INVERSE PROBABILITY TILTING WITH HETEROGENEOUS AVERAGE TREATMENT EFFECTS

JEFFREY P. COHEN AND KE YANG

Abstract. We extend the Inverse Probability Tilting (IPT) estimator by developing a nonparametric, general IPT (GIPT) estimator that allows for average treatment effect (ATE) heterogeneity and addresses "missing data" problems. GIPT re-weights twice: using propensity scores that equate moments across treated (and untreated) sub-samples with the entire sample, as in IPT; and also, down-weighting observations far from each target point. This allows for heterogeneous ATE estimates. Monte Carlo simulations validate the strong small sample performance of GIPT. Among many possible applications of GIPT, we demonstrate how a severe storm leading to an extended water-boil advisory, imposed much longer on sub-sections of Metro-Vancouver Canada (the "treatment"), impacted individual commercial property prices (the ATEs) differently.

Keywords: Inverse probability tilting; missing data; storm impacts on real estate; spatial dependence.

JEL Codes: C31, R3

Date: 02/24/2019.

Jeffrey P. Cohen is an associate professor at University of Connecticut, School of Business, Department of Finance, Jeffrey.Cohen@business.uconn.edu; Ke Yang is an associate professor at University of Hartford, Barney School of Business, kyang@hartford.edu. We acknowledge the helpful comments on earlier versions of this paper from seminar participants at SUNY-Albany, University of Missouri-Columbia, Rutgers University, and University of Cincinnati. We also appreciate feedback from participants at the 2018 NY Camp Econometrics XIII. Bryan Graham and Wes Wilson also provided helpful comments. Any remaining errors are our own.

1. Introduction

Two increasingly popular areas of focus in recent applied statistical research are average treatment effect (ATE) heterogeneity, and missing data problems. Our methodology generalizes an existing approach and addresses both of these issues in one framework.

One set of approaches to missing data problems in general settings is propensity score approaches. There is an extensive body of literature on Inverse Probability Weighting (IPW), as in Rosenbaum and Rubin (1983) and followed by Imbens (2004) and Wooldridge (2007). More recently, Graham et al (2012) developed an approach called Inverse Probability Tilting (IPT), which imposes a balance between the treated and control groups while estimating the ATE.

Some recent research has focused on specific types of missing data problems and some have addressed them with propensity score approaches. For instance, Abrevaya and Donald (2017) consider a situation where some observations on an explanatory variable are missing, and they develop an estimator to handle this problem.

One objective of this paper is to incorporate a second adjustment for missing data problems as a part of the estimation strategy in our application. Specifically, we generalize the IPT estimator to allow for re-weighting based on heterogeneity, in addition to a propensity score approach for the missing data problem. The attractive features of IPT that we describe below have prompted us to explore this generalization of IPT. This type of additional adjustment is important in the context of many treatment effect problems, because the ATEs can be different across various target points.

In particular, the issue of ATE heterogeneity has received some recent attention. While one advantage of IPT is that it leads to a unique treatment effect for each observation, it may also be useful to consider heterogeneity in a nonparametric framework that could lead to different ATEs across individual observations or target points. In such situations, an approach to deal with the missing data problem while allowing for heterogeneity in ATEs across target points is desirable. Thus, we demonstrate how a general version of the IPT approach that considers heterogeneity in the data can address the missing data problem while at the same time allowing for ATEs to vary across observations.

One potential application of our estimator is property sales, where a treatment is imposed on some properties in a geographic region, but neither on others in the same region nor upon any properties in a neighboring region. With this particular missing data problem, the researcher knows what price a treated property sold for, but does not know how much the same property would have sold for if it had been untreated. The ATEs might vary across different locations. While this is the specific application that we consider in this paper in order to demonstrate the implementation of our estimator, there are many other potential applications of our generalization of IPT, in contexts where there is heterogeneity in the data and a treatment is imposed on units at some target points, but not in others.

A health and development economics application of IPT is described in Graham et al. (2011). That application considers the impact of a cash transfer program, which was provided to some households but not others, on household calorie consumption in 42 rural Nicaraguan communities in the early 2000s. This type of application would be an excellent candidate for our generalization of IPT that would allow for heterogeneous average treatment effects in different villages.

In the remainder of this paper, we first motivate one type of missing data problem (although our estimator can be applicable to a broad range of other missing data problems). Next we explain the generalizations to the IPT estimator that allow for ATE heterogeneity, and the adjustments to the propensity score weights made to allow for more distant observations to be down-weighted relative to more close observations. We call this generalization a GIPT estimator (representing "General Inverse Probability Tilting"). We prove consistency and asymptotic normality of GIPT in Appendix A. We describe the computation process of the GIPT estimator, then provide some Monte Carlo evidence to demonstrate that the estimator performs well. We apply this GIPT estimator to the case of how commercial property prices in the metro-Vancouver, BC Canada region may be impacted differently, shortly before versus after a storm leading to an extended water-boil advisory that is imposed on some parts of the region for much longer than other areas. Finally, we discuss potential future extensions to this approach and summarize our findings.

2. Motivation

Consider the following problem as one particular type of missing data problem. First, suppose one is interested in analyzing a data set on units that are in various locations throughout a particular geographic region, to determine the ATEs at each location in the region shortly after versus shortly before a random "event". The treatment area may be confined to specific parts of a particular metro area, for instance, which we call the subject area. The "untreated" observations are a set of units that are in some parts of the metro area, before and after the "event" as well as a set of those within the subject area before the "event" . Then, we can estimate the effect of being in the treatment sub-sample opposed to the non-treated sub-sample.

But in empirical applications, the researcher does not know what the treatment outcome would have been if a particular unit had been untreated. Also, the researcher does not observe the outcome for units in the treated group if each unit had been untreated. These two situations are the missing data problem that we consider in this paper. In these cases, in order to obtain valid treatment effects, one can re-weight the data with propensity scores. There are several approaches to accomplishing this. One is an Inverse Probability Weighting (IPW) approach, which has received extensive attention in the literature (see, e.g., Rosenbaum and Rubin (1983), and Imbens (2004), among others), and is one approach to obtain the propensity score weighting parameter. An attractive alternative is the IPT approach, as in Graham et al (2012), which generates separate tilting parameter estimates for the treated and untreated samples, and imposes a balance between the treated and control groups when estimating the ATE. There are alternative missing data approaches that have been proposed by others, such as random forests (Wager and Athey, 2017), and some methods closely related to IPT (e.g., Imai and Ratkovic, 2014, and Hainmueller, 2012), among others. A comparison of many of these approaches is presented in Frölich et al (2017), however there is no known analysis of GIPT in the literature. Allowing for ATE heterogeneity in missing data problems in the context of IPT is one contribution of our paper.[1]

Specifically, the IPT and IPW approaches do not allow for heterogeneity in the ATEs and the tilting parameters. If the target points of observations are varied, this could be an important consideration in many particular applications. It may be helpful to re-weight a second time, to consider heterogeneity across target points. This is common in the non-parametric estimation literature, specifically, with an approach called Locally Weighted Regressions (LWR), also commonly referred to as Geographically Weighted Regressions (GWR), as in Brunsdon et al (1996). McMillen and Redfearn (2010) describe LWR and present an application. However, no known work has incorporated this type of estimation into an IPT framework.

3. Approach

3.1. **Model.** Suppose that there are N units, indexed by i = 1, . . . ,N, viewed as drawn randomly from a large population. We postulate the existence for each unit of a pair of potential outcomes, $Y_i(0)$ for the outcome under the control treatment and $Y_i(1)$ for the outcome under the active treatment. Let $X_i = \{X_i^1, L_i\}$. Each unit has a vector of covariates, pretreatment variables or exogenous variables, X_i^1, and vector of covariates L_i that may consist of a subset of X_i^1 and/or a set of variables not included as part of X_i^1 (such as geographic coordinates). Each unit is exposed to a single treatment; $D_i = 0$ if unit i is untreated and $D_i = 1$ if unit i receives the active treatment. We therefore observe for each unit the triple (D_i, Y_i, X_i), where Y_i is the realized outcome:

$$Y_i \equiv Y_i(D_i) = \begin{cases} Y_i(0) & if \quad D_i = 0, \\ Y_i(1) & if \quad D_i = 1. \end{cases}$$

Distributions of (D_i, Y_i, X_i) refer to the distribution induced by the random sampling from the population. We follow the potential outcomes of Neyman (1923) and Rubin (1974), assuming

[1] Other recent contributions to this ATE heterogeneity literature have included Allcott (2015), Hsu et al (2018), and Hotz et al (2005).

the existence of potential outcomes, $Y(1)$ and $Y(0)$, corresponding respectively to the outcome the subject at a specific target point would have experienced with or without treatment. Then we can define the average treatment effect (ATE) at l as

$$\gamma(l) = \mathbb{E}[Y(1) - Y(0)|L = l].$$

In practice, however, one only observes

$$Y_i = (1 - D_i)Y_i(0) + D_iY_i(1)$$

i.e., only $Y_i(1)$ for actively treated units or $Y_i(0)$ for untreated units are observed at any given target point. First, we make the following assumption:

Assumption 1. *(Unconfoundedness)* $\{Y(1), Y(0)\} \perp D|X$.

This assumption effectively implies that we can treat nearby observations as having come from a randomized experiment. It follows immediately that the ATE at target point l, $\gamma(l)$, is given as:

$$\gamma(L = l) = E\left[E\left[Y|D = 1, X\right] - E\left[Y|D = 0, X\right]|L = l\right]$$

or equivalently

$$\gamma(L = l) = E\left[\frac{DY}{p(X)} - \frac{(1 - D)Y}{1 - p(X)}|L = l\right] \tag{3.1}$$

where $p(X) = P\left[D = 1|X = x\right] = E[D_i|X_i = x]$ is the propensity score function that prescribes the conditional probability of receiving treatment at x (which is a generalization of the setup in Rosenbaum and Rubin, 1983). As this propensity score function is generally unknown, many earlier methods on average treatment effect estimation differ in how they estimate $p(X)$ using, e.g., variants of maximum likelihood approaches, such as the Inverse Probability Weighting

(IPW) estimator that we describe in the next section, and then the estimate of $p(X)$ implies an ATE.

3.2. **General Inverse Probability Tilting Estimator (GIPT).** Rosenbaum and Rubin (1983) proposed the Inverse Probability Weighting ATE estimator by first replacing the $p(X)$ with a maximim likelihood estimator, then averaging over sample points. The Rosenbaum and Rubin (1983) setup implicitly assumes no variation in the ATE across observations. Graham et al (2012) proposed an alternative method by estimating the propensity score function that imposes a balance across the treated and control groups with a particular estimator consisting of two separate tilting parameters, one for each observation in the treatment group and another for observations in the control group. We incorporate target point specific weights into the IPT estimator from Graham et al (2012), in the following way. Our method of estimating the target point specific average treatment effects is based on a generalization of the IPT estimator proposed by Graham et al (2012) and relies upon the following assumptions 2 through 9 below, in addition to Assumption 1 above (the unconfoundedness assumption).

