD2000	Population density	2.58
LAGPINC	Spatial lag of per capita income	2.38

Table 6 Regression results between the estimated CATE and covariates of interests provided by the 'best_linear_projection' function in the *grf* package

Variable	Estimate	Std. Error	t-value	Pr(> t)
(Intercept)	-0.823	1.148	-0.717	0.474
LGCO2C00	-3.55E-04	0.004	-0.087	0.931
POPD2000	4.260	5.829	0.731	0.465
WHTP2000	-0.109	0.157	-0.695	0.487
BLKP2000	-0.269	0.282	-0.956	0.339
HSPP2000	-0.030	0.186	-0.160	0.873
U18P2000	-0.049	0.113	-0.432	0.666
O65P2000	0.002	0.071	0.030	0.976
AUTO2000	0.825	0.268	3.079	0.002
TRAN2000	0.879	0.393	2.238	0.026
BWLK2000	0.015	0.543	0.027	0.979
TT302000	-0.210	0.198	-1.065	0.287
PINC2000	6.41E-07	9.14E-07	0.702	0.483
AGED2000	-0.021	0.039	-0.534	0.593
AVBL2000	-1.01E-05	3.13E-05	-0.323	0.747
LAGCO2C17	-0.010	0.005	-1.862	0.063
LAGTT30	0.210	0.221	0.949	0.343
LAGPOPD	-8.729	7.460	-1.170	0.242
LAGWHTP	0.186	0.343	0.542	0.588
LAGBLKP	0.298	0.502	0.593	0.553
LAGHISPP	0.089	0.357	0.249	0.804
LAGU18P	-0.280	0.262	-1.068	0.286
LAGO65P	-0.114	0.117	-0.975	0.330
LAGPINC	-2.56E-06	1.94E-06	-1.318	0.188
LAGAGED	-0.041	0.091	-0.455	0.649
LAGAVBL	-1.23E-05	2.80E-05	-0.438	0.661
LAGTRANS	0.247	1.313	0.188	0.851
LAGAUTO	0.312	1.025	0.305	0.761
LAGBWLK	-2.594	1.244	-2.084	0.037
LAGTREAT	0.050	0.060	0.836	0.404

Funding Open Access funding provided by the IReL Consortium.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Anselin L, Rey S (2014) *Modern spatial econometrics in practice: a guide to GeoDa, GeoDaSpace and PySAL*. GeoDa Press, Chicago
- Athey S, Imbens G (2016) Recursive partitioning for heterogeneous causal effects. *Proc Natl Acad Sci* 113(27):7353–7360
- Athey S, Wager S (2019) Estimating treatment effects with causal forests: an application. *Obs Stud* 5:36–51
- Athey S, Tibshirani J, Wager S (2019) Generalized random forests. *Ann Stat* 47(2):1148–1178
- Bailey L, Mokhtarian PL, Little A (2008) The broader connection between public transportation, energy conservation and greenhouse gas reduction. ICF International, Fairfax
- Baylis K, Ham A (2015) How important is spatial correlation in randomized controlled trials?. In 2015 AAEA & WAEA joint annual meeting, July 26–28, agricultural and applied economics association & western agricultural economics association, San Francisco
- Boarnet MG, Bostic RW, Rodnyansky S, Burinskiy E, Eisenlohr A, Jamme H-T, Santiago-Bartolomei R (2020) Do high income households reduce driving more when living near rail transit? *Transp Res Part D Transp Environ*. <https://doi.org/10.1016/j.trd.2020.102244>
- Braveman OA, Arkin E, Proctor D, Kauh T, Holm N (2022) Systemic and structural racism: definitions, examples, health damages, and approaches to dismantling. *Health Aff*. <https://doi.org/10.1377/hlthaff.2021.01394>
