Spatial Sorting, Agglomeration Economies, and Travel Cost Endogeneity in Recreation Demand Models*

Jacob T. Bradt †

June 2022

Abstract

Conventional recreation demand models assume that travel cost is exogenous. In reality, the costs that individuals face when choosing a recreation site to visit are the result of a spatial sorting equilibrium, which may be influenced by access to outdoor recreation sites and their environmental amenities. I explore the bias introduced by ignoring the potential for recreation sites and their attributes to determine travel cost in conventional recreation demand models and provide an instrumental variables approach to accounting for this endogeneity problem. I demonstrate the importance of accounting for the spatial non-uniformity of recreation sites and residences in a series of numerical simulations, finding that the instrumental variables approach ensures coverage of true parameter values. I implement the approach in a nationwide model of demand for overnight campground reservations as a function of price, water quality, and other observable site attributes. I find that not correcting for travel cost endogeneity via the instrumental variables approach nearly doubles estimates of consumers willingness-to-pay for improvements in water quality. This highlights the importance of relaxing the assumption of exogenous travel costs in real world applications.

*I thank Joe Aldy, Richard Zeckhauser, Frank Pinter, Jim Stock, and Devesh Raval as well as seminar and conference participants at the 2021 Association of Environmental and Resource Economists (AERE) Summer Conference and Harvard for invaluable comments and discussion. I gratefully acknowledge funding from the Vicki Norberg Bohm Fellowship. I am responsible for all remaining errors.

†Harvard University, John F. Kennedy School of Government; jbradt@g.harvard.edu.

1 Introduction

Outdoor recreation is an industry of both intrinsic and instrumental value. Private spending on recreational activities represents 2.1 percent of United States GDP annually (Bureau of Economic Statistics, 2019) and participation in outdoor recreation activities has increased in recent decades, with visitation to outdoor sites administered by the National Park Service rising by 16 percent over the period from 2010-2019 (National Park Service, 2020).

In addition to being an economically-meaningful industry in its own right, outdoor recreation provides insight into individuals' interactions with the natural environment. Environmental conditions at outdoor recreation sites—air quality on mountaintops, water quality at fishing sites, flora at campsites—generate substantial non-market value to visitors, affecting the quality of visitors' experience. As a result, decisions of which outdoor recreation sites to visit reveal how individuals' value the natural amenities of those sites. Studying consumer demand for outdoor recreation therefore not only sheds light on a consequential industry, but also allows us to understand the value of important environmental qualities which do not trade in market settings and whose value therefore proves difficult to quantify. Understanding these values has important implications for—among other public and private sector activities—the provision of public goods, the development of regulations, and litigation of environmental damages.

I re-examine a key assumption that underlies empirical applications of recreation demand estimation, namely, that individuals' proximity to recreation sites is exogenous to their recreation decisions. Recreation demand estimation is a common tool in the literature examining the willingness-to-pay (WTP) for changes to non-market, environmental amenities, such as water or air quality. The logic underlying this approach stems from Hotelling (1947)'s simple insight that consumption of an outdoor recreation site's amenities requires the agent to incur the cost of a trip to that site. The cost of traveling to a given site therefore serves as an implicit price for visits to that site, and site visits serve as quantities demanded. With data on travel costs and visitation patterns for different recreation sites, it is therefore possible to estimate models of demand for site visits. Moreover, it is possible to define recreation sites as bundles of attributes—including, for example, the types of outdoor activities supported, environmental qualities, and physical amenities, among others—so with data on these attributes across sites or over time, researchers can estimate how these amenities affect recreation demand. Estimating recreation demand as a function of travel cost and site attributes, including environmental amenities, allows for the valuation of changes in environmental quality.

Researchers studying recreation demand often implicitly assume that price—defined as

individual-specific travel cost—is exogenous in the visitation decision. However, there is ample reason to believe that correlation between travel cost, which is a direct function of individuals' choice of residence location, and unobserved characteristics may exist. The standard random utility models used to study recreation demand implicitly assume all characteristics of recreation sites are observed, a non-trivial assumption considering the large set of factors which may influence recreation decisions. While this alone does not call into question the assumption of price exogeneity, the potential for preferences for outdoor recreation and the consumption of environmental amenities to enter into the decision of residence location suggests that non-random correlations between travel costs and unobserved characteristics may exist. For example, if anglers choose their permanent residence based on attributes of neighboring fishing sites or climbers choose to reside near high-value rock climbing locations, then estimates of individuals' responsiveness to price—the travel cost parameter—will be biased. This has important implications for the estimation of WTP or the measurement of welfare under different policy counterfactuals using standard recreation demand models.

There is ample evidence to suggest that individuals do indeed sort—i.e., select their residence location—based on natural amenities and proximity to outdoor recreation opportunities. Work in the regional science and demography literature documents a decades-long trend of in-migration to areas with high environmental qualities in the rural US (Hjerpe et al., 2020; Rickman and Rickman, 2011). Recent empirical work in the spatial and urban economics literature also documents Tiebout (1956)-like sorting on preferences for spatial characteristics (Bayer and Timmins, 2007; Klaiber and Phaneuf, 2010), including climate and environmental qualities (Albouy et al., 2016; Bayer et al., 2009). Moreover, preferences for environmental qualities entering the choice of residence location is an implicit assumption of one of the other main techniques used to value non-market environmental amenities, the hedonic property framework of Rosen (1974). Indeed, much of this literature finds non-trivial capitalization of environmental amenities in housing prices, suggesting that individuals' do consider these factors in choosing where to live (Bishop et al., 2020).

While there is strong evidence that individuals may sort on preferences for outdoor recreation use values, there is also expansive evidence that households may sort away from high value recreation sites. In particular, there is a large literature that documents that agglomeration economies draw individuals towards urban centers (Glaeser, 2010). These urban centers are likely located far from remote, high value recreation sites, resulting in a non-random spatial distribution of household locations and recreational sites. While this form of "negative sorting"—i.e., sorting away from recreation sites—is fundamentally different from the "positive sorting"—i.e., sorting towards recreation sites—that I discuss above, both arise due to non-random forces resulting in a non-uniform distribution of residences and recreation

sites. Failure to account for this non-uniformity, regardless of underlying cause, can result in biased estimates of individuals' sensitivity to travel cost in recreation demand estimation.

To address the potential for sorting to impact recreation demand estimation, I adapt a standard econometric approach that accounts for endogeneity in discrete choice models to the recreation demand context. Specifically, I outline how a two-stage control function approach to recreation demand estimation can mitigate concerns of bias introduced by travel cost endogeneity. This approach, first introduced by Heckman (1978), is widely applied in other contexts, including the management literature (Petrin and Train, 2010; Villas-Boas and Winer, 1999). The approach is analogous to the two-stage least squares estimator in linear models: in the first stage, travel cost is regressed in a linear model on a set of instruments which plausibly satisfy instrument relevance and an exclusion restriction and the residuals from this regression are included in estimation of the non-linear discrete choice model of site choice in the second stage. I demonstrate the effectiveness of this approach in Monte Carlo simulations as well as a nationwide model of recreation trip demand as a function of price, water quality, and other observable site attributes. Given the implications for the design of environmental policy, the assessment of environmental damages, and other important phenomena, my objective in applying this approach to estimate a nationwide recreation demand model is to quantify how ignoring travel cost endogeneity can bias estimates of WTP and the welfare effects of changes in water quality in a real-world setting.

Monte Carlo simulations reveal that employing a two-stage control function approach effectively eliminates bias in estimates of individuals' travel cost sensitivity, substantially outperforming more standard indicator variable approaches to account for unobservable factors. For example, in one set of numerical simulations, I find that the control function approach reduces mean squared error by 80% over an approach that relies on site-specific constants to control for unobservable factors. Using a control function approach in the context of a nationwide model of campsite demand, I find non-trivial differences in parameter estimates and calculated welfare changes under different environmental policy counterfactuals when accounting for the non-uniform distribution of residences and recreation sites. In particular, I estimate that not correcting for endogeneity via a control function nearly doubles estimates of consumers' WTP for improvements in water quality. While on an individual-level, a change in marginal WTP on the order of several dollars appears small, this difference can lead to drastically different conclusions when aggregated over large populations as is often done in regulatory impact analysis or the estimation of environmental damages.