Assumption 2. *(Random Sampling). $\{D_i, X_i, Y_{1i}\}_{i=1}^N$ is an independently and identically distributed random sequence. We observe D, X, and $Y = DY_1$ for each sampled unit.*

Assumption 3. *(Identification) For some known $K \times 1$ vector of functions $\Phi(Y, X, \gamma)$,*

$$E(\Phi(Y, X, \gamma)) = 0$$

with (i) $E(\Phi(Y, X, \gamma)) \neq 0$ for all $\gamma \neq \gamma_0$, $\gamma \in \Theta \subset \mathbb{R}^K$, and Θ compact with $\gamma_0 \in int(\Theta)$,(ii) $|\Phi(Y, X, \gamma)| \leq c(Y, X)$ for all Y, X with $c(\cdot)$ a non-negative function and $\mathbb{E}(c(Y, X)) < \infty$, (iii) $\Phi(Y, X, \gamma)$ is continuous on Θ for each Y, X and continuously differentiable in a neighborhood of γ_0, (iv) $\mathbb{E}[\|\Phi(Y, X, \gamma)\|^2] < \infty$, and (v) $\mathbb{E}[sup_{\gamma \in \Theta} \|\nabla_\gamma \Phi(Y, X, \gamma)\|] < \infty$.

Assumption 4. *(Strong Overlap) $p(X) = P[D = 1|X = x]$ is bounded away from 0 and 1 over $\aleph$, the support of X.*

Assumption 5. *There is a continuous function $\delta_0(\cdot)$ and compact, known vector $r(X)$ of linearly independent functions of X, and known function $G(\cdot)$such that (i) $G(\cdot)$ is strictly increasing, continuously differentiable, and maps into the unit interval with $\lim_{\nu \to -\infty} G(\nu) = 0$ and $\lim_{\nu \to \infty} G(\nu) = 1$, (ii) $p(x) = G\left(r(w(l)x^1)'\delta_0(l)\right)$ for all $x \in \aleph$, and (iii) $G(r(w(l)x^1)'\delta_0(l))$ is bounded away from 0 and 1 for $\delta_0(\cdot)$ and $x \in \aleph$.*

GIPT is a kernel based estimator.[2] The following additional regularity assumptions are needed for the GIPT estimator to have desirable large sample properties. Assumptions 6 through 8 are analogous to assumptions made by Abrevaya and Donald (2017).

Assumption 6. *(Distribution of X): the Support χ of the k-dimensional covariate X is a Cartesian product of compact intervals, and the density of X, $f(X)$ are $p-$times continuously differentiable over χ.*

Assumption 7. *(Kernels): $K(\cdot)$ is a kernel of order s, is symmetric around zero, is equal to zero outside $\prod_{i=1}^{k}[-1,1]$, integrate to 1 and is continuously differentiable.*

Assumption 8. *(Bandwidths): The bandwidth b satisfies the following conditions as $N \to \infty$: $b \to 0$ and $log(N)/(Nb^{k+s}) \to 0$.*

Assumption 8 implies that b is a nuisance parameter.[3]

[2] There is a large literature on geographically weighted regressions (GWR), which is essentially a form of weighted least squares and is a commonly used kernel estimator in spatial studies to allow for geographic heterogeneity in regression parameters. In other words, this approach leads to the possibility of different marginal effects at each target point. The basic idea behind GWR is to assign higher weights to observations near the target point when calculating a point specific estimate. The measure of distance between observations has a natural geographic interpretation in spatial modeling. The GWR approach is readily extended to Maximum-Likelihood Estimation (MLE) methods as well. While a typical MLE procedure chooses estimates to maximize the log-likelihood function, the geographically weighted version of MLE estimates a pseudo log-likelihood function, where the log-likelihood function depends on the functional form of the regression model. See McMillen and McDonald (2004), for more details.

[3] There is a large literature on kernel and bandwidth selection in nonparametric estimation. For kernel selection, McMillen and Redfearn (2010) indicate that the results tend to be robust with respect to the specific functional form of the kernel, but more sensitive to the bandwidth. Silverman (1986) proposes a "rule of thumb" bandwidth, while others such as McMillen and Redfearn (2010) propose variations of cross validation techniques. In the context of GIPT, we describe our bandwidth selection process below, which was somewhat different in the Monte Carlo simulations than with the empirical application of GIPT.

In developing the GIPT estimator, we modify equation (A.22) in Graham et al (2012)[4] by incorporating kernel weights and a bandwidth parameter. If the researcher believes that the potential outcome function $G(\cdot)$ is a non-parametric function, then we could transform both $t(\cdot)$ and D_i with some kernel weights[5]. More specifically, suppose one is interested in the first m moments (however, the choice of number of moments to be included is described in more detail in footnote 3 below). Then, we denote $\tau(\hat{w}_i(l)x_i^1) = [1, \hat{w}_i(l)x_i^1, (\hat{w}_i(l)x_i^1)^2, \cdots (\hat{w}_i(l)x_i^1)^m,]'$, as a column vector where the weight $\hat{w}_i(l) = \left[K\left(\frac{d_i(l)}{b}\right)\right]^{1/2}$, with $K(\cdot)$ being the Gaussian kernel, b being the bandwidth parameter, m is the number of moments included, and $d_i(l)$ being the distance between observations i and target point $L = l$. This setup amounts to a non-parametric specification of the tilting parameters, $\delta^0(l)$ and $\delta^1(l)$, as defined following Assumption 9 below.

Assumption 9. *(Moment Conditional Expectation Function Model): For some unique matrix* Π^* *and vector of linear independent functions* $\tau^*(w_i x_i^1)$ *with a constant in the first row, we have*

$$E(\Phi(y, \gamma_0(l) \mid X) = \Pi^* \tau^*(w_i x_i^1))$$

Graham et al (2012) describe the implications for overfitting the propsensity score depending on the requirements of their Assumption 3.1 (analogous to our Assumption 9).[6]

Analogous to equation (5) and (6) in Graham et al (2012), when our assumptions 1 through 8 hold, then at each target point l we have the following just-identified unconditional moment

[4] In Graham et al (2012), they compute separate tilting parameters for the treatment and control groups by solving an optimization problem that imposes a balance between the two groups. Among their assumptions includes variants of our assumptions 1 through 5, but the location variable, l, is not included in their vector of X. See our assumption 5 below for more details.

[5] In the case $G(\cdot)$ is a non-parametric function, a naive way to estimate treatment effect heterogeneity is to estimate, e.g. using IPT, the conditional effects for each different location, $L = l$. Our proposed method is conceptually more appealing because $G(\cdot)$ at $L = l$ is estimated using observations not only at location l, but also observation in the surroundings. In addition, with our method, researchers can control how to use nearby observations through the choice of kernel function and bandwidth. Furthermore, the spatial dependence is also accounted for in our proposed method, through assigning higher weights to observations closer to the target location and lower weights further away.

[6] Graham et al (2012) indicate that their Assumption 3.1 has implications for whether the propensity score needs to include additional moments (when r(X) is contained within t*(X)) or when the opposite is true where a replacement is made to "eliminate any overidentifying restrictions." In other words, IPT "overfits the propensity score if Assumption 3.1 requires us to do so..." (Graham et al, 2012). In our problem with GIPT, the analogous carries through to GIPT, depending on what Assumption 9 requires us to do.

problem:

$$\mathbb{E}\left[\frac{\hat{w}_i(l)\cdot D_i}{G\left(\tau(\hat{w}_i(l)\cdot x_i^1)'\delta_0(l)\right)}\Phi(X,Y,\gamma_0)\right]=0, \tag{3.2}$$

$$\mathbb{E}\left[\left(\frac{\hat{w}_i(l)\cdot D_i}{G\left(\tau(\hat{w}_i(l)\cdot x_i^1)'\delta_0(l)\right)}-1\right)\tau(\hat{w}_i(l)\cdot x_i^1)\right]=0. \tag{3.3}$$

Our GIPT estimator chooses $\hat{\beta}_{GIPT}(l)=\left[\hat{\gamma}'(l),\hat{\delta}'_{GIPT}(l)\right]'$ at each target point l to solve the sample analogue of the above two equations, i.e.

$$\frac{1}{N}\sum_{i=1}^{N}\left[\frac{\hat{w}_i(l)\cdot D_i}{G\left(\tau(\hat{w}_i(l)\cdot x_i^1)'\hat{\delta}_{GIPT}(l)\right)}\Phi(X_i,Y_i,\hat{\gamma}_{GIPT})\right]=0, \tag{3.4}$$

$$\frac{1}{N}\sum_{i=1}^{N}\left[\left(\frac{\hat{w}_i(l)\cdot D_i}{G\left(\tau(\hat{w}_i(l)\cdot x_i^1)'\hat{\delta}_{GIPT}(l)\right)}-1\right)\tau(\hat{w}_i(l)\cdot x_i^1)\right]=0. \tag{3.5}$$

Theorem 1. *Given the missing data model defined by assumptions 1 through 8, then at each target point l for $\hat{\gamma}_{GIPT}(l)$, the solution to equation (3.4) (i) $\hat{\gamma}_{GIPT}(l)$ is a consistent estimator of $\gamma_0(l)$ and (ii) $\sqrt{N}(\hat{\gamma}_{GIPT}(l)-\gamma_0(l))\xrightarrow{D}\mathcal{N}(0,\mathcal{I}(\gamma_0(l))^{-1})$.*

Proof: See Appendix A.

Equation (3.5) is solved first, separately for each target point l, and the resulting tilting parameter estimates for each target point l are plugged into equation (3.4) for each target point l to obtain the estimate of the ATE at each target point, l.[7]

For computational simplicity, G is often assumed to take the Logit functional form, that is, $G(v)=exp(v)/[1+exp(v)]$, and $\phi_v=1/G(v)$. Let $h=0,1$ denote the treatment status of each individual with “1” for treatment group and “0” for control group. Then to compute $\tilde{\delta}^h(l)$, for each target point the GIPT estimator solves the following optimization problem separately for each target point l, adapted from equation (A.22) of Graham et al. (2012) to incorporate multiple target points:[8]

[7] While it might be desirable to test restrictions among the ATEs, this is not a straightforward issue to implement.

[8] While there is no required minimum number of target points, the problem is only interesting when there is more than one target point, as otherwise there will not be any ATE heterogeneity.

For each target point, l, choose $\delta^h(l)$ to

$$\max \mathtt{L}(\delta^h(l)) = (1/N){\textstyle\sum_i} D_i^h \hat{w}_i(l)\phi^h(\tau(\hat{w}_i(l)x_i^1)'\delta^h(l)) - (1/N){\textstyle\sum_i}\tau(\hat{w}_i(l)x_i^1)'\delta^h(l)$$

where D_i^h is the treatment dummy for group h and ϕ^h are specific to group h. In this case where there is a control group and one treatment group, then the notation for these dummies can equivalently be reduced to $(1-D)$ and (D), respectively.