- Butts K (2021) Difference-in-differences estimation with spatial spillovers. <https://arxiv.org/abs/2105.03737> [econ]
- Calthorpe P (1993) *The next American metropolis: ecology, community, and the American dream*. Princeton Architectural Press, New York
- Card D, Krueger AB (1994) Minimum wages and employment: a case study of the fast-food industry in New Jersey and Pennsylvania. *Am Econ Rev* 84(4):772–793
- Credit K (2018) Transit-oriented economic development: the impact of light rail on new business starts in the Phoenix. *AZ Region Urban Stud* 55(13):2838–2862. <https://doi.org/10.1177/0042098017724119>
- Credit K (2022) Spatial models or random forest? evaluating the use of spatially explicit machine learning methods to predict employment density around new transit stations in Los Angeles. *Geogr Anal* 54(1):58–83
- Cui Y, Kosorok MR, Sverdrup E, Wager S, Zhu R (2023) Estimating heterogenous treatment effects with right-censored data via causal survival forests. <https://arxiv.org/abs/2001.09887v5> [stat.ME]
- Davis JMV, Heller SB (2020) Rethinking the benefits of youth employment programs: the heterogeneous effects of summer jobs. *Rev Econ Stat* 102(4):664–677
- Deines JM, Wang S, Lobell DB (2019) Satellites reveal a small positive yield effect from conservation tillage across the US Corn Belt. *Environ Res Lett*. <https://doi.org/10.1088/1748-9326/ab503b>
- Dorie V, Hill J, Shalit U, Scott M, Cervone D (2019) Automated versus do-it-yourself methods for causal inference: lessons learned from a data analysis competition. *Stat Sci* 34(1):43–68
- Ewing R, Cervero R (2010) Travel and the built environment: a meta-analysis. *J Am Plann Assoc* 76(3):265–294
- Ewing R, Hamidi S (2014) Longitudinal analysis of transit's land use multiplier in Portland (OR). *J Am Plann Assoc* 80(2):123–137. <https://doi.org/10.1080/01944363.2014.949506>
- Farbmacher H, Kögel H, Spindler M (2021) Heterogeneous effects of poverty on attention. *Labour Econ*. <https://doi.org/10.1016/j.labeco.2021.102028>
- Gately CK, Hutyra LR, Wing IS (2015) Cities, traffic, and CO₂: a multidecadal assessment of trends, drivers, and scaling relationships. *Proc Natl Acad Sci* 112(16):4999–5004
- Gelman A, Hill J (2006) *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, Cambridge
- Glaeser EL, Kahn ME (2010) The greenness of cities: carbon dioxide emissions and urban development. *J Urban Econ* 67:404–418
- Gudipudi R, Fluschnik T, Ros AGC, Walther C, Kropp JP (2016) City density and CO₂ efficiency. *Energy Policy* 91:352–361
- Guerra, E., Cervero, R., & Tischler, D. (2011). The half-mile circle: Does it best represent transit station catchments?. Working Paper: Institute of Transportation Studies, University of California, Berkeley.