The remaining sections of the paper are organized as follows. Section 2 discusses the relevant literature. Section 3 presents a discrete choice model of recreation site decisions with endogenous travel costs and outlines the control function approach to account for this

endogeneity. Section 4 presents evidence on the effectiveness of the control function approach from Monte Carlo simulations and Section 5 describes the empirical setting and data on campground reservations that I use as an application of this correction. Section 6 presents the results from this empirical application and Section 7 concludes.

2 Related Literature

This paper relates to several broad literatures. The first is the expansive literature using recreation demand models to value non-market environmental amenities. Early empirical implementations of the travel cost logic—first introduced by Hotelling (1947)—estimates single site demand, mostly using zonal aggregate data on individuals' travel costs (Ward and Loomis, 1986). Later work notes the importance of accounting for substitution across recreation sites, relying on McFadden (1974)'s random utility maximization (RUM) framework to model agents' choice among a discrete set of potential sites (Phaneuf and Smith, 2005). First applied to the context of recreation demand by Hanemann (1978), RUM models are the dominant approach for describing consumer preferences for recreation and are used to value water quality changes (Abidoye and Herriges, 2012; Abidoye et al., 2012; Egan et al., 2009; Smith et al., 1986), fish abundance (Kling and Thomson, 1996; Parsons et al., 2000; Shaw and Ozog, 1999), beach width (Parsons et al., 1999), and a host of physical site amenities (Hicks and Strand, 2000).

Several studies recognize the potential for the endogeneity of site attributes in recreation demand models.Murdock (2006) develops an estimation procedure that accounts for unobserved site-specific attributes in the RUM model using alternative-specific constants while also allowing for inference on time-invariant, observed site characteristics. In describing potential sources of endogeneity in the travel cost variable, Murdock (2006) discusses the non-uniform distribution of residences and recreation locations. Abidoye et al. (2012) similarly construct a random utility model with a full set of alternative-specific constants to account for unobserved site attributes; however, their approach differs from Murdock (2006) in that they employ a Bayesian approach to estimating model parameters describing preferences for time-invariant attribtes. Noting that congestion—a key site attribute, which is often omitted in recreation demand modelling—is endogenously determined by individuals' site visitation decision, Timmins and Murdock (2007) use an instrumental variables approach to account for this endogeneity in a revealed preference context. von Haefen and Phaneuf (2008) demonstrate that, when available, stated preference data can be combined with revealed preference data to identify site quality effects on behavior in the presence of unobservable site and user characteristics.

I build on recent work on endogenous attributes in random utility models of recreation demand by illustrating residential sorting as a source of endogeneity and demonstrating a simple fix to account for this issue in estimation. To account for travel cost endogeneity without fully modeling the residential sorting process, I rely on a simple, yet flexible control function approach. While the alternative-specific constant approach of Murdock (2006) is feasible in many settings, the control function approach for which I advocate is computationally and conceptually straightforward and easily allows for inference on site-specific attributes. Moreover, the approach of relying on alternative-specific constants does not mitigate the particular source of endogeneity in question: since the form of travel cost endogeneity arises over individual decision makers rather than sites, allowing for mean valuations of sites through alternative-specific constants leaves meaningful residual variation which may influence parameter estimates. I formalize this point in Section 3.3.

This paper also relates to a similarly expansive literature examining the capitalization of non-market environmental amenities in housing prices. Most of the results examining the price effects of environmental qualities use the hedonic property framework of Rosen (1974) to estimate capitalization in home prices. The hedonic framework has been applied to value proximity to hazardous waste sites (Greenstone and Gallagher, 2008), changes in air quality (Bajari et al., 2012; Bento et al., 2014), proximity to shale gas wells (Muehlenbachs et al., 2015), flood risk (Hallstrom and Smith, 2005), and water quality (Keiser and Shapiro, 2019). Generally, these studies find evidence in favor of capitalization of environmental amenities: residential transactions appear to account for a home's exposure to environmental amenities, both positive (i.e., amenities) and negative (i.e., disamenities). These general findings suggest that environmental qualities—the recreation site attributes of interest in most applications of recreation demand modelling—do indeed play a role in determining individuals' permanent residence location.

Several papers do consider both recreation site choices and residence locations when valuing non-market environmental amenities. Phaneuf et al. (2008) point out that conventional hedonic property studies estimating WTP for non-market environmental amenities may not fully capture the set values that homeowners derive from non-market environmental amenities. In particular, Phaneuf et al. (2008) argue that recreational use values are not fully incorporated in valuations of environmental amenities derived from conventional hedonic analyses. The authors present a theoretical model which motivates a two-stage revealed preference model in which a recreation demand model is first estimated as a function of the environmental quality of interest, and the resulting estimates of marginal welfare gains from changes in environmental amenities are then incorporated into a standard hedonic property model. Phaneuf et al. (2008) apply this conceptual model to study ecosystem services deliv-

ered by a watershed, finding that accounting for recreational use values in a hedonic property model meaningfully increases estimates of welfare derived from the presence of a watershed. Kuwayama et al. (2020) apply the approach of Phaneuf et al. (2008) to estimates WTP for water quality improvements in Tampa Bay, FL. I build on this literature by demonstrating the importance of accounting for not only recreation demand in models of residence location choice, but also residence location choices in models of recreation demand.

3 A Model of Recreation Demand with Endogenous Travel Cost

This section presents a standard discrete choice model of demand for recreation stes and outlines a method for accounting for the potential endogeneity of travel cost due to non-random residential sorting. I begin by presenting a discrete choice model of site selection before introducing the main endogeneity concern. Finally, I present the two-stage control function approach.

3.1 Discrete Choice Model of Site Selection

The basic RUM hypothesis assumes that individuals select the alternative yielding the highest level of utility when facing a well-defined choice set (McFadden, 1974). Let u_{ijt} denote the conditional utility received by individual i when selecting alternative $k \in \{1, ..., J\}$ on choice occasion t. The individual selects alternative $j \in \{1, ..., J\}$ if and only if $u_{ijt} > u_{ikt} \forall k \neq j$. Let $y_{ijt} = 1$ if individual chooses alternative j and $y_{ijt} = 0$ otherwise, i.e.

$$y_{ijt} = \begin{cases} 1 & u_{ijt} > u_{ikt} \forall k \neq j \\ 0 & \text{otherwise} \end{cases}$$

Since it is not possible to observe all factors influencing individual site selection decisions, conditional utility is parameterized as a function of observable individual- and alternative-specific attributes, $\boldsymbol{X_{ijt}}$, and some residual term, ε_{ijt}, which is known to the individual when making their decision, but unobserved by the econometrician. In particular, individual i's conditional utility from visiting recreation site j on choice occasion t is a function of observed attributes $\boldsymbol{X'_{ijt}} = (\boldsymbol{q'_{jt}}, \boldsymbol{x'_{jt}}, c_{ijt})$, where $\boldsymbol{q_{jt}}$ is a vector of environmental qualities, $\boldsymbol{x_{jt}}$ is a vector of observed site-specific attributes and c_{ijt} is person- and site-specific travel cost, and a residual term ε_{ijt}:

$$u_{ijt} = \underbrace{\boldsymbol{q'_{jt}}\boldsymbol{\beta_i^q} + \boldsymbol{x'_{jt}}\boldsymbol{\beta_i^x} - c_{ijt}\alpha_i^c}_{v_{ijt}(\boldsymbol{X}_{ijt};\theta_i)} + \varepsilon_{ijt} \tag{1}$$

The standard approach to fully specify the model given in Equation 1 is to make an assumption on the distribution of the idiosyncratic shocks to preferences, i.e., the residual term ε_{ijt}. While there are several different distributional assumptions made in the recreation demand literature, the most common of these is that ε_{ij} is distributed Type 1 Extreme Value (T1EV) across the population, and is iid across individuals and sites, which corresponds to the logit model. This has a number of desirable properties, including the fact that each individual's resulting choice probability for the different alternatives has a simple, closed-form solution.

