Online Appendix:
Causal Inference for Spatial Treatments

Michael Pollmann

February 14, 2023

1 Examples of papers studying spatial treatments

Table OA1 lists examples of papers studying spatial treatments. The outcomes or outcomes units mentioned in the table are either directly studied in each paper or are closely related to the question studied. The list is meant to help the reader map empirical objects into the framework of this paper and to illustrate the breadth of topics involving spatial treatments. Not all of these papers had precise location data on treatments and/or outcome units, but such data could in principle be collected in all instances. Dell and Olken (2020) is the only example on this list explicitly considering counterfactual treatment locations.[1] The theory in the present paper derives standard errors complementing the p-values of randomization tests of the sharp null reported in the original paper.

Table OA1: Examples of papers studying spatial treatments, and outcomes or outcomes units that are either directly studied in each paper or are closely related to the question studied.

paper	spatial treatment	outcome / outcome units
Aliprantis and Hartley (2015)	public housing demolition	crime in local neighborhoods
Athey et al. (2018)	restaurant opening	utility of consumers
Buchmueller et al. (2006)	hospital closure	mortality of residents
Cohen and Dupas (2010)	subsidized bed nets sold at hospitals	adoption of bed nets in local communities
Currie et al. (2015)	toxic plant opening and closing	house prices, infant health
Dell and Olken (2020)	site of historic sugar mill	economic development of nearby towns
Diamond and McQuade (2019)	low income housing projects	house prices
Di Tella and Schargrodsky (2004)	police presence in city blocks	number of car thefts
Duflo (2001)	school construction	educational attainment in nearby villages
Ellickson and Grieco (2013)	Wal-Mart entry	entry, exit of competitors
Feyrer et al. (2017)	fracking site	income of local residents
Greenstone and Moretti (2003)	large manufacturing plant entry	property values, labor earnings of residents
Greenstone et al. (2010)	large manufacturing plant entry	TFP of other plants
Jia (2008)	Wal-Mart entry	profit/exit of small discount stores
Keiser and Shapiro (2019)	wastewater treatment plants	commercial & recreational value of rivers
Linden and Rockoff (2008)	sex offenders moving in	house prices
Miguel and Kremer (2004)	deworming administered at schools	worm prevalence in local population
Oates (1969)	(spending on) local public goods	property values
Seim (2006)	video store entry	effect on local competitors
Siegfried and Zimbalist (2000)	sport stadiums	local businesses, property values
Stock (1991)	toxic waste cleanup	property values

[1] There are other empirical studies considering counterfactual treatment locations, but to the best of my knowledge none include statistical theory allowing design-based inference.

2 Setup and training of neural networks for finding counterfactual locations

The implementation of estimation based on the unconfoundedness assumption proposed here relies on estimates of the probability of treatment at any location in a region conditional on all the features of the region. The probability of treatment across space resembles the spatial distribution of treatment locations $\mathcal{S}_j \sim G(Z_j)$, where Z_j are the characteristics of region j, potentially relative locations of all individuals in the region as well as moments of their covariates. One could then use the estimated $\hat{G}$ to inform the treatment probabilities at each point in the region as inputs in the estimators proposed in this paper.

In practice, it is typically sufficient to find a finite number of candidate treatment locations that offer a plausible counterfactual to the realized treatment locations. With a continuous distribution across space, a simple approximation of the assignment process such as independent assignment can lead to unrealistic assignments that are considered in computing standard errors. More complex assignment processes for continuous distributions may instead be analytically intractable. In addition, computationally, it is often impractical to use a continuous distribution G because the weight of individual i when estimating effects at distance d would depend on the integral of the noisy $\hat{G}$ along a ring with radius d around her location, r_i, for each of the typically many individuals $i \in \mathbb{I}$. Instead, I recommend finding a finite number of candidate locations. The average across these finitely many candidate locations approximates the strategy based on the complete distribution G, setting $\hat{G}$ to exactly zero for many of the implausible locations.

I propose taking draws $\mathcal{S}_j \sim G(Z_j)$ to obtain candidate treatment locations, where $G(Z_j)$ is estimated implicitly. Perhaps surprisingly, recent machine learning methods achieve good results at this task, despite the difficulty of estimating G itself. Specifically, I recommend a formulation similar to generative adversarial networks (Goodfellow et al., 2014); see Liang (2018) and Singh et al. (2018) on the relationship between generative adversarial networks and density estimation. Most closely related to this paper, Athey et al. (2019) use generative adversarial networks to draw artificial observations from the distribution that generated the (real) sample, for use in Monte Carlo simulations.

Generative adversarial methods for drawing $\mathcal{S}_j \sim G(Z_j)$ are based on iteration between two steps. First, a generator generates draws $\tilde{\mathcal{S}}_j \sim \tilde{G}(Z_j)$, where $\tilde{G}$ is an implicit estimate of the density maintained by the generator in the current iteration. Second, a discriminator receives as input either counterfactual locations proposed by the generator, $\tilde{\mathcal{S}}_j \mid Z_j$, or real treatment locations, $\mathcal{S}_j \mid Z_j$, and guesses whether its input is real. Both the generator and the discriminator are highly flexible models (typically neural networks) designed for their given tasks. The discriminator is trained by taking (stochastic) gradient descent steps in the direction that improves discrimination between real and counterfactual locations. The generator is trained by taking (stochastic) gradient descent steps in the direction that leads to fooling the discriminator into classifying counterfactual locations as real.

Effectively, the output of such models is a set of counterfactual candidate treatment locations $\tilde{\mathcal{S}}_j \mid Z_j$ that are indistinguishable (to the discriminator) from real treatment locations $\mathcal{S}_j \mid Z_j$. With a sufficiently flexible discriminator, the process is similar to matching.[2] If a proposed candidate location $\tilde{\mathcal{S}}_j$ is noticeably different from all real treatment locations $\mathcal{S}$, a flexible discriminator will learn to reject $\tilde{\mathcal{S}}_j$. In contrast, synthetic control-type methods (cf. Abadie et al., 2010) would average multiple candidate locations, for instance, $\tilde{\mathcal{S}}_a$ and $\tilde{\mathcal{S}}_b$, to create a synthetic counterfactual for a real treatment location $\mathcal{S}_j$. If $\tilde{\mathcal{S}}_a$ and $\tilde{\mathcal{S}}_b$ individually differ from all real treatment locations $\mathcal{S}$, the discriminator will reject them despite their average resembling $\mathcal{S}_j$.

Intuitively, the goal is to find "false positives:" Occasions when the discriminator fails to reject a counterfactual location suggested by the generator. Discriminator networks do not necessarily make binary predictions but may give a continuous activation score that indicates how likely a location is to be real. In practice, I recommend matching on the activation score, rather than taking all locations with high activation scores because some *real* treatment locations may have low activation scores. Matching on the activation score helps find suitable counterfactual locations resembling each real location. Such locations are likely to be decent matches for the real treatment locations because they must share features of realized locations to achieve these comparable activation scores.

I discuss how to tune generic machine learning methods to find suitable candidate treatment locations in

[2]Standard matching methods, however, are unlikely to perform well due to high dimensional covariates that describe spatial data, such as relative spatial locations between many individuals as well as their characteristics.

social science applications. I recommend four high-level implementation choices in adapting these methods. First, the discretization of geographic space into a fine grid for tractability. Second, *convolutional* neural networks capture the idea that spatial neighborhoods matter in a parsimonious way. Third, incorporating the adversarial task of the discriminator into a classification task for the generator substantially simplifies training. Fourth, data augmentation (rotation, mirroring, shifting) for settings where absolute locations and orientation are irrelevant.

Discretization To tractably summarize the relative spatial locations of individuals and treatment locations, I recommend discretizing geographic space into a fine grid. Discretization provides an approximation that is particularly tractable for the convolutional neural networks recommended below. In principle, future improvements to, for instance, Capsule Neural Networks (Hinton et al., 2011) or other novel methods, may replace convolutional neural networks as the preferred architecture and eliminate the need for discretization.

For each grid cell, one can include a count of individuals with residence in the cell, potentially separately for individuals with different values of covariates, as well as average covariate values of the individuals in the cell or other moments of their covariates. Based on the architecture of convolutional neural networks, suggested below, it is typically not necessary to also pre-compute covariates describing the neighborhood of each cell. The convolutional neural network can compute such neighborhood averages if they help predict the outcome (here, whether a location is likely to be treated). If the grid is very fine, discretization retains almost all meaningful information about relative locations. For instance, in the application of this paper, each grid cell has size 0.025mi $\times$ 0.025mi (approximately 40m $\times$ 40m). The discretized grid creates a three-dimensional array: The first two dimensions determine spatial location, and the third dimension enumerates the different covariates that are summarized. Rather than taking the spatial dimensions to be entire regions, I recommend using smaller (square) areas within a region such that the probability of treatment in the approximate center of the area is plausibly only affected by individuals and covariates within the area.

Convolutional neural networks Convolutional neural networks have been particularly successful at image recognition (Krizhevsky et al., 2012). In image recognition, the input is a 3D array: a 2D grid of pixels, with a third dimension given by multiple RGB color channels. For spatial treatments, the input also is a 3D array: the 2D spatial grid, with a third dimension given by the covariates as described above.

Convolutional steps in neural networks generally retain the shape of the 2D grid, but the value of each neuron is a function of the covariates (or neurons) of the previous step not just at the same grid cell, but also the covariates (or neurons) at neighboring grid cells. Figure OA1 illustrates this aspect of the convolution operation. Importantly, convolutional layers average the neighborhoods of grid cells at any point in the grid with the same weights. Reusing parameters across points in space makes convolutional layers substantially more parsimonious than fully connected layers, and allows the neural network to capture neighborhood patterns appearing in different parts of a region in a unified way.

In particular, I recommend using at least two convolutions with reasonably large spatial reach. Consider the application in this paper, where grocery stores are spatial treatments and restaurants are outcome units with foot traffic as the outcome variable. The first convolution allows each grid cell to see the covariates of grid cells around it. In the application of this paper, the output of the first convolution for a particular grid cell may be: "There are 3 grocery stores nearby, 4 competing restaurants very close, and 10 restaurants within walking distance." The second convolution then uses the information on such neighborhoods to determine whether treatment is likely in a grid cell: "If there are many grid cells nearby (in all directions) containing restaurants or grocery stores facing much competition, this location is probably in the center of a shopping area and reasonably likely to contain another grocery store." Intuitively, the first convolution may measure what is important to the restaurants, while the second convolution translates how that is important for the treatment location choice.

Adversarial Classification Generative adversarial networks (Goodfellow et al., 2014) are oftentimes difficult to train despite recent advances such as networks with Wasserstein-type criterion function (Arjovsky and Bottou, 2017; Arjovsky et al., 2017). The difficulty arises because the training of the generator and discriminator networks needs to be sufficiently balanced such that both improve. For instance, if the discriminator early on becomes (close to) perfect at discriminating between the proposals of the generator

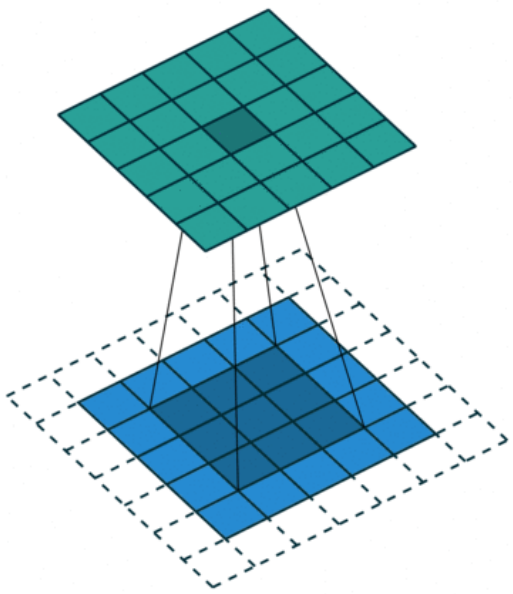


Figure OA1: Convolutions in a neural network allow the prediction of a candidate location in a grid cell to depend on the characteristics of neighboring grid cells (up to a user-specified distance). These models remain parsimonious by requiring the *same* "neighborhood scan" to be performed for each grid cell.

and the real treatment locations, the gradient for the generator is relatively flat (little improvement in any direction) and hence the generator fails to improve. Similarly, if the discriminator is insufficiently flexible, even poor proposals by the generator may pass, such that the false positives are not necessarily similar to the real treatment locations.

In contrast, convolutional neural networks for image classification are much easier to train, and, in this case, can be adapted to the same task. Hence, I recommend setting up the problem of finding candidate treatment locations as a classification task. Specifically, the convolutional neural network takes a 3D input array and "classifies" it into, say, 101 categories, where categories correspond either to the $10 \times 10 = 100$ grid cells in the center of the input area, or an additional "no missing treatment location" category. The distinction from other generation tasks is that here the set of possible outputs is relatively small, for instance, the 101 categories described above. In contrast, in image generation, there are infinitely many possible images that could be generated.

To retain the adversarial nature of the task, I propose simultaneously training the classification on three sets of data and adding a final fully connected layer. The three sets of data are as follows: First, areas with at least one real treatment location, but with one treatment location removed. The correct classification of such input data is into the category corresponding to the grid cell where the treatment location was removed. Second, areas with at least one real treatment location, but without any treatment location removed. The correct classification of such input data is into the no missing treatment location category. Third, areas without treatment locations. These areas are also correctly classified as not missing any treatment location. The output of the convolutional layers is a prediction for each grid cell, of whether it is missing a treatment location. A final fully connected layer combines the location-specific predictions into the categories mentioned above: one category corresponding to each of the central grid cells, plus one category to no missing treatment location.

This neural network architecture balances two tasks: a generative task of picking the correct location if a treatment location is missing, predominantly performed by the convolutional layers; and a discriminatory task of deciding whether a treatment location is missing at all, predominantly performed by the final fully connected layer. This structure retains the attractive interpretation of generative adversarial networks but is substantially easier to train. It also resembles denoising autoencoders (cf. Vincent et al., 2008), where the

removal of a real treatment location represents noise added to the input, with the autoencoder trained to remove the noise, here meaning to add the removed real treatment location. The idea of using the second and third sets of training examples without missing treatment location has precedents in the literature on adversarial examples and adversarial training (see Biggio et al., 2013; Szegedy et al., 2013).

The setup as an adversarial task, as well as the prediction of categories, additionally is beneficial because it generates draws near the local *modes* rather than the *mean* of the treatment location distribution (cf. Goodfellow, 2016; Lotter et al., 2016). The importance of sampling from the mode rather than the mean of the location distribution becomes clear in a simple example. Suppose all areas contain three possible locations in one-dimensional space: 1, 2, and 3. For instance, 2 may be the city center, while 1 and 3 are suburbs on either side of the city. In the data, if a region is treated, treatment always occurs in the suburbs; at either location 1 or location 3, each with probability 0.5. However, estimating the likely location of the treatment with the familiar mean squared error loss function will estimate the mean of the treatment location distribution, predicting treatment at location 2. In contrast, the adversarial loss function as well as loss functions used for classification tasks are minimized by predicting either 1 or 3 because these categories are most likely to correspond to the correct location.[3] In contrast, location 2 is rejected as a candidate treatment location because treatment is never observed at such a location.

Data Augmentation Data augmentation serves two closely related purposes. First, rotating, mirroring, and shifting input areas produces additional, albeit dependent, observations but preserves all relative distances. Additional observations are helpful because training neural networks requires a large number of training samples. Second, these transformations effectively regularize the parameters of the estimated model. One can choose transformations that induce equivariance to rotation, mirroring, and shifts as appropriate for the particular setting. For instance, in many applications in the social sciences, North-South and East-West orientation are irrelevant on a small scale; only the relative distances matter.[4] Suppose there is an individual who visits a business to the North of her home because it is on the way to work in the North. If the whole space was rotated counterclockwise by 90 degrees, the individual equally visits the same business now to the West as it is still on the way to work, now also rotated to be to the West of her home. In image classification, the use of data augmentation is common and associated with a reduction in overfitting and greater generalizability of the learned models (Yaeger et al., 1996; Simard et al., 2003; Krizhevsky et al., 2012).

Shifting the entire grid has two additional desirable effects: First, imposing a continuous shift of the grid relative to covariates renders the exact discretization less relevant. The *average* (across draws from the shift distribution) distance in grid cells between two observations becomes directly proportional to their actual distance. Second, the location of an observation *within* a grid cell is no longer fixed. Shifting within-cell location is attractive because the classification is not informative of whether the candidate treatment location is at the center or towards the edge of a grid cell. With a continuous shift of the observations, the center of the grid cell points to different absolute locations depending on the shift. One can then average over several realizations of the shift to reduce the influence of the particular translation of grid cells to absolute locations.

Two notable alternatives or complements to data augmentation in the machine learning literature are spatial transformer networks and imposing equivariance directly on the parameters. First, spatial transformer networks (Jaderberg et al., 2015) estimate a rotation or other transformation that makes the subsequent classification task as easy as possible. Second, recent work considers imposing the desired equivariance property on the convolution kernel or adding layers to the network that effectively average the appropriate kernel coefficients (Cohen and Welling, 2016; Dieleman et al., 2016; Gens and Domingos, 2014; Dudar and Semenov, 2018; Dzhezyan and Cecotti, 2019). However, research suggests that data augmentation and other regularization techniques already achieve the first-order gains implied by these properties (Srivastava et al., 2014; Kauderer-Abrams, 2017; Yang et al., 2019). One can also inspect the models to assess the implied degree of equivariance (Goodfellow et al., 2009; Zeiler and Fergus, 2014; Lenc and Vedaldi, 2015).

[3] In general adversarial networks, one input to the network is white noise. This noise effectively chooses between the different local modes of the distribution. In the setup as a classification task proposed here, data augmentation, as described below, plays a similar role.

[4] Applications in environmental economics are notable exceptions if, for instance, wind direction is relevant. In such cases, rotation hinders the ability of the model to capture patterns due to, for instance, wind consistently blowing from one direction, and may require the inclusion of wind direction in estimation. The choice of appropriate data augmentation is therefore application specific.

Table OA2: The initial sample of all possibly relevant businesses consists of all businesses in the SafeGraph "point of interest" data with location within six miles from one of the five cities or with a zip code given in the table.

city	latitude	longitude
South San Francisco	37.653540	-122.416866
Burlingame	37.584103	-122.366083
Belmont	37.516493	-122.294191
Menlo Park	37.451967	-122.177993
Mountain View	37.389389	-122.083210
ZIP codes:		
94002, 94005, 94010, 94014, 94015, 94016, 94019, 94020, 94022, 94024, 94025, 94027, 94028, 94030, 94032, 94035, 94037, 94040, 94041, 94042, 94043, 94044, 94061, 94062, 94063, 94064, 94065, 94066, 94070, 94080, 94083, 94085, 94086, 94087, 94089, 94101, 94102, 94104, 94105, 94110, 94112, 94114, 94117, 94121, 94124, 94127, 94128, 94129, 94130, 94131, 94132, 94133, 94134, 94169, 94192, 94301, 94303, 94304, 94305, 94306, 94309, 94401, 94402, 94403, 94404, 94497, 94530, 94538, 94555, 94603, 95014, 95015, 95051, 95054, 95101, 95112		

3 Implementation details for the empirical application

3.1 Data Processing

I use the July 2021 release of SafeGraph's data for the year 2020. In this release of the data, SafeGraph applies its current algorithm to the data it collected in 2020, and updates its data sets attributing smartphone pings to businesses. In this paper, I focus on businesses in the San Francisco Bay Area, specifically in the Peninsula and South Bay between South San Francisco and Sunnyvale, see Figure 2 in the main text. To create the initial sample of all possibly relevant businesses for which SafeGraph has recorded data, I keep all businesses that either lie within six miles from several points throughout the Bay Area or have a SafeGraph-determined ZIP code falling within a list of relevant ZIP codes, see Table OA2.