The first order condition for this optimization problem is:

$$\partial(\mathtt{L}(\delta^h(l)))/\partial\delta^h(l) = (1/N){\textstyle\sum_i} D_i^h \hat{w}_i(l)\tau(\hat{w}_i(l)x_i^1)'\phi_\delta^h(\cdot) \; - (1/N){\textstyle\sum_i}\tau(\hat{w}_i(l)x_i^1)' = 0,$$

and the second order condition is:

$$\partial^2(\mathtt{L}(\delta^h(l)))/(\partial\delta^h(l))^2 = (1/N)\sum_i D_i^h \hat{w}_i(l)\tau(\hat{w}_i(l)x_i^1)''\phi_{\delta\delta}^h(\cdot)$$

Graham et al (2012) show for IPT that $\phi_{\delta\delta}^h(\cdot) < 0$ (see their equation A.21), so that ($\mathtt{L}$) is strictly concave. It follows here that concavity holds for GIPT at each target point, l.

When the treatment status is denoted by h, where $h = 0$ is the control group and $h = 1$ is the treatment group, it is reasonably straightforward to solve the optimization problem above (analogous to equation A.22 in Graham et al, 2012) for $\tilde{\delta}^h(l)$ for all l. The GIPT estimator will lead to separate parameter estimates of $\tilde{\delta}^h(l)$, $l = 1, \cdots, N$. In contrast, the IPT estimator includes a single estimate of $\tilde{\delta}^h(l)$, for all l.

Our GIPT discussion below closely parallels parts of the IPT approach of Graham et al (2012). When there is one treatment group and one control group, then let N_1 and N_0 denote the number of treated units and untreated units, respectively. First, for the unit at target point $L = l$ in the treatment group, the GIPT estimator of δ, denoted by $\tilde{\delta}^1$, is a solution to:

$$\text{(3.6)} \qquad \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{\hat{w}_i(l)\cdot D_i}{G\left(\tau(\hat{w}_i(l)\cdot x_i^1)'\tilde{\delta}^1(l)\right)} - 1\right\}\tau(\hat{w}_i(l)\cdot x_i^1) = 0,$$

where, given assumptions 5 and 9, $G\left(\tau(\hat{w}_i(l)\cdot x_i^1)'\tilde{\delta}^1(l)\right) = p(x)$ for all $x \in \mathbb{X}$ and some δ_1, $\tau(\hat{w}_i(l)\cdot x_i^1)$ is a $1+M$ column vector of known functions of X with a constant as its first element, and $\tilde{\delta}^1$ is a vector of estimates of δ_1. Following the logic of Graham et al (2012), the propensity score for the i^{th} unit in the treated sample can be written as:

$$\tilde{\pi}_i^1(l) = \frac{1}{N}\frac{\hat{w}_i(l)}{G\left(\tau(\hat{w}_i(l)\cdot x_i^1)'\tilde{\delta}^1(l)\right)}, \quad i = N_0+1, N_0+2, \cdots, N. \tag{3.7}$$

These two equations imply:

$$\sum_{i=N_0+1}^{N_1} \tilde{\pi}_i^1(l)\cdot\tau(\hat{w}_i(l)\cdot x_i^1) = \frac{1}{N}\sum_{i=1}^{N}\tau(\hat{w}_i(l)\cdot x_i^1). \tag{3.8}$$

Second, for the target point $L = l$ in the untreated group, the GIPT estimator of δ^0, denoted as $\tilde{\delta}^0(l)$, is the solution to:

$$\frac{1}{N}\sum_{i=1}^{N}\left\{\frac{\hat{w}_i(l)\cdot(1-D_i)}{1-G\left(\tau(\hat{w}_i(l)\cdot x_i^1)\tilde{\delta}^0(l)\right)} - 1\right\}\tau(\hat{w}_i(l)\cdot x_i^1) = 0, \quad i = 1, \cdots, N_0. \tag{3.9}$$

Similarly, the propensity score for the i^{th} unit in the control sample can be written as:

$$\tilde{\pi}_i^0(l) = \frac{1}{N}\frac{\hat{w}_i(l)}{1-G\left(\tau(\hat{w}_i(l)\cdot x_i^1)'\tilde{\delta}^0(l)\right)}. \tag{3.10}$$

These two equations imply:

$$\sum_{i=1}^{N_0}\tilde{\pi}_i^0(l)\cdot\tau(\hat{w}_i(l)\cdot x_i^1) = \frac{1}{N}\sum_{i=1}^{N}\tau(\hat{w}_i(l)\cdot x_i^1). \tag{3.11}$$

In words, equation (3.8) states that after twice reweighting the moments of x_i^1 across treated units, once with the propensity score parameters and once with the kernel weights, this equals the (kernel weighted) moments of x_i^1 over the entire sample. An analogous relationship for the untreated sample and the entire sample is in equation (3.11).

The GIPT ATE estimate for the unit at target point $L = l$ is given by

$$\tilde{\gamma}^{GIPT}(l) = \sum_{i=N_0+1}^{N} \tilde{\pi}_i^1(l) Y_i - \sum_{i=1}^{N_0} \tilde{\pi}_i^0(l) \cdot Y_i \qquad (3.12)$$

where $\tilde{\pi}_i^1(l)$ and $\tilde{\pi}_i^0(l)$ are target point dependent and defined by (3.7) and (3.10).

With GIPT we estimate an ATE for each target observation. In footnote 21 of the Appendix of Graham et al (2012), they describe the process for obtaining the overall ATE that is based on the single treatment effect for each observation. Our approach to obtaining the ATE for each target observation is similar to the overall ATE generation process outlined by Graham et al (2012), but we modify the moments condition using $\tau(\hat{w}_i(l) x_i^1)$ instead of $t(x)$. With GIPT, we obtain a very representative estimate of the ATE by generating an ATE estimate for each target point, rather than generating one treatment effect for each target point and using these to calculate one overall ATE. Assumptions 1 through 8 are satisfied in our Monte Carlo study below, and in many applications that consist of randomized treatments. In applications where there are multiple target points, we would expect GIPT would lead to a precise estimate of the ATE at each target point, and in turn, the overall average of the ATEs may have lower bias than the estimated ATE from IPT. On the other hand, in applications where the dimension of l is large, the performance of estimates of ATE might be affected negatively, as we can expect from any nonparametric estimator, especially when the sample size is small.

We next perform Monte Carlo simulations to demonstrate that the GIPT estimator performs well in small samples.

4. Monte Carlo Study

In our specific Monte Carlo Study we consider (without loss of generality) a model with heterogeneity in the geographic locations of observations. We first denote the two-dimensional vector, $l_i = [l_i^1, l_i^2]$. In this Monte Carlo study we generate our response variables, y_i, from the following causal model and selection model:

$$y_i = \beta_0(l_i) + DT_i \cdot DS_i \cdot \beta_1(l_i) + x \cdot \beta_2(l_i) + u_i, \tag{4.1}$$

$$DS_i = \begin{cases} 1 \quad for \quad l_i^1 + 0.25 \times l_i^2 > 1.25 \\ 0 \quad for \quad l_i^1 + 0.25 \times l_i^2 \leq 1.25 \end{cases}, \ i = 1, \cdots, N \tag{4.2}$$

$$DT_i = \begin{cases} 1 \quad for \quad i > N/2 \\ 0 \quad for \quad i \leq N/2 \end{cases}, \ i = 1, \cdots, N \tag{4.3}$$

where (4.1) is the causal model that produces the response variable y_i, (4.2) and (4.3) is the selection model that produces the treatment group. If DS_i equals 1, this indicates that the unit is in the location where some observations are treated and units with 0 will be in the control group. Also, DT_i is a dummy such that a value of 1 indicates an observation is only possibly treated shortly after an unexpected event. Therefore, the treated sample will be comprised of the observations for which $D_i = DT_i \times DS_i = 1$; in other words, the treated sample consists of those units for which both $DS_i = 1$ and $DT_i = 1$. The vector $l_i = [l_i^1, l_i^2]$ is a two-dimensional location vector generated from a bi-variate uniform distribution between $[0, 2]$, u_i is i.i.d. following a standard normal distribution; x_i^1 is a random variable generated from the normal distribution $N[0, 3]$, and v_i is i.i.d from the standard normal distribution. Additionally, for simplicity we set $\beta_0(l_i) = 0$ and $\beta_2(l_i) = 0.2$, and $\beta_1(l_i)$, our main interest in the estimation, is a variant of a bi-variate standard normal density function:

$$\beta_1(l_i) = \tfrac{1}{2\pi} \exp\left(\tfrac{(l_i^1)^2 + (l_i^2)^2}{2}\right).$$

Note that this data generating process - as given in (4.1) (4.2) and (4.3) - is designed to meet the assumptions discussed in Section 3. First, the distribution of the outcome, Y, is independent of the treatment status ("unconfoundedness"); Second, $\{Y_i, X_i, D_i\}_{i=1}^{N}$are i.i.d. (the “random sampling" assumption). Third, $\mathbb{P}(D_i = 1|Y, X) = \mathbb{P}(D_i = 1|X)$ (The "missing at random" assumption). Finally, $\mathbb{P}(D_i = 1|X = x) = \mathbb{P}(D_i = 1) > 0$, as D_i and X are independent in these data generating processes (The "strong overlap" assumption). The Gaussian kernel choice satisfies the symmetry assumption and the bandwidth will be determined to satisfy Assumption 8 (it is a nuisance parameter).

We use two different sample sizes, $N = 300$ and $N = 600$, as the number of individuals. This model is estimated with a variant of difference-in-differences[9] (hereafter denoted quasi-DID), IPT and GIPT as defined in section 2. For the GIPT estimator, the optimal bandwidth for each sample size is calculated through a grid search of 8 different bandwidths. For a grid of b values, the average squared error, $ASE(b) = \frac{1}{N}\sum_{i=1}^{N}\left\{\tilde{\gamma}_j^{GIPT} - \gamma_j^{GIPT}\right\}^2$, is computed for 100 replications and then averaged to estimate the mean ASE (MASE). The function $MASE(b)$ is then compared over the grid values of b. The optimal bandwidth, b_{MASE}, is chosen to be the value of b that yields the minimum MASE value. One optimal bandwidth is obtained for each sample size for the GIPT estimator. For the $N = 300$ sample, the optimal bandwidth is determined to be 0.85, and for the $N = 600$ sample the optimal bandwidth is 0.75. Next, using the optimal bandwidth for each sample size, we perform 500 iterations for each sample size, and then compute the average bias and ASE for each. The average bias and ASEs are reported in Table 1. In addition, in Figure 1 we also plot the distributions, with histogram and estimated density, of the ASE results from the 500 repetitions on each estimator with two different sample sizes.