Note that Equation 1 allows the coefficients on the additive, observed components of utility, $\boldsymbol{\theta_i}' = (\boldsymbol{\beta_i^{q\prime}}, \boldsymbol{\beta_i^{x\prime}}, \alpha_i^c)$, to vary across individuals i in the population. To make this assumption tractable, it is typically assumed that this individual-level heterogeneity follows some parameterized distribution, $f(\theta)$. This specification allows for heterogeneity in preferences for the different observable attributes and, when combined with the assumption that the error term is distributed T1EV, is referred to in the literature as the mixed or random parameters logit. Allowing for individual-level preference heterogeneity has several benefits over a standard logit model without individual-level parameters. First, it results in unrestricted substitution patterns. In the standard logit framework, two alternatives with equivalent choice probabilities will have the same substitution patterns. This property is undesirable in many contexts, the recreation demand context included: just because two sites have similar probabilities of being visited in the data does not mean that individuals are equally likely to substitute towards them as a result of a change in travel cost or environmental quality. Moreover, allowing for individual-level preference heterogeneity enables the model to capture the important variation in preferences for recreation and environmental amenities which drive the sorting process giving rise to the potential for endogeneity of the travel cost variable.

A common assumption in the literature estimating mixed logit models is that the individual-level parameters are normally distributed, $\theta_i \overset{iid}{\sim} \mathcal{N}(\mu, \Sigma)$. This means that means and standard deviations are estimated for each normally-distributed coefficient with individual-level heterogeneity, thereby providing information on the distribution of preferences for different observed attributes in the population. The value in making such a parameteric assumption is that rather than the infeasible objective of estimating the full set of individual-specific parameters, θ_i, the objects of interest for estimation are now the parameters of the distribution of the individual-level parameters. This greatly reduces the dimensionality of the model and makes estimation tractable.

Taking the common assumption of an extreme-value error term, $\varepsilon_{ijmt} \overset{iid}{\sim}$ T1EV, it is possible to specify the closed-form choice probabilities in this model. In particular, the

probability that individual i chooses site j on choice occasion t is:

$$p_{ijt} = Pr(j \in \arg\max_{k \in \mathcal{C}} u_{ikt}) = \int \frac{\exp(v_{ijt}(\boldsymbol{X}_{ijt}; \theta))}{\sum_{k \in \mathcal{C}} \exp(v_{ikt}(\boldsymbol{X}_{ijt}; \theta))} f(\theta) d\theta \tag{2}$$

where $\mathcal{C} = \{1, ..., J\}$ is the choice set. This random parameters logit probability is a weighted average of the logit formula evaluated at different values of the parameters, θ, with the weights given by the density $f(\theta)$. Estimation proceeds via simulated maximum likelihood, where the choice probability given by Equation 2 is approximated by the average across a large number of draws from, for example, $\theta_i \overset{iid}{\sim} \mathcal{N}(\mu, \Sigma)$ and the simulated log likelihood is defined as:

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{j=1}^{J} y_{ijt} \log(\hat{p}_{ijt}(\theta)) \tag{3}$$

The maximum simulated likelihood estimate is the value of θ that maximizes Equation 3.

3.2 Endogenous Travel Cost

Consider the issue of the endogeneity of the travel cost parameter. This arises due to non-zero correlation between the travel cost and unobserved site characteristics. Such correlation may be due to individuals who like certain sites or site attributes choosing to locate nearer to those sites, or perhaps due to the potential for high-value, more ecologically pristine sites to be located further away from population centers. The important point is that, though this endogeneity is the result of unobservable site characteristics, it arises over decision makers rather than over sites or groups of sites. Thus, a dummy variable approach is not feasible to fully account for the endogeneity concern in this setting: it is not possible to estimate constants for each decision-maker, since the constants would be infinity (for the chosen alternative) and negative infinity (for the non-chosen alternatives), perfectly predicting the choices and leaving no information for the estimation of target model parameters.

To be more precise about the nature of the endogeneity problem, consider Equation 1. Re-writing the residual as $\varepsilon_{ijt} = \zeta_{ijt} + \tilde{\varepsilon}_{ijt}$, where ζ_{ijt} is the unobserved component of the residual with which travel cost is correlated and $\tilde{\varepsilon}_{ijt}$ is iid extreme value, gives the following specification of individual i's utility from alternative j on choice occasion t:

$$u_{ij} = v_{ijt}(\theta_i) + \underbrace{\zeta_{ijt} + \tilde{\varepsilon}_{ijt}}_{\varepsilon_{ijt}} \tag{4}$$

The nature of this endogeneity problem—individuals making residence decisions based on

recreation sites—is such that c_{ijt} is endogenous, whereas $(\boldsymbol{q'_{jt}}, \boldsymbol{x'_{jt}})$ are exogenous. Let the endogenous travel cost be expressed as a function of observed instruments, z_{ijt}, and unobserved factors, μ_{ijt}:

$$c_{ijt} = w(z_{ijt}; \gamma) + \mu_{ijt} \tag{5}$$

where $w(\cdot)$ is some function of the observed instruments parameterized by γ and $z_{ijt} \perp \mu_{ijt}, \zeta_{ijt}$, but $\mu_{ijt} \not\perp \zeta_{ijt}$. The correlation between μ_{ijt} and ζ_{ijt} implies that c_{ijt} and ζ_{ijt} are correlated, which is the motivating endogeneity problem.

3.3 A Control Function Approach to Account for Travel Cost Endogeneity

Assume that μ_{ijt} and ζ_{ijt} are jointly normal with the same covariance matrix for all alternatives j. Correlation arises in this construction because unobserved attributes affect utility as well as travel costs, thereby entering both ζ_{ijt} and μ_{ijt}. Decompose ζ_{ijt} into its mean conditional on μ_{ijt} and deviations around this mean: $\zeta_{ijt} = \mathbb{E}[\zeta_{ijt}|\mu_{ijt}] + \tilde{\zeta}_{ijt}$ where $\tilde{\zeta}_{ijt}$ is normal with zero mean and constant variance. This error component can be expressed as: $\tilde{\zeta}_{ijt} = \sigma\eta_{ijt}$, where η_{ijt} is standard normal. The conditional expectation $\mathbb{E}[\zeta_{ijt}|\mu_{ijt}]$ is a function of μ_{ijt} and can be approximated using a control function:

$$\mathbb{E}[\zeta_{ijt}|\mu_{ijt}] = CF(\mu_{ijt}; \lambda) \tag{6}$$

where λ parameterizes the control function. The simplest assumption is that $CF(\mu_{ijt}; \lambda) = \lambda\mu_{ijt}$. Substituting the conditional mean and deviations into Equation 4 gives:

$$u_{ij} = v_{ijt}(\theta_i) + \lambda\mu_{ijt} + \sigma\eta_{ijt} + \tilde{\varepsilon}_{ijt} \tag{7}$$

The model described by Equation 7 is estimated in two steps. First, Equation 5 is estimated. This is a regression with the endogenous travel cost variable as the dependent variable and the exogenouss instruments, z_{ijt} as the explanatory variables. A simple linear functional form is assumed for $w(\cdot)$, such that the instruments enter Equation 5 additively and the parameters γ are estimated by ordinary least squares. The residuals for this regression provide estimates of μ_{ijt}: $\hat{\mu}_{ijt} = c_{ijt} - w(z_{ijt}; \hat{\gamma})$.

In the second step, we assume that $\tilde{\varepsilon}_{ijt}$ is distributed T1EV and the choice model described by Equation 7 is estimated with $\hat{\mu}_{ijt}$ entering the control function. The resulting second stage estimation routine is effectively a mixed logit over the new error components η_{ijt} as well as the random elements of θ_i and with $\hat{\mu}_{ijt}$ entering as an additional additively separable term in the utility function. Thus, the second stage of the estimation routine is

analogous to that described in Section 3.1.

Thus, with a set of parametric and distributional assumptions on the nature of the correlation between travel cost and the unobserved, individual- and site-specific factor, it is possible to account for the endogeneity problem and recover unbiased parameter estimates. While the linear assumption on the distribution of ζ_{ijt} conditional on μ_{ijt} may appear strong, it is possible in practice to allow for more flexible specifications of the control function in Equation 6. This approach of course relies on consistent estimates of $\hat{\mu}_{ijt}$ in the first step, which requires valid instruments, z_{ijt}. In particular, instruments must satisfy the following relatively standard assumptions:

1. Instrument relevance: $Cov(c_{ijt}, z_{ijt}) \neq 0$
2. Instrument exogeneity: $Cov(z_{ijt}, \varepsilon_{ijt}) = 0$

In Section 6, I provide examples of empirical instruments which plausibly satisfy these assumptions in a nationwide model of demand for campground reservations and which may generalize to other recreation demand contexts.