To define the units of interest and ensure high-quality data for this application, I take three additional steps in processing the data. First, I determine the grocery and convenience stores that I consider "treatments" in this paper. Second, I manually set the location of each of these treatments to correspond to the main entrance of the store. Third, I check and de-duplicate restaurant location data to restrict to real restaurants that were likely to be open in early 2020.

Based on SafeGraph's "point of interest" data, I find 167 unique grocery and convenience store (treatment) locations that were open in 2020 in the interior of the study area. Starting from the sample defined above, I define the possible businesses of interest as those within 3 miles of Burlingame, 5 miles of Belmont, 5.5 miles of Menlo Park, or 2.95 miles of Mountain View, with the city locations as in Table OA2. Focusing on grocery stores in the interior of the study area guarantees that the full sample includes data on all businesses that are within different distances of interest from the grocery stores. To find locations consumers typically visit to purchase groceries, I start with all businesses with 4-digit NAICS code 4451 (grocery and convenience stores) assigned by SafeGraph, and then add all Costco, Target, and Walmart stores (which SafeGraph classifies as general merchandise stores, 4523), for a total of 313 stores. Of these stores, I exclude 28 stores that SafeGraph determines to have closed permanently before the COVID-19 pandemic (in or before February 2020; there were no further grocery store closures until July as recorded by SafeGraph), as well as 1 store that SafeGraph determines to have opened only in November 2020. For the remaining 284 stores, I verify manually that they fit my definition of grocery or convenience store. I exclude 100 stores;

primarily convenience stores that are part of gas stations, delis, and food producers and importers/exporters that are incorrectly classified as grocery stores by SafeGraph's algorithm. I confirm, based on newspaper articles, Yelp entries, and Google Street View imagery, that another 17 grocery stores were either not open in 2020 (closed before or opened after) or were duplicate entries in the data set. Overall, I consider 167 treatment locations; 139 locations are labeled as grocery (or general merchandise) stores by SafeGraph, with the remaining 28 labeled as convenience stores by SafeGraph.

For the 167 grocery and convenience stores in the sample, I manually determine the latitude and longitude of the main entrance, which serves two related purposes. First, the main entrance and exit is the relevant location to measure distances to or from for trip sequencing: If a consumer considers visiting a coffee shop before or after a grocery store, the additional distance she has to travel is based on the front door of the grocery store, not a location in the interior. Second, placing the location of grocery stores at their main entrances typically reduces the differences between taking straight-line distance (as in this paper) and walking distance (likely the economically relevant distance metric) between grocery stores and restaurants. When the grocery store location is instead placed in the interior of the store, restaurants that are *behind* the grocery store can appear closer than restaurants that are next door. Hence, placing the location of the grocery store at its front entrance improves the interpretability of estimates by distance. The latitude and longitude given in the SafeGraph data instead reflect "the general center of the business,"[5] typically in the interior of the store. I use Google Maps satellite as well as Street View imagery to locate the main entrances of all grocery stores. For about three-quarters of the grocery and convenience stores, the difference in locations is less than 20 meters. The largest differences in locations (of around 70 meters) occur for a handful of particularly large Costco, Safeway, Target, and Walmart stores.

I audit the data on restaurant (outcome unit) locations in three steps. First, I de-duplicate observations by checking for similarity of business names between any two businesses with locations within 50 meters of each other according to SafeGraph data. To detect duplicates based on name similarity, I focus my attention on businesses with high relative Levenshtein distance. This distance measures the minimum number of character edits needed to make the names of the two businesses equal, relative to the length of the longer business name. Most duplicates I detect are clear typos in the name of one of the observations, and some are abbreviations of business names that I verify to indeed describe the same business using Google Maps and Street View data. Second, I audit the SafeGraph location data by comparing the latitude and longitude in the SafeGraph "point of interest" data to the latitude and longitude obtained by searching for the business name and street address (also from the "point of interest" data) on Google Maps. This analysis confirms the high quality of the SafeGraph location data. Randomly inspecting the locations of a few dozen restaurants in more detail, I find that neither the SafeGraph nor the Google maps locations are systematically closer to the entrance of the restaurants. Given the much smaller size (area) of restaurants compared to grocery stores, as well as the much greater number of restaurants, I do not manually record the latitudes and longitudes of their entrances. Third, I focus on businesses that were reliably assigned visits by SafeGraph. I restrict the non-grocery store sample to businesses for which SafeGraph reported at least 7 visits in each of the four weeks starting in January 2020. This step excludes businesses that were not open at the time, not properly assigned visits by SafeGraph's algorithm, or are too small to reliably measure visits for, but retains 95-97.5% of all *visits* (depending on the week) in the SafeGraph data. Importantly, I take each of the three steps without knowledge of which businesses are, in the later analysis, considered treated or control.

3.2 Convolutional Neural Network

I use a convolutional neural network (CNN) to identify plausible counterfactual locations. First, I specify the input into the training of the CNN. Second, I describe the architecture of the CNN. Third, I use the trained CNN to predict many plausible counterfactual locations, followed by additional matching steps, to select the final counterfactual locations used in the analysis.

I project the latitude and longitude of all businesses into two-dimensional Cartesian space using the NAD83 (2011) projection, EPSG:6419 California zone 3. This projection gives the location in meters East and North relative to a point near the San Francisco Bay Area. In applications where data come from different

[5]SafeGraph documentation, `https://docs.safegraph.com/docs/core-places#section-latitude-longitude` accessed on July 29, 2021.

Table OA3: Number of businesses by 4-digit NAICS code that are in the larger neighborhoods forming the input into the convolutional neural network. The number of grocery stores exceeds 167 here because additional grocery stores that are not in the interior of the main study area are included in these larger neighborhoods.

NAICS code	description	# unique businesses
7225	Restaurants and Other Eating Places	1975
7139	Other Amusement and Recreation Industries	606
7121	Museums, Historical Sites, and Similar Institutions	409
8131	Religious Organizations	324
6111	Elementary and Secondary Schools	265
6244	Child Day Care Services	264
4451	Grocery Stores	244
4471	Gasoline Stations	182
any	–	7845

regions, the researcher should choose the appropriate projection for each region to ensure the accuracy of relative distances within regions.

The CNN learns to predict treatment locations in the areas around prespecified locations: real grocery store locations and semi-randomly chosen locations. The semi-randomly chosen locations, together with the real grocery store locations, are meant to cover the areas in which counterfactual locations could plausibly occur. I start with the locations of all businesses for which the nearest grocery store is between 0.2 miles and 2 miles away. The areas around businesses even closer to a grocery store are already included in the consideration set by including the area of that grocery store. Next, I jitter these locations by adding independent shocks from a normal distribution with mean ± 0.0004 and standard deviation 0.0001 to their latitudes and longitudes, where the sign of the mean is independently drawn to be $+1$ or -1 for each location and coordinate. This step ensures that the *center* of each area does not fall exactly onto a business because real grocery store locations never exactly coincide with the locations of other businesses. Finally, to avoid including an area multiple times, I detect all pairs of jittered locations that are within 100 meters of one another. I drop locations that are listed "first" (in the arbitrary order based on the row numbers of the businesses the location is based on) in any such pair. The areas around both the resulting $1,900$ semi-random locations and the 167 real grocery store locations are used as input to the CNN.

The CNN predictions are based on observable characteristics describing small 2D grid cells around the prespecified locations. Each grid cell covers an area of $0.025\text{mi} \times 0.025\text{mi}$ (approximately $40\text{m} \times 40\text{m}$). I use the count of businesses by 4-digit NAICS code for the codes given in Table OA3 as observable characteristics of each grid cell. That is, a cell covering two gasoline stations, one car dealership, and no other businesses, will have "covariate value" 2 for the covariate indicating industry group 4471 (gasoline stations) and 3 for the covariate indicating "any" industry, with the remaining covariates at 0 because there is no separate covariate for the relatively rare car dealerships (NAICS code 4411, less than 100 in the study area).

Each input observation to the CNN consists of one of the 2,067 areas described above. The covariates of the 2D grid are separate "channels" constituting a 3D tensor for each such observation. Each area consists of 50×50 grid cells. All coordinates within an area are jointly shifted, rotated, and mirrored randomly using independent uniform distributions for each of the three operations. The maximum absolute shift is such that the original center is placed within one of the central 10×10 grid cells. Hence, there are at least another 20 grid cells (0.5mi) of "padding" on all sites of the original center until the edge of the area.

The CNN consists of 4 sequential 2D convolutions and a final linear (fully connected) layer yielding $10 \times 10 + 1 = 101$ outputs. I use 2D instance normalization and leaky rectified linear activation for all neurons in the CNN, and replication padding to ensure the output of each convolution has the same spatial dimension as the input. The first convolution takes the 9 input channels (eight specific industries and one for any industry) and convolves it with a kernel size of 5 (considering the 5×5 grid cells centered around a given grid cell) into 18 channels. This layer can "smooth" the input such that the hard borders between grid cells due to discretization become less relevant. The increase in the number of channels allows the neural network

to learn a larger number of non-linearities. The second convolution takes the 18 channels of the previous layer and convolves them with a kernel size of 21 with a stride of 2 into 36 channels, such that each grid cell can view grid cells up to 20 cells away in any direction but skipping every other cell for parsimony. This layer allows each grid cell to learn about its neighborhood up to even relatively large distances (approximately $20 \times 0.025\text{mi} = 0.5\text{mi}$). The third convolution takes the 36 channels of the previous layer and convolves them with a kernel size of 5 into 36 channels, again allowing some smoothing across grid cells to counteract the skipping of every other grid cell of the previous layer. The fourth convolution takes the 36 channels of the previous layer and convolves them with a kernel size of 21 with a stride of 2 into a single channel. Intuitively, this layer forces a single prediction for each grid cell based on the large neighborhood (up to 20 cells away in any direction). The final layer linearly combines the 50×50 grid cells of the single channel of the previous layer into 101 "categories" that constitute the predictions of whether and where an additional grocery store may be located.

The 101 categories correspond to the central 10×10 grid as well as one category indicating a prediction of no additional grocery store. I train the CNN on batches consisting of 64 observations (areas). Half (32) of the observations are areas around a real grocery store, but with that grocery store removed from the input channel count of grocery stores per grid cell. The random shift and rotation of the input are such that this removed grocery store could have been in any of the central 10×10 grid cells. For these observations, the prediction maximizing the cross-entropy loss is the category corresponding to the cell that the grocery store has been removed from. All other categories are equal in terms of loss and worse than the correct category, which trains the CNN to identify the mode, rather than the average, location. A quarter (16) of the observations are areas around real grocery stores with no grocery store removed. The correct classification of such observations is into the category corresponding to "no missing grocery store" instead of any of the 10×10 grid cells. The last quarter (16) of the observations of each batch are areas around the semi-random prespecified locations. Their correct classification is also the category corresponding to "no missing grocery store."

After training, I evaluate the areas of the prespecified locations for possible grocery store locations according to the CNN. In this step, I input batches consisting of 32 observations into the trained CNN. In each batch, 4 observations are areas around real grocery stores: 2 have the grocery store removed from the input, while 2 do not have the grocery store removed. An additional 28 observations are areas around the semi-random prespecified locations. The trained neural network calculates predictions for $5,000$ batches. Predictions for observations with removed grocery stores allow me to learn the activation scores of real grocery store locations. The remaining observations yield possible counterfactual locations.

I find good matches for real grocery store locations among the possible counterfactual locations in two steps. In the first step, I find for each real grocery store location possible counterfactual locations with similar CNN activation. Specifically, I take each prediction for a removed real grocery store separately (there are multiple such predictions for each real grocery store under different random shifts, rotation, and mirroring), and match in descending order of activation, with replacement, within the possible counterfactual locations (excluding the prediction category for "no missing grocery store"). I repeat the same matching process (matching with replacement using the complete set of possible counterfactual locations) using relative activation within neighborhood-observation, corresponding to the cross-entropy loss function. Taking the union of these matches, I obtain $19,857$ possible locations that the CNN evaluated as similar to a real grocery store location under at least one shift, rotation, and mirroring. I drop 43 of these locations that are closer to the nearest real grocery store than two thirds of the minimum distance between any two real grocery stores. In the second step, I use propensity score matching to pick the final counterfactual locations among the $19,814$ remaining locations. I estimate a propensity score model using the real and possible counterfactual locations as observations in a logistic regression. There are three sets of regressors: 1) the numbers of restaurants in distance bins of width 0.025 miles from the location, up to a distance of 0.2 miles; 2) the average number of grocery stores near the restaurants in each bin broken out for each bin into similar bins of distance from the restaurant; 3) the total number of businesses (of any industry) in distance bins of width 0.25 miles, up to a distance of 1 mile. I match, with replacement, each grocery store location to the possible counterfactual location with the closest estimated propensity scores. The final sample consists of 162 counterfactual locations and the 167 real grocery store locations.

For the final sample of real grocery stores and most plausible counterfactual locations, I estimate a propensity score to analyze the sample as a quasi-experiment conditional on these locations and propensity

scores. The propensity score estimation uses the same regressors as the estimated propensity score for matching. The inverse probability weighting estimator only uses this propensity score to weight the "control" observations (restaurants near counterfactual locations) because the average treatment effect on the treated (ATT) estimator does not require reweighting of the "treated" observations (restaurants near real grocery stores). The primary purpose of re-estimating the propensity score is to balance exposure to grocery stores appropriately between treated and control restaurants. When estimating the average effect of one marginal grocery store on restaurants at a distance d, the treated and control restaurants at that distance indeed differ on average by one grocery store at distance d, and have similar average exposure to grocery stores at other distances as Figure 3 in the main text illustrates. By selecting the counterfactual locations from the CNN predictions based on the relative locations of other businesses in the area, these locations and propensity score weights also balance exposure to other businesses in the neighborhood as shown in Figure 4 of the main text.

4 General weights and covariance across distances

Define the estimator

$$\tilde{\tau}(d) \equiv \mu_t(d) - \mu_c(d) + \frac{\sum_{j=1}^{J} \mathcal{W}_j \sum_{s\in\mathbb{S}_j} \mathbb{1}\{\mathcal{S} \ni s\} \sum_{i\in\mathbb{I}_j} w_i(s,d)(\mathcal{Y}_i - \mu_t(d))}{\sum_{j=1}^{J} \pi_j \sum_{s\in\mathbb{S}_j} \pi_j(s) \sum_{i\in\mathbb{I}_j} w_i(s,d)}$$
$$- \frac{\sum_{j=1}^{J} \frac{1-\mathcal{W}_j}{1-\pi_j} \pi_j \sum_{s\in\mathbb{S}_j} \pi_j(s) \sum_{i\in\mathbb{I}_j} w_i(s,d)(\mathcal{Y}_i - \mu_c(d))}{\sum_{j=1}^{J} \pi_j \sum_{s\in\mathbb{S}_j} \pi_j(s) \sum_{i\in\mathbb{I}_j} w_i(s,d)}$$

where

$$\mu_t(d) \equiv \frac{\sum_{j=1}^{J} \pi_j \sum_{s\in\mathbb{S}_j} \pi_j(s) \sum_{i\in\mathbb{I}_j} w_i(s,d) Y_i(s)}{\sum_{j=1}^{J} \pi_j \sum_{s\in\mathbb{S}_j} \pi_j(s) \sum_{i\in\mathbb{I}_j} w_i(s,d)}$$
$$\mu_c(d) \equiv \frac{\sum_{j=1}^{J} \pi_j \sum_{s\in\mathbb{S}_j} \pi_j(s) \sum_{i\in\mathbb{I}_j} w_i(s,d) Y_i(0)}{\sum_{j=1}^{J} \pi_j \sum_{s\in\mathbb{S}_j} \pi_j(s) \sum_{i\in\mathbb{I}_j} w_i(s,d)}$$

for arbitrary non-stochastic weights $w_i(s,d)$.

The theorem below describes the covariance of the estimator at different distances under Assumptions 1 and 3 from the main text combined with either Assumption 2 from the main text or Assumption OA1 defined here:

Assumption OA1 (Independent Assignment Across Regions)**.** *Treatments are assigned across regions independently. For $j \neq j'$,*

$$\mathcal{W}_j \perp\!\!\!\perp \mathcal{W}_{j'}$$

with marginal treatment probabilities $\Pr(\mathcal{W}_j = 1) \equiv \pi_j$.
Conditional on treatment in region j, assignment to a particular location within the region is independent of assignment in other regions j'. For all $s \in \mathbb{S}_j$ and $s' \in \mathbb{S}_{j'}$ with $j \neq j'$: $\mathbb{1}\{\mathcal{S} \ni s\} \perp\!\!\!\perp \mathbb{1}\{\mathcal{S} \ni s'\} \mid \mathcal{W}_j = 1, \mathcal{W}_{j'} = 1$.

Theorem OA1. *Under Assumptions 1 and 3, with either Assumption 2 ($\mathfrak{C} = 1$) or Assumption OA1 ($\mathfrak{C} = 0$), the covariance of the estimator $\tilde{\tau}$ at distances d and d' is*

$$\begin{aligned}\operatorname{cov}(\tilde{\tau}(d), \tilde{\tau}(d')) =& \frac{J-1}{J} \frac{\tilde{V}_t^{location}(d,d')}{J_t} + \frac{1}{2}\frac{\mathfrak{C}}{J}\Big(\frac{\tilde{V}_t^{region}(d \mid d')}{J_t} + \frac{\tilde{V}_t(d' \mid d)}{J_t}\Big) \\ &+ \frac{J-1+\mathfrak{C}}{J} \frac{\tilde{V}_c^{region}(d,d')}{J_c} + \frac{1}{2}\frac{J-1+\mathfrak{C}}{J}\Big(\frac{\tilde{V}_c^{region}(d \mid d')}{J_c} + \frac{\tilde{V}_c^{region}(d' \mid d)}{J_c}\Big) \\ &- \frac{1}{2}\frac{J-1+\mathfrak{C}}{J}\Big(\frac{\tilde{V}_{ct}^{region}(d,d')}{J} + \frac{\tilde{V}_{ct}^{region}(d',d)}{J}\Big) + \frac{1}{2}\frac{\tilde{V}_{tt}^{region}(d,d',\mathfrak{C})}{J}\end{aligned}$$

where

$$\tilde{V}_t^{location}(d,d') = \frac{1}{J-1}\sum_{j=1}^{J}\frac{\pi_j}{\pi}\sum_{s\in\mathbb{S}_j}\pi_j(s)\frac{Y_j^t(s,d)}{\bar{n}(d)}\frac{Y_j^t(s,d')}{\bar{n}(d')}$$

$$\tilde{V}_t^{region}(d \mid d') = \frac{1}{J-1}\sum_{j=1}^{J}\frac{\pi_j}{\pi}\Big(\sum_{s\in\mathbb{S}_j}\pi_j(s)\frac{Y_j^t(s,d)}{\sqrt{\bar{n}(d)\bar{n}(d')}}\Big)^2$$

$$\tilde{V}_c^{region}(d,d') = \frac{1}{J-1}\sum_{j=1}^{J}\pi_j\big(\frac{\pi_j}{\pi}\big)^2\frac{1-\pi}{1-\pi_j}\frac{Y_j^c(d)}{\bar{n}(d)}\frac{Y_j^c(d')}{\bar{n}(d')}$$

$$\tilde{V}_c^{region}(d \mid d') = \frac{1}{J-1}\sum_{j=1}^{J}(1-\pi)\big(\frac{\pi_j}{\pi}\big)^2\Big(\frac{Y_j^c(d)}{\sqrt{\bar{n}(d)\bar{n}(d')}}\Big)^2$$

$$\tilde{V}_{ct}^{region}(d,d') = \frac{1}{J-1}\sum_{j=1}^{J}\big(\frac{\pi_j}{\pi}\big)^2\Big(\sum_{s\in\mathbb{S}_j}\pi_j(s)\frac{Y_j^t(s,d)-Y_j^c(d')}{\sqrt{\bar{n}(d)\bar{n}(d')}}\Big)^2$$

$$\tilde{V}_{tt}^{region}(d,d',\mathfrak{C}) = \frac{1}{J-1}\sum_{j=1}^{J}\big(\frac{\pi_j}{\pi}\big)^2\Big(\sum_{s\in\mathbb{S}_j}\pi_j(s)\frac{(Y_j^t(s,d)-Y_j^t(s,d'))}{\sqrt{\bar{n}(d)\bar{n}(d')}}\Big)^2\cdot\frac{J-1-\mathfrak{C}\frac{1-\pi_j}{\pi_j}}{J}$$

with

$$Y_j^t(s,d) \equiv \sum_{i\in\mathbb{I}_j} w_i(s,d)(Y_i(s)-\mu_t(d)) \qquad Y_j^c(d) \equiv \sum_{s\in\mathbb{S}_j}\pi_j(s)\sum_{i\in\mathbb{I}_j} w_i(s,d)(Y_i(0)-\mu_c(d))$$

$$\bar{n}(d) \equiv \frac{1}{J}\sum_{j=1}^{J}\frac{\pi_j}{\pi}\sum_{s\in\mathbb{S}_j}\pi_j(s)\sum_{i\in\mathbb{I}_j} w_i(s,d)$$

and $\pi \equiv \frac{1}{J}\sum_{j=1}^{J}\pi_j$, $J_t \equiv \pi J$, *and* $J_c \equiv (1-\pi)J$ *correspond to the average region treatment probability and expected number of treated and control regions. If they are constant (as under Assumption 2), the definitions coincide with the notation of the main text.*