Since some preliminary finite sample experimental evidence on the performance of the IPT estimator is already available (Graham et al, 2012), we are primarily interested in the performance of the GIPT relative to estimators that do not account for geographic variation. There are general regularities that are evident. As expected, increases in the sample size reduce the ASE for all estimators, suggesting that the estimators under study converge with sample size.

[9] We describe this as a variant of DID because we assume that we do not have multiple observations for the same target point.

Across both sample sizes, the IPT estimator performs at least as well as the quasi-DID estimator, in both ASE and average bias. Improvement of GIPT, as measured by MASE, over IPT and quasi-DID, ranges from 49% for $N = 300$ to 57% for $N = 600$. The key implication of these results is that in situations where geographic variation is an important factor in the data, the proposed GIPT estimator provides a simple but effective way to account for it. The ASE distribution plots in Figure 1 indicate a similar pattern. For each of the three estimators, increases in the sample size from 300 to 600 generally shift the ASE distribution towards zero. When the three estimators are compared with each other for the same sample size, the ASE distribution of GIPT are much closer to zero than that of the other two estimators.

We also plot the GIPT estimated ATEs based on our simulations, in Figures 2b and 3b (separately for $N = 300$ and $N = 600$, respectively). The corresponding true ATEs for these samples are plotted in Figures 2a and 3a, respectively. In comparing the GIPT ATEs against the corresponding true ATEs, it is apparent that as the sample size increases from $N = 300$ to $N = 600$, the GIPT ATEs more closely approximate the true ATEs. This implies that GIPT is a consistent estimator of the true ATEs as the sample size increases.

5. Application: Commercial Real Estate Prices in the Vancouver, BC Metro Area

Similar to the purpose of the IPT application in Graham et al (2012), our application is intended to illustrate the GIPT method as applied to a particular dataset and problem. The metro-Vancouver area was hit with a series of major storms in November, 2006, which led to severe mudslides that caused contaminated storm runoff to enter the water supply (Evans, 2007). Some parts of the metro area were required to boil water for an extended period of 10 days longer (i.e., 12 days total) than the rest of the metro area (CBC News, 2006). This impacted restaurants, coffee shops, and other water-dependent businesses (Dowd, 2006). The affected area included the City of Vancouver, while the adjacent City of Richmond (and many other parts of the metro area) had the advisory lifted on the second day. This may have been a type of information shock, which could influence the probabilities of similar advisories from future storms. We examine how sale prices for properties that sold within several months after this advisory in a section of Vancouver (the treated sample) were affected differently from other properties sold in the same section of Vancouver several months before the advisory and properties that sold in nearby parts of Richmond before and after the advisory (the control sample). Thus, our identification strategy relies upon an unexpected event (the extended water boil advisory) that affects some geographic areas but not others. We have a missing data issue with this data set, because we know what properties in the control group sold for, but we do not know what these properties would have sold for if they had been in the treatment group. Thus, a propensity score type of approach would be desirable. Meanwhile, there are clear differences in the geographic locations of properties in our sample. It is of interest to determine empirically how the effects of such a shock might be absorbed differently into property values across locations. Therefore, we consider three different approaches in this application, a variant of quasi-DID; IPT; and GIPT.

There is a literature that examines the effects of a storm on property values, including Bin et al (2013), Atreya and Czajkowski (2016), and others. None of this literature, however, considers the missing data problem in the same context or with the same approach as we are addressing it here. Also, most of the other studies in the literature focus on residential property values, while our application examines the commercial property value impacts (which is important in our context because many businesses in our sample are water dependent). Finally, we study the impacts of the storm using a quasi-experiment of the effects of a water boil advisory that was imposed on some areas of the metro area, including the City of Vancouver, for much longer than others. Therefore, we can examine the differential impacts of the water boil advisory on treated versus control areas, shortly before versus shortly after the advisory.

In the real estate finance and investments literature (e.g., Ling and Archer, 2017), a commercial property's value or sale price can be approximated by the ratio of its net operating income (NOI) to the capitalization rate (i.e., cap rate). In some cities, such as New Orleans, a major storm (i.e., a hurricane) such as Katrina led to property destruction as well as major disruption in abilities of businesses to operate for an extended period of time. In theory, if there is an event that alters an investor's estimate of basic long term risk, then such an event is often accompanied by an increase in the cap rate. In New Orleans, this increased risk likely led to a higher cap rate, due to the possibilities of repeat storm events in the future, which lowered the value of commercial properties. The storm also lowered the properties' NOI due to lost revenues, etc. People may have revised their estimate of New Orleans' vulnerability because of rising sea levels, eroded barrier marshes, etc. Although the impacts of the storm in Vancouver may have been somewhat different, this 12 day extended water boil advisory in the city of Vancouver caused major disruption of some business operations, especially for those that were water-oriented such as supermarkets, restaurants, day care facilities, etc (Dowd, 2006; CBC News, 2006). Such a disruption can be expected to lead to greater long-term risk of a repeat event for all properties; and/or lost revenues or additional insurance costs, for instance, for certain businesses that are water dependent. These financial losses can be expected to impact their NOI, which translates into an effect on property values and in turn, the sale prices of many properties. But other commercial property sale prices may not be affected, perhaps because they may not be as water dependent.

When we are estimating the ATE of the extended water boil advisory on the price per square foot of living area for commercial properties, the lot size (building area plus land area) of the property is expected to be negatively correlated with the NOI (and in turn, the total sale price). This is due to the fact that a larger lot size requires higher expenses for lawn maintenance and snow removal, for instance. But the effect of lot size on the price per square foot of living area may be either positive or negative. A larger lot size may or may not lead to economies of scale that are inherent in the maintenance of a commercial building. Greater economies of scale are expected to lead to higher NOI and therefore a higher price per square foot of the overall property. There also may be particularly strong price effects for older properties, or properties that have not been renovated recently. These older properties may be expected to rent for less, need more repairs, and require more to upkeep due to unanticipated issues resulting from the age of the property. This can also be expected to factor into the NOI for a property. In other words, an older property, or one that has not been renovated recently, should have a lower NOI than a similar, nearby property that has been renovated recently. Therefore, it is important to use the lot size and the effective age as a proxies for NOI, especially since we do not have direct estimates of NOI in our dataset. The effective age is the number of years between the year of most recent sale and the last major renovation of a property. Properties that were renovated in the year in which they were most recently sold have an effective age of 0. Similarly, properties that have never been renovated have an effective age equal to the actual age of the property. In our model specifications, we use as the control variable the interaction term of lot size (in thousand square feet) and the effective age of the property (in years). For reasons described above, these two variables are the two best proxies for NOI that we have available to us. Also, in the IPT and GIPT specifications, when we try to include two separate quasi-DID for these two variables, using the first two moments of each, the model is unable to solve. We are interested in the ATE from the extended water boil advisory, and we desire to control for the lot size and effective age as proxies for NOI but are not directly interested in their marginal effects. Therefore, using the interaction term enables us to control for both of these factors as proxies for NOI. Finally, Graham et al (2011) and Anderson (1982) suggest interaction terms be included in these types of propensity score models. So for all of these reasons, we use the first two moments of the interaction term in the IPT and GIPT specifications. Obviously, for consistency across specifications, we use the interaction term in the quasi-DID model as well. The impact of a change in cap rate associated with long-term risk due to the storm is reflected in the treatment effect dummy. Property owners are expected to adjust their forecasts of long-term risk after the storm, and this is reflected by the treatment effect estimate. One would expect property owners in different locations to have different forecasts of long-term risk, and therefore we might expect heterogeneity in the ATE estimates.

Also, in these types of treatment effect studies it is recommended to exclude observations in a buffer zone of properties that are excluded from the analysis (see, for instance, Angrist and Pischke (2009)). Therefore, we restrict our attention to a section of the metro area where some observations are in the City of Vancouver (which was subject to the water boil advisory for 12 days after the storm) and others in nearby parts of the neighboring City of Richmond (which had the water boil advisory lifted after one day). We avoid including properties outside of this buffer zone, e.g., in the central business district of Vancouver, where there are potentially many other confounding factors. Our focus on properties in the City of Richmond near the Vancouver border allows for a buffer zone consisting in properties in the western part of Richmond. We focus on a period of several months before, and several months after the 12 day water boil advisory which occurred for the City of Vancouver in November 2006. The choice of this time period allows for a buffer in the temporal dimension. We end our sample in August 2007 because we want to avoid the effects of the recession that started in late-2007, and we begin in January 2006 because we want to avoid other events that might have impacted property values before 2006 (thus, creating a temporal buffer beyond several months around the date of the storm).

In our data set, there are 96 commercial sales observations in the selected neighborhoods between January 2006 and August 2007 for which there are also data on sale price, square footage, lot size and the effective age. Figure 4 shows the locations of our sample of 96 commercial properties that sold (as arms-length transactions) in parts of the City of Vancouver and City of Richmond between January 2006 and August 2007. These data are from the BC Assessment database, which were purchased from Landcor.

Descriptive statistics are presented in Table 2. The average commercial property sold for approximately C$ 215 per square foot, had a lot size of about 35,000 square feet, had an effective age of 38.76 years (i.e., there were 38.76 years since the last major renovation), and 26 percent of the observations were in the treatment group (i.e., in the City of Vancouver - opposed to the City of Richmond - and sold after the extended water boil advisory was imposed on the City of Vancouver).

We first estimate the following variant of a quasi-DID model: $Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + e$, where Y_i is price per square foot for property i, X_i is the product of the lot size and the effective age. We assume that e is an i.i.d. error term with mean 0 and constant variance, and $E(e_i e_j) = 0$ for $i \neq j$. $D_i = 1$ for those properties in our data set that sold between November 2006 and August 2007 (i.e., after the extended water boil advisory), inside the City of Vancouver; and $D_i = 0$ for properties that sold in the City of Richmond before and after the advisory, and those properties that sold in the City of Vancouver before the advisory. The regression coefficient β_2 is the "treatment effect" of locating in the City of Vancouver after the storm.

The second model is IPT. We consider the first 2 moments so that $t(x) = [1, X, X^2]$, and X is the product of the lot size and effective age, and Y is the sale price per square foot. We reweight the X's so that the sample mean and variance of X in the treated sub-sample (and separately, in the untreated sub-sample) equals the entire sample mean and variance of X. We utilize the same data set as we used for the quasi-DID estimation. We calculate the ATE using IPT.

Finally, we estimate the GIPT model, with Gaussian kernel weights given as

$$\hat{w}_i(l) = \left[exp(-0.5 * (d_i(l)/b)^2)\right]^{1/2}, \tag{5.1}$$

where $d_i(l)$ is the Euclidean distance between property i and location l, and b is a bandwidth parameter. We explain the bandwidth determination in more detail below. In the GIPT model, we consider the first two moments and use $\tau(\hat{w}_i(l)X) = [1, \hat{w}_i(l)X_i^1, (\hat{w}_i(l)X_i^1)^2]$ for each target point, l. In this context, we are re-weighting by including distance weights in the propensity score weighted averages of X_i^1 so that the re-weighted mean and variance of X_i^1 for the treated sample equals the re-weighted mean and variance for the entire sample.