4 Monte Carlo Evidence

To demonstrate the effectiveness of the two-stage control function, I implement this approach on simulated data and compare the results to those I obtain without the correction. Given that the current best practice in the literature to account for unobservables in recreation demand models is to estimate site-specific constants, I also implement this approach and compare results to those obtained with the control function (Lupi et al., 2020).

I compare estimates obtained from these three separate approaches across two data generating processes. In the first data generating process, I assume that individual i's indirect utility from alternative j follows

$$u_{ij} = x_j + x_{ij} - 2c_{ij} - \xi_{ij} + \varepsilon_{ij}$$

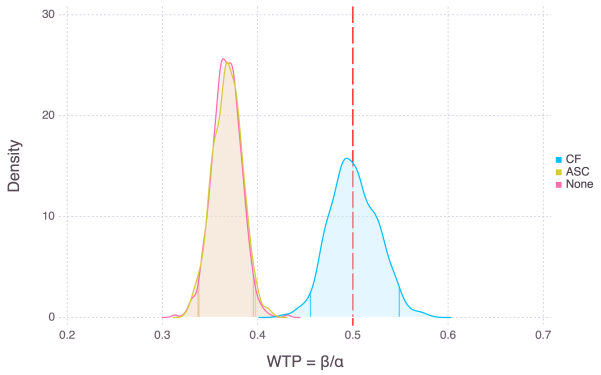
where

$$c_{ij} = 5 + 0.5z_{ij} + \underbrace{0.5\xi_{ij} + \tilde{\mu}_{ij}}_{=\mu_{ij}}$$

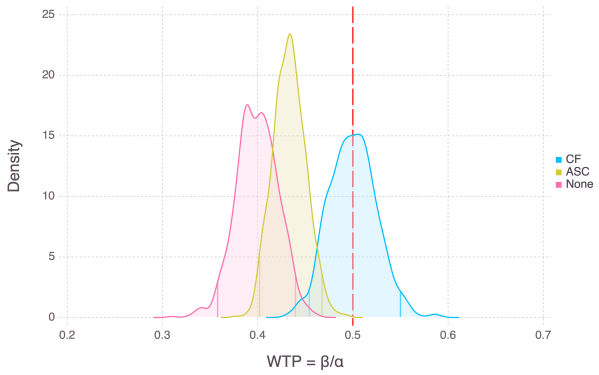
and $x_j, x_{ij}, z_{ij}, \xi_{ij} \sim \mathcal{U}[-3, 3]$, $\tilde{\mu}_{ij} \sim \mathcal{N}(0, 1)$, and ε_{ij} ∼T1EV. This model assumes that the unobservable term, ξ_{ij}, is fully idiosyncratic across individuals and sites. The travel cost endogeneity arises from the fact that c_{ij} is a linear function of ξ_{ij}. The target of estimation is the ratio of parameters on the exogenous, idiosyncratic observable, x_{ij}, and the

Figure 1. Estimates of WTP from Monte Carlo Simulations

(a) Estimates with Idiosyncratic Unobservable



(b) Estimates with Correlated Unobservable



Notes: Distributions of estimates of the willingness to pay (i.e., ratios parameters on the exogenous and endogenous, individual-alternative specific variables) from 1000 Monte Carlo simulations using the control function approach (blue), an alternative-specific constant approach to account for unobservables (yellow), and no endogeneity correction (red). The left panel (a) includes estimates from Monte Carlo simulations with fully idiosyncratic unobservables. The right panel (b) includes estimates from Monte Carlo simulations with unobservables that have an idiosyncratic component and an alternative-specific component. The vertical red lines show the true value of WTP.

endogenous, observable travel cost term, c_{ij}. By construction, the true value of this target object—effectively a measure of willingness to pay for x_{ij}—is 0.5.

I estimate this WTP across 1000 simulated samples using three estimators, a standard logit discrete choice model, a logit discrete choice model with alternative specific constants, and a logit discrete choice model with a control function. In each sample, I assume that there are 1000 individeuals and 10 alternatives. As shown in panel (a) of Figure 1, estimates of the control function approach across all 1000 samples vastly outperform those from the estimator with no endogeneity correction as well as those from the estimator with alternative specific constants. As shown in Table 1, I can reject the null hypothesis of equivalence between the WTP estimate and the true value based on a two-sided t-test in the case of the no endogeneity correction and alternative-specific constant estimators; however, I cannot reject equivalence in the case of the control function correction.

I test each of these three estimators on an additional data generating process that introduces correlation in unobservables across alternatives. In particular, rather than assuming that ξ_{ij} is drawn independently across individuals and sites, I assume that $\xi_{ij} = \tilde{\xi}_j + \tilde{\xi}_{ij}$, where each component is drawn from a uniform distribution over -3.0 to 3.0. As shown in panel (b) of Figure 1, introducing correlation in the unobservable term across alternatives improves

Table 1. Willingness-to-pay Estimates from Monte Carlo Simulations

Endogeneity Correction:	None	ASC	CF
Idiosyncratic ξ_{ij}			
Average	0.367	0.368	0.501
Bias	-0.133	-0.132	0.001
MSE	0.018	0.018	0.001
t-test	-278.047	-268.575	1.280
Idiosyncratic ξ_{ij} *w/ correl. over* j			
Average	0.399	0.433	0.500
Bias	-0.101	-0.067	-0.000
MSE	0.011	0.005	0.001
t-test	-146.338	-123.315	-0.243

Notes: Average willingness-to-pay estimates from Monte Carlo simulation. Table also reports statistical bias, mean squared error, and a t-test of equivalence to the true WTP estimate for each estimator.

the performance of the estimator with alternative-specific constants: statistical bias using this estimator is reduced by 40% relative to the case with a fully idiosyncratic unobservable. However, I can still reject the null hypothesis of equivalence between the alternative-specific constant estimates and the true WTP value as shown in Table 1. The control function approach continues to perform well when introducing correlation in unobservables across sites, with minimal bias.

5 Empirical Application: Demand for Federal Campsites

I explore the real-world performance of the two-stage control function approach to address endogeneity in recreation demand models. The setting that I use to do so is the market for campsite reservations on federally-managed land in the contiguous US I obtain data on site location, availability, and other physical attributes as well as the universe of reservations made at federally-managed campgrounds from the Recreation Information Database (RIDB), a central repository for information about federal recreational opportunities. The records on historical reservations available through RIDB are unique relative to other administrative data sources in not only their exhaustiveness, but also their inclusion of (limited) geographic data on the permanent residence of individuals.

I combine the reservation and campground data with a number of climatological variables and data on water quality from a nationwide network of monitors to allow for the modeling of campsite demand as a function of oberved site attributes, climate, and water quality. I select Water quality as the environmental quality of interest as it is well-studied in the recreation

Table 2. Summary Statistics for Estimation Sample

	Mean	St. Dev.
Travel Cost ($100s)	6.542	1.952
Dissolved Oxygen (mg/L)	7.472	1.568
Climate:		
Precipitation (deviation from mean)	−0.000	63.667
Dew Point (deviation from mean)	0.000	6.260
Temperature (deviation from mean)	0.000	3.958
Campsite Attributes:		
% Flush Toilet	0.290	0.391
% Shower	0.421	0.367
Campsite Activities:		
% Biking	0.564	0.496
% Climing	0.127	0.333
% Fishing	0.827	0.378
% Hiking	0.864	0.343
% Swimming	0.264	0.441

Notes: Summary statistics for variables currently used in estimating recreation demand.

demand literature, providing a helpful point of comparison for the proposed approach herein. Moreover, water pollution abatement is a major environmental policy area: since the passage of the 1972 Clean Water Act, public and private spending on water pollution abatement has totaled over $1 trillion, or $100 per person-year (Keiser and Shapiro, 2019).