$\tilde{V}_t^{\text{location}}(d,d')$ is the pseudo-covariance of treated potential outcomes at distances d and d' from the same location. $\tilde{V}_t^{\text{region}}(d)$ is the pseudo-variance of the within-region average treated potential outcomes at distance d from any location. $\tilde{V}_c^{\text{region}}(d,d')$ is the pseudo-covariance of within-region average control potential outcomes at distance d and d'. $\tilde{V}_c^{\text{region}}(d)$ is the pseudo-variance of the within-region average control potential outcomes at distance d from any location. $\tilde{V}_{ct}^{\text{region}}(d,d')$ is similar to a pseudo-variance of treatment effects but contrasts treated potential outcomes at distance d with control potential outcomes at distance d'. $\tilde{V}_{tt}^{\text{region}}(d,d')$ is the pseudo-variance of the within-region average difference of treated potential outcomes at distances d and d'. If π_j is constant across j and $d=d'$, then $\tilde{V}_c^{\text{region}}(d)$ and $\tilde{V}_c^{\text{region}}(d,d')$ can be combined and except for the final term of $\tilde{V}_{tt}^{\text{region}}(d,d')$, all region-level probabilities π_j in the formulas above cancel with their average, π. $\tilde{V}_{tt}^{\text{region}}(d,d')=0$ mechanically when $d=d'$. $\bar{n}(d)$ is the approximate simple region-average of the within-region expected number of effective (weighted) treated individuals. Note that even $\tilde{V}_t^{\text{region}}(d \mid d')$ and $\tilde{V}_c^{\text{region}}(d \mid d')$ divide by the geometric mean of the approximate average effective number of treated individuals at distance d and d' despite squaring only potential outcomes at distance d.

Proof:

Rewrite the estimator

$$\tilde{\tau}(d) = \mu_t(d)-\mu_c(d) + \frac{\sum_{j=1}^{J}\mathcal{W}_j\sum_{s\in\mathbb{S}_j}\mathbb{1}\{\mathcal{S}\ni s\}\sum_{i\in\mathbb{I}_j}w_i(s,d)(Y_i(s)-\mu_t(d))}{\sum_{j=1}^{J}\pi_j\sum_{s\in\mathbb{S}_j}\pi_j(s)\sum_{i\in\mathbb{I}_j}w_i(s,d)}$$
$$-\frac{\sum_{j=1}^{J}\frac{1-\mathcal{W}_j}{1-\pi_j}\pi_j\sum_{s\in\mathbb{S}_j}\pi_j(s)\sum_{i\in\mathbb{I}_j}w_i(s,d)(Y_i(0)-\mu_c(d))}{\sum_{j=1}^{J}\pi_j\sum_{s\in\mathbb{S}_j}\pi_j(s)\sum_{i\in\mathbb{I}_j}w_i(s,d)}$$

by replacing the realized outcome $\mathcal{Y}_i$ by the potential outcome $Y_i(s)$ or $Y_i(0)$ corresponding to the treatment status that is selected by the indicators $\mathcal{W}_j\mathbb{1}\{\mathcal{S}\ni s\}$ and $1-\mathcal{W}_j$.

Let $\mathcal{T}_j(s) \equiv \mathcal{W}_j \mathbb{1}\{\mathcal{S} \ni s\}$. Then $\mathcal{W}_j = \sum_{s\in\mathbb{S}_j} \mathcal{T}_j(s)$. Substituting $\mathcal{T}_j(s)$, $Y_j^t(s,d)$, and $Y_j^c(d)$:

$$\tilde{\tau}(d) = \mu_t(d) - \mu_c(d) + \frac{\sum_{j=1}^J \sum_{s\in\mathbb{S}_j} \mathcal{T}_j(s) Y_j^t(s,d)}{\sum_{j=1}^J \pi_j \sum_{s\in\mathbb{S}_j} \pi_j(s) \sum_{i\in\mathbb{I}_j} w_i(s,d)} - \frac{\sum_{j=1}^J \frac{1-\sum_{s\in\mathbb{S}_j}\mathcal{T}_j(s)}{1-\pi_j}\pi_j Y_j^c(d)}{\sum_{j=1}^J \pi_j \sum_{s\in\mathbb{S}_j} \pi_j(s) \sum_{i\in\mathbb{I}_j} w_i(s,d)}.$$

Combining terms multiplied by the stochastic $\mathcal{T}_j(s)$,

$$\tilde{\tau}(d) = \mu_t(d) - \mu_c(d) + \frac{\sum_{j=1}^J \sum_{s\in\mathbb{S}_j} \mathcal{T}_j(s)(Y_j^t(s,d) + \frac{\pi_j}{1-\pi_j} Y_j^c(d))}{\sum_{j=1}^J \pi_j \sum_{s\in\mathbb{S}_j} \pi_j(s) \sum_{i\in\mathbb{I}_j} w_i(s,d)} - \frac{\sum_{j=1}^J \frac{\pi_j}{1-\pi_j} Y_j^c(d)}{\sum_{j=1}^J \pi_j \sum_{s\in\mathbb{S}_j} \pi_j(s) \sum_{i\in\mathbb{I}_j} w_i(s,d)}.$$

In a design-based analysis, only $\mathcal{T}_j(s)$ is stochastic in the expression above. For ease of notation, define

$$m(d,d') \equiv \left(\sum_{j=1}^J \pi_j \sum_{s\in\mathbb{S}_j} \pi_j(s) \sum_{i\in\mathbb{I}_j} w_i(s,d)\right)\left(\sum_{j=1}^J \pi_j \sum_{s\in\mathbb{S}_j} \pi_j(s) \sum_{i\in\mathbb{I}_j} w_i(s,d')\right).$$

Then

$$\begin{aligned}\operatorname{cov}(\tilde{\tau}(d),\tilde{\tau}(d')) &= \operatorname{cov}\left(\frac{\sum_{j=1}^J \sum_{s\in\mathbb{S}_j} \mathcal{T}_j(s)(Y_j^t(s,d) + \frac{\pi_j}{1-\pi_j} Y_j^c(d))}{\sum_{j=1}^J \pi_j \sum_{s\in\mathbb{S}_j} \pi_j(s) \sum_{i\in\mathbb{I}_j} w_i(s,d)}, \frac{\sum_{j=1}^J \sum_{s\in\mathbb{S}_j} \mathcal{T}_j(s)(Y_j^t(s,d') + \frac{\pi_j}{1-\pi_j} Y_j^c(d'))}{\sum_{j=1}^J \pi_j \sum_{s\in\mathbb{S}_j} \pi_j(s) \sum_{i\in\mathbb{I}_j} w_i(s,d')}\right) \\ &= \sum_{j=1}^J \sum_{s\in\mathbb{S}_j} \sum_{j'=1}^J \sum_{s'\in\mathbb{S}_{j'}} \operatorname{cov}(\mathcal{T}_j(s), \mathcal{T}_{j'}(s')) \frac{\left(Y_j^t(s,d) + \frac{\pi_j}{1-\pi_j} Y_j^c(d)\right)\left(Y_{j'}^t(s',d') + \frac{\pi_{j'}}{1-\pi_{j'}} Y_{j'}^c(d')\right)}{m(d,d')}.\end{aligned}$$

The covariances are straightforward to derive because $\mathcal{T}_j(s)$ are Bernoulli random variables. $\Pr(\mathcal{T}_j(s) = 1) = \pi_j\pi_j(s)$, and $\Pr(\mathcal{T}_j(s) = 1 \text{ and } \mathcal{T}_j(s') = 1) = 0$ for $s \neq s'$ by Assumption 3. Under independent assignment, $\operatorname{cov}(\mathcal{T}_j(s), \mathcal{T}_{j'}(s')) = 0$ for $j \neq j'$. If instead the number of treated regions is fixed and the region-level probability of treatment is constant (Assumption 2), then one can obtain by the law of total probability

$$E(\mathcal{T}_j(s)\mathcal{T}_{j'}(s')) = \pi_j\pi_j(s)\pi_{j'}(s) E(\mathcal{W}_{j'} \mid \mathcal{W}_j = 1)$$

where $E(\mathcal{W}_{j'} \mid \mathcal{W}_j = 1) = \frac{\pi J - 1}{J-1}$ because if j is treated, then $\pi J - 1$ of the other $J-1$ regions are treated, all with equal probability, such that simple algebra yields

$$E(\mathcal{T}_j(s)\mathcal{T}_{j'}(s')) - E(\mathcal{T}_j(s))E(\mathcal{T}_{j'}(s')) = \frac{\pi(1-\pi)}{J-1}\pi_j(s)\pi_j(s').$$

Hence,

$$\operatorname{cov}(\mathcal{T}_j(s), \mathcal{T}_{j'}(s')) = \begin{cases} \pi_j\pi_j(s)(1-\pi_j\pi_j(s)) & \text{if } j = j', s = s' \\ -\pi_j^2\pi_j(s)\pi_j(s') & \text{if } j = j', s \neq s' \\ -\mathfrak{C}\frac{\pi(1-\pi)}{J-1}\pi_j(s)\pi_{j'}(s') & \text{if } j \neq j' \end{cases}$$

where $\mathfrak{C} = 1$ for Assumption 2 and $\mathfrak{C} = 0$ for Assumption OA1.

Then

$$\begin{aligned}&\sum_{j=1}^J \sum_{s\in\mathbb{S}_j} \sum_{j'=1}^J \sum_{s'\in\mathbb{S}_{j'}} \operatorname{cov}(\mathcal{T}_j(s), \mathcal{T}_{j'}(s')) \frac{\left(Y_j^t(s,d) + \frac{\pi_j}{1-\pi_j} Y_j^c(d)\right)\left(Y_{j'}^t(s',d') + \frac{\pi_{j'}}{1-\pi_{j'}} Y_{j'}^c(d')\right)}{m(d,d')} \\ =&\sum_{j=1}^J \sum_{s\in\mathbb{S}_j} \pi_j\pi_j(s)(1-\pi_j\pi_j(s)) \frac{\left(Y_j^t(s,d) + \frac{\pi_j}{1-\pi_j} Y_j^c(d)\right)\left(Y_j^t(s,d') + \frac{\pi_j}{1-\pi_j} Y_j^c(d')\right)}{m(d,d')} \\ &- \sum_{j=1}^J \sum_{s\in\mathbb{S}_j} \sum_{s'\in\mathbb{S}_j} \mathbb{1}\{s \neq s'\}\pi_j^2\pi_j(s)\pi_j(s') \frac{\left(Y_j^t(s,d) + \frac{\pi_j}{1-\pi_j} Y_j^c(d)\right)\left(Y_j^t(s',d') + \frac{\pi_j}{1-\pi_j} Y_j^c(d')\right)}{m(d,d')} \\ &- \mathfrak{C}\sum_{j=1}^J \sum_{s\in\mathbb{S}_j} \sum_{j=1}^J \sum_{s'\in\mathbb{S}_{j'}} \mathbb{1}\{j \neq j'\}\frac{\pi(1-\pi)}{J-1}\pi_j(s)\pi_{j'}(s') \frac{\left(Y_j^t(s,d) + \frac{\pi_j}{1-\pi_j} Y_j^c(d)\right)\left(Y_{j'}^t(s',d') + \frac{\pi_{j'}}{1-\pi_{j'}} Y_{j'}^c(d')\right)}{m(d,d')}.\end{aligned}$$

Adding and subtracting the $s = s'$ term from the second summation and combining the added term with the first summation:

$$
\begin{aligned}
=&\sum_{j=1}^{J}\sum_{s\in\mathbb{S}_j}\pi_j\pi_j(s)\frac{\big(Y_j^t(s,d)+\frac{\pi_j}{1-\pi_j}Y_j^c(d)\big)\big(Y_j^t(s,d')+\frac{\pi_j}{1-\pi_j}Y_j^c(d')\big)}{m(d,d')}\\
&-\sum_{j=1}^{J}\sum_{s\in\mathbb{S}_j}\sum_{s'\in\mathbb{S}_j}\pi_j^2\pi_j(s)\pi_j(s')\frac{\big(Y_j^t(s,d)+\frac{\pi_j}{1-\pi_j}Y_j^c(d)\big)\big(Y_j^t(s',d')+\frac{\pi_j}{1-\pi_j}Y_j^c(d')\big)}{m(d,d')}\\
&-\mathfrak{C}\sum_{j=1}^{J}\sum_{s\in\mathbb{S}_j}\sum_{j=1}^{J}\sum_{s'\in\mathbb{S}_{j'}}\mathbb{1}\{j\neq j'\}\frac{\pi(1-\pi)}{J-1}\pi_j(s)\pi_{j'}(s')\frac{\big(Y_j^t(s,d)+\frac{\pi_j}{1-\pi_j}Y_j^c(d)\big)\big(Y_{j'}^t(s',d')+\frac{\pi_{j'}}{1-\pi_{j'}}Y_{j'}^c(d')\big)}{m(d,d')}.
\end{aligned}
$$

Similarly adding and subtracting the $j = j'$ term of the third summation, as well as using that $\pi \equiv \pi_j$ for all j under Assumption 2, yields

$$
\begin{aligned}
=&\sum_{j=1}^{J}\sum_{s\in\mathbb{S}_j}\pi_j\pi_j(s)\frac{\big(Y_j^t(s,d)+\frac{\pi_j}{1-\pi_j}Y_j^c(d)\big)\big(Y_j^t(s,d')+\frac{\pi_j}{1-\pi_j}Y_j^c(d')\big)}{m(d,d')}\\
&-\sum_{j=1}^{J}\sum_{s\in\mathbb{S}_j}\sum_{s'\in\mathbb{S}_j}(\pi_j^2\pi_j(s)\pi_j(s')-\mathfrak{C}\frac{\pi_j(1-\pi_j)}{J-1}\pi_j(s)\pi_j(s'))\frac{\big(Y_j^t(s,d)+\frac{\pi_j}{1-\pi_j}Y_j^c(d)\big)\big(Y_j^t(s',d')+\frac{\pi_j}{1-\pi_j}Y_j^c(d')\big)}{m(d,d')}\\
&-\mathfrak{C}\sum_{j=1}^{J}\sum_{s\in\mathbb{S}_j}\sum_{j=1}^{J}\sum_{s'\in\mathbb{S}_{j'}}\frac{\pi(1-\pi)}{J-1}\pi_j(s)\pi_{j'}(s')\frac{\big(Y_j^t(s,d)+\frac{\pi}{1-\pi}Y_j^c(d)\big)\big(Y_{j'}^t(s',d')+\frac{\pi}{1-\pi}Y_{j'}^c(d')\big)}{m(d,d')}.
\end{aligned}
$$

The third summation consists of products that are separable in j and j'. Substituting for $Y_j^t(s,d)$ and $Y_j^c(d)$ and refactoring the summation yields a factor $\sum_{j=1}^{J}\pi\sum_{s\in\mathbb{S}_j}\pi_j(s)(Y_j^t(s,d)+\frac{\pi}{1-\pi}Y_j^c(d))=0$:

$$
\begin{aligned}
\sum_{j=1}^{J}\pi\sum_{s\in\mathbb{S}_j}\pi_j(s)Y_j^t(s,d)&=\sum_{j=1}^{J}\pi\sum_{s\in\mathbb{S}_j}\pi_j(s)\sum_{i\in\mathbb{I}_j}w_i(s,d)(Y_i(s)-\mu_t(d))=0\\
\sum_{j=1}^{J}\pi\sum_{s\in\mathbb{S}_j}\pi_j(s)\frac{\pi}{1-\pi}Y_j^c(d)&=\frac{\pi}{1-\pi}\sum_{j=1}^{J}\pi\Big(\sum_{s\in\mathbb{S}_j}\pi_j(s)\Big)\sum_{s\in\mathbb{S}_j}\pi_j(s)\sum_{i\in\mathbb{I}_j}w_i(s,d)(Y_i(0)-\mu_c(d))=0
\end{aligned}
$$

by the definitions of $\mu_t(d)$ and $\mu_c(d)$ and because $\sum_{s\in\mathbb{S}_j}\pi_j(s)=1$. Hence the third summation in the variance expression is equal to 0.