We present the quasi-DID and IPT results in Tables 3 and 4. First, with quasi-DID the treatment dummy, D_i, has a coefficient estimate of $\beta_2 = -49.97$, implying that the typical commercial property in the treated sample sold for approximately C$ 49.97 less per square foot than the typical property in the control sample. However, β_2, the ATE estimate, is highly insignificant (t-statistic=-1.24). With IPT, the ATE is C$ -50.37 (with t-statistic of -2.13), indicated by the coefficient "ate gamma" in Table 4.

With the GIPT approach, we first must determine the appropriate bandwidth. We first contemplate a "Rule of Thumb" bandwidth, as in Silverman (1986). However, this criterion requires normality of the distances data in order for it to be applicable. An informal examination of the locations of the properties in Figure 2 indicate that it is inconclusive as to whether the distances have a normal distribution. Therefore, without evidence of normality of these distances data, we estimate bandwidths in the range of 0.03 and somewhat higher and lower, moving up and down in units of 0.01. Bandwidths smaller than 0.03 cause difficulties in the GIPT estimations that preclude it from solving for many of the target points. We choose the smallest of these bandwidths, h=0.03, for which the GIPT estimations solve with ease. This bandwidth choice allows for the maximum amount of variation in the parameter estimates. In fact, as we experimented with increasing the bandwidth above h=0.03, the variation in the ATE estimates from GIPT across observations decreases dramatically, in general approaching the ATE estimate from IPT for the higher bandwidths. This result is expected, as with a higher bandwidth there are more observations receiving positive weight than with a lower bandwidth, so the GIPT ATE estimates with the higher bandwidths closely approximate the IPT ATE estimate.[10]

In terms of the GIPT assumptions that we describe in section 3.4 above, we propose that our data set and application satisfy these assumptions.[11] We rely on an identification strategy that considers properties that sold inside and outside of the water boil advisory zone, in a reasonably short time frame before versus after the extended water boil advisory date. For unconfoundedness, we assume that nearby properties have come from a randomized experiment, as the treatment does not depend on the price of the property. Specifically, properties that are close to each other do not necessarily have the same treatment status, as can be seen in Figure 4. Properties on the south side of the Frasier River are in Richmond (untreated), while those just to the north are in Vancouver (some of which are treated). Also, any given pair of properties in Vancouver that are close to each other are not necessarily both treated, because some of the nearby properties in the City of Vancouver sold shortly before the advisory and were therefore untreated. For our control variables, the interaction of effective age and lot size, it is reasonable

[10] For illustrative purposes in this empirical application, we select the smallest bandwidth for which the GIPT model is still able to solve. However, a more formal approach would be to follow an algorithm for the bandwidth selection, such as bootstrap bandwidth selection or Mean Average Squared Error (MASE) methods.

[11] While in this specific application it may not be straightforward to demonstrate that some of the GIPT assumptions are satisfied, our purposes are primarily illustrative of an application of the GIPT technique, as was the case for the IPT application in Graham et al (2012).

to assume that nearby observations have no impact on the value of these two variables at a particular target point. We have data missing at random, as we know what properties sold for at their location but not what they would have sold for at other locations. We also assume for the purposes of demonstrating how to implement the GIPT estimator that we have strong overlap. We also postulate that assumptions 5, 6, 8 and 9 hold, and since we use the Gaussian kernel, the symmetric kernel distribution (assumption 7) is satisfied. We next estimate the ATEs for all target points, l, using the GIPT estimator that we have developed in this paper. Figure 5 shows the ranges of ATEs of the metro-Vancouver area with the locations of the sample of commercial properties that sold in the period of our sample. This range is C$ 8.08 to approximately C$ -61.90 , but the former ATE has a relatively large standard error and is statistically insignificant. Most of the larger ATEs are statistically significant (P-value<0.05). Figure 4 demonstrates the locations of properties with ATE that have P-value<0.05. We take the mean of all of the 96 ATEs (which we denote as the "AATE"), in Table 5. The AATE equals approximately C$ -38.38, while the mean of the standard errors is C$ 22.54. In general, the properties with the most negative and significant ATEs are located in the central and south areas of Richmond and central Vancouver, while those with statistically insignificant ATEs are in east Vancouver.

While the ATE from quasi-DID and IPT are statistically insignificant, with GIPT we find that most of the 96 observations have negative ATEs, but 85 out of the 96 observations have statistically significant ATEs (with P-value<0.05). Thus, using GIPT enables us to unmask which specific locations would be significantly impacted by the storm related water boil advisory and which would not. Interestingly, many of the properties with significantly negative ATEs are concentrated in 5 distinct neighborhoods of Richmond (which did not experience the 12-day extended water boil advisory).

Within each of these 5 neighborhoods of Richmond, at least one (and sometimes several) of the properties in our sample are in a water-intensive industry. For instance, in a neighborhood around Horseshoe Way in the southern part of Richmond, there is a company that manufactures liquid cleaning products and health/beauty products. Nearby there is a recycling center and a millworks production company. While we expect the ATE of the liquid product manufacturing company property to be affected by an extended water boil advisory, the ATE of the other two companies properties in the same neighborhood are likely to be impacted by their proximity to the liquid product manufacturing company property. About 0.5 km south of this neighborhood is another cluster of properties with large negative, and statistically significant ATEs, including one where there is a company that processes fish products for use as fresh and preserved bait; nearby there is a produce market that undoubtedly relies on water to clean its produce; and an event planning company. In this situation, the fish products company and produce market may have a strong impact on the ATE of the event planning company due to its close proximity. Approximately 3 km north of this neighborhood (10011 Blundell Road in Richmond), there is a daycare facility with a statistically significant ATE, which was formerly a convenience market and the daycare moved into the space subsequent to the storm. The property may have had a negative, statistically significant ATE because the property relies daily on clean water for the children and staff to wash hands, dishes, etc, and if it had been in the treated group, this would have been expected to lower the value of the property. On the other hand, there is a daycare facility in Vancouver (3165 Kingsway, Vancouver) with a statistically insignificant ATE, which may be somewhat surprising, although perhaps this facility relies more on hand sanitizer and other less water-intensive ways to keep its students clean. A more plausible explanation is the fact that at this address there is also a lighting store that is likely not water intensive, so the presence of this store may offset the effect on the property's overall ATE from the daycare. Approximately 2 km to the northwest of the daycare in Richmond is a restaurant/bakery, and an office building. In this case, the restaurant/bakery clearly would be impacted by an extended water boil advisory, while the ATE of the office building may be impacted due to the proximity to the restaurant/bakery. Finally, approximately 0.5 km north of the restaurant/bakery there is a cluster of 4 other properties that have statistically significant (negative) ATEs. These include a large shopping plaza with restaurants, a coffee shop, doctor's offices, a drug store, and other offices. Very close to this shopping plaza is an automobile repair garage, a dermatology office, and an office building. It is likely that the water dependency of many of the businesses in the shopping plaza is one explanation for a significantly negative ATE for that property, while the significantly negative ATEs for the other nearby properties may be at least in part determined by proximity to the shopping plaza.[12]

Finally, one might argue that a fuzzy regression discontinuity framework could be appropriate for this particular problem, as in Angrist and Pischke (2009). But this is not the case in our specific application. The propensity score,

$$p(x) = Pr(D_i = 1|X_i = x) = E[D_i|X_i = x],$$

does not necessarily jump at any particular value of x. There are both large and small lot sizes in our sample of properties in Richmond and Vancouver, and also there are both old and new properties in both cities as well (as required by the strong overlap assumption of IPT). Therefore, our X, the interaction term of lot size and effective age, does not have a natural jump point in the probability of treatment at any specific value of x. In future work, it may be of interest to explore how to address potential fuzzy regression discontinuity in the context of IPT and GIPT, for specific applications where at particular values of x there is a natural jump point in the propensity score.

[12] One might conjecture that some of the differences in ATEs in the treated area (in the City of Vancouver after the boil water advisory) versus the control area (in the City of Richmond before the boil water advisory, and both Richmond and Vancouver before the advisory) may be due to differences in property tax rates in the two cities in these two years. We informally examined the property tax rates in these two cities in 2006 and 2007, and found that the 2006 base rate in Richmond for class 6 properties (commercial) was C$ 22.38361 per thousand dollars of assessed values. There were some additional add-ons for sewer debt, which ranged between C$ 0.23300 and C$ 0.28300 in 2006, implying a total tax rate of approximately C$ 22.64 per thousand dollars of assessed value. There is an additional parking tax for Richmond properties with parking, at a rate of C$ 0.78 per square meter of parking spaces. The 2007 tax rate in Vancouver for Class 6 properties (commercial) was C$ 24.87171. Therefore, there is a difference of approximately C$ 2.23 per thousand dollars of assessed value. Assuming this differential is expected to persist indefinitely into the future (i.e., an infinite time horizon), and a discount rate of 5%, this implies a difference of C$ 2.23*(1+0.05)/0.05 over the life of the property, or a total expected property tax differential of C$ 46.83 per thousand dollars of assessed value. We assume the sale price of a property is highly correlated with its assessed value. Then, if the ATE is C$ -45 for a property that sold in Richmond before the water boil advisory in 2006, for instance, then C$ 2.10 of this C$ -45, or less than 5% of the ATE, can be attributed to expected differences in property taxes in the two jurisdictions in the two years.

6. Conclusion/Discussion

We develop a GIPT estimator that allows for ATE heterogeneity across target points, and we prove consistency and asymptotic normality, as well as demonstrate the desirable small sample performance with Monte Carlo simulations. We demonstrate the use of this GIPT estimator in an application of how a major storm that leads to an extended water boil advisory in some areas impacts property prices differently in a major Canadian metro area. The GIPT estimator can be a useful technique to generate ATEs for each target location, and re-weight with propensity scores when there is missing data. As we show in our application and in our simulation study, there are several benefits, as well as some potential limitations, of using the GIPT approach in these types of applications. One advantage of GIPT is that we are able to generate heterogeneous ATE estimates for each target point. We can also test for the statistical significance of each of the ATEs. The average of the ATE's, or the AATE, is one way of summarizing this information over all target points, if so desired. In our specific application, one may be particularly interested in the ATE estimates that are statistically significant, in order to determine where remediation should be undertaken to try to prevent similar damage to the water supply in the future. There are many other potential missing data problem applications of the GIPT estimator where it would be desirable to generate heterogeneous ATEs.

Another advantage of using GIPT in applied settings, as demonstrated by our Monte Carlo simulations, is that the bias and average squared errors of the GIPT estimator appears to be lower than the bias for the quasi-DID and IPT estimators. Even when there is heterogeneity in the ATE estimates, GIPT is a more computationally intensive procedure and in some cases this may diminish its feasibility, especially in very large samples. However, there are approaches to address this issue in the nonparametric estimation literature, including limiting the number of target points to obtain a representative sample of ATE estimates. We have also addressed the important issue of bandwidth selection, which is crucial for each specific context of a given empirical application and Monte Carlo simulations when using the GIPT framework. As we have demonstrated in our application, the GIPT approach can extract important information about which individual observations have statistically significant ATEs, and it allows for heterogeneity in the magnitudes of the ATEs across space.