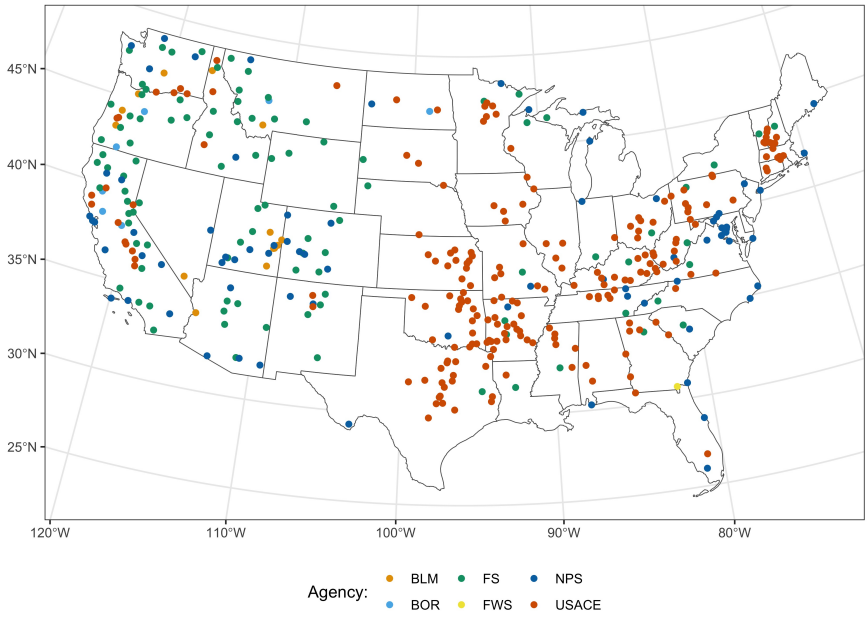
5.1 Campground Attributes and Reservations

I collect information on campground attributes from RIDB for over 3000 campgrounds on federally-managed recreation areas in the contiguous US. Campground data include the number of campsites at each campground; the set of recreation activities (e.g., biking, hiking, or climbing) listed as associated with the area (e.g., National Park or National Forest) in which the campground is located; and the type of facility amenities (e.g., the availability of showers or flush toilets) at each campground.

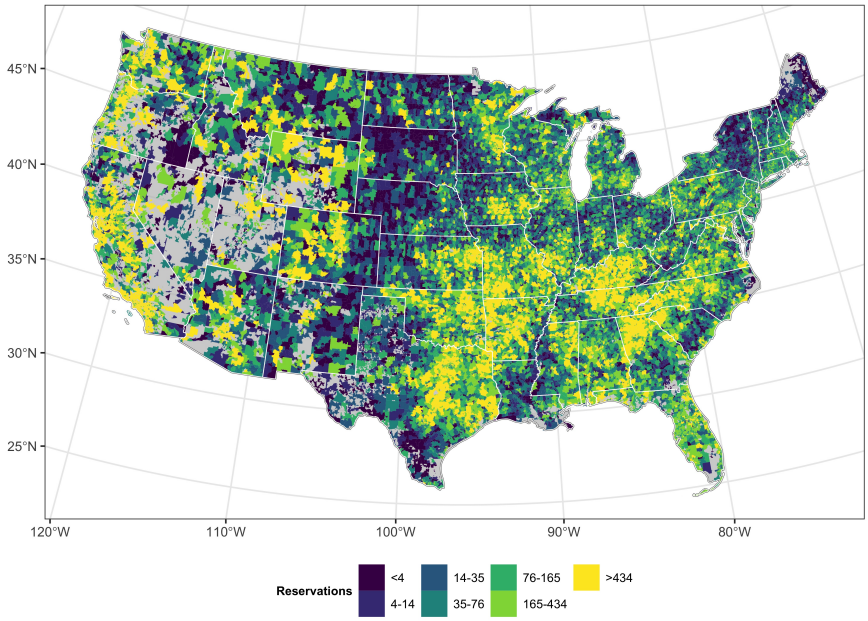
Several approaches are available to ease the computational burden of estimating a recreation demand model with a discrete choice set exceeding 3000 alternatives. The first possible approach is to estimate a discrete choice model on a subset of, typically randomly selected, alternatives (Train, 2009). Under certain modeling assumptions, estimation on a subset of alteratives can be accomplished without inducing inconsistency. While this approach is appealing in that it does not require additional assumptions, it does add computational complexity over the alternative solution, which is particularly appealing in this setting. The second available approach is geographic aggregation. In many discrete choice settings, the de-

Figure 2. Maps of Federal Campgrounds and Reservation Origins

(a) Federal recreation areas with reservable campgrounds



(b) Cumulative reservations by ZIP code of origin



Notes: These figures depict maps of (a) federal recreation areas with reservable campgrounds (b) associated reservations by ZIP code of origin (2016-2018). Data are from the Recreation Information Database (RIDB), a centralized repository for data on the universe of federal campsites and associated campsite reservations in the US

cision of how best to aggregate individual alternatives based on geography can prove difficult, with results potentially non-robust to the aggregation decision. In the case of campground demand on federally-managed land, this is arguably not the case: campgrounds are explicitly linked to larger recreation areas, making this outer nest a suitable candidate for the aggregation of alternative campgrounds. This aggregation is also intuitive: it is plausible that individuals first choose to visit Grand Canyon National Park and then, having decided to visit this area, select a campsite among the available alternatives. Campers primarily experience environmental amenities when visiting a campground through the broader outdoor recreation area in which the site is located.

For these reasons, I aggregate the individual camper's decision problem to the level of the recreation area, resulting in 408 total alternatives within the contiguous US in the final choice set. Figure 2 maps the alternatives in the final choice set. I aggregate data on campground facility amenities to the recreation area-level, with attributes weighted by the number of campsites within each campground. Thus, the resulting recreation area-level attribute describing, for example, the availability of flush toilets can be interpreted as the share of campsites within the recreation area with access to this amenity. Similarly, recreation area-level activity variables can be interpreted as the share of campsites within a recreation area listing a given activity.

I also acquire data on the universe of historical reservations made at federal campgrounds for the years 2016-2018 from RIDB. These data include the campsite reserved; the start and end dates of the campsite reservation; the date on which the reservation is made; the home billing ZIP code of the reserving individual; the number of individuals in the party; and the price of the reservation. I drop reservations made by individuals from home locations outside the contiguous US, including international visits and visits from Hawaii, Alaska, and US territories. There are over 2 million unique campsite reservations included in each year in these data.[1] I use the reserved campsite to identify the campground at which the individual makes a reservation, which in turn defines which of the 408 recreation areas in the final choice set the individual visits. The innovation of these publicly-available administrative data is their inclusion of geographic information on the origin of campers in the form of a billing ZIP code. While this is a relatively coarse geographic unit, it still allows for the application of the travel cost logic in this setting. Figure 2 maps cumulative campsite reservations by

[1]The data also include unique reservation identifiers, which help to link reservations made by the same individual in a single transaction. This allows for the identification of multi-site trips made by a single individual, which would otherwise be treated as separate trips by different individuals. This does present a challenge when individuals make campsite reservations to different recreation areas. The current analysis randomly selects one of the recreation areas visited by an individual; however, future analysis may attempt to more robustly model the intensive margin of trip duration in a way that accounts for trips that combine visits to different recreation areas.

ZIP code of origin for the years 2016-2018.

5.2 Expected Travel Cost

The cost of traveling to a recreation site is an integral component to the travel cost model. For each point of origin in the data, it is necessary to calculate the cost of traveling to all sites in the choice set, regardless of whether any particular individual visits those sites. Given the nationwide scale of the current empirical setting, an additional complicating factor is the likelihood that individuals may select different modes of transportation depending on the recreation site alternative in question. Since data on travel mode are unavailable, it is not possible to directly model the choice of whether to fly or drive—the two most likely modes or transport—to each recreation site. Following the approach of English et al. (2018), I calculate expected travel cost from each point of origin to each destination in the choice set as a weighted average of driving and flying costs, where the weights are based on the observed shares of flying versus driving in a nationwide telephone survey conducted by English et al. (2018), who estimate a nationwide model of demand for visits to Gulf Coast beaches.