Next, expand the products of potential outcomes:

$$
\begin{aligned}
&\big(Y_j^t(s,d)+\frac{\pi_j}{1-\pi_j}Y_j^c(d)\big)\big(Y_j^t(s',d')+\frac{\pi_j}{1-\pi_j}Y_j^c(d')\big)\\
=&Y_j^t(s,d)Y_j^t(s',d')+\frac{\pi_j}{1-\pi_j}Y_j^t(s,d)Y_j^c(d')+\frac{\pi_j}{1-\pi_j}Y_j^t(s',d')Y_j^c(d)+\big(\frac{\pi_j}{1-\pi_j}\big)^2Y_j^c(d)Y_j^c(d').
\end{aligned}
$$

Dropping the third summation that equals zero of the variance formula, substituting these four products, and

simplifying the variance formula by combining terms with identical products of potential outcomes yields

$$\begin{aligned}
\operatorname{cov}(\tilde{\tau}(d),\tilde{\tau}(d')) =& \sum_{j=1}^{J}\pi_j\frac{\sum_{s\in\mathbb{S}_j}\pi_j(s)Y_j^t(s,d)Y_j^t(s,d')}{m(d,d')} \\
&+(1+\frac{\mathfrak{C}}{J-1})\sum_{j=1}^{J}\pi_j^2\frac{\big(\sum_{s\in\mathbb{S}_j}\pi_j(s)Y_j^t(s,d)\big)\big(Y_j^c(d')\big)}{m(d,d')} \\
&+(1+\frac{\mathfrak{C}}{J-1})\sum_{j=1}^{J}\pi_j^2\frac{\big(\sum_{s\in\mathbb{S}_j}\pi_j(s)Y_j^t(s,d')\big)\big(Y_j^c(d)\big)}{m(d,d')} \\
&+(1+\frac{\mathfrak{C}}{J-1})\sum_{j=1}^{J}\frac{\pi_j^3}{1-\pi_j}\frac{Y_j^c(d)Y_j^c(d')}{m(d,d')} \\
&-\sum_{j=1}^{J}(\pi_j^2-\mathfrak{C}\frac{\pi_j(1-\pi_j)}{J-1})\frac{\big(\sum_{s\in\mathbb{S}_j}\pi_j(s)Y_j^t(s,d)\big)\big(\sum_{s\in\mathbb{S}_j}\pi_j(s)Y_j^t(s,d')\big)}{m(d,d')}.
\end{aligned}$$

Based on the binomial formula, $ab=\frac{1}{2}(a^2+b^2-(a-b)^2)$, so

$$\big(\sum_{s\in\mathbb{S}_j}\pi_j(s)Y_j^t(s,d)\big)\big(Y_j^c(d')\big)=\frac{1}{2}\left(\big(\sum_{s\in\mathbb{S}_j}\pi_j(s)Y_j^t(s,d)\big)^2+\big(Y_j^c(d')\big)^2-\big(\sum_{s\in\mathbb{S}_j}\pi_j(s)Y_j^t(s,d)-Y_j^c(d')\big)^2\right)$$

$$\begin{aligned}
\big(\sum_{s\in\mathbb{S}_j}\pi_j(s)Y_j^t(s,d)\big)\big(\sum_{s\in\mathbb{S}_j}\pi_j(s)Y_j^t(s,d')\big) =& \frac{1}{2}\left(\big(\sum_{s\in\mathbb{S}_j}\pi_j(s)Y_j^t(s,d)\big)^2+\big(\sum_{s\in\mathbb{S}_j}\pi_j(s)Y_j^t(s,d')\big)^2\right. \\
&\left.-\big(\sum_{s\in\mathbb{S}_j}\pi_j(s)(Y_j^t(s,d)-Y_j^t(s,d'))\big)^2\right).
\end{aligned}$$

Hence

$$\begin{aligned}
\operatorname{cov}(\tilde{\tau}(d),\tilde{\tau}(d')) =& \sum_{j=1}^{J}\pi_j\frac{\sum_{s\in\mathbb{S}_j}\pi_j(s)Y_j^t(s,d)Y_j^t(s,d')}{m(d,d')} \\
&+\frac{1}{2}\frac{\mathfrak{C}}{J-1}\sum_{j=1}^{J}\pi_j\frac{\big(\sum_{s\in\mathbb{S}_j}\pi_j(s)Y_j^t(s,d)\big)^2+\big(\sum_{s\in\mathbb{S}_j}\pi_j(s)Y_j^t(s,d')\big)^2}{m(d,d')} \\
&+(1+\frac{\mathfrak{C}}{J-1})\sum_{j=1}^{J}\frac{\pi_j^3}{1-\pi_j}\frac{Y_j^c(d)Y_j^c(d')}{m(d,d')} \\
&+\frac{1}{2}(1+\frac{\mathfrak{C}}{J-1})\sum_{j=1}^{J}\pi_j^2\frac{\big(Y_j^c(d)\big)^2+\big(Y_j^c(d')\big)^2}{m(d,d')} \\
&-\frac{1}{2}(1+\frac{\mathfrak{C}}{J-1})\sum_{j=1}^{J}\pi_j^2\frac{\big(\sum_{s\in\mathbb{S}_j}\pi_j(s)Y_j^t(s,d)-Y_j^c(d')\big)^2+\big(\sum_{s\in\mathbb{S}_j}\pi_j(s)Y_j^t(s,d')-Y_j^c(d)\big)^2}{m(d,d')} \\
&+\frac{1}{2}\sum_{j=1}^{J}(1-\mathfrak{C}\frac{1-\pi_j}{\pi_j(J-1)})\pi_j^2\frac{\big(\sum_{s\in\mathbb{S}_j}\pi_j(s)(Y_j^t(s,d)-Y_j^t(s,d'))\big)^2}{m(d,d')}.
\end{aligned}$$

Theorem OA1 follows directly by defining and factorizing the terms given in the theorem.

5 Aggregate Effects

The aggregate effect of a single treatment on all affected individuals is of importance for cost-benefit and welfare analyses. In this section, I propose estimators of aggregate effects that build on the estimators of

individual-level effects of the previous section.

In experiments with spatial treatments, there are two units of observation: outcome individuals and spatial treatments. The treatment effects discussed in the main part of the paper are average effects per outcome individual. The aggregate treatment effects of this section are average effects per spatial treatment.

Suppose the researcher is interested in the aggregate effect that a single treatment location has on all affected individuals. Define the estimand

$$\tau^{agg} \equiv \frac{\sum_{j=1}^{J} \sum_{s \in \mathbb{S}_j} \pi_j \pi_j(s) w_j(s) \sum_{i \in \mathbb{I}_j} \tau_i(s)}{\sum_{j=1}^{J} \pi_j \sum_{s \in \mathbb{S}_j} \pi_j(s) w_j(s)}$$

where, as before, $\tau_i(s) = Y_i(s) - Y_i(0)$ is the effect of treatment location s on individual i. The aggregate treatment effect sums the $\tau_i(s)$ across individuals i and averages them across candidate treatment locations s, with weights $w_j(s)$.

In this section, I focus on the average aggregate treatment effect on the treated, τ^{AATT}, which uses weights $w_j(s) = 1$. The estimand places larger weight on the effects of treatment locations that are more likely to be realized. The estimand τ^{AATT} therefore answers the question: What is the expected aggregate effect of a treatment location under the observed policy of assigning treatments to locations?

One can estimate the aggregate effect τ^{AATT} by aggregating outcomes at the region-level:

$$\hat{\tau}^{AATT,1} \equiv \frac{1}{\sum_{j=1}^{J} \mathcal{W}_j} \sum_{j=1}^{J} \mathcal{W}_j Y_j - \frac{1}{\sum_{j=1}^{J} \frac{(1-\mathcal{W}_j)\pi_j}{1-\pi_j}} \sum_{j=1}^{J} \frac{(1-\mathcal{W}_j)\pi_j}{1-\pi_j} Y_j$$

where $Y_j \equiv \sum_{i \in \mathbb{I}_j} Y_i$. $\hat{\tau}^{AATT,1}$ is the inverse probability weighting estimator of an average treatment effect on the treated, where the outcome variable of interest is the sum of the outcomes of all individuals in a region. When there is a single candidate treatment location per region, standard results from the literature on experiments with individual-level treatments apply (cf. Imbens, 2004), with regions taking the role of individuals.

Estimators based on region-aggregate outcomes are likely to have large variance. Each region-aggregate outcome is the sum of outcomes of individuals in the region. If there is substantial variance in the number of individuals per region and outcomes are positive, the aggregate outcome of regions with many individuals can be substantially larger than the aggregate outcome of smaller regions. For instance, suppose that the number of individuals per region is Poisson distributed with mean n, and individual-level outcomes are i.i.d. within and across regions, with mean μ and variance σ^2. Then region-aggregate outcomes have variance $n \cdot (\sigma^2 + \mu^2)$ by the law of total variance. Hence, aggregate potential outcomes have large variance, which leads to a large variance of the estimator (cf. Imbens, 2004).

Variation in region sizes generates a large variance of the region-aggregate estimator $\hat{\tau}^{AATT,1}$ in two ways. First, if there is variance in the number of individuals per region, then in finite samples, some treatment assignments will be such that there are more individuals in treated regions than in control regions.[6] Suppose outcomes are positive and constant, such that all individuals have the same (positive) value for the outcome. Then the treatment effect estimate $\hat{\tau}^{AATT,1}$ in such a sample is positive and sensitive to the scale of the outcome value. Hence, the estimator $\hat{\tau}^{AATT,1}$ can have a large variance even when there is *no* variance in potential outcomes. Second, variation in region sizes increases the variance in a sampling-of-regions thought experiment. Even if the average individual-level treatment effect was known, needing to estimate the number of times the effect is realized on average per region can increase the variance substantially. The design-based variances considered in this paper condition on the individuals in the sample. With a known number of individuals and a known individual-level average treatment effect, it is possible to form an estimator of aggregate treatment effects with a design-based variance equal to zero, in contrast to the variance results for the estimator $\hat{\tau}^{AATT,1}$ above.

I recommend an estimator of average aggregate effects that reduces the variance by building on the estimators of average individual-level effects at a distance d. Let

$$\hat{\tau}^{AATT,2} \equiv \sum_{d \in \mathbb{D}} \tilde{n}(d) \hat{\tau}(d)$$

[6] Stratification in the experimental design or analysis is an alternative solution to this problem. However, when the number of regions is small or moderate, stratification may not be practical or sufficient to resolve this issue.

where $\tilde{n}(d)$ is the average number of individuals at distance $d \pm h$ from candidate treatment locations:

$$\tilde{n}(d) = \frac{\sum_{j=1}^{J} \sum_{s \in \mathbb{S}_j} \pi_j \pi_j(s) \sum_{i \in \mathbb{I}_j} w_i(s,d)}{\sum_{j=1}^{J} \sum_{s \in \mathbb{S}_j} \pi_j \pi_j(s)}$$

using the same distance bins for both $\hat{\tau}(d)$ and $\bar{n}(d)$, $w_i(s,d) = \mathbb{1}\{|d(s,r_i) - d| \leq h\}$. Here, the choice of distance bin (instead of a different kernel) is essential. The set of distances $\mathbb{D}$ contains the midpoints of the bins that partition the full space into distance bins. For instance, if one uses distance bins $[0,1]$, $(1,2]$, $\ldots$, $(9,10]$ for a treatment that is known not to have effects past a distance of 10 miles, then $\mathbb{D} = \{0.5, 1.5, \ldots, 9.5\}$ and $h = 0.5$.

The theoretical properties of the estimator $\hat{\tau}^{AATT,2}$ follow from those of $\hat{\tau}(d)$ in Theorem 1 of the main text, and the covariance across distances as given in Theorem OA1.

Theorem OA2. *Under Assumptions 1, 3, and 2, the estimator $\hat{\tau}^{AATT,2}$ has an approximate finite population distribution over the assignment distribution with*

(i) unbiasedness: $E(\hat{\tau}^{AATT,2}) \approx \tau^{AATT}$

(ii) variance:

$$\begin{aligned}
\operatorname{var}\Big(\hat{\tau}^{AATT,2}\Big) \approx \sum_{d \in \mathbb{D}} \tilde{n}(d)^2 \Big(\frac{J-1}{J} \frac{\tilde{V}_t^{location}(d)}{J_t} + \frac{\tilde{V}_c^{region}(d)}{J_c} + \frac{1}{J} \frac{\tilde{V}_t^{region}(d)}{J_t} - \frac{\tilde{V}_{ct}^{region}(d)}{J} \Big) \\
+ 2 \sum_{d \in \mathbb{D}} \sum_{d' \in \mathbb{D}, d' \neq d} \tilde{n}(d)\tilde{n}(d') \Big(\frac{J-1}{J} \frac{\tilde{V}_t^{location}(d,d')}{J_t} + \frac{\tilde{V}_c^{region}(d,d')}{J_c} \\
+ \frac{1}{J} \frac{\tilde{V}_t^{region}(d,d')}{J_t} - \frac{\tilde{V}_{ct}^{region}(d,d')}{J} \Big)
\end{aligned}$$

where

$$\begin{aligned}
\tilde{V}_t^{location}(d,d') &\equiv \frac{1}{\bar{n}(d) \cdot \bar{n}(d') \cdot (J-1)} \sum_{j=1}^{J} \sum_{s \in \mathbb{S}_j} \pi_j(s) \Bigg(\Big(\sum_{i \in \mathbb{I}_j} \mathbb{1}\{|d(s,r_i) - d| \leq h\} (Y_i(s) - \mu_t(d)) \Big) \\
&\qquad \cdot \Big(\sum_{i \in \mathbb{I}_j} \mathbb{1}\{|d(s,r_i) - d'| \leq h\} (Y_i(s) - \mu_t(d')) \Big) \Bigg) \\
\tilde{V}_c^{region}(d,d') &\equiv \frac{1}{\bar{n}(d) \cdot \bar{n}(d') \cdot (J-1)} \sum_{j=1}^{J} \Bigg(\Big(\sum_{s \in \mathbb{S}_j} \pi_j(s) \sum_{i \in \mathbb{I}_j} \mathbb{1}\{|d(s,r_i) - d| \leq h\} (Y_i(0) - \mu_c(d)) \Big) \\
&\qquad \cdot \Big(\sum_{s \in \mathbb{S}_j} \pi_j(s) \sum_{i \in \mathbb{I}_j} \mathbb{1}\{|d(s,r_i) - d'| \leq h\} (Y_i(0) - \mu_c(d')) \Big) \Bigg) \\
\tilde{V}_t^{region}(d,d') &\equiv \frac{1}{\bar{n}(d) \cdot \bar{n}(d') \cdot (J-1)} \sum_{j=1}^{J} \Bigg(\Big(\sum_{s \in \mathbb{S}_j} \pi_j(s) \sum_{i \in \mathbb{I}_j} \mathbb{1}\{|d(s,r_i) - d| \leq h\} (Y_i(s) - \mu_t(d)) \Big) \\
&\qquad \cdot \Big(\sum_{s \in \mathbb{S}_j} \pi_j(s) \sum_{i \in \mathbb{I}_j} \mathbb{1}\{|d(s,r_i) - d'| \leq h\} (Y_i(s) - \mu_t(d')) \Big) \Bigg) \\
\tilde{V}_{ct}^{region}(d,d') &\equiv \frac{1}{\bar{n}(d) \cdot \bar{n}(d') \cdot (J-1)} \sum_{j=1}^{J} \Bigg(\Big(\sum_{s \in \mathbb{S}_j} \pi_j(s) \sum_{i \in \mathbb{I}_j} \mathbb{1}\{|d(s,r_i) - d| \leq h\} \big(Y_i(s) - Y_i(0) - (\mu_t(d) - \mu_c(d))\big) \Big) \\
&\qquad \cdot \Big(\sum_{s \in \mathbb{S}_j} \pi_j(s) \sum_{i \in \mathbb{I}_j} \mathbb{1}\{|d(s,r_i) - d'| \leq h\} \big(Y_i(s) - Y_i(0) - (\mu_t(d') - \mu_c(d'))\big) \Big) \Bigg)
\end{aligned}$$

and $\bar{n}(d)$, $\mu_t(d)$, and $\mu_c(d)$ are defined as in Theorem 1.

Proof: *The variance result follows from Theorems 1 and OA1. For approximate unbiasedness, note that* $\tilde{n}(d)$ *is non-stochastic, hence by Theorem 1 and the definition of* $\tilde{n}(d)$

$$\begin{aligned}
E(\hat{\tau}^{AATT,2}) &= \sum_{d\in\mathbb{D}} \tilde{n}(d) E(\hat{\tau}(d)) \\
&\approx \sum_{d\in\mathbb{D}} \tilde{n}(d) \frac{\sum_{j=1}^{J} \sum_{s\in\mathbb{S}_j} \pi_j \pi_j(s) \sum_{i\in\mathbb{I}_j} \mathbb{1}\{|d(s,r_i)-d| \leq h\} \tau_i(s)}{\sum_{j=1}^{J} \sum_{s\in\mathbb{S}_j} \pi_j \pi_j(s) \sum_{i\in\mathbb{I}_j} \mathbb{1}\{|d(s,r_i)-d| \leq h\}} \\
&= \frac{\sum_{j=1}^{J} \sum_{s\in\mathbb{S}_j} \pi_j \pi_j(s) \sum_{d\in\mathbb{D}} \sum_{i\in\mathbb{I}_j} \mathbb{1}\{|d(s,r_i)-d| \leq h\} \tau_i(s)}{\sum_{j=1}^{J} \sum_{s\in\mathbb{S}_j} \pi_j \pi_j(s)} \\
&= \frac{\sum_{j=1}^{J} \sum_{s\in\mathbb{S}_j} \pi_j \pi_j(s) \sum_{i\in\mathbb{I}_j} \tau_i(s)}{\sum_{j=1}^{J} \sum_{s\in\mathbb{S}_j} \pi_j \pi_j(s)} \\
&= \tau^{AATT}
\end{aligned}$$

Remark 1. The optimal choice of distance bins (and bandwidths) remains an open question. If individuals are distributed uniformly across space, equal-width rings with larger radii have a larger area and hence contain more individuals. In practice, in densely populated areas, smaller bins may be preferable, and under suitable sequences of populations (infill asymptotics *and* growing number of regions), it may be possible to allow $h \to 0$ and $|\mathbb{D}| \to \infty$. Generally, in the formula above, additional distance bins decrease the (squared) weights $\tilde{n}(d)$ at the cost of increasing variances $\text{var}\big(\hat{\tau}(d)\big)$.

6 Estimator when only the nearest realized location matters

The identification argument in the proof of Theorem 3 suggests the estimator

$$\begin{aligned}
\hat{\tau}_{\text{nearest}}(d) \equiv & \frac{\sum_{s\in\mathbb{S}} \mathbb{1}\{\mathcal{S} \ni s\} \sum_{i\in\mathbb{I}} \frac{\mathcal{N}_i(s)}{\Pr(\mathcal{N}_i(s)=1|\mathcal{S}\ni s)} w_i(s,d) \mathcal{Y}_i}{\sum_{s\in\mathbb{S}} \mathbb{1}\{\mathcal{S} \ni s\} \sum_{i\in\mathbb{I}} \frac{\mathcal{N}_i(s)}{\Pr(\mathcal{N}_i(s)=1|\mathcal{S}\ni s)} w_i(s,d)} \\
& - \frac{\sum_{s\in\mathbb{S}} \frac{\mathbb{1}\{\mathcal{S}\not\ni s\}}{1-\pi_s} \pi_s \sum_{i\in\mathbb{I}} \frac{\mathcal{N}_i(0)}{\Pr(\mathcal{N}_i(0)=1|\mathcal{S}\not\ni s)} w_i(s,d) \mathcal{Y}_i}{\sum_{s\in\mathbb{S}} \frac{\mathbb{1}\{\mathcal{S}\not\ni s\}}{1-\pi_s} \pi_s \sum_{i\in\mathbb{I}} \frac{\mathcal{N}_i(0)}{\Pr(\mathcal{N}_i(0)=1|\mathcal{S}\not\ni s)} w_i(s,d)}
\end{aligned}$$

where $\mathcal{N}_i(s)$ is an indicator for s being the nearest realized treatment location to i, and $\mathcal{N}_i(0)$ is an indicator for no treatment location within d_0 of i being realized:

$$\begin{aligned}
\mathcal{N}_i(s) &= \mathbb{1}\{\mathcal{S} \ni s\} \prod_{s'\in\mathbb{S}\setminus\{s\}} (1 - \mathbb{1}\{\mathcal{S} \ni s'\})^{\mathbb{1}\{d(s',r_i)<d(s,r_i)\}} \\
\mathcal{N}_i(0) &= \prod_{s\in\mathbb{S}} (1 - \mathbb{1}\{\mathcal{S} \ni s\})^{\mathbb{1}\{d(s,r_i)<d_0\}}
\end{aligned}$$

and the (conditional) probabilities of these events are, under independent assignment,

$$\begin{aligned}
\Pr(\mathcal{N}_i(s) = 1 \mid \mathcal{S} \ni s) &= \prod_{s'\in\mathbb{S}\setminus\{s\}} (1-\pi_s)^{\mathbb{1}\{d(s',r_i)<d(s,r_i)\}} \\
\Pr(\mathcal{N}_i(0) = 1 \mid \mathcal{S} \not\ni s) &= \frac{1}{1-\pi_s} \prod_{s'\in\mathbb{S}} (1-\pi_{s'})^{\mathbb{1}\{d(s,r_i)<d_0\}}.
\end{aligned}$$

It is straightforward to show that $E(\hat{\tau}_{\text{nearest}}(d)) \approx \tau(d)$ and the approximate variance of the estimator can be derived analogously to the previous results.

If the event $\mathcal{N}_i(0)$ is rare, the variance of the $\hat{\tau}_{\text{nearest}}(d)$ will likely be large. The difficulty lies in estimating the weighted mean of $Y_i(0)$. Additive separability allows identifying this mean from differences in exposure, but Assumption 7 only allows using individuals who are unexposed to the treatment (within distance d_0).

Under Assumption 7, the estimator, therefore, tends to use drastically fewer observations, increasing the variance.