Clearly, there are advantages to both the IPT and GIPT approaches to addressing the missing data problem in generating heterogeneous estimates of ATE's. There is also evidence that GIPT is superior to quasi-DID. GIPT performs much better than quasi-DID in our Monte Carlo simulations, and this is to be expected, in part because quasi-DID ignores the missing data problem.

In future work, it would be of interest to consider modifying the GIPT framework to contexts where there is a balanced panel dataset (e.g., space-time), to address a broader array of applied missing data problems. Such an extension could also contribute to the literature on ATE heterogeneity by allowing for the possibility that the ATE could vary over target points and also over a long period of time. This may first necessitate extension of the regular IPT framework to a balanced panel data setting, as well as generating Monte Carlo evidence to validate the performance of the approach.

References

[1] Abrevaya, J., & Donald, S. G. (2017). A GMM approach for dealing with missing data on regressors and instruments. Review of Economics and Statistics, 99(4), 657-662.

[2] Allcott, H. (2015). Site selection bias in program evaluation. The Quarterly Journal of Economics, 130(3), 1117-1165.

[3] Anderson, J. A. (1982). 7 Logistic discrimination. Handbook of statistics, 2, 169-191.

[4] Angrist, J.D. & J.S. Pischke. (2009). Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press.

[5] Atreya, A., & Czajkowski, J. (2016). Graduated flood risks and property prices in Galveston County. Real Estate Economics.

[6] Bin, O., & Landry, C. E. (2013). Changes in implicit flood risk premiums: Empirical evidence from the housing market. Journal of Environmental Economics and management, 65(3), 361-376.

[7] Bitler, M. P., Gelbach, J. B., & Hoynes, H. W. Forthcoming. Can Variation in Subgroups' Average Treatment Effects Explain Treatment Effect Heterogeneity? Evidence from a Social Experiment. Review of Economics and Statistics.

[8] Brunsdon, C., Fotheringham, A. S., & Charlton, M. E. (1996). Geographically weighted regression: a method for exploring spatial nonstationarity. Geographical analysis, 28(4), 281-298.

[9] CBC News. 2006. Greater Vancouver boil-water advisory lifted. November 27. http://www.cbc.ca/news/canada/british-columbia/greater-vancouver-boil-water-advisory-lifted-1.584398 (accessed on 7/22/2017).

[10] Dowd, A. 2006. Water Warning Leaves Vancouver High and Dry. The Star Online, November 18. http://www.thestar.com.my/news/world/2006/11/18/water-warning-leaves-vancouver-high-and-dry_1/ (Accessed on 7/24/2017).

[11] Evans, G.M. 2007. Taking our water supply for granted. BC Medical Journal, 49(2): 62.

[12] Frölich, M., Huber, M., & Wiesenfarth, M. (2017). The finite sample performance of semi-and nonparametric estimators for treatment effects and policy evaluation. Computational Statistics & Data Analysis, 115, 91-102.

[13] Graham, B.S., C. Campos de Xavier Pinto, D. Egel. (2012). "Inverse Probability Tilting for Moment Condition Models with Missing Data," Review of Economic Studies, 79: 1053-1079.

[14] Graham, Bryan S., C. Campos de Xavier Pinto, D. Egel. (2011). " Inverse Probability Tilting Estimation of Average Treatment Effects in Stata." The Stata Journal, pp. 1-16.

[15] Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. Political Analysis, 20(1), 25-46.

[16] Hotz, V. J., Imbens, G. W., & Mortimer, J. H. (2005). Predicting the efficacy of future training programs using past experiences at other locations. Journal of Econometrics, 125(1-2), 241-270.

[17] Hsu, Y. C., Huber, M., & Lai, T. C. (2018). Nonparametric estimation of natural direct and indirect effects based on inverse probability weighting. Journal of Econometric Methods, 8(1).

[18] Imbens, G.W. (2004). "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review," Review of Economics and Statistics, 86: 4-29.

[19] Imai, K., & Ratkovic, M. (2014). Covariate balancing propensity score. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 76(1), 243-263.

[20] Ling, D., & Archer, W. (2017). Real estate principles: A value approach. 5th edition. McGraw-Hill Higher Education.

[21] McMillen, D.P. and J. F. McDonald (2004). "Locally Weighted Maximum Likelihood Estimation: Monte Carlo Evidence and an Application," in Luc Anselin, Raymond J.G.M. Florax, and Sergio J. Rey (eds.), Advances in Spatial Econometrics. New York: Springer, 225- 239.

[22] McMillen, D.P. and C. Redfearn (2010). "Estimation and Hypothesis Testing for Nonparametric Hedonic House Price Functions," Journal of Regional Science 50, 712-733.

[23] Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1), 41-55.

[24] Silverman, B. W. (1986). Density estimation for statistics and data analysis. Monographs on Statistics and Applied Probability, London: Chapman and Hall, 1986.

[25] Wager, S., & Athey, S. (2017). Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association, (just-accepted).

[26] Wooldridge, J. M. (2007). Inverse probability weighted estimation for general missing data problems. Journal of Econometrics, 141(2), 1281-1301.

Appendix A. Proof of Consistency and Asymptotic Normality of the GIPT average treatment effects estimator

First, given Assumption 8, it follows that the bandwidth b is a "nuisance parameter".[13] Next, given that the elements of the vector l are i.i.d., then $\hat{w}_i(l)$ is i.i.d., since a function of an i.i.d. series is i.i.d. Given that w depends only on l and b, we know that w is i.i.d. Note that since our random sampling assumption is that both X and D are i.i.d., we know that $w \cdot X$ and $w \cdot D$ are both i.i.d., since the product of two i.i.d. series is i.i.d. Thus, following the reasoning of Graham et al. (2012), who appeal to Wooldridge (2007), $\hat{\gamma}_{GIPT}$ is a consistent estimator of $\gamma_0(l)$ at any target point l.

Given the above i.i.d. discussion, the asymptotic normality of $\hat{\gamma}_{GIPT}$ follows from Theorem 6.1 of Newey and McFadden (1994), as described by Graham et al. (2012).[14] Let $\beta = (\gamma', \delta')'$, the $K + 1 + M \times 1$ moment vector and derivative matrix equal

$$m_i(\beta) = \left[\begin{array}{c} \frac{\hat{w}_i D_i}{G_i(\delta)} \Phi_i(\gamma) \\ \left(\frac{\hat{w}_i D_i}{G_i(\delta)} - 1 \right) \tau_i \end{array} \right],$$

where $\hat{w}_i = \hat{w}_i(l)$ and

$$M_i(\beta) = \left[\begin{array}{cc} \frac{\hat{w}_i D_i}{G_i(\delta)} \frac{\partial \Phi_i(\gamma)}{\partial \gamma}, & \frac{\hat{w}_i D_i}{G_i(\delta)} \frac{G_{1i}(\delta)}{G_i(\delta)} \Phi_i(\gamma) \tau_i' \\ 0, & \frac{\hat{w}_i D_i}{G_i(\delta)} \frac{G_{1i}(\delta)}{G_i(\delta)} \Phi_i(\gamma) \tau_i \tau_i' \end{array} \right].$$

The subscript "0" denotes the true value. First consider the case where Assumptions 1-8 hold. Let $M = \mathbf{E}\left[M_i(\beta_0)\right]$ and $\Omega = \mathrm{E}\left[m_i(\beta_0) m_i(\beta_0)'\right]$, then $\sqrt{N}\left[\hat{\gamma}_{GIPT}(l) - \gamma_0(l)\right] \xrightarrow{D} \mathrm{N}(0, \Delta_{\mathrm{o}})$for $\Delta_0 = \{(M'\Omega^{-1}M)^{-1}\}_{1:k,1:k}$, where $A_{1:k,1:k}$ is the upper left $K \times K$ block of A. The covariance of $m_i = m_i(\beta_0)$ equals

$$\Omega = \left[\begin{array}{cc} \mathbb{E}\left[\hat{w}_i^2 \frac{\Phi \Phi'}{G_i} \right], & E_0 \\ E_0' & F_0 \end{array} \right]$$

[13] We use a fi?rst stage estimator for the bandwidth that converges at the "correct rate" (to be a nuisance parameter) under the regularization assumption imposed in Assumption 8.

[14] Double robustness of GIPT follows from the proof of consistency in this appendix; however it is not straightforward to prove local efficiency of GIPT.

with

$$E_0 = \mathbb{E}\left[\frac{\hat{w}_i^2 \Phi \left[1 - \frac{G}{\hat{w}_i}\right]\tau'}{G}\right], \qquad F_0 = \mathbb{E}\left[\frac{\hat{w}_i^2 \left[1 - \frac{2G}{\hat{w}_i}\right]\tau\tau'}{G}\right].$$

The population mean of $M_i = M_i(\beta_0)$ is

$$M = \left[\begin{array}{cc} \mathbb{E}\left[\hat{w}_i \frac{\partial \Phi_i(\gamma)}{\partial \gamma}\right], & -\mathbb{E}\left[\hat{w}_i \frac{G_{1i}(\delta)}{G_i(\delta)} \Phi\tau'\right] \\ 0, & -\mathbb{E}\left[\hat{w}_i \frac{G_{1i}(\delta)}{G_i(\delta)} \tau\tau'\right] \end{array}\right].$$

So the limiting sampling variance for $\sqrt{N}\left[\hat{\gamma}_{GIPT}(l) - \gamma_0(l)\right]$ equals $M^{-1}\Omega M^{-1'}$.