Let c_{ij} represent the round-trip travel cost to individual i of traveling to recreation area j.[2] Let ρ_{ij} represent the probability of flying to site j. Expected travel cost is therefore calculated as:

$$\mathbb{E}[c_{ij}] = (1 - \rho_{ij})c_{ij}^{d} + \rho_{ij}c_{ij}^{f} \tag{8}$$

where c_{ij}^{d} and c_{ij}^{f} are the round-trip costs of driving and flying from individual i's home ZIP code to recreation area destination j, respectively. Round-trip driving costs are calculated as a function of round-trip driving distance, d_{ij}; round-trip driving time, t_{ij}; the number of trip nights, l_i; and party size, n_i:

$$c_{ij}^{d} = \frac{1}{n}\left(p^{d}d_{ij} + p_{j}^{l}l_{i}\right) + p_{i}^{t}t_{ij} \tag{9}$$

where p^{d} is the per-mile cost of driving, p_{j}^{l}, is the average per night cost of a site reservation made at a given recreation area, and p_{i}^{t} is the per-hour opportunity cost of time.[3] The resulting c_{ij}^{d} is therefore the round-trip, per-person cost of driving from individual i's origin ZIP code to the destination j, inclusive of driving costs, lodging costs, and the opportunity

[2]Since individuals' home geography is defined at the ZIP code level, c_{ij} is in practice estimated for each unique ZIP code origin-recreation area destination combination in the data, resulting in over 14 million total route cost calculations.

[3]Note that I calculate each individual i's travel costs using the relevant prices and costs in the period in which they make their visit; however, for notational simplicity, the time subscripts are excluded in this subsection.

Table 3. Flying Choice Probabilities from English et al. (2018)

		Pr(flying|dist, HH income, HH size)			
	HH Income:	≤$70k	>$70k	≤$70k	>$70k
Driving dist. (mi.)	HH Size:	≤2	≤2	>2	>2
$[0, 250]$		0.000	0.000	0.000	0.000
$(250, 500]$		0.000	0.030	0.000	0.000
$(500, 1000]$		0.168	0.338	0.056	0.201
$(1000, 1500]$		0.736	0.788	0.443	0.784
$(1500, \infty]$		0.842	0.880	0.842	0.880

Notes: Flying choice probabilities reported as a function of one-way driving distance (miles), household income, and household size (taken from English et al. (2018)).

cost of travel time.

I calculate round-trip driving distances and times using the Open Source Routing Machine (OSRM) and OpenStreetMap road network data. OSRM offers a high performance routing engine to find the shortest route between two destinations using the OpenStreetMap road network. The per mile driving cost, p^d, is based on analysis by the American Automobile Association (AAA) and includes per-mile costs of gasoline, maintenance, and depreciation for an average passenger vehicle in the US[4] The number of trip nights for each individual i, l_i, is taken directly from the reservations data, which provides the start and end dates for each reservation.[5] I calculate reservation costs per-night directly from the reservation data and directly observe party size in the reservation data. The final component of driving cost, $p_i^t t_{ij}$, represents the opportunity cost of time, where p_i is taken to be one-third the median household income in individual i's ZIP code of residence divided by 2080 hours-worked per year as is standard practice in the recreation demand literature (Lupi et al., 2020).

I calculate round-trip flying costs for each origin-destination pair, c_{ij}^f, by first identifying the three nearest airports to each origin and destination.[6] I then calculate total flight costs

[4]AAA publishes per-mile cost estimates inclusive of fuel and maintenance (inclusive of tire costs) as well as depreciation costs as a function of miles traveled for the average passenger vehicle on an annual basis. Per-mile depreciation costs are calculated as the difference in average depreciation between driving 5000 miles more and 5000 miles less than AAA's baseline scenario of 15,000 annual vehicle miles traveled for the average passenger vehicle. English et al. (2018) employ a similar approach. The resulting average driving cost per-mile ranges for the years 2008 to 2018 ranges from a low of around $0.19 in 2009 to a high of around $0.27 in 2018.

[5]This assumes that individuals only stay at federal campgrounds during their trip, a plausibly tenuous assumption. Future work will seek to estimate trip duration as a function of trip distance. Moreover, this assumes that individuals' trip duration would be constant across alternatives, with counterfactual trips of equal length to the trip actually taken. Again, modeling trip duration as a function of distance traveled may be a more reasonable approach to estimating this cost in the future.

[6]I only consider airports with a minimum of 100,000 enplanements per year to remove small, regional airports and airfields which mostly service general aviation flights and are unlikely to be used by most

for each of the 9 potential origin-destination airport combinations and select the least cost route as the flight cost for that ZIP code-recreation area pair. Flying costs include: (1) round-trip airfare from the origin to the destination airport; (2) round-trip travel time as in Equation 9 above; and (3) lodging costs as in Equation 9 above. I calculate airfare for each route for each quarter-year in the 2016-2018 sample using the 30th percentile fare for round-trip tickets from the Airline Origin and Destination Survey (DB1B), which is a 10% sample of all airline tickets collected by the Bureau of Transportation Statistics.[7] I assign a baggage fee of $50 to all fares other than those on JetBlue and Southwest Airlines. I use data on the miles flown and number of segments for all ticketed itineraries on each origin-destination route to calculate flying times, assuming an average cruising speed of 550 nautical miles per hour, an average layover time of 2 hours, and that travelers arrive at the departing airport 2 hours before their flight. I estimate the hourly opportunity cost of time as in Equation 9.

The final component necessary to calculate expected travel costs given by Equation 8 is the probability of flying, ρ_{ij}. I model flying probabilities as a function of one-way driving distance, household income, and household size: $\rho_{ij} = f(d_{ij}, w_i, s_i)$. I estimate this probability using data collected by English et al. (2018), who implement a nationwide representative telephone survey of visitors to Gulf Coast beaches. English et al. (2018) average travel mode choices for discrete one-way travel distance intervals over a set of household types based on an income and size threshold and report the corresponding flying probabilities. Table 3 replicates English et al. (2018)'s Table 2.[8] I generate the probability of an individual i falling into each of the four household types reported in Table 3 for each year in the sample using 5-Year American Community Survey data on household income and household size at the ZIP code level. I then generate the resulting flight probability for each ZIP code-recreation site pair as the probability weighted average of the flying probabilities for the relevant one-way driving distance interval reported in Table 3.

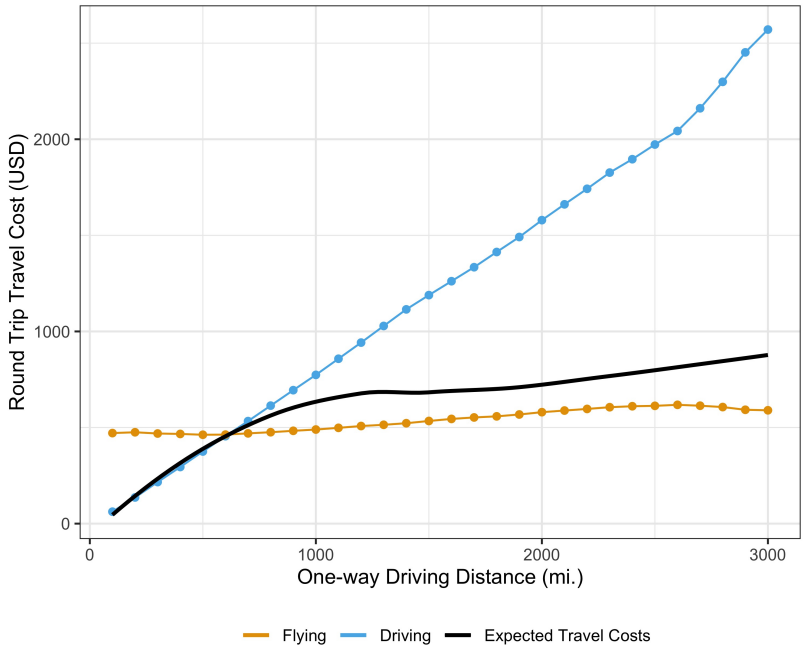
Figure 3 shows the resulting flying, driving, and expected travel costs as a function of one-way driving distance. The figure shows that driving costs increase monotonically with one-way driving distance, starting from around $50 for trips under 100 miles and reaching over $2500 for trips exceeding 3000 miles. Flying costs are initially higher than driving costs, but are on average cheaper than driving costs for one-way driving distances over 700 miles, increasing relatively modestly with distance. Expected travel costs initially follow driving

vacation travelers.