There are, effectively, two options for addressing this issue. First, the researcher can impose additional structure. As discussed, under, for instance, Assumption 5, alternative estimators with likely smaller variance are feasible. Other assumptions more in the spirit of Assumption 7 may be conceivable. Second, the researcher can change the target of estimation. Minor improvements in the variance are possible by choosing weights $w_i(s,d)$ including a factor $\Pr(\mathcal{N}_i(s) = 1 \mid \mathcal{S} \ni s)$ or, for interpretation likely less attractively, $\Pr(\mathcal{N}_i(0) = 1 \mid \mathcal{S} \not\ni s)$. More substantial gains arise by changing the estimand to not rely on treatment effects $\tau_i(s)$ but instead build on $\tau_i(s \mid S_i(s))$ for some $S_i(s) \subset \{s' \in \mathbb{S} : \ d(s, r_i) \leq d(s', r_i)\}$. More research is needed to develop recommendations for the choice of $S_i(s)$ with desirable interpretation and inferential properties.

7 Variance Estimation

The variance in Theorem 1 depends on four variances: $\tilde{V}_t^{\text{location}}(d)$, $\tilde{V}_c^{\text{region}}(d)$, $\tilde{V}_t^{\text{region}}(d)$, and $\tilde{V}_{ct}^{\text{region}}(d)$. The first two variances are straightforward to estimate, as given below. The third variance, $\tilde{V}_t^{\text{region}}(d)$ cannot be estimated directly, but is bounded by $\tilde{V}_t^{\text{location}}(d)$. Alternatively, it can be approximated as discussed below. The fourth variance, the variance of treatment effects $\tilde{V}_{ct}^{\text{region}}(d)$, is generally not identified,[7] but since it appears negatively in the overall variance, it can be dropped resulting in a conservative estimator of the variance (cf. Imbens and Rubin, 2015, ch. 6).

A natural estimator of $\tilde{V}_t^{\text{location}}(d)$ is

$$\hat{V}_t^{\text{location}}(d) \equiv \frac{\sum_{j=1}^J \sum_{s\in\mathbb{S}_j} \mathcal{W}_j \mathbb{1}\{\mathcal{S}_j = s\}\Big(\sum_{i\in\mathbb{I}_j} w_i(s,d)(Y_i - \bar{Y}_t(d))\Big)^2}{(J_t - 1)\Big(\frac{1}{J_t}\sum_{j=1}^J \sum_{s\in\mathbb{S}_j} \mathcal{W}_j \mathbb{1}\{\mathcal{S}_j = s\} \sum_{i\in\mathbb{I}_j} w_i(s,d)\Big)^2}$$

which takes the average squared difference from the mean over those individuals who are treated at distance d. Note that while one can calculate $\bar{n}(d)$ exactly, it is likely preferable in practice to use the average number of individuals near treated locations in the sample, which more accurately reflects the averaging in the numerator.

Similarly, a natural estimator of $\tilde{V}_c^{\text{region}}(d)$ is

$$\hat{V}_c^{\text{region}}(d) \equiv \frac{\sum_{j=1}^J (1-\mathcal{W}_j)\Big(\sum_{s\in\mathbb{S}_j} \pi_j(s) \sum_{i\in\mathbb{I}_j} w_i(s,d)(Y_i - \bar{Y}_c(d))\Big)^2}{(J_c - 1)\Big(\frac{1}{J_c}\sum_{j=1}^J (1-\mathcal{W}_j) \sum_{s\in\mathbb{S}_j} \pi_j(s) \sum_{i\in\mathbb{I}_j} w_i(s,d)\Big)^2}.$$

Using $\hat{V}_t^{\text{location}}(d)$ as a conservative estimator of $\tilde{V}_t^{\text{region}}(d)$, the conservative estimator for the variance of the estimator $\hat{\tau}(d)$ is

$$\widehat{\text{var}}_{\text{conservative}}(\hat{\tau}(d)) \equiv \frac{\hat{V}_t^{\text{location}}(d)}{J_t} + \frac{\hat{V}_c^{\text{region}}(d)}{J_c}.$$

If there is reason to believe that there is substantial variance within regions (rather than across regions), it may be preferable to approximate $\tilde{V}_t^{\text{region}}(d)$ directly rather than estimate it conservatively with $\hat{V}_t^{\text{location}}(d)$. Specifically, consider forming the estimator

$$\hat{V}_c^{\text{location}}(d) \equiv \frac{\sum_{j=1}^J (1-\mathcal{W}_j) \sum_{s\in\mathbb{S}_j} \pi_j(s)\Big(\sum_{i\mathbb{I}_j} w_i(s,d)(Y_i - \bar{Y}_c(d))\Big)^2}{(J_c - 1)\Big(\frac{1}{J_c}\sum_{j=1}^J (1-\mathcal{W}_j) \sum_{s\in\mathbb{S}_j} \pi_j(s) \sum_{i\in\mathbb{I}_j} w_i(s,d)\Big)^2}$$

[7] For the variance of treatment effects to be zero (constant treatment effects), the distributions of treated and control must be identical up to a location shift. More generally, the covariance of treated and control potential outcomes is partially identified from the marginal variances. Heckman et al. (1997) use the Fréchet-Hoeffding inequality to form bounds on the variance of treatment effects. Aronow et al. (2014) use the same bounds to improve the Neyman (1923, 1990, cf. Imbens and Rubin (2015)) variance estimator.

which is analogous to $\hat{V}_t^{\text{location}}(d)$ but for control, rather than treated, regions. Under some assumptions, for instance, constant additive or constant multiplicative treatment effects,

$$\frac{\tilde{V}_t^{\text{region}}(d)}{\tilde{V}_t^{\text{location}}(d)} = \frac{\tilde{V}_c^{\text{region}}(d)}{\tilde{V}_c^{\text{location}}(d)}$$

where $\tilde{V}_c^{\text{location}}(d)$ is the appropriate population analogue, such that a plausible estimator for $\tilde{V}_t^{\text{region}}(d)$ is

$$\hat{V}_t^{\text{region}}(d) = \hat{V}_t^{\text{location}}(d) \frac{\hat{V}_c^{\text{region}}(d)}{\hat{V}_c^{\text{location}}(d)}.$$

This estimator uses that the ratio of within-region and across-region variances of treated and control potential outcomes are approximately similar in most settings where the effect of the treatment (and its heterogeneity) is small relative to other sources of variance in the outcome. When the equality of ratios is not exact, deviations can lead to either conservative or anti-conservative estimates of $\tilde{V}_t^{\text{region}}(d)$. In practice, even if the estimator $\hat{V}_t^{\text{region}}(d)$ is not conservative, the variance estimator

$$\widehat{\text{var}}_{\text{approx}}(\hat{\tau}(d)) \equiv \frac{J-1}{J} \frac{\hat{V}_t^{\text{location}}(d)}{J_t} + \frac{\hat{V}_c^{\text{region}}(d)}{J_c} + \frac{1}{J} \frac{\hat{V}_t^{\text{region}}(d)}{J_t}$$

likely is still conservative for $\text{var}(\hat{\tau}(d))$ by the omission of the variance of treatment effects term.

Comparison of the conservative variance estimate, $\widehat{\text{var}}_{\text{conservative}}(\hat{\tau}(d))$, and the variance estimate using the approximation, $\widehat{\text{var}}_{\text{approx}}(\hat{\tau}(d))$, can serve as a plausible benchmark for the benefits any refinements of estimators of $\tilde{V}_t^{\text{region}}(d)$ can plausibly yield. In practice, since $\tilde{V}_t^{\text{region}}(d)$ receives weight $\frac{1}{J}$ relative to the other variances, the difference is likely to be small.

8 Parametric Estimators

I discuss issues in imposing parametric assumptions on the decay of treatment effects over distance from treatment and estimation by least squares regression. First, I show how to impose a parametric model on the individual-level effects at different distances. Second, I show how to estimate aggregate effects based on such a model.

Linear parametric models for the decay of average treatment effects over distance from treatment take the form

$$\tau(d) = \sum_k \beta_k \tilde{\lambda}_k(d)$$

where $\tilde{\lambda}_k$ are known functions of distance, and β_k are coefficients to be estimated.

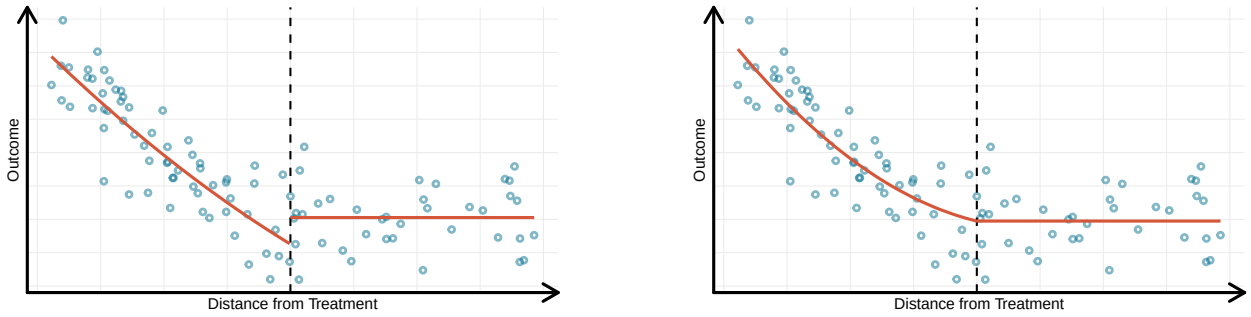
In many settings, one needs to impose a distance after which the treatment has no effect, even within a region, to obtain reasonable estimates from parametric models. Assumption 6 in the main text formalizes this assumption. Without such a restriction, any simple functional form for $\tilde{\lambda}$ will typically offer a poor approximation for at least some distances d from treatment.

One can improve the approximation to the treatment effect at short distances by using functions that only fit the treatment effect pattern up to the maximum distance d_0:

$$\tau(d) = \sum_k \beta_k \lambda_k(d) \mathbb{1}\{d \leq d_0\}.$$

Relatively simple functions λ_k may well approximate the average treatment effects at distances $d \in (0, d_0)$. The assumption resembles a "bet on sparsity" (Hastie et al., 2001): If treatment effects are negligible at distances longer than d_0, the estimators proposed below will likely perform well. If treatment effects are not negligible even at long distances, then no (parametric) estimator will perform well.

For instance, one can impose a linear functional form on the treatment effect decay by choosing $\lambda_1(d) = 1$, $\lambda_2(d) = d$. The coefficient β_2 then measures the rate of decay, while β_1 measures the effect of the treatment on individuals right by the treatment location. A quadratic functional form is imposed by $\lambda_1(d) = 1$,



(a) Unrestricted estimator (b) Restricted estimator

Figure OA2: Illustration of imposing continuous effects in distance at d_0. The figure shows a scatter plot of outcomes against distance from treatment with an unrestricted and restricted quadratic fit superimposed.

$\lambda_2(d) = d$, $\lambda_3(d) = d^2$. In principle, the analysis in this section can be extended also to functional forms that are non-linear in the parameters, such as exponential decay of treatment effects with an unknown rate of decay, $\tau(d) = \exp(-\beta d)$.

To estimate the parameter β, suppose initially that there is only a single candidate treatment location in each region. Then one can define the distance of individual i from the candidate treatment location uniquely as d_i, irrespective of realized treatment. Then estimate the *weighted* linear regression

$$Y_i = \sum_k \beta_k \Big(\mathcal{W}_{j(i)} \lambda_k(d_i, x_i) \mathbb{1}\{d_i \leq d_0\} \Big) + h(d_i) + \epsilon_i$$

with ATT inverse probability weights (weight 1 on the treated, weight $\pi_j/(1-\pi_j)$ on the control). The functions $\lambda_k(d_i, x_i)$ can depend on individual characteristics x_i to allow for heterogeneity in effects, such as separate λ_k for distinct groups of individuals.

The function h models the average control potential outcomes at each distance from candidate treatment locations. For semiparametric estimation, specify the treatment effect decay (λ) parametrically, and estimate h nonparametrically, as a partially linear model (e.g. Robinson, 1988). Here, I instead focus on parametric linear estimation, which imposes known parametric functions λ and h and estimates their coefficients:

$$Y_i = \alpha_0 + \sum_k \beta_k \Big(\mathcal{W}_{j(i)} \lambda_k(d_i, x_i) \mathbb{1}\{d_i \leq d_0\} \Big) + \sum_\ell \gamma_\ell \Big(h_\ell(d_i) \mathbb{1}\{d_i \leq d_0\} \Big) + \epsilon_i.$$

The same caveat about setting a maximum distance applies also to h. Since there is no interest in effects at distances larger than d_0, the constant α_0 captures the mean outcome for individuals at these larger distances.

In practice, one typically not only wants to impose a zero treatment effect after distance d_0, but a treatment effect that tends to zero continuously at d_0.[8] To this end, estimate the linear regression with transformed covariates

$$Y_i = \alpha_0 + \sum_k \beta_k \Big(\mathcal{W}_{j(i)} (\lambda_k(d_i, x_i) - \lambda_k(d_0, x_i)) \mathbb{1}\{d_i \leq d_0\} \Big) + \sum_\ell \gamma_\ell \Big(h_\ell(d_i) \mathbb{1}\{d_i \leq d_0\} \Big) + \epsilon_i$$

which imposes the restriction $\tau(d_0) = \sum_k \beta_k \lambda_k(d_0) = 0$. Figure OA2 illustrates what it means to impose this restriction. In panel (a), without the restriction, the estimated treatment effect will jump to 0 discontinuously at d_0. Imposing the restriction in panel (b), the estimated treatment effect is continuous also at d_0. The restriction generally reduces the variance of the estimator, in particular for estimating aggregate effects, as discussed below. In practice, most functional forms for λ imply not just a zero effect after distance d_0, but also a non-zero effect at distances slightly shorter than d_0.

[8] In principle, one could additionally impose higher-order smoothness such as differentiability at d_0. However, higher-order smoothness generally requires more complicated functional forms λ to retain sufficient flexibility at shorter distances. In practice, more complicated functional forms likely negate any improvements in precision.

The same parametric functional form can be imposed to estimate the average aggregate effects of the treatment. Under the parametric model, the average aggregate treatment effect on the treated is

$$\tau^{AATT} = \frac{1}{J} \sum_i \sum_k \beta_k (\lambda_k(d_i, x_i) - \lambda_k(d_0, x_i)) \mathbb{1}\{d_i \leq d_0\}$$

Solving for β_1 and substituting the resulting expression in the regression specification above, one obtains the one-step regression specification

$$\begin{aligned} Y_i = \alpha_0 + \tau^{AATT} & \Big(\mathcal{W}_{j(i)} \frac{(\lambda_1(d_i, x_i) - \lambda_1(d_0, x_i)) \mathbb{1}\{d_i \leq d_0\}}{\frac{1}{J} \sum_{i'} (\lambda_1(d_{i'}, x_{i'}) - \lambda_1(d_0, x_{i'})) \mathbb{1}\{d_{i'} \leq d_0\}} \Big) \\ & + \sum_k \beta_k \Bigg(\Big(\mathcal{W}_{j(i)} (\lambda_k(d_i, x_i) - \lambda_k(d_0, x_i)) \mathbb{1}\{d_i \leq d_0\} \Big) \\ & \qquad - \Big(\mathcal{W}_{j(i)} (\lambda_k(d_i, x_i) - \lambda_k(d_0, x_i)) \mathbb{1}\{d_i \leq d_0\} \\ & \qquad \cdot \frac{\frac{1}{J} \sum_{i'} (\lambda_k(d_{i'}, x_{i'}) - \lambda_k(d_0, x_{i'})) \mathbb{1}\{d_{i'} \leq d_0\}}{\frac{1}{J} \sum_{i'} (\lambda_1(d_{i'}, x_{i'}) - \lambda_1(d_0, x_{i'})) \mathbb{1}\{d_{i'} \leq d_0\}} \Big) \Bigg) \\ & + \sum_\ell \gamma_\ell \Big(h_\ell(d_i) \mathbb{1}\{d_i \leq d_0\} \Big) + \epsilon_i \end{aligned}$$

where the coefficient on the first (transformed) covariate is the estimate of the average aggregate treatment effect. The transformed covariates are readily computed by realizing they are equal to the original covariates multiplied or shifted by average covariates. The average here is taken across all regions, both treated and untreated, such that this estimate has similarly attractive properties as the nonparametric estimator $\hat{\tau}^{AATT,2}$ above, in leveraging that the *number* of individuals near candidate treatment locations are available irrespective of assignment.

When there is more than one candidate treatment location per region, augment the regression approach as follows. The variable d_i is not uniquely defined, since there are multiple "distances from candidate treatment locations" for individuals. Suppose individual i in a control region ($\mathcal{W}_{j(i)} = 0$) is 1 mile away from one candidate treatment location and 5 miles away from a different candidate treatment location. Then i should be used to estimate the control mean $h(d)$ for the two distances $d = 1$ and $d = 5$. One can therefore duplicate observation i. Specifically, if individual i is in a region with $|\mathbb{S}_{j(i)}|$ candidate treatment locations, then include i $|\mathbb{S}_{j(i)}|$ times in the regression. Each version of i uses the distance d_i to a different candidate treatment location. Observations in control regions at their distance relative to candidate locations s then receive ATT inverse probability weights $\pi_j \pi_j(s)/(1 - \pi_j)$ to ensure $E(\epsilon_i | d_i = d, x_i = x) = 0$.

Simulations (not reported) suggest that standard errors clustered at the region level (cf. Liang and Zeger, 1986) provide a reasonable, but perhaps conservative, estimate of the variance of these estimators. One can derive formal results along the lines of Abadie et al. (2020, 2017). When there is a single candidate treatment location per region and a single distance of interest, the spatial setting considered here coincides with the setting of clustered assignment of Abadie et al. (2017), and hence their results and interpretation of Liang and Zeger (1986) clustered standard errors follow immediately. Refinements of Liang and Zeger (1986) clustered standard errors may be possible following Abadie et al. (2020) for the non-clustered setting using "attributes." In the spatial setting, such attributes are readily available in the form of the *number* of units near candidate treatment locations. Effectively, one can form a tighter bound on the variance of treatment effects using these attributes by exploiting heterogeneous treatment effects and appealing to the law of total variance to maintain that the estimator is still conservative for the true variance. For parametric models of the treatment effect by distance, one needs to extend the analysis of Abadie et al. (2017) to include (multiple) continuous regressors that are deterministic functions of the binary, randomly assigned, treatment. With multiple candidate treatment locations per region, one further needs to extend the binary treatment to a multi-valued (but still discrete) treatment.

9 Variance in Single Region Settings

Write the infeasible estimator as:

$$\begin{aligned}\tilde{\tau} &= \mu_t - \mu_c + \frac{\sum_{s\in\mathbb{S}} \mathbb{1}\{\mathcal{S} \ni s\} \sum_{i\in\mathbb{I}} w_i(s,d)(\mathcal{Y}_i - \mu_t)}{\sum_{s\in\mathbb{S}} \pi_s \sum_{i\in\mathbb{I}} w_i(s,d)} - \frac{\sum_{s\in\mathbb{S}} \mathbb{1}\{\mathcal{S} \not\ni s\} \frac{\pi_s}{1-\pi_s} \sum_{i\in\mathbb{I}} w_i(s,d)(\mathcal{Y}_i - \mu_c)}{\sum_{s\in\mathbb{S}} \pi_s \sum_{i\in\mathbb{I}} w_i(s,d)} \\ &= \mu_t - \mu_c + \frac{\sum_{s\in\mathbb{S}} \mathbb{1}\{\mathcal{S} \ni s\} \sum_{i\in\mathbb{I}} w_i(s,d)(\mathcal{Y}_i - \mu_t) - \sum_{s\in\mathbb{S}} \mathbb{1}\{\mathcal{S} \not\ni s\} \frac{\pi_s}{1-\pi_s} \sum_{i\in\mathbb{I}} w_i(s,d)(\mathcal{Y}_i - \mu_c)}{\sum_{s\in\mathbb{S}} \pi_s \sum_{i\in\mathbb{I}} w_i(s,d)}\end{aligned}$$

where, for brevity, I suppress the dependence of μ on d throughout.