- Retrieved from: <http://www.its.berkeley.edu/sites/default/files/publications/UCB/2011/VWP/UCB-ITS-VWP-2011-5.pdf>
- Gulen H, Jens CE, Page TB (2019) An application of causal forest in corporate finance: How does financing affect investment?. Working Paper, available at SSRN: <https://ssrn.com/abstract=3583685>
- Hagenauer J, Omrani H, Helbich M (2019) Assessing the performance of 38 machine learning models: the case of land consumption rates in Bavaria, Germany. *Int J Geogr Inf Sci* 33:1399–1419
- Hawkins DM (2004) The problem of overfitting. *J Chem Inf Comput Sci* 44(1):1–12
- Hill JF (2011) Bayesian nonparametric modeling for causal inference. *J Comput Graph Stat* 20(1):217–240. <https://doi.org/10.1198/jcgs.2010.08162>
- Hoffman I, Mast E (2019) Heterogeneity in the effect of federal spending on local crime: evidence from causal forests. *Reg Sci Urban Econ*. <https://doi.org/10.1016/j.regsciurbeco.2019.103463>
- Holtzclaw J (2000) Does a mile in a car equal a mile on a train? Exploring public transit's effectiveness in reducing driving. Retrieved from: <http://sierraclub.typepad.com/files/transitleveragereport-forholtzclaw.pdf>. Accessed 10 Aug 2021
- Huber M, Steinmayr A (2019) A framework for separating individual-level treatment effects from spillover effects. *J Bus Econ Stat*. <https://doi.org/10.1080/07350015.2019.1668795>
- Huntington-Klein N, Arenas A, Beam E, Bertoni M, Bloem JR, Burli P, Chen N, Greico P, Ekpe G, Pugatch T, Saavedra M, Stopnitzky Y (2020). The influence of hidden researcher decisions in applied microeconomics. IZA institute of labor economics. Discussion Paper No. 13233. Retrieved from: <https://docs.iza.org/dp13233.pdf>
- Imai K, Ratkovic M (2013) Estimating treatment effect heterogeneity in randomized program evaluation. *Ann Appl Stat* 7(1):443–470
- IPCC (2018) Global warming of 1.5 °C: An IPCC Special Report on the impacts of global warming of 1.5 °C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty. In: Masson-Delmotte V, Zhai P, Pörtner HO, Roberts D, Skea J, Shukla PR, Pirani A, Moufouma-Okia W, Péan C, Pidcock R, Connors S, Matthews JBR, Chen Y, Zhou X, Gomis MI, Lonnoy E, Maycock T, Tignor M, Waterfield T (eds) International Panel on Climate Change
- Jones C, Kammen DM (2013) Spatial distribution of U.S. household carbon footprints reveals suburbanization undermines greenhouse gas benefits of urban population density. *Environ Sci Technol* 48:895–902
- Knaus MC, Lechner M, Strittmatter A (2018) Machine learning estimation of heterogeneous causal effects: empirical Monte Carlo evidence. <https://arxiv.org/abs/1810.13237>
- Kolak M, Anselin L (2020) A spatial perspective on the econometrics of program evaluation. *Int Reg Sci Rev* 43(1–2):128–153
- Kosfeld R, Mitze T, Rode J, Wälde K (2021) The Covid-19 containment effects of public health measures: a spatial difference-in-differences approach. *J Reg Sci* 61(4):799–825
- Krzywinski M, Altman N (2017) Classification and regression trees. *Nat Methods* 14:757–758
- Kuhn M, Johnson K (2013) Over-fitting and model tuning. In: Applied predictive modeling. Springer, New York. 1007/978-1-4614-6849-3_4
- Künzel SR, Sekhon JS, Bickel PJ, Yu B (2019) Metalearners for estimating heterogeneous treatment effects using machine learning. *Proc Nat Acad Sci (PNAS)* 116(10):4156–4165
- Lagakos SW (2006) The challenge of subgroup analyses-reporting without distorting. *N Engl J Med* 354:1667–1669
- Lee S, Lee B (2014) The influence of urban form on GHG emissions in the U.S. household sector. *Energy Policy* 68:534–549
- LeSage JP, Pace RK (2009) Introduction to spatial econometrics. CRC Press, New York
- Matthay EC, Hagan E, Gottlieb LM, Tan ML, Vlahov D, Adler NE, Glymour MM (2020) Alternative causal inference methods in population health research: evaluating tradeoffs and triangulating evidence. *SSM Popul Health*. <https://doi.org/10.1016/j.ssmph.2019.100526>
- Mitchell LE, Lin JC, Bowling DR, Pataki DE, Strong C, Schauer AJ, Bares R, Bush SE, Stephens BB, Mendoza D, Mallia D, Holland L, Gurney KR, Ehleringer JR (2017) Long-term urban carbon dioxide observations reveal spatial and temporal dynamics related to urban characteristics and growth. *Proc Natl Acad Sci* 115(12):2912–2917
- Mohammad S, Graham D, Melo P, Anderson RJ (2013) A meta-analysis of the impact of rail projects on land and property values. *Transp Res Part A* 50:158–170

- Newman P, Kenworthy J (1999) Sustainability and cities: overcoming automobile dependence. Island Press, Washington, D.C.