[7]Following English et al. (2018), since 40% of fares on most routes are for business and are typically higher fares, I assume that the 30th percentile fare from the DB1B data represents the median non-business fare.

[8]Intuitively, English et al. (2018) estimate that flying probabilities are increasing in one-way driving distance and household income and decreasing in household size.

Figure 3. Expected Travel Costs by One-way Itinerary Driving Distance



Notes: Average flying, driving, and expected travel cost by one-way driving distance for the years 2016-2018. Average total travel costs are constructed across 100-mile distance bins across all origin-destination pairs in the data.

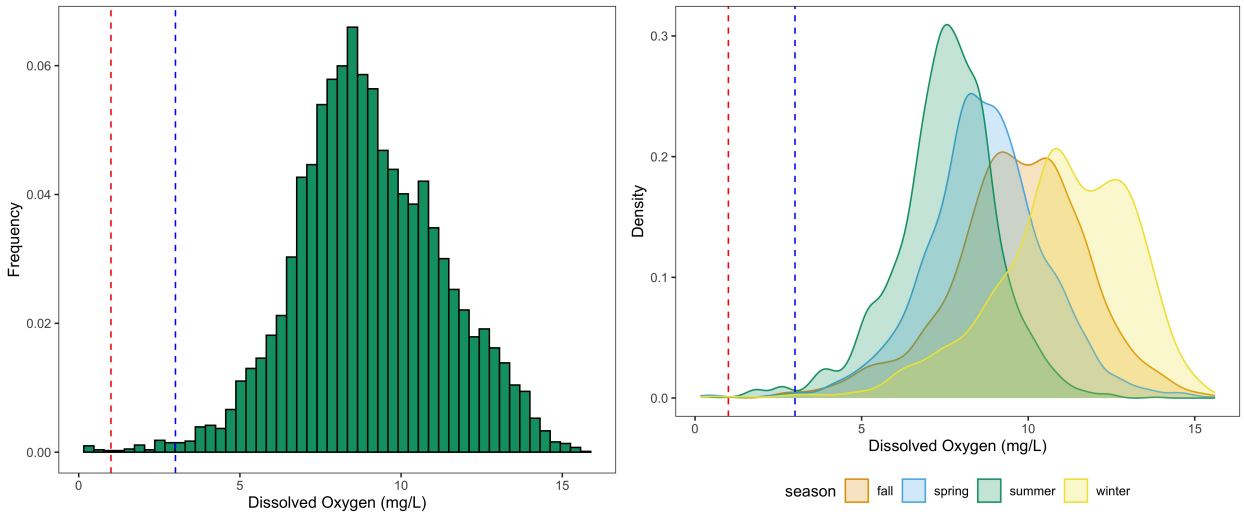
costs quite closely, but are closer to flying costs for distances over 700 miles, which makes sense given the probabilities reported in Table 3.

5.3 Climate Attributes

I obtain daily average temperature, precipitation, and dew point time series from the PRISM Climate Group in 4-km resolution grids for the contiguous US over the 2016-2018 period. These data are constructed from monitoring station data by the PRISM Climate Group using a climatology-aided interpolation process, which allows for the observation of weather variables across a broader set of geographies than those in close proximity to monitoring stations. I obtain similar gridded data on 30-year climate normals—long-term averages—for the period of 1981-2010 from the PRISM Climate Group for each of these three weather variables. I use the climate normals to calculate daily deviations from the 1981-2010 climate normals for the period 2016-2018. I then average these deviations to the month-by-year level and assign them to recreation areas in the choice set by taking the month-year deviation for each climatological variable at the centroid of each recreation area.

Figure 4. Empirical Distribution of Water Quality Measure

(a) Full distribution of DO values **(b)** Seasonal distributions of DO values



Notes: Distribution of DO values at water quality monitors linked to recreation areas in the choice set over the period 2016-2018. (a) shows the full distribution, while (b) shows the distribution of measurements by season. Higher values are associated with better water quality. The dotted blue and red lines show the values below which waterbodies are said to be hypoxic and anoxic, respectively.

5.4 Water Quality

I take water quality data from the Water Quality Portal (WQP), which offers the most comprehensive spatial and geographic coverage of water quality data in the US, combining data collected by over 400 state, federal, tribal and local agencies. There are several broad indicators which feature in the literature conducting recreation demand and hedonic analysis of water quality: dissolved oxygen (DO), fecal coliform, total suspended solids, pH, Secchi depth, and harmful algal concentrations, to name a few. One of the more common measures is DO (Keiser and Shapiro, 2019). Given its prevalence in the literature and the fact that it is a primary indicator of nutrient pollution, one of the more common forms of water pollution, I uses DO as the main measure of water quality in my analysis.

DO is a simple measure of the amount of oxygen that is present in water. All aquatic animals need DO to breathe, so DO is considered a direct indicator of an aquatic resource's abilty to support aquatic life. Low levels of oxygen (hypoxia) or no measurable oxygen (anoxia) can occur when excess organic materials, such as large algal blooms, are decomposed by microorganisms, which consume large quantities of DO in the decomposition process. The underlying cause of hypoxia in waterways is the discharge of large quantities of nutrients often found in fertilizers, animal waste, and other sources prone to runoff, which promote

the growth excess organic material (e.g., algae) and hence contribute to oxygen depletion. In addition to being a good measure of the health of aquatic life, DO is a relatively salient water quality indicator as hypoxia and anoxia can result in noticeable impairment to waterways, such as algal blooms. Indeed, in addition to being unpleasant, certain forms of algal blooms which might be associated with low DO levels can be noxious to humans.

To assign dissolved oxygen measurements to recreation areas, I identify water quality monitors that fall within 5 km. of the boundary for each recreation area. I calculate weighted monthly averages of DO measurements for each recreation area, with each DO measurement weighted by its inverse distance to the centroid of the recreation area. Figure 4 shows the total and seasonal resulting distribution of monthly average DO measurements. As Figure 4 demonstrates, the concentration of DO is inversely related to water temperature, with the lowest levels observed in the summer and spring, and the highest levels observed in the winter and fall.

5.5 Defining the Information Set

Another novelty of the RIDB reservation data relative to the standard survey data used in recreation demand modeling (most of which is collected onsite) is the inclusion of the date on which reservations are made. This is unique in that—due to data limitations—standard applications of the recreation demand model assume that individuals decision problem and consumption of the recreation experience happen concurrently, which may be true in certain instances, but is certainly false in others (particularly when considering visits to large, high-value natural areas such as National Parks). The inclusion of the date of booking allows for the accurate definition of individuals' information sets at the point at which decisions about recreation site visits are being made.

In the results reported in Section 6, I assume that time-varying site attributes—the climate and water quality variables—enter individuals' decision problem based on values in the month of the site visit in the previous year. For example, I assume that an individual making a reservation in May to visit a campsite at a given recreation area in July will consider the water quality conditions at the same time (July) of the previous year. Future work will test alternative ways of incorporating time-varying site attributes into the agent's information set at the time of making a reservation, including observed time-varying attributes at the time of visiting the site (the standard assumption in the literature) and at the time of making the reservation.

Table 4. First Stage Results

	Travel Cost
Crude Price × Driving Distance	0.00003***
	(0.00003, 0.00003)
Average Commute Time	0.012***
	(0.012, 0.012)
Exogenous Variables	✓
Observations	9,660,970
R^2	0.522

Notes: Results from first stage control function regression. Note: p values and 95% confidence intervals are calculated using robust standard errors; *$p < 0.1$, **$p < 0.5$, ***$p < 0.01$. Exogenous variables include non-travel cost variables listed in Table 2.