Define exposure mappings (Aronow and Samii, 2017) based on Assumption 6 as follows. $\mathbb{M}_i \equiv 2^{\{s\in\mathbb{S}:\, d(s,r_i)\le d_0\}}$ is the set of all possible ways in which treatment can be assigned to those locations that possibly affect i. With slight abuse of notation, denote i's potential outcome under exposure $m \in \mathbb{M}_i$ by $Y_i(m)$. Let the random variable $\mathcal{M}_m^i$ be the indicator for whether exposure m of individual i is realized. Then $\mathcal{Y}_i = \sum_{m\in\mathbb{M}_i} \mathcal{M}_m^i Y_i(m)$. Denote the marginal and joint probabilities of exposures by $\pi_m^i \equiv \Pr(\mathcal{M}_m^i = 1)$ and $\pi_{m,m'}^{i,i'} \equiv \Pr(\mathcal{M}_m^i = 1 \text{ and } \mathcal{M}_{m'}^{i'} = 1)$. Let

$$\mathcal{T}_s^a \equiv \begin{cases} 1 & \text{if } a = t \text{ and } \mathcal{S} \ni s \\ 1 & \text{if } a = c \text{ and } \mathcal{S} \not\ni s \\ 0 & \text{otherwise} \end{cases}$$

be an indicator for the events $\mathcal{S} \ni s$ $(a = t)$ and $\mathcal{S} \not\ni s$ $(a = c)$.

For the variance of the estimator, note that only the numerator of the ratio in the definition of $\tilde{\tau}$ is stochastic. Using the definitions above, rewrite the numerator:

$$\begin{aligned}&\sum_{s\in\mathbb{S}} \mathbb{1}\{\mathcal{S} \ni s\} \sum_{i\in\mathbb{I}} w_i(s,d)(\mathcal{Y}_i - \mu_t) - \sum_{s\in\mathbb{S}} \mathbb{1}\{\mathcal{S} \not\ni s\} \frac{\pi_s}{1-\pi_s} \sum_{i\in\mathbb{I}} w_i(s,d)(\mathcal{Y}_i - \mu_c) \\ =&\sum_{s\in\mathbb{S}} \sum_{a\in\{c,t\}} \mathcal{T}_s^a \Big(\mathbb{1}\{a = t\} \sum_{i\in\mathbb{I}} w_i(s,d)(\mathcal{Y}_i - \mu_t) - \mathbb{1}\{a = c\} \frac{\pi_s}{1-\pi_s} \sum_{i\in\mathbb{I}} w_i(s,d)(\mathcal{Y}_i - \mu_c) \Big) \\ =&\sum_{i\in\mathbb{I}} \sum_{m\in\mathbb{M}_i} \sum_{s\in\mathbb{S}} \sum_{a\in\{c,t\}} \mathcal{M}_i^m \mathcal{T}_s^a \Big(\mathbb{1}\{a = t\} w_i(s,d)(Y_i(m) - \mu_t) - \mathbb{1}\{a = c\} \frac{\pi_s}{1-\pi_s} w_i(s,d)(Y_i(m) - \mu_c) \Big)\end{aligned}$$

where, importantly, only $\mathcal{M}_i^m \mathcal{T}_{s,a}$ is stochastic. For ease of notation, define

$$\begin{aligned}\tilde{Y}_i^{s,a}(m) &\equiv \mathbb{1}\{a = t\} w_i(s,d)(Y_i(m) - \mu_t) - \mathbb{1}\{a = c\} \frac{\pi_s}{1-\pi_s} w_i(s,d)(Y_i(m) - \mu_c) \\ &= \Big(-\frac{\pi_s}{1-\pi_s} \Big)^{\mathbb{1}\{a=c\}} w_i(s,d)(Y_i(m) - \mu_a)\end{aligned}$$

where, for brevity, I suppress the dependence of $\tilde{Y}$ on d throughout.

Then

$$\begin{aligned}&\operatorname{var}\Big(\sum_{i\in\mathbb{I}} \sum_{m\in\mathbb{M}_i} \sum_{s\in\mathbb{S}} \sum_{a\in\{c,t\}} \mathcal{M}_i^m \mathcal{T}_s^a \tilde{Y}_i^{s,a}(m) \Big) \\ =&\sum_{i\in\mathbb{I}} \sum_{m\in\mathbb{M}_i} \sum_{s\in\mathbb{S}} \sum_{a\in\{c,t\}} \sum_{i'\in\mathbb{I}} \sum_{m'\in\mathbb{M}_{i'}} \sum_{s'\in\mathbb{S}} \sum_{a'\in\{c,t\}} \operatorname{cov}(\mathcal{M}_i^m \mathcal{T}_s^a, \mathcal{M}_{i'}^{m'} \mathcal{T}_{s'}^{a'}) \tilde{Y}_i^{s,a}(m) \tilde{Y}_{i'}^{s',a'}(m') \\ =&\sum_{i\in\mathbb{I}} \sum_{m\in\mathbb{M}_i} \sum_{s\in\mathbb{S}} \sum_{a\in\{c,t\}} \operatorname{var}(\mathcal{M}_i^m \mathcal{T}_s^a) \tilde{Y}_i^{s,a}(m)^2 \\ &+\sum_{i\in\mathbb{I}} \sum_{m\in\mathbb{M}_i} \sum_{s\in\mathbb{S}} \sum_{a\in\{c,t\}} \sum_{s'\in\mathbb{S}} \sum_{a'\in\{c,t\}} \mathbb{1}\{s \ne s' \text{ or } a \ne a'\} \operatorname{cov}(\mathcal{M}_i^m \mathcal{T}_s^a, \mathcal{M}_i^m \mathcal{T}_{s'}^{a'}) \tilde{Y}_i^{s,a}(m) \tilde{Y}_i^{s',a'}(m) \\ &+\sum_{i\in\mathbb{I}} \sum_{m\in\mathbb{M}_i} \sum_{s\in\mathbb{S}} \sum_{a\in\{c,t\}} \sum_{m'\in\mathbb{M}_i} \sum_{s'\in\mathbb{S}} \sum_{a'\in\{c,t\}} \mathbb{1}\{m \ne m'\} \operatorname{cov}(\mathcal{M}_i^m \mathcal{T}_s^a, \mathcal{M}_i^{m'} \mathcal{T}_{s'}^{a'}) \tilde{Y}_i^{s,a}(m) \tilde{Y}_i^{s',a'}(m') \\ &+\sum_{i\in\mathbb{I}} \sum_{m\in\mathbb{M}_i} \sum_{s\in\mathbb{S}} \sum_{a\in\{c,t\}} \sum_{i'\in\mathbb{I}} \sum_{m'\in\mathbb{M}_i} \sum_{s'\in\mathbb{S}} \sum_{a'\in\{c,t\}} \mathbb{1}\{i \ne i'\} \operatorname{cov}(\mathcal{M}_i^m \mathcal{T}_s^a, \mathcal{M}_{i'}^{m'} \mathcal{T}_{s'}^{a'}) \tilde{Y}_i^{s,a}(m) \tilde{Y}_{i'}^{s',a'}(m').\end{aligned} \tag{OA1}$$

Define

$$\pi_{i,s}^{m,a} \equiv \Pr(\mathcal{M}_i^m \mathcal{T}_s^a = 1) \qquad \pi_{i,s,i',s'}^{m,a,m',a'} \equiv \Pr(\mathcal{M}_m^i \mathcal{T}_s^a = 1 \text{ and } \mathcal{M}_{m'}^{i'} \mathcal{T}_{s'}^{a'} = 1)$$

such that $\text{cov}(\mathcal{M}_i^m \mathcal{T}_s^a, \mathcal{M}_{i'}^{m'} \mathcal{T}_{s'}^{a'}) = \pi_{i,s,i',s'}^{m,a,m',a'} - \pi_{i,s}^{m,a} \pi_{i',s'}^{m',a'}$ and $\text{var}(\mathcal{M}_i^m \mathcal{T}_s^a) = \pi_{i,s}^{m,a}(1 - \pi_{i,s}^{m,a})$.

Initially consider the first two (lines of) summations in the final expression in Equation (OA1), which each have a single summation over i and m. Substituting the (co-) variances and then the definitions of $\tilde{Y}_i^{s,a}(m)$ yields

$$\begin{aligned}
&\sum_{i\in\mathbb{I}} \sum_{m\in\mathbb{M}_i} \sum_{s\in\mathbb{S}} \sum_{a\in\{c,t\}} \text{var}(\mathcal{M}_i^m \mathcal{T}_s^a) \tilde{Y}_i^{s,a}(m) \\
&\quad + \sum_{i\in\mathbb{I}} \sum_{m\in\mathbb{M}_i} \sum_{s\in\mathbb{S}} \sum_{a\in\{c,t\}} \sum_{s'\in\mathbb{S}} \sum_{a'\in\{c,t\}} \mathbb{1}\{s \neq s' \text{ or } a \neq a'\} \text{cov}(\mathcal{M}_i^m \mathcal{T}_s^a, \mathcal{M}_i^m \mathcal{T}_{s'}^{a'}) \tilde{Y}_i^{s,a}(m) \tilde{Y}_i^{s',a'}(m) \\
=&\sum_{i\in\mathbb{I}} \sum_{m\in\mathbb{M}_i} \sum_{s\in\mathbb{S}} \sum_{a\in\{c,t\}} \pi_{i,s}^{m,a}(1 - \pi_{i,s}^{m,a}) \tilde{Y}_i^{s,a}(m)^2 \\
&\quad + \sum_{i\in\mathbb{I}} \sum_{m\in\mathbb{M}_i} \sum_{s\in\mathbb{S}} \sum_{a\in\{c,t\}} \sum_{s'\in\mathbb{S}} \sum_{a'\in\{c,t\}} \mathbb{1}\{s \neq s' \text{ or } a \neq a'\} \big(\pi_{i,s,i,s'}^{m,a,m,a'} - \pi_{i,s}^{m,a} \pi_{i,s'}^{m,a'}\big) \tilde{Y}_i^{s,a}(m) \tilde{Y}_i^{s',a'}(m) \qquad \text{(OA2)} \\
=&\sum_{i\in\mathbb{I}} \sum_{m\in\mathbb{M}_i} \sum_{s\in\mathbb{S}} \sum_{a\in\{c,t\}} \pi_{i,s}^{m,a} w_i(s,d)(Y_i(m) - \mu_a)^2 \cdot \Big((1 - \pi_{i,s}^{m,a}) \Big(\frac{\pi_s}{1 - \pi_s}\Big)^{2 \cdot \mathbb{1}\{a=c\}} w_i(s,d)\Big) \\
&\quad + \sum_{i\in\mathbb{I}} \sum_{m\in\mathbb{M}_i} \sum_{s\in\mathbb{S}} \sum_{a\in\{c,t\}} \sum_{s'\in\mathbb{S}} \sum_{a'\in\{c,t\}} \mathbb{1}\{s \neq s' \text{ or } a \neq a'\} \big(\pi_{i,s,i,s'}^{m,a,m,a'} - \pi_{i,s}^{m,a} \pi_{i,s'}^{m,a'}\big) (-1)^{\mathbb{1}\{a \neq a'\}} \\
&\qquad \cdot \Big(\frac{\pi_s}{1 - \pi_s}\Big)^{\mathbb{1}\{a=c\}} \Big(\frac{\pi_{s'}}{1 - \pi_{s'}}\Big)^{\mathbb{1}\{a'=c\}} w_i(s,d) w_i(s',d)(Y_i(m) - \mu_a)(Y_i(m) - \mu_{a'}).
\end{aligned}$$

Next, consider the summations in the third ($m \neq m'$) and fourth ($i \neq i'$) lines of the final expression in Equation (OA1). Separate these summations based on whether $\mathcal{M}_i^m \mathcal{M}_{i'}^{m'} = 0$ with probability 1, such that $\pi_{i,i'}^{m,m'} = 0$. For any given treatment assignment, only the potential outcome corresponding to a single exposure of each individual is observed. Hence, for $m \neq m'$, $\mathcal{M}_i^m \mathcal{M}_i^{m'} = 0$ with probability 1, and, by definition, $\pi_{i,s,i,s'}^{m,a,m',a'} = 0$ irrespective of s, s', a, a'. Similarly, even when $i \neq i$, $\mathcal{M}_i^m \mathcal{M}_{i'}^{m'} = 0$ with probability 1 for some i, m, i', m' if there is at least one candidate treatment location that can affect both i and i' and m and m' correspond to different assignments for such a location. Then, by definition, $\pi_{i,i'}^{m,m'} = 0$ and also $\pi_{i,s,i',s'}^{m,a,m',a'} = 0$. Hence,

$$\begin{aligned}
&\sum_{i\in\mathbb{I}} \sum_{m\in\mathbb{M}_i} \sum_{s\in\mathbb{S}} \sum_{a\in\{c,t\}} \sum_{m'\in\mathbb{M}_i} \sum_{s'\in\mathbb{S}} \sum_{a'\in\{c,t\}} \mathbb{1}\{m \neq m'\} \text{cov}(\mathcal{M}_i^m \mathcal{T}_s^a, \mathcal{M}_i^{m'} \mathcal{T}_{s'}^{a'}) \tilde{Y}_i^{s,a}(m) \tilde{Y}_i^{s',a'}(m') \\
&\quad + \sum_{i\in\mathbb{I}} \sum_{m\in\mathbb{M}_i} \sum_{s\in\mathbb{S}} \sum_{a\in\{c,t\}} \sum_{i'\in\mathbb{I}} \sum_{m'\in\mathbb{M}_i} \sum_{s'\in\mathbb{S}} \sum_{a'\in\{c,t\}} \mathbb{1}\{i \neq i'\} \text{cov}(\mathcal{M}_i^m \mathcal{T}_s^a, \mathcal{M}_{i'}^{m'} \mathcal{T}_{s'}^{a'}) \tilde{Y}_i^{s,a}(m) \tilde{Y}_{i'}^{s',a'}(m') \\
=&-\sum_{i\in\mathbb{I}} \sum_{m\in\mathbb{M}_i} \sum_{s\in\mathbb{S}} \sum_{a\in\{c,t\}} \sum_{m'\in\mathbb{M}_i} \sum_{s'\in\mathbb{S}} \sum_{a'\in\{c,t\}} \mathbb{1}\{m \neq m'\} \pi_{i,s}^{m,a} \pi_{i,s'}^{m',a'} \tilde{Y}_i^{s,a}(m) \tilde{Y}_i^{s',a'}(m') \\
&\quad - \sum_{i\in\mathbb{I}} \sum_{m\in\mathbb{M}_i} \sum_{s\in\mathbb{S}} \sum_{a\in\{c,t\}} \sum_{i'\in\mathbb{I}} \sum_{m'\in\mathbb{M}_i} \sum_{s'\in\mathbb{S}} \sum_{a'\in\{c,t\}} \mathbb{1}\{i \neq i'\} \mathbb{1}\{\pi_{i,i'}^{m,m'} = 0\} \pi_{i,s}^{m,a} \pi_{i',s'}^{m',a'} \tilde{Y}_i^{s,a}(m) \tilde{Y}_{i'}^{s',a'}(m') \\
&\quad + \sum_{i\in\mathbb{I}} \sum_{m\in\mathbb{M}_i} \sum_{s\in\mathbb{S}} \sum_{a\in\{c,t\}} \sum_{i'\in\mathbb{I}} \sum_{m'\in\mathbb{M}_i} \sum_{s'\in\mathbb{S}} \sum_{a'\in\{c,t\}} \mathbb{1}\{i \neq i'\} \mathbb{1}\{\pi_{i,i'}^{m,m'} > 0\} \big(\pi_{i,s,i',s'}^{m,a,m',a'} - \pi_{i,s}^{m,a} \pi_{i',s'}^{m',a'}\big) \tilde{Y}_i^{s,a}(m) \tilde{Y}_{i'}^{s',a'}(m').
\end{aligned}$$

The first line equals exactly the "missing" $i = i'$ terms of the second line because $\pi_{i,i}^{m,m'} = 0$ if and only if $m \neq m'$. Combining these lines, it is then convenient to treat cases $a = a'$ and $a \neq a'$ separately because the sign of the terms multiplying potential outcomes $Y_i(m)Y_{i'}(m')$ differs across the two cases such that they need to be bounded differently (in estimation because the potential outcomes cannot be observed simultaneously

for conflicting exposures). The expression above, therefore, equals

$$
\begin{aligned}
=&-\sum_{i\in\mathbb{I}}\sum_{m\in\mathbb{M}_i}\sum_{s\in\mathbb{S}}\sum_{a\in\{c,t\}}\sum_{i'\in\mathbb{I}}\sum_{m'\in\mathbb{M}_i}\sum_{s'\in\mathbb{S}}\mathbb{1}\{\pi_{i,i'}^{m,m'}=0\}\pi_{i,s}^{m,a}\pi_{i',s'}^{m',a}\tilde{Y}_i^{s,a}(m)\tilde{Y}_{i'}^{s',a}(m')\\
&-2\sum_{i\in\mathbb{I}}\sum_{m\in\mathbb{M}_i}\sum_{s\in\mathbb{S}}\sum_{i'\in\mathbb{I}}\sum_{m'\in\mathbb{M}_i}\sum_{s'\in\mathbb{S}}\mathbb{1}\{\pi_{i,i'}^{m,m'}=0\}\pi_{i,s}^{m,t}\pi_{i',s'}^{m',c}\tilde{Y}_i^{s,t}(m)\tilde{Y}_{i'}^{s',c}(m')\\
&+\sum_{i\in\mathbb{I}}\sum_{m\in\mathbb{M}_i}\sum_{s\in\mathbb{S}}\sum_{a\in\{c,t\}}\sum_{i'\in\mathbb{I}}\sum_{m'\in\mathbb{M}_i}\sum_{s'\in\mathbb{S}}\sum_{a'\in\{c,t\}}\mathbb{1}\{i\neq i'\}\mathbb{1}\{\pi_{i,i'}^{m,m'}>0\}\big(\pi_{i,s,i',s'}^{m,a,m',a'}-\pi_{i,s}^{m,a}\pi_{i',s'}^{m',a'}\big)\tilde{Y}_i^{s,a}(m)\tilde{Y}_{i'}^{s',a'}(m').
\end{aligned}
\tag{OA3}
$$

Substituting for $\tilde{Y}_i^{s,a}(m)$, the products $\tilde{Y}_i^{s,a}(m)\tilde{Y}_{i'}^{s',a'}(m')$ are

$$
\begin{aligned}
\tilde{Y}_i^{s,a}(m)\tilde{Y}_{i'}^{s',a}(m')&=\Big(\frac{\pi_s}{1-\pi_s}\frac{\pi_{s'}}{1-\pi_{s'}}\Big)^{\mathbb{1}\{a=c\}}w_i(s,d)w_{i'}(s',d)(Y_i(m)-\mu_a)(Y_{i'}(m')-\mu_a)\\
\tilde{Y}_i^{s,t}(m)\tilde{Y}_{i'}^{s',c}(m')&=-\frac{\pi_{s'}}{1-\pi_{s'}}w_i(s,d)w_{i'}(s',d)(Y_i(m)-\mu_t)(Y_{i'}(m')-\mu_c),
\end{aligned}
$$

and using the first and second binomial formulas:

$$
\begin{aligned}
-(Y_i(m)-\mu_a)(Y_{i'}(m')-\mu_a)&=\frac{1}{2}(Y_i(m)-\mu_a)^2+\frac{1}{2}(Y_{i'}(m')-\mu_a)^2-2(\frac{Y_i(m)+Y_{i'}(m')}{2}-\mu_a)^2\\
2(Y_i(m)-\mu_t)(Y_{i'}(m')-\mu_c)&=(Y_i(m)-\mu_t)^2+(Y_{i'}(m')-\mu_c)^2-\big((Y_i(m)-Y_{i'}(m'))-(\mu_t-\mu_c)\big)^2.
\end{aligned}
$$