- Nikparvar B, Thill J-C (2021) Machine learning of spatial data. *Int J Geo Inf*. <https://doi.org/10.3390/ijgi10090600>
- Parker AB, Naylor CD (2000) Subgroups, treatment effects, and baseline risks: some lessons from major cardiovascular trials. *Am Heart J* 139:952–961
- Pollmann M (2020) Causal inference for spatial treatments. [arXiv: 2011.00373v1](https://arxiv.org/abs/2011.00373v1)
- Robins JM, Rotnitzky A (1995) Semiparametric efficiency in multivariate regression models with missing data. *J Am Stat Assoc* 90(429):122–129
- Roth J, Sant'Anna PHGC, Bilinski AI, Poe J (2022) What's trending in difference-in-differences? A synthesis of the recent economics literature. Retrieved from: https://www.jonathandroth.com/assets/files/DiD_Review_Paper.pdf
- Rubin D (1978) Bayesian inference for causal effects: the role of randomization. *Ann Stat* 6:34–58
- Rubin D (1980) Randomization analysis of experimental data: the fisher randomization test comment. *J Am Stat Assoc* 75(371):591–593
- Rubin DB (2001) Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Serv Outcomes Res Method* 2:169–188. <https://doi.org/10.1023/A:1020363010465>
- Rubin D (2005) Causal inference using potential outcomes: design, modeling, decisions. *J Am Stat Assoc* 100(469):322–331
- Shalit, U., Johansson, F. D., & Sontag, D. (2017). Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning*: 3076–3085.
- Strittmatter A (2019) What is the value added by using causal machine learning methods in a welfare experiment evaluation? <https://arxiv.org/abs/1808.00943v2>
- Tibshirani J, Athey S, Sverdrup E, Wager S (2022a) The GRF algorithm. Retrieved from: <https://grf-labs.github.io/grf/REFERENCE.html#causal-forests>
- Tibshirani J, Athey S, Sverdrup E, Wager S (2022b) Estimate a rank-weighted average treatment effect (RATE). Retrieved from: https://grf-labs.github.io/grf/reference/rank_average_treatment_effect.html
- Tobler W (1970) A computer movie simulating urban growth in the Detroit region. *Econ Geogr* 46:234–240
- US Census Bureau (2022) Fastest-growing cities are still in the west and south. Release Number CB22–90. Retrieved from: <https://www.census.gov/newsroom/press-releases/2022/fastest-growing-cities-population-estimates.html#:~:text=Top%20Places%20for%20Population%20Growth,in%20less%20than%20seven%20years.>
- VanderWeele TJ, Hernán MA (2013) Causal inference under multiple versions of treatment. *J Causal Inference* 1(1):1–20. <https://doi.org/10.1515/jci-2012-0002>
- Wager S, Athey S (2018) Estimation and inference of heterogeneous treatment effects using random forests. *J Am Stat Assoc* 113(523):1228–1242
- Werner CM, Brown BB, Tribby CP, Tharp D, Flick K, Miller HJ, Smith KR, Jensen W (2017) Evaluating the attractiveness of a new light rail extension: testing simple change and displacement change hypotheses. *Transp Policy* 45:15–23
- Yadlowsky S, Fleming S, Shah N, Brunskill E, Wager S (2021) Evaluating treatment prioritization rules via rank-weighted average treatment effects. <https://arxiv.org/abs/2111.07966> [stat.ME]
- Yoshida T, Seya H (2021) Spatial prediction of apartment rent using regression-based and machine learning-based approaches with a large dataset. <https://arxiv.org/abs/2107.12539>
- Zhao F, Chow LF, Li MT, Ubaka I, Gan A (2003) Forecasting transit walk accessibility: regression model alternative to buffer. *Transp Res Rec* 1835:34–41

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.