When Assumption 5 does not hold, but Assumption 1- 4 and 6 - 9 hold, let$\beta_\star = (\gamma_0', \delta_\star')'$, where $\delta_\star'$ is the pseudo-true propensity score parameter. Let $G_\star = G(\tau(w(l)x^1)'\delta_\star)$,etc. Now,

$$\Omega_\star = \left[\begin{array}{cc} \mathbb{E}\left(\frac{\hat{w}_i^2 p_0(x)\Phi\Phi'}{G_\star^2}\right), & \mathbb{E}\left[\left(\frac{\hat{w}_i^2 p_0(x)\Phi}{G_\star^2} - \frac{\hat{w}_i p_0(x)\Phi}{G_\star}\right)\tau'\right] \\ \mathbb{E}\left[\left(\frac{\hat{w}_i^2 p_0(x)\Phi'}{G_\star^2} - \frac{\hat{w}_i p_0(x)\Phi}{G_\star}\right)\tau\right], & \mathbb{E}\left[\left(\frac{\hat{w}_i^2 p_0(x)}{G_\star^2} - \frac{2\hat{w}_i p_0(x)\Phi}{G_\star} + 1\right)\tau\tau'\right] \end{array}\right]$$

and

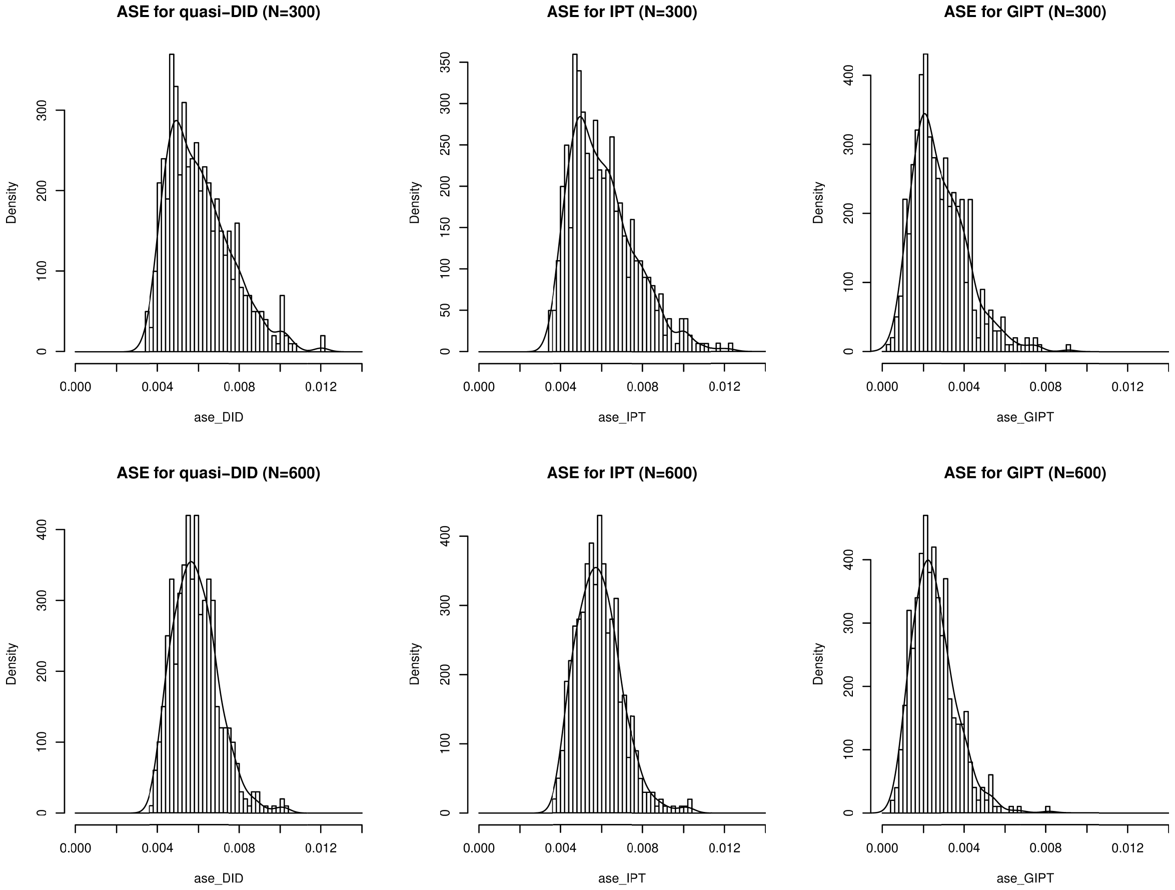
$$M_\star = \left[\begin{array}{cc} \mathbb{E}\left[\frac{\hat{w}_i p_0(x)}{G_\star}\frac{\partial \Phi_i(\gamma)}{\partial \gamma}\right], & -\mathbb{E}\left[\frac{\hat{w}_i p_0(x)}{G_\star}\frac{G_{1\star}}{G_\star}\Phi\tau'\right] \\ 0, & -\mathbb{E}\left[\frac{\hat{w}_i p_0(x)}{G_\star}\frac{G_{1\star}}{G_\star}\tau\tau'\right] \end{array}\right],$$

so that $\Delta_0 = \left\{\left(M_\star' \Omega_\star^{-1} M_\star\right)^{-1}\right\}_{1:k,1:k}$.

Consistent variance-covariance matrix estimation. If assumptions 1-4 and 6-8 hold, as well as either 5 or 9 hold (or both 5 and 9 hold), as well as additional regularity conditions, then the assymptotic variance at each target point l of $\hat{\gamma}_{GIPT}$ may be consistently estimated by $\hat{\Delta} = \left\{\left(\hat{M}'\hat{\Omega}^{-1}\hat{M}\right)^{-1}\right\}_{1:k,1:k}$ with $\hat{M} = \sum_{i=1}^N M_i(\hat{\beta})/N$, $\hat{\Omega} = \sum_{i=1}^N m_i(\hat{\beta}) m_i(\hat{\beta})'/N$.

Appendix B. Figures

Figure 1: Simulation Results on Average Squared Errors (ASE) Distributions From quasi-DID, IPT and GIPT [15] [16]



[15] In the ASE for GIPT (N=300) plot, an outlier value (maximum) is dropped for the convenience of plotting.

[16] Observe that 1.Increases in sample sizes reduce ASE for all estimators; 2. Across both sample sizes, ASE distributions from GIPT are closer to left and narrower, compared to that of quasi-DID and IPT, suggesting that the GIPT estimator out performs both quasi-DID and IPT.

Figures 2 and 3: Simulations Scatter-plots - The True ATEs and the GIPT Estimates