Substituting these equations sequentially into the first and second line of Equation (OA3):

$$
\begin{aligned}
=&-\sum_{i\in\mathbb{I}}\sum_{m\in\mathbb{M}_i}\sum_{s\in\mathbb{S}}\sum_{a\in\{c,t\}}\sum_{i'\in\mathbb{I}}\sum_{m'\in\mathbb{M}_i}\sum_{s'\in\mathbb{S}}\mathbb{1}\{\pi_{i,i'}^{m,m'}=0\}\pi_{i,s}^{m,a}\pi_{i',s'}^{m',a}\Big(\frac{\pi_s}{1-\pi_s}\frac{\pi_{s'}}{1-\pi_{s'}}\Big)^{\mathbb{1}\{a=c\}}\\
&\qquad\cdot w_i(s,d)w_{i'}(s',d)(Y_i(m)-\mu_a)(Y_{i'}(m')-\mu_a)\\
&+2\sum_{i\in\mathbb{I}}\sum_{m\in\mathbb{M}_i}\sum_{s\in\mathbb{S}}\sum_{i'\in\mathbb{I}}\sum_{m'\in\mathbb{M}_i}\sum_{s'\in\mathbb{S}}\mathbb{1}\{\pi_{i,i'}^{m,m'}=0\}\pi_{i,s}^{m,t}\pi_{i',s'}^{m',c}\frac{\pi_{s'}}{1-\pi_{s'}}\\
&\qquad\cdot w_i(s,d)w_{i'}(s',d)(Y_i(m)-\mu_t)(Y_{i'}(m')-\mu_c)\\
&+\sum_{i\in\mathbb{I}}\sum_{m\in\mathbb{M}_i}\sum_{s\in\mathbb{S}}\sum_{a\in\{c,t\}}\sum_{i'\in\mathbb{I}}\sum_{m'\in\mathbb{M}_i}\sum_{s'\in\mathbb{S}}\sum_{a'\in\{c,t\}}\mathbb{1}\{i\neq i'\}\mathbb{1}\{\pi_{i,i'}^{m,m'}>0\}\big(\pi_{i,s,i',s'}^{m,a,m',a'}-\pi_{i,s}^{m,a}\pi_{i',s'}^{m',a'}\big)\tilde{Y}_i^{s,a}(m)\tilde{Y}_{i'}^{s',a'}(m')\\
=&\sum_{i\in\mathbb{I}}\sum_{m\in\mathbb{M}_i}\sum_{s\in\mathbb{S}}\sum_{a\in\{c,t\}}\sum_{i'\in\mathbb{I}}\sum_{m'\in\mathbb{M}_i}\sum_{s'\in\mathbb{S}}\mathbb{1}\{\pi_{i,i'}^{m,m'}=0\}\pi_{i,s}^{m,a}\pi_{i',s'}^{m',a}\Big(\frac{\pi_s}{1-\pi_s}\frac{\pi_{s'}}{1-\pi_{s'}}\Big)^{\mathbb{1}\{a=c\}}\\
&\qquad\cdot w_i(s,d)w_{i'}(s',d)\Big(\frac{1}{2}(Y_i(m)-\mu_a)^2+\frac{1}{2}(Y_{i'}(m')-\mu_a)^2-2\big(\frac{Y_i(m)+Y_{i'}(m')}{2}-\mu_a\big)^2\Big)\\
&+\sum_{i\in\mathbb{I}}\sum_{m\in\mathbb{M}_i}\sum_{s\in\mathbb{S}}\sum_{i'\in\mathbb{I}}\sum_{m'\in\mathbb{M}_i}\sum_{s'\in\mathbb{S}}\mathbb{1}\{\pi_{i,i'}^{m,m'}=0\}\pi_{i,s}^{m,t}\pi_{i',s'}^{m',c}\frac{\pi_{s'}}{1-\pi_{s'}}\\
&\qquad\cdot w_i(s,d)w_{i'}(s',d)\Big((Y_i(m)-\mu_t)^2+(Y_{i'}(m')-\mu_c)^2-\big((Y_i(m)-Y_{i'}(m'))-(\mu_t-\mu_c)\big)^2\Big)\\
&+\sum_{i\in\mathbb{I}}\sum_{m\in\mathbb{M}_i}\sum_{s\in\mathbb{S}}\sum_{a\in\{c,t\}}\sum_{i'\in\mathbb{I}}\sum_{m'\in\mathbb{M}_i}\sum_{s'\in\mathbb{S}}\sum_{a'\in\{c,t\}}\mathbb{1}\{i\neq i'\}\mathbb{1}\{\pi_{i,i'}^{m,m'}>0\}\big(\pi_{i,s,i',s'}^{m,a,m',a'}-\pi_{i,s}^{m,a}\pi_{i',s'}^{m',a'}\big)\tilde{Y}_i^{s,a}(m)\tilde{Y}_{i'}^{s',a'}(m')
\end{aligned}
$$

Splitting the summations into some that square single potential outcomes and others that square averages

or differences of potential outcomes:

$$
\begin{aligned}
=&\sum_{i\in\mathbb{I}}\sum_{m\in\mathbb{M}_i}\sum_{s\in\mathbb{S}}\sum_{a\in\{c,t\}}\pi_{i,s}^{m,a}\Big(\frac{\pi_s}{1-\pi_s}\Big)^{\mathbb{1}\{a=c\}}w_i(s,d)(Y_i(m)-\mu_a)^2\\
&\qquad\cdot\sum_{i'\in\mathbb{I}}\sum_{m'\in\mathbb{M}_i}\sum_{s'\in\mathbb{S}}\mathbb{1}\{\pi_{i,i'}^{m,m'}=0\}\pi_{i',s'}^{m',a}\Big(\frac{\pi_{s'}}{1-\pi_{s'}}\Big)^{\mathbb{1}\{a=c\}}w_{i'}(s',d)\\
&-2\sum_{i\in\mathbb{I}}\sum_{m\in\mathbb{M}_i}\sum_{s\in\mathbb{S}}\sum_{a\in\{c,t\}}\sum_{i'\in\mathbb{I}}\sum_{m'\in\mathbb{M}_i}\sum_{s'\in\mathbb{S}}\mathbb{1}\{\pi_{i,i'}^{m,m'}=0\}\pi_{i,s}^{m,a}\pi_{i',s'}^{m',a}\Big(\frac{\pi_s}{1-\pi_s}\frac{\pi_{s'}}{1-\pi_{s'}}\Big)^{\mathbb{1}\{a=c\}}\\
&\qquad\cdot w_i(s,d)w_{i'}(s',d)\Big(\frac{Y_i(m)+Y_{i'}(m')}{2}-\mu_a\Big)^2\\
&+\sum_{i\in\mathbb{I}}\sum_{m\in\mathbb{M}_i}\sum_{s\in\mathbb{S}}\sum_{a\in\{c,t\}}\pi_{i,s}^{m,a}\Big(\frac{\pi_s}{1-\pi_s}\Big)^{\mathbb{1}\{a=c\}}w_i(s,d)(Y_i(m)-\mu_a)^2\\
&\qquad\cdot\sum_{i'\in\mathbb{I}}\sum_{m'\in\mathbb{M}_i}\sum_{s'\in\mathbb{S}}\sum_{a'\in\{c,t\}\setminus\{a\}}\mathbb{1}\{\pi_{i,i'}^{m,m'}=0\}\pi_{i',s'}^{m',a'}\Big(\frac{\pi_{s'}}{1-\pi_{s'}}\Big)^{\mathbb{1}\{a'=c\}}w_{i'}(s',d)\\
&-\sum_{i\in\mathbb{I}}\sum_{m\in\mathbb{M}_i}\sum_{s\in\mathbb{S}}\sum_{i'\in\mathbb{I}}\sum_{m'\in\mathbb{M}_i}\sum_{s'\in\mathbb{S}}\mathbb{1}\{\pi_{i,i'}^{m,m'}=0\}\pi_{i,s}^{m,t}\pi_{i',s'}^{m',c}\frac{\pi_{s'}}{1-\pi_{s'}}\\
&\qquad\cdot w_i(s,d)w_{i'}(s',d)\big((Y_i(m)-Y_{i'}(m'))-(\mu_t-\mu_c)\big)^2\\
&+\sum_{i\in\mathbb{I}}\sum_{m\in\mathbb{M}_i}\sum_{s\in\mathbb{S}}\sum_{a\in\{c,t\}}\sum_{i'\in\mathbb{I}}\sum_{m'\in\mathbb{M}_i}\sum_{s'\in\mathbb{S}}\sum_{a'\in\{c,t\}}\mathbb{1}\{i\neq i'\}\mathbb{1}\{\pi_{i,i'}^{m,m'}>0\}\big(\pi_{i,s,i',s'}^{m,a,m',a'}-\pi_{i,s}^{m,a}\pi_{i',s'}^{m',a'}\big)\tilde{Y}_i^{s,a}(m)\tilde{Y}_{i'}^{s',a'}(m')
\end{aligned}
\tag{OA4}
$$

Finally, combine the results in Equations (OA1), (OA2), and (OA4), and substitute $\tilde{Y}_i^{s,a}(m)$. Then

$$
\begin{aligned}
&\operatorname{var}\Big(\sum_{i\in\mathbb{I}}\sum_{m\in\mathbb{M}_i}\sum_{s\in\mathbb{S}}\sum_{a\in\{c,t\}}\mathcal{M}_i^m\mathcal{T}_s^a\tilde{Y}_i^{s,a}(m)\Big)\\
=&\sum_{i\in\mathbb{I}}\sum_{m\in\mathbb{M}_i}\sum_{s\in\mathbb{S}}\sum_{a\in\{c,t\}}\pi_{i,s}^{m,a}w_i(s,d)(Y_i(m)-\mu_a)^2\cdot\Big((1-\pi_{i,s}^{m,a})\Big(\frac{\pi_s}{1-\pi_s}\Big)^{2\cdot\mathbb{1}\{a=c\}}w_i(s,d)\Big)\\
&+\sum_{i\in\mathbb{I}}\sum_{m\in\mathbb{M}_i}\sum_{s\in\mathbb{S}}\sum_{a\in\{c,t\}}\sum_{s'\in\mathbb{S}}\sum_{a'\in\{c,t\}}\mathbb{1}\{s\neq s'\text{ or }a\neq a'\}\big(\pi_{i,s,i,s'}^{m,a,m,a'}-\pi_{i,s}^{m,a}\pi_{i,s'}^{m,a'}\big)(-1)^{\mathbb{1}\{a\neq a'\}}\\
&\qquad\cdot\Big(\frac{\pi_s}{1-\pi_s}\Big)^{\mathbb{1}\{a=c\}}\Big(\frac{\pi_{s'}}{1-\pi_{s'}}\Big)^{\mathbb{1}\{a'=c\}}w_i(s,d)w_i(s',d)(Y_i(m)-\mu_a)(Y_i(m)-\mu_{a'})\\
&+\sum_{i\in\mathbb{I}}\sum_{m\in\mathbb{M}_i}\sum_{s\in\mathbb{S}}\sum_{a\in\{c,t\}}\pi_{i,s}^{m,a}w_i(s,d)(Y_i(m)-\mu_a)^2\\
&\qquad\cdot\Big(\frac{\pi_s}{1-\pi_s}\Big)^{\mathbb{1}\{a=c\}}\sum_{i'\in\mathbb{I}}\sum_{m'\in\mathbb{M}_i}\sum_{s'\in\mathbb{S}}\mathbb{1}\{\pi_{i,i'}^{m,m'}=0\}\pi_{i',s'}^{m',a}\Big(\frac{\pi_{s'}}{1-\pi_{s'}}\Big)^{\mathbb{1}\{a=c\}}w_{i'}(s',d)\\
&-2\sum_{i\in\mathbb{I}}\sum_{m\in\mathbb{M}_i}\sum_{s\in\mathbb{S}}\sum_{a\in\{c,t\}}\sum_{i'\in\mathbb{I}}\sum_{m'\in\mathbb{M}_i}\sum_{s'\in\mathbb{S}}\mathbb{1}\{\pi_{i,i'}^{m,m'}=0\}\pi_{i,s}^{m,a}\pi_{i',s'}^{m',a}\Big(\frac{\pi_s}{1-\pi_s}\frac{\pi_{s'}}{1-\pi_{s'}}\Big)^{\mathbb{1}\{a=c\}}\\
&\qquad\cdot w_i(s,d)w_{i'}(s',d)\Big(\frac{Y_i(m)+Y_{i'}(m')}{2}-\mu_a\Big)^2\\
&+\sum_{i\in\mathbb{I}}\sum_{m\in\mathbb{M}_i}\sum_{s\in\mathbb{S}}\sum_{a\in\{c,t\}}\pi_{i,s}^{m,a}w_i(s,d)(Y_i(m)-\mu_a)^2\\
&\qquad\cdot\Big(\frac{\pi_s}{1-\pi_s}\Big)^{\mathbb{1}\{a=c\}}\sum_{i'\in\mathbb{I}}\sum_{m'\in\mathbb{M}_i}\sum_{s'\in\mathbb{S}}\sum_{a'\in\{c,t\}\setminus\{a\}}\mathbb{1}\{\pi_{i,i'}^{m,m'}=0\}\pi_{i',s'}^{m',a'}\Big(\frac{\pi_{s'}}{1-\pi_{s'}}\Big)^{\mathbb{1}\{a'=c\}}w_{i'}(s',d)\\
&-\sum_{i\in\mathbb{I}}\sum_{m\in\mathbb{M}_i}\sum_{s\in\mathbb{S}}\sum_{i'\in\mathbb{I}}\sum_{m'\in\mathbb{M}_i}\sum_{s'\in\mathbb{S}}\mathbb{1}\{\pi_{i,i'}^{m,m'}=0\}\pi_{i,s}^{m,t}\pi_{i',s'}^{m',c}\frac{\pi_{s'}}{1-\pi_{s'}}\\
&\qquad\cdot w_i(s,d)w_{i'}(s',d)\big((Y_i(m)-Y_{i'}(m'))-(\mu_t-\mu_c)\big)^2\\
&+\sum_{i\in\mathbb{I}}\sum_{m\in\mathbb{M}_i}\sum_{s\in\mathbb{S}}\sum_{a\in\{c,t\}}\sum_{i'\in\mathbb{I}}\sum_{m'\in\mathbb{M}_i}\sum_{s'\in\mathbb{S}}\sum_{a'\in\{c,t\}}\mathbb{1}\{i\neq i'\}\mathbb{1}\{\pi_{i,i'}^{m,m'}>0\}\big(\pi_{i,s,i',s'}^{m,a,m',a'}-\pi_{i,s}^{m,a}\pi_{i',s'}^{m',a'}\big)\\
&\qquad\cdot\Big(-\frac{\pi_s}{1-\pi_s}\Big)^{\mathbb{1}\{a=c\}}\Big(-\frac{\pi_{s'}}{1-\pi_{s'}}\Big)^{\mathbb{1}\{a'=c\}}w_i(s,d)w_{i'}(s',d)(Y_i(m)-\mu_a)(Y_{i'}(m')-\mu_{a'}).
\end{aligned}
$$

The first, third, and fifth summations all contain $\pi_{i,s}^{m,a}w_i(s,d)(Y_i(m)-\mu_a)^2$ post-multiplied by different factors. Hence, they can be combined.

Recall that the denominator used in $\tilde{\tau}$ equals $\sum_{s\in\mathbb{S}}\pi_s\sum_{i\in\mathbb{I}}w_i(s,d)$. Define

$$\bar{n}(d)\equiv\frac{1}{\mathbb{S}}\sum_{s\in\mathbb{S}}\pi_s\sum_{i\in\mathbb{I}}w_i(s,d).$$

Then

$$\begin{aligned}
\mathrm{var}(\tilde{\tau}) &= \frac{1}{|\mathbb{S}|^2}\,\mathrm{var}\Big(\sum_{i\in\mathbb{I}}\sum_{m\in\mathbb{M}_i}\sum_{s\in\mathbb{S}}\sum_{a\in\{c,t\}}\mathcal{M}_i^m\mathcal{T}_s^a\tilde{Y}_i^{s,a}(m)\Big)/\bar{n}(d)^2 \\
=&\frac{1}{|\mathbb{S}|}\Bigg(\frac{1}{|\mathbb{S}|}\sum_{a\in\{c,t\}}\sum_{s\in\mathbb{S}}\sum_{i\in\mathbb{I}}\sum_{m\in\mathbb{M}_i}\pi_{i,s}^{m,a}\frac{w_i(s,d)}{\bar{n}(d)}v_{i,s}^{m,a}(d)(Y_i(m)-\mu_a)^2 \\
&+\frac{1}{|\mathbb{S}|}\sum_{s\in\mathbb{S}}\sum_{s'\in\mathbb{S}}\sum_{i\in\mathbb{I}}\sum_{m\in\mathbb{M}_i}\sum_{i'\in\mathbb{I}}\sum_{m'\in\mathbb{M}_i}\sum_{a\in\{c,t\}}\sum_{a'\in\{c,t\}}\mathbb{1}\{i\neq i'\}\mathbb{1}\{\pi_{i,i'}^{m,m'}>0\}\big(\pi_{i,s,i',s'}^{m,a,m',a'}-\pi_{i,s}^{m,a}\pi_{i',s'}^{m',a'}\big) \\
&\qquad\cdot\Big(-\frac{\pi_s}{1-\pi_s}\Big)^{\mathbb{1}\{a=c\}}\Big(-\frac{\pi_{s'}}{1-\pi_{s'}}\Big)^{\mathbb{1}\{a'=c\}}\frac{w_i(s,d)w_{i'}(s',d)}{\bar{n}(d)^2}(Y_i(m)-\mu_a)(Y_{i'}(m')-\mu_{a'}) \\
&+\frac{1}{|\mathbb{S}|}\sum_{s\in\mathbb{S}}\sum_{s'\in\mathbb{S}}\sum_{i\in\mathbb{I}}\sum_{m\in\mathbb{M}_i}\sum_{a\in\{c,t\}}\sum_{a'\in\{c,t\}}\mathbb{1}\{s\neq s'\text{ or }a\neq a'\}\big(\pi_{i,s,i,s'}^{m,a,m,a'}-\pi_{i,s}^{m,a}\pi_{i,s'}^{m,a'}\big) \\
&\qquad\cdot\Big(-\frac{\pi_s}{1-\pi_s}\Big)^{\mathbb{1}\{a=c\}}\Big(-\frac{\pi_{s'}}{1-\pi_{s'}}\Big)^{\mathbb{1}\{a'=c\}}\frac{w_i(s,d)w_i(s',d)}{\bar{n}(d)^2}(Y_i(m)-\mu_a)(Y_i(m)-\mu_{a'}) \\
&-\frac{2}{|\mathbb{S}|}\sum_{a\in\{c,t\}}\sum_{s\in\mathbb{S}}\sum_{s'\in\mathbb{S}}\sum_{i\in\mathbb{I}}\sum_{m\in\mathbb{M}_i}\sum_{i'\in\mathbb{I}}\sum_{m'\in\mathbb{M}_i}\mathbb{1}\{\pi_{i,i'}^{m,m'}=0\}\pi_{i,s}^{m,a}\pi_{i',s'}^{m',a}\Big(\frac{\pi_s}{1-\pi_s}\frac{\pi_{s'}}{1-\pi_{s'}}\Big)^{\mathbb{1}\{a=c\}} \\
&\qquad\cdot\frac{w_i(s,d)w_{i'}(s',d)}{\bar{n}(d)^2}\Big(\frac{Y_i(m)+Y_{i'}(m')}{2}-\mu_a\Big)^2 \\
&-\frac{1}{|\mathbb{S}|}\sum_{s\in\mathbb{S}}\sum_{s'\in\mathbb{S}}\sum_{i\in\mathbb{I}}\sum_{m\in\mathbb{M}_i}\sum_{i'\in\mathbb{I}}\sum_{m'\in\mathbb{M}_i}\mathbb{1}\{\pi_{i,i'}^{m,m'}=0\}\pi_{i,s}^{m,t}\pi_{i',s'}^{m',c}\frac{\pi_{s'}}{1-\pi_{s'}} \\
&\qquad\cdot\frac{w_i(s,d)w_{i'}(s',d)}{\bar{n}(d)^2}\big((Y_i(m)-Y_{i'}(m'))-(\mu_t-\mu_c)\big)^2\Bigg).
\end{aligned}\tag{OA5}$$

where

$$\begin{aligned}
v_{i,s}^{m,a}(d)\equiv&\Big(\frac{\pi_s}{1-\pi_s}\Big)^{\mathbb{1}\{a=c\}}\Big((1-\pi_{i,s}^{m,a})\Big(\frac{\pi_s}{1-\pi_s}\Big)^{\mathbb{1}\{a=c\}}\frac{w_i(s,d)}{\bar{n}(d)} \\
&+\sum_{i'\in\mathbb{I}}\sum_{m'\in\mathbb{M}_i}\sum_{s'\in\mathbb{S}}\sum_{a'\in\{c,t\}}\mathbb{1}\{\pi_{i,i'}^{m,m'}=0\}\pi_{i',s'}^{m',a'}\Big(\frac{\pi_{s'}}{1-\pi_{s'}}\Big)^{\mathbb{1}\{a'=c\}}\frac{w_{i'}(s',d)}{\bar{n}(d)}\Big).
\end{aligned}$$

Define

$$\begin{aligned}
\tilde{V}_a(d)\equiv&\frac{1}{|\mathbb{S}|}\sum_{s\in\mathbb{S}}\sum_{i\in\mathbb{I}}\sum_{m\in\mathbb{M}_i}\pi_{i,s}^{m,a}\frac{w_i(s,d)}{\bar{n}(d)}v_{i,s}^{m,a}(d)(Y_i(m)-\mu_a)^2 \\
\tilde{V}_\times(d)\equiv&\frac{1}{|\mathbb{S}|}\sum_{s\in\mathbb{S}}\sum_{s'\in\mathbb{S}}\sum_{i\in\mathbb{I}}\sum_{m\in\mathbb{M}_i}\sum_{i'\in\mathbb{I}}\sum_{m'\in\mathbb{M}_i}\sum_{a\in\{c,t\}}\sum_{a'\in\{c,t\}}\Bigg(\mathbb{1}\{i\neq i'\text{ or }s\neq s'\text{ or }a\neq a'\} \\
&\cdot\mathbb{1}\{\pi_{i,i'}^{m,m'}>0\}\big(\pi_{i,s,i',s'}^{m,a,m',a'}-\pi_{i,s}^{m,a}\pi_{i',s'}^{m',a'}\big)\Big(-\frac{\pi_s}{1-\pi_s}\Big)^{\mathbb{1}\{a=c\}}\Big(-\frac{\pi_{s'}}{1-\pi_{s'}}\Big)^{\mathbb{1}\{a'=c\}} \\
&\cdot\frac{w_i(s,d)w_{i'}(s',d)}{\bar{n}(d)^2}(Y_i(m)-\mu_a(d))(Y_{i'}(m')-\mu_{a'}(d))\Bigg) \\
\tilde{V}_{aa}(d)\equiv&\frac{2}{|\mathbb{S}|}\sum_{a\in\{c,t\}}\sum_{s\in\mathbb{S}}\sum_{s'\in\mathbb{S}}\sum_{i\in\mathbb{I}}\sum_{m\in\mathbb{M}_i}\sum_{i'\in\mathbb{I}}\sum_{m'\in\mathbb{M}_i}\mathbb{1}\{\pi_{i,i'}^{m,m'}=0\}\pi_{i,s}^{m,a}\pi_{i',s'}^{m',a} \\
&\cdot\Big(\frac{\pi_s}{1-\pi_s}\frac{\pi_{s'}}{1-\pi_{s'}}\Big)^{\mathbb{1}\{a=c\}}\frac{w_i(s,d)w_{i'}(s',d)}{\bar{n}(d)^2}\Big(\frac{Y_i(m)+Y_{i'}(m')}{2}-\mu_a\Big)^2
\end{aligned}$$

$$\tilde{V}_{ct}(d) \equiv \frac{1}{|\mathbb{S}|}\sum_{s\in\mathbb{S}}\sum_{s'\in\mathbb{S}}\sum_{i\in\mathbb{I}}\sum_{m\in\mathbb{M}_i}\sum_{i'\in\mathbb{I}}\sum_{m'\in\mathbb{M}_i}\mathbb{1}\{\pi_{i,i'}^{m,m'}=0\}\pi_{i,s}^{m,t}\pi_{i',s'}^{m',c}\frac{\pi_{s'}}{1-\pi_{s'}}$$
$$\cdot\frac{w_i(s,d)w_{i'}(s',d)}{\bar{n}(d)^2}\big((Y_i(m)-Y_{i'}(m'))-(\mu_t-\mu_c)\big)^2$$

Then $\text{var}(\tilde{\tau}(d)) = \frac{1}{|\mathbb{S}|}(\tilde{V}_t(d)+\tilde{V}_c(d)+\tilde{V}_\times(d)-\tilde{V}_{tt}(d)-\tilde{V}_{cc}(d)-\tilde{V}_{ct}(d))$ as stated in the theorem.

10 Quality of Approximations in Simulations

I assess the quality of the approximation $\hat{\tau}\approx\tilde{\tau}$ in the setting with separate regions through simulations. The design-based results of this paper describe re-assignment of treatment within a fixed and finite population. I first simulate, for a single such population, this thought experiment and assess the quality of the approximation, estimator, and coverage of confidence intervals based on the estimated approximate variance. Then, I repeat the simulations for other such populations and report summary statistics of the assessments across populations.

I simulate populations of candidate treatment locations with their probabilities and individuals with their potential outcomes as follows. In each population, there are 100 regions. The treatment probability of each region is 0.3, such that in each assignment exactly 30 regions are treated and 70 regions are in the control group. In each region, there are 3 candidate treatment locations. The treatment probability of each candidate location, conditional on treatment in the region, is drawn i.i.d. from the Uniform$(0,1)$ distribution and then normalized such that the probabilities sum to one within each region. The number of individuals at the distance of interest $(d\pm h)$ from each candidate location is equal to 1 plus an i.i.d. draw from the Poisson distribution with mean 19, for a mean number of individuals equal to 20 with variance 19. Potential outcomes in the absence of treatment are i.i.d. normal with mean 0 and variance 1. Treatment effects (at the distance of interest) are i.i.d. normal with mean 1 and variance 0.5.

The treatment assignment in a given population follows Assumptions 2 and 3. For each population, I simulate 1,000 treatment assignments to approximate the design distribution. I calculate the estimators $\hat{\tau}$ and $\tilde{\tau}$, as well as the Horvitz-Thompson estimator $\hat{\tau}_{HT}$ that is defined similar to $\hat{\tau}$ with denominators replaced by their expected values (or, identically, $\tilde{\tau}$ without the centering). I compute the bias relative to the population-specific ATT. I compute the variance of each of the three estimators as well their correlations. Finally, I also compute how frequently confidence intervals formed by taking $\hat{\tau}\pm 1.96\cdot\text{se}$ cover the ATT of the population, where se is the square root of the proposed feasible variance estimator.

I repeat these simulations for 1,000 such populations and report summary statistics (mean and quantiles) of the population-specific statistics in Table OA4. The first column shows the 0.01 quantile across these populations, the second column shows the mean, the third column shows the median, and the fourth column shows the 0.99 quantile. The first two rows show the absolute bias (multiplied by 100) of the estimators $\hat{\tau}$ and $\hat{\tau}_{HT}$. Note that the Horvitz-Thompson estimator is exactly unbiased over the design distribution for each population. Hence, any bias for $\hat{\tau}_{HT}$ is due to a difference between the distribution of the 1,000 assignments simulated for each population and the true design distribution. While the recommended estimator, $\hat{\tau}$, is not necessarily exactly unbiased over the design distribution, its bias in these simulations is small enough such that it is similar to the simulation noise evident in the non-zero bias of $\hat{\tau}_{HT}$.

Rows 3 and 4 show the design-based variance of $\hat{\tau}$ and $\hat{\tau}_{HT}$ relative to the design-based variance of the infeasible $\tilde{\tau}$. For 98% of simulated populations, the design-based variance of $\hat{\tau}$ is within 1% of the design-based variance of $\tilde{\tau}$. At least in these simulations, using the variance of $\tilde{\tau}$ in place of the variance of $\hat{\tau}$ is innocuous. The Horvitz-Thompson estimator, in contrast, has larger design-based variance for all simulated populations. Its variance in these simulations is typically 60% than the variance of $\tilde{\tau}$. The reason the variances of $\hat{\tau}$ and $\tilde{\tau}$ are so similar is that in any given sample the two estimators are extremely close. In fact, even for the populations where the two estimators are the *least* alike, their correlation is rounded to 1 (row 5). The correlation of $\hat{\tau}_{HT}$ and $\tilde{\tau}$ is also high, but, nevertheless, noticeably lower (row 6).

Row 6 shows that the confidence intervals using the estimated variance have close to nominal coverage in all populations. Coverage may be below nominal levels for three reasons. First, in finite samples, the estimated variance differs from the true variance. Second, the normal approximation may be inaccurate in

Table OA4: Summary statistics of design-based properties across simulated populations.

	0.01 quantile	mean	median	0.99 quantile
absolute bias $\hat{\tau}$	0	0.13	0.11	0.41
absolute bias $\hat{\tau}_{HT}$	0	0.16	0.13	0.52
$\text{var}(\hat{\tau})/\text{var}(\tilde{\tau})$	0.99	1	1	1.01
$\text{var}(\hat{\tau}_{HT})/\text{var}(\tilde{\tau})$	1.33	1.61	1.6	1.95
$\text{cor}(\hat{\tau}, \tilde{\tau})$	1	1	1	1
$\text{cor}(\hat{\tau}_{HT}, \tilde{\tau})$	0.71	0.79	0.79	0.85
coverage 95% CI	0.92	0.94	0.94	0.96

finite samples. Third, there are small differences between the estimators $\hat{\tau}$ and $\tilde{\tau}$. In these simulations, two reasons unambiguously push the estimated variance to exceed the true variance. First, there is treatment effect heterogeneity and the variance of treatment effects term $\tilde{V}_{ct}^{\text{region}}$ cannot be estimated. Second, there are multiple candidate treatment locations in all regions, such that $\tilde{V}_{t}^{\text{region}}$ cannot be estimated and must instead be bounded by the larger $\tilde{V}_{t}^{\text{location}}$. On net, these factors balance out to close to nominal coverage for all populations in these simulations.

Overall, the simulation results support the recommendations of this paper to use the estimator $\hat{\tau}$ and to do inference using estimates of the variance of the infeasible estimator $\tilde{\tau}$.

References

Abadie, A., S. Athey, G. W. Imbens, and J. M. Wooldridge (2017). When should you adjust standard errors for clustering? *NBER Working Paper Series* (24003).

Abadie, A., S. Athey, G. W. Imbens, and J. M. Wooldridge (2020). Sampling-based vs. design-based uncertainty in regression analysis. *Econometrica 88*(1), 265–296.

Abadie, A., A. Diamond, and J. Hainmueller (2010). Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American Statistical Association 105*(490), 493–505.

Aliprantis, D. and D. Hartley (2015). Blowing it up and knocking it down: The local and city-wide effects of demolishing high concentration public housing on crime. *Journal of Urban Economics 88*, 67–81.

Arjovsky, M. and L. Bottou (2017). Towards principled methods for training generative adversarial networks. *arXiv preprint arXiv:1701.04862*.

Arjovsky, M., S. Chintala, and L. Bottou (2017). Wasserstein gan. *arXiv preprint arXiv:1701.07875*.

Aronow, P. M., D. P. Green, and D. K. K. Lee (2014). Sharp bounds on the variance in randomized experiments. *The Annals of Statistics 42*(3), 850–871.

Aronow, P. M. and C. Samii (2017). Estimating average causal effects under general interference, with application to a social network experiment. *The Annals of Applied Statistics 11*(4), 1912–1947.

Athey, S., D. Blei, R. Donnelly, F. Ruiz, and T. Schmidt (2018). Estimating heterogeneous consumer preferences for restaurants and travel time using mobile location data. *AEA Papers and Proceedings 108*, 64–67.

Athey, S., G. W. Imbens, J. Metzger, and E. M. Munro (2019). Using wasserstein generative adversarial networks for the design of monte carlo simulations. *NBER Working Paper Series* (26566).

Biggio, B., I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and F. Roli (2013). Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, Berlin, Heidelberg, pp. 387–402. Springer.

Buchmueller, T. C., M. Jacobson, and C. Wold (2006). How far to the hospital? the effect of hospital closures on access to care. *Journal of Health Economics 25*(4), 740–761.

Cohen, J. and P. Dupas (2010). Free distribution or cost-sharing? evidence from a randomized malaria prevention experiment. *The Quarterly Journal of Economics 125*(1), 1–45.

Cohen, T. S. and M. Welling (2016). Group equivariant convolutional networks. In *International conference on machine learning*, pp. 2990–2999. PMLR.

Currie, J., L. Davis, M. Greenstone, and R. Walker (2015). Environmental health risks and housing values: evidence from 1,600 toxic plant openings and closings. *American Economic Review 105*(2), 678–709.

Dell, M. and B. A. Olken (2020). The development effects of the extractive colonial economy: The dutch cultivation system in java. *The Review of Economic Studies 87*(1), 164–203.

Di Tella, R. and E. Schargrodsky (2004). Do police reduce crime? estimates using the allocation of police forces after a terrorist attack. *American Economic Review 94*(1), 115–133.

Diamond, R. and T. McQuade (2019). Who wants affordable housing in their backyard? an equilibrium analysis of low-income property development. *Journal of Political Economy 127*(3), 1063–1117.

Dieleman, S., J. De Fauw, and K. Kavukcuoglu (2016). Exploiting cyclic symmetry in convolutional neural networks. In *International conference on machine learning*, pp. 1889–1898. PMLR.

Dudar, V. and V. Semenov (2018). Use of symmetric kernels for convolutional neural networks. In *XVIII International Conference on Data Science and Intelligent Analysis of Information*, pp. 3–10. Springer.

Duflo, E. (2001). Schooling and labor market consequences of school construction in indonesia: Evidence from an unusual policy experiment. *American Economic Review 91*(4), 795–813.

Dzhezyan, G. and H. Cecotti (2019). Symnet: Symmetrical filters in convolutional neural networks. *arXiv preprint arXiv:1906.04252*.

Ellickson, P. B. and P. L. E. Grieco (2013). Wal-mart and the geography of grocery retailing. *Journal of Urban Economics 75*, 1–14.

Feyrer, J., E. T. Mansur, and B. Sacerdote (2017). Geographic dispersion of economic shocks: Evidence from the fracking revolution. *American Economic Review 107*(4), 1313–1334.

Gens, R. and P. M. Domingos (2014). Deep symmetry networks. In *Advances in Neural Information Processing Systems*, Volume 27, pp. 2537–2545.

Goodfellow, I. (2016). Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*.

Goodfellow, I., H. Lee, Q. Le, A. Saxe, and A. Ng (2009). Measuring invariances in deep networks. In *Advances in Neural Information Processing Systems*, Volume 22, pp. 646–654.

Goodfellow, I., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems*, Volume 27, pp. 2672–2680.

Greenstone, M., R. Hornbeck, and E. Moretti (2010). Identifying agglomeration spillovers: Evidence from winners and losers of large plant openings. *Journal of Political Economy 118*(3), 536–598.

Greenstone, M. and E. Moretti (2003). Bidding for industrial plants: Does winning a 'million dollar plant' increase welfare? *NBER Working Paper Series* (9844).

Hastie, T. J., R. J. Tibshirani, and J. H. Friedman (2001). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer Series in Statistics. Springer.

Heckman, J. J., J. Smith, and N. Clements (1997). Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts. *The Review of Economic Studies 64*(4), 487–535.

Hinton, G. E., A. Krizhevsky, and S. D. Wang (2011). Transforming auto-encoders. In T. Honkela, W. Duch, M. Girolami, and S. Kaski (Eds.), *Artificial Neural Networks and Machine Learning – ICANN 2011*, Volume 6791 of *Lecture Notes in Computer Science*, Berlin, Heidelberg, pp. 44–51. Springer.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics 86*(1), 4–29.

Imbens, G. W. and D. B. Rubin (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. New York, NY: Cambridge University Press.

Jaderberg, M., K. Simonyan, A. Zisserman, and K. Kavukcuoglu (2015). Spatial transformer networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, Volume 28, pp. 2017–2025.

Jia, P. (2008). What happens when wal-mart comes to town: An empirical analysis of the discount retailing industry. *Econometrica 76*(6), 1263–1316.

Kauderer-Abrams, E. (2017). Quantifying translation-invariance in convolutional neural networks. *arXiv preprint arXiv:1801.01450*.

Keiser, D. A. and J. S. Shapiro (2019). Consequences of the clean water act and the demand for water quality. *The Quarterly Journal of Economics 134*(1), 349–396.

Krizhevsky, A., I. Sutskever, and G. E. Hinton (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, Volume 25, pp. 1097–1105.

Lenc, K. and A. Vedaldi (2015). Understanding image representations by measuring their equivariance and equivalence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 991–999.

Liang, K.-Y. and S. L. Zeger (1986). Longitudinal data analysis using generalized linear models. *Biometrika 73*(1), 13–22.

Liang, T. (2018). On how well generative adversarial networks learn densities: Nonparametric and parametric results. *arXiv preprint arXiv:1811.03179*.

Linden, L. and J. E. Rockoff (2008). Estimates of the impact of crime risk on property values from Megan's laws. *American Economic Review 98*(3), 1103–1127.

Lotter, W., G. Kreiman, and D. Cox (2016). Unsupervised learning of visual structure using predictive generative networks. *arXiv preprint arXiv:1511.06380*.

Miguel, E. and M. Kremer (2004). Worms: identifying impacts on education and health in the presence of treatment externalities. *Econometrica 72*(1), 159–217.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Roczniki Nauk Rolniczych Tom X*, 1–51. [in Polish].

Neyman, J. (1990). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science 5*(4), 465–472. [translated by D. M. Dabrowska and T. P. Speed].

Oates, W. E. (1969). The effects of property taxes and local public spending on property values: An empirical study of tax capitalization and the tiebout hypothesis. *Journal of Political Economy 77*(6), 957–971.

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica 56*(4), 931–954.

Seim, K. (2006). An empirical model of firm entry with endogenous product-type choices. *The RAND Journal of Economics 37*(3), 619–640.

Siegfried, J. J. and A. Zimbalist (2000). The economics of sports facilities and their communities. *Journal of Economic Perspectives 14*(3), 95–114.

Simard, P. Y., D. Steinkraus, and J. C. Platt (2003). Best practices for convolutional neural networks applied to visual document analysis. In *Seventh International Conference on Document Analysis and Recognition*, Volume 3, pp. 958–963. IEEE.

Singh, S., A. Uppal, B. Li, C.-L. Li, M. Zaheer, and B. Poczos (2018). Nonparametric density estimation under adversarial losses. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, Volume 31, pp. 10225–10236.

Srivastava, N., G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov (2014). Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research 15*(1), 1929–1958.

Stock, J. H. (1991). Nonparametric policy analysis: an application to estimating hazardous waste cleanup benefits. In W. A. Barnett, J. Powell, and G. Tauchen (Eds.), *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, Chapter 3, pp. 77–98. Cambridge University Press.

Szegedy, C., W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Vincent, P., H. Larochelle, Y. Bengio, and P.-A. Manzagol (2008). Extracting and composing robust features with denoising autoencoders. In A. McCallum and S. Roweis (Eds.), *25th International Conference on Machine Learning*, pp. 1096–1103.

Yaeger, L., R. Lyon, and B. Webb (1996). Effective training of a neural network character classifier for word recognition. In *Advances in Neural Information Processing Systems*, Volume 9, pp. 807–816.

Yang, F., Z. Wang, and C. Heinze-Deml (2019). Invariance-inducing regularization using worst-case transformations suffices to boost accuracy and spatial robustness. *arXiv preprint arXiv:1906.11235*.

Zeiler, M. D. and R. Fergus (2014). Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pp. 818–833. Springer.