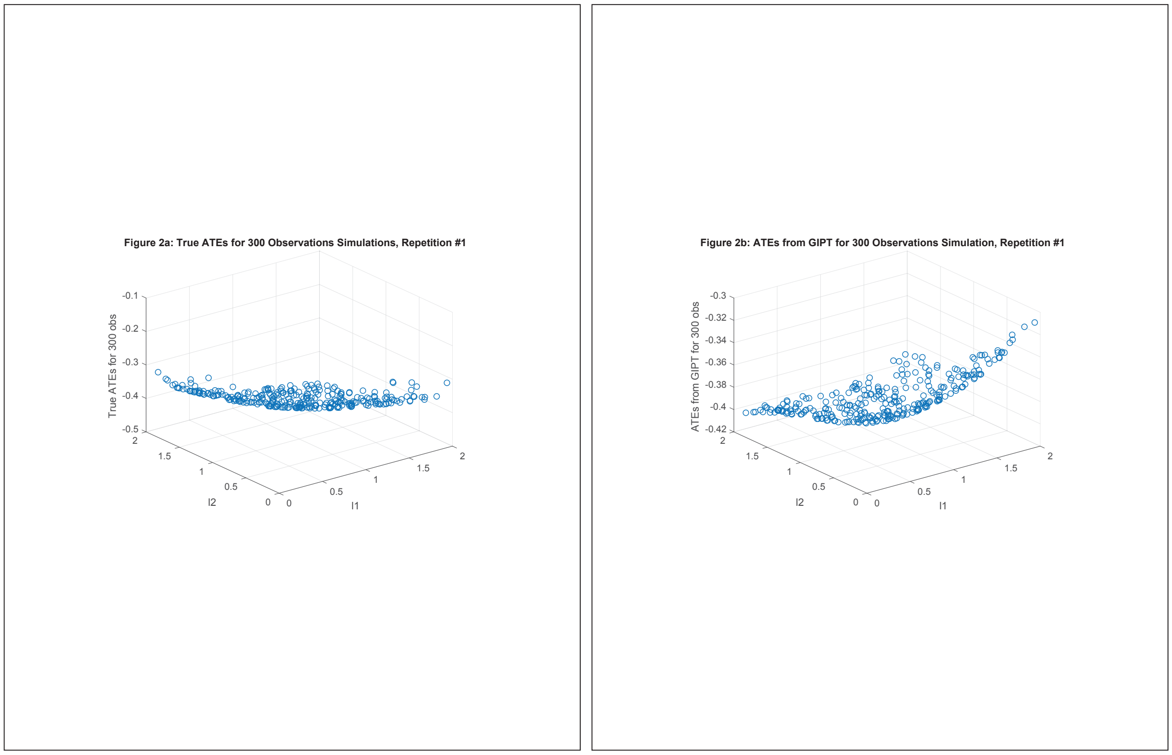


Figure 2a: True ATEs for 300 Observations Simulations, Repetition #1

Figure 2b: ATEs from GIPT for 300 Observations Simulation, Repetition #1

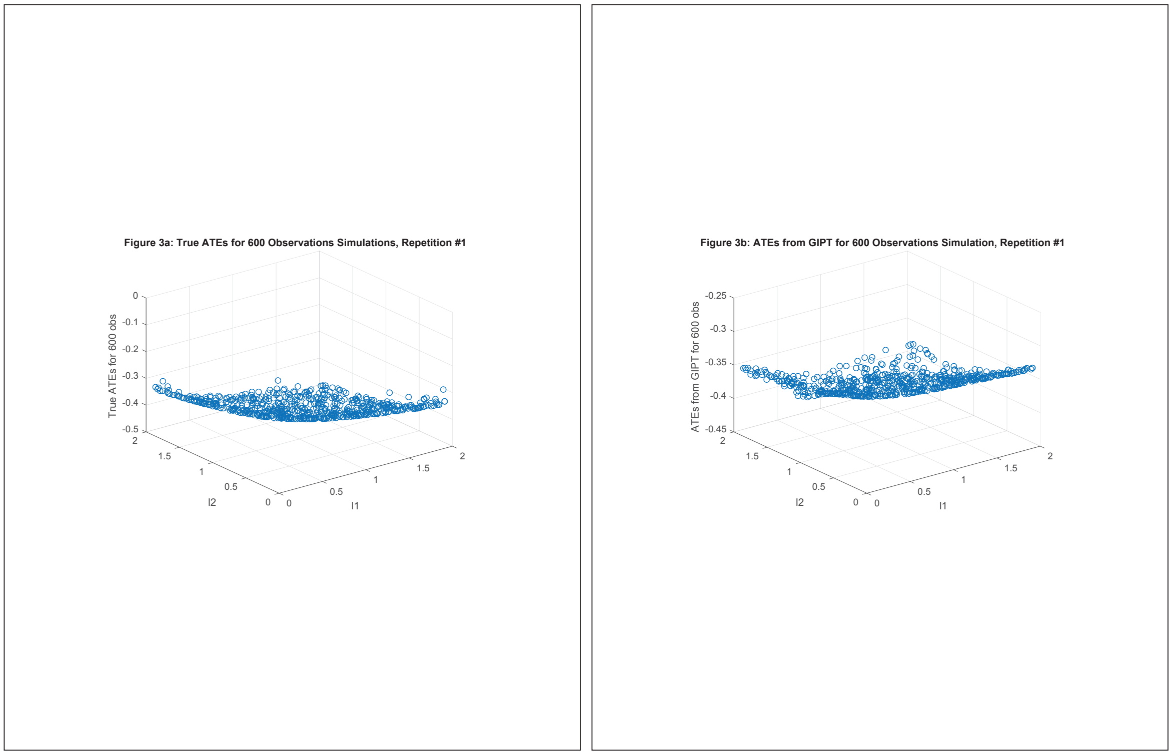
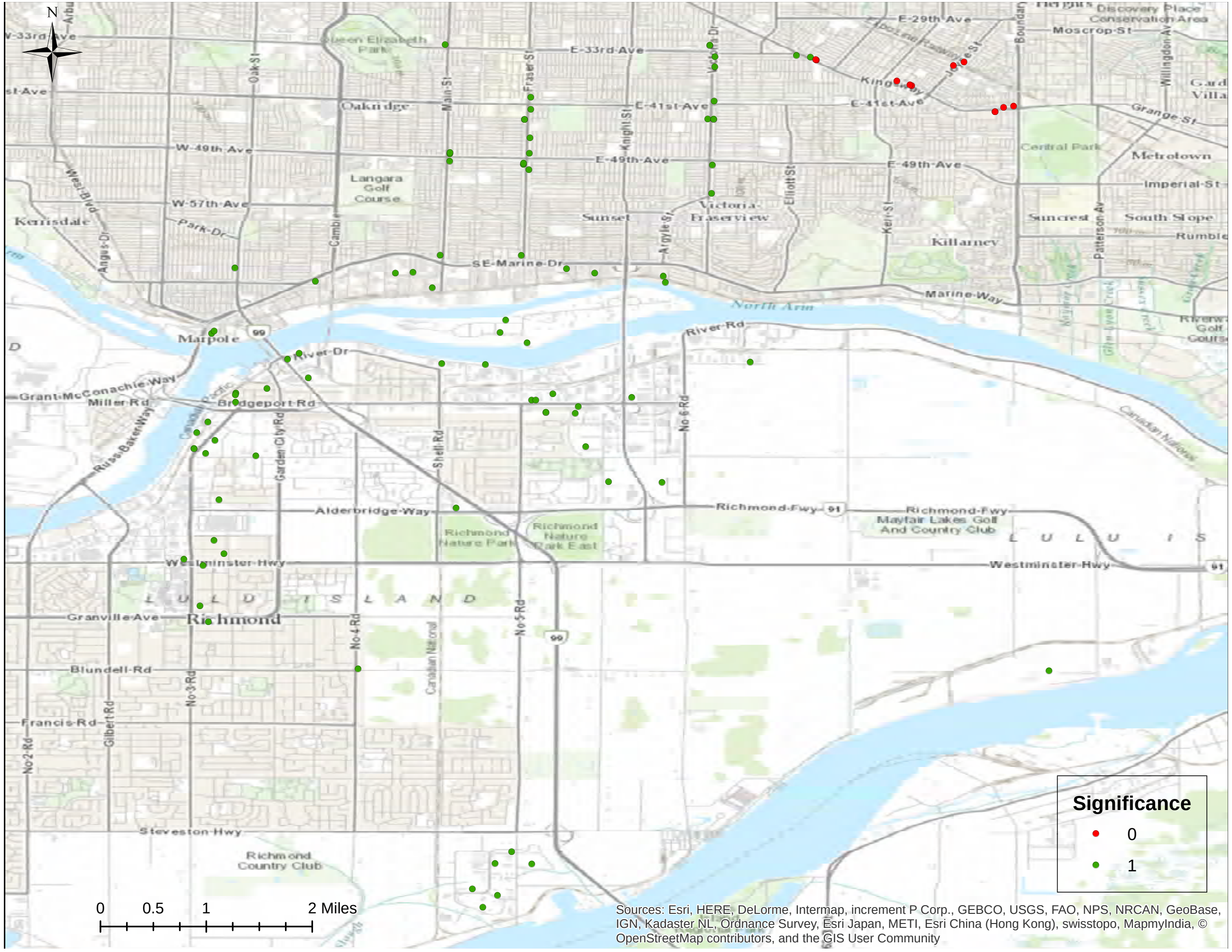


Figure 3a: True ATEs for 600 Observations Simulations, Repetition #1

Figure 3b: ATEs from GIPT for 600 Observations Simulation, Repetition #1

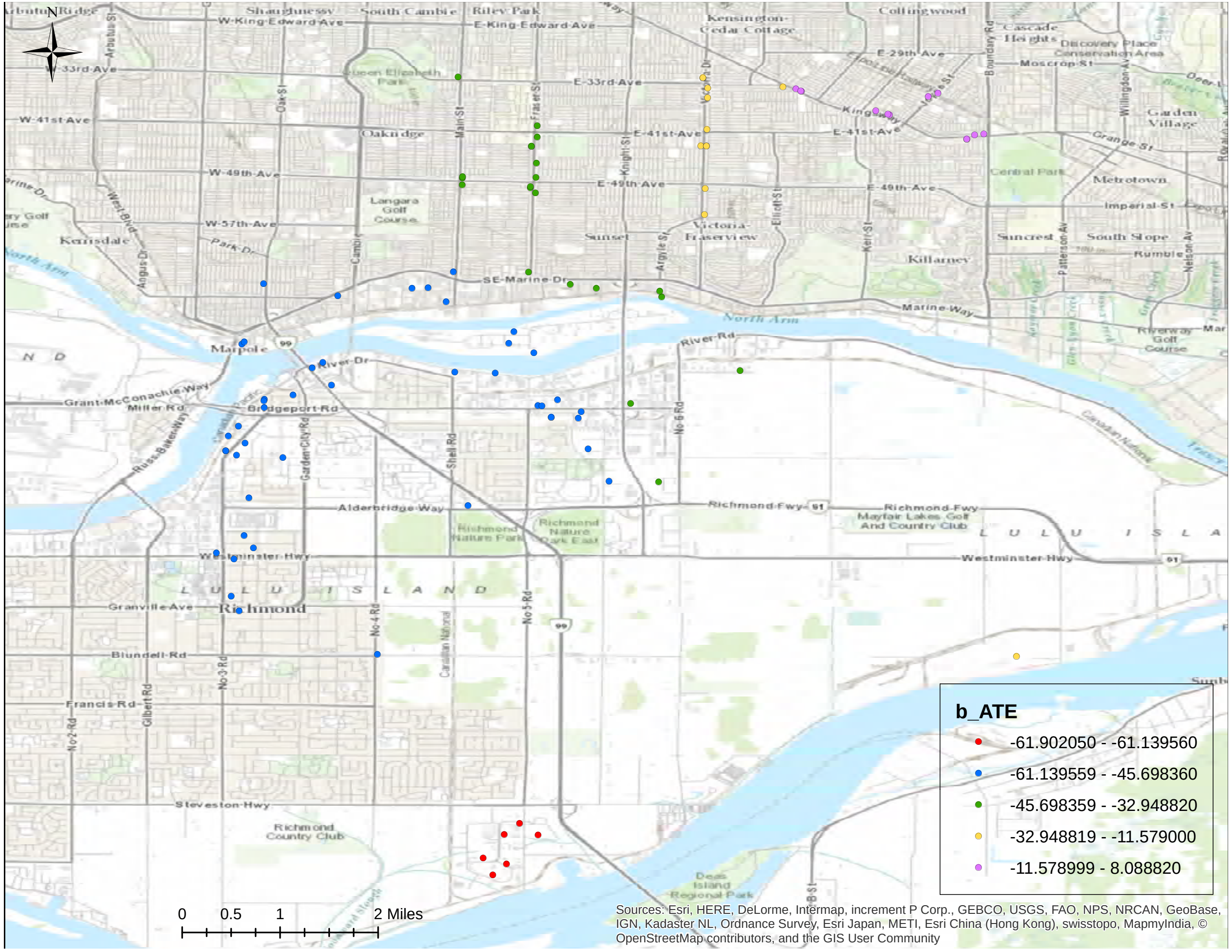
Figure 4: Properties with Statistically Significant ATE from GIPT Estimations[17]



Sources: Esri, HERE, DeLorme, Intermap, increment P Corp., GEBCO, USGS, FAO, NPS, NRCAN, GeoBase, IGN, Kadaster NL, Ordnance Survey, Esri Japan, METI, Esri China (Hong Kong), swisstopo, MapmyIndia, © OpenStreetMap contributors, and the GIS User Community

[17] Significance=1 if P<0.05, Significance= 0 o.w.

Figure 5: ATE Values from GIPT Estimations



Sources: Esri, HERE, DeLorme, Intermap, increment P Corp., GEBCO, USGS, FAO, NPS, NRCAN, GeoBase, IGN, Kadaster NL, Ordnance Survey, Esri Japan, METI, Esri China (Hong Kong), swisstopo, MapmyIndia, © OpenStreetMap contributors, and the GIS User Community

Appendix C. Tables

Table 1: Simulation Results - Small Sample Performances for GIPT, IPT and quasi-DID[18]

	GIPT	IPT	quasi-DID
		Sample Size = 600	
Bias	.0074211	-.0408718	-.0408698
ASE	.0025323	.0058796	.0058782
		Sample Size = 300	
Bias	.0015103	-.0410462	-.0410958
ASE	.0031035	.0060545	.0060559

[18] The bandwidth used for GIPT is 0.75 with N=600 and 0.85 with N=300. See section 4 for more details for bandwidth selection algorithm.

Table 2: Descriptive Statistics, Vancouver Application

	(1)				
	mean	sd	min	max	count
sale price per square ft	215.9012	169.2998	20.60159	1128.099	96
Effective Age	38.76042	12.05142	9	70	96
Lotsize(thous sqft)	34.92404	49.13306	2.76459	246.88	96
Treatment Dummy	.2604167	.4411657	0	1	96

Table 3: Quasi-DID Model Results, Vancouver Application

	(1)
	sale price per square ft
ATE	-49.97
	(-1.24)
[effective age]x[lotsize(thous square feet)]	-0.0467*
	(-1.69)
([effective age]x[lotsize(thous square feet)])^2	0.00000289
	(1.04)
Constant	270.3***
	(8.94)
R-sq	0.069
N	96

t statistics in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 4: Inverse Probability Tilting Estimation Results, Vancouver Application

	(1)
delta1	
[effective age]x[lotsize(thous square feet)]	-0.000775**
	(-1.98)
([effective age]x[lotsize(thous square feet)])^2	9.42e-08**
	(2.21)
Constant	-0.641**
	(-1.97)
delta0	
[effective age]x[lotsize(thous square feet)]	-0.00198
	(-1.35)
([effective age]x[lotsize(thous square feet)])^2	0.000000229
	(1.43)
Constant	-0.0760
	(-0.13)
ate	
gamma	-50.37**
	(-2.13)
Observations	96

t statistics in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 5: General Inverse Probability Tilting Results, bandwidth=0.03

	(1)				
	mean	sd	min	max	count
ATE	-38.38539	18.17553	-61.90205	8.08882	96
Standard Errors of ATE	22.54444	1.739582	18.68966	26.44306	96