6 Empirical Application: Results

Implementing the control function approach outlined in Section 3.3 in this empirical setting requires the construction of valid instruments, z_{ijt}. Instruments should have the desirable property of being correlated with travel cost, but exogenous to the unobserved characteristic, ζ_{ijt}. The first instrument that I use in this setting is the price of a barrel of crude oil interacted with one-way driving distance to each site. This is clearly correlated with travel cost, but is likely exogenous to unobserved characteristics. While this assumption of instrument exogeneity is not testable, it is unlikely that crude prices interacted with driving distance is correlated with the unobserved characteristics as those factors which determine world oil prices are plausibly different from those which would enter both individuals' residence and recreation decisions. The second instrument that I use in this setting is average commuting time at the ZIP code level.

I use residuals from the control function regression reported in Table 4 in the second stage estimation of the choice model described by Equation 7. As a point of comparison, I also estimate the choice model described by Equation 1, i.e., the analogous model without a control function correction. This represents the relevant point of comparison for the control function approach as this form of choice model is akin to the status quo in most recreation demand contexts which does not explicitly address the form of travel cost endogeneity under study. I model random parameters (i.e., individual-level heterogeneity) for the travel cost and water quality variables. In addition to estimating the choice models given by Equations 1 and 7, I estimate versions of these choice models with no individual-level parameters in order to demonstrate the importance of accounting for individual-level preference heterogeneity on the parameters of interest.

For the sake of simplicity and exposition and to ease computational burdens, I subset the main data to include all reservations made in July 2018. The July 2018 estimation sample includes a total of 87,906 distinct visits to 110 different recreation areas.

Table 5 reports parameter estimates from the estimation of four distinct choice models. Columns (1) and (2) report parameter estimates for two choice models which do not account for the endogeneity of travel cost, with the choice model estimated in column (1) not allowing for individual-level heterogeneity in the parameters on travel cost and DO and the choice model estimated in column (2) allowing individual-level heterogeneity on these parameters. Thus, column (1) reports estimates of population mean parameters whereas column (2) reports estimated means and standard deviations for the distributions of these parameters, which I assume to be normally distributed in the population. I estimate the random parameters logit reported in column (2) using 100 draws from the standard normal distribution. Examining the first two columns of Table 5, the estimated parameters have the desired signs: the estimated parameter(s) on travel cost is negative, implying that all else equal, individuals prefer sites which are cheaper to visit, and the estimated parameter(s) on DO is positive, implying that all else equal, individuals prefer sites with higher water quality values. Comparing parameter estimates across columns (1) and (2), accounting for preference heterogeneity also appears important: the standard deviation term for travel costs is large and statistically significant, implying substantial variation in price sensitivity within the population is necessary to justify the observed site visitation decisions. Interestingly, the standard deviation term on the DO variable is statistically-indistinguishable from zero (and in fact is negative).[9]

Comparing parameter estimates between the two choice models which account for travel cost endogeneity, reported in columns (3) and (4) of Table 5, and those which do not account for this endogeneity shows the importance of the control function approach. The t-statistic on the λ parameter in columns (3) and (4) provides a simple test of travel cost endogeneity. Moreover, I estimate smaller travel cost parameters (i.e., more negative, larger in magnitude) when accounting for travel cost endogeneity than those reported in columns (1) and (2), suggesting that the endogeneity biases these parameter estimates towards zero when left unaccounted for.

To demonstrate the importance of accounting for the non-uniform distribution of residences and recreation sites to the welfare analysis of non-market environmental amenities,

[9]This could be driven by the relatively few number of simulation draws used or the methods used to construct the DO measure. The negative standard deviation estimate is interesting as these standard deviations should be strictly non-negative. Future results will use a bounded optimization routine to ensure non-negative standard deviation estimates, which may result in qualitatively different estimates of this standard deviation.

Table 5. Discrete Choice Model Estimates

	No Endogeneity Correction		Control Function	
	(1) Cond. Logit	(2) R.P. Logit	(3) Cond. Logit	(4) R.P. Logit
Travel Cost				
α_c	-0.900^{***}	-1.342^{***}	-0.950^{***}	-1.724^{***}
	$(-0.904, -0.895)$	$(-1.354, -1.330)$	$(-0.961, -0.938)$	$(-1.741, -1.706)$
σ_c		0.866^{***}		0.972^{***}
		$(0.853, 0.879)$		$(0.959, 0.986)$
DO				
β_d	0.096^{***}	0.107^{***}	0.096^{***}	0.109^{***}
	$(0.089, 0.102)$	$(0.098, 0.115)$	$(0.089, 0.102)$	$(0.100, 0.118)$
σ_d		-0.009		0.004
		$(-0.033, 0.014)$		$(-0.015, 0.023)$
λ			0.079^{***}	0.500^{***}
			$(0.062, 0.096)$	$(0.483, 0.517)$
Controls	✓	✓	✓	✓
Mean $\widehat{MWTP}$	10.616	10.301	10.097	6.071
Visitors	87,906	87,906	87,827	87,827
Sites	110	110	110	110
Log Likelihood	$-239{,}961.8$	$-233{,}040.3$	$-239{,}654.8$	$-230{,}925.6$

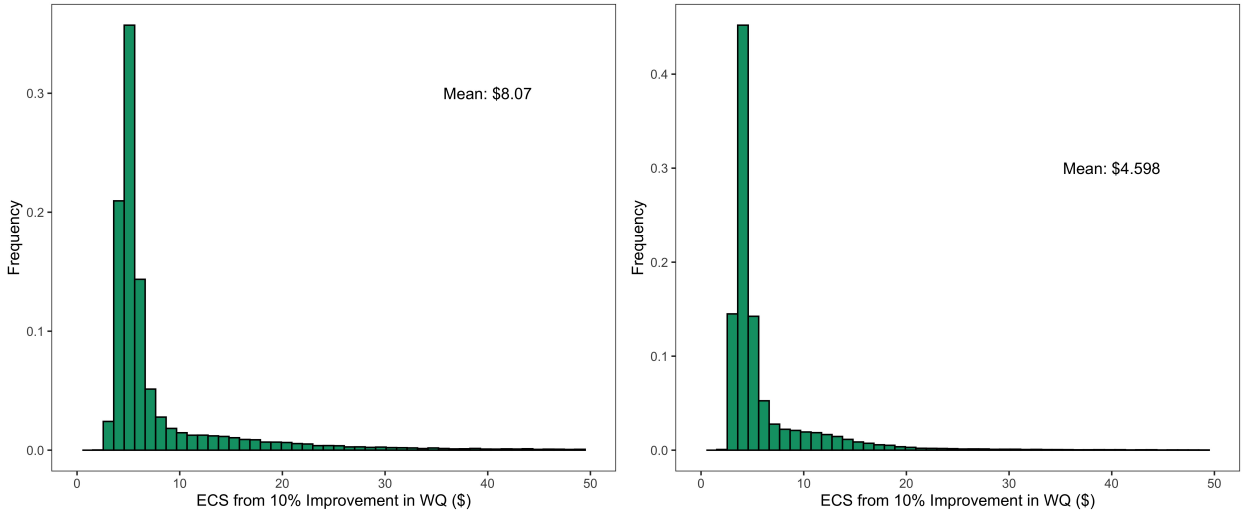
Notes: Parameter estimates for choice models with individual-level parameters on travel cost and dissolved oxygen (columns 2 and 4) are estimated using 100 simulation draws. Travel cost is in terms of 100s of dollars and dissolved oxygen is reported in mg/L. Mean marginal willingness to pay for dissolved oxygen is calculated as a function of the estimated travel cost and dissolved oxygen (DO) parameters and is reported in terms of dollars per mg/L. λ is the linear control function parameter on the resideuals from the regression reported in Table 4. Note: p values and 95% confidence intervals are calculated using bootstrapped standard errors; $^{*}p < 0.1$, $^{**}p < 0.5$, $^{***}p < 0.01$. Control variables include remaining variables listed in Table 2. Results use campground reservation data for the period of July 2018. A total of 110 unique recreation areas (of a total of 408 in the choice set) are visited during this period.

I calculate the marginal willingness to pay (MWTP) for changes in DO as well as partial equilibrium welfare estimates under a simple counterfactual for each choice model estimated in Table 5. To calculate a monetary estimate of the MWTP for DO, I divide the parameter estimate for DO by the parameter estimate on the travel cost variable, which I can interpret as the marginal utility of income. In the case of the two conditional logit choice models which do not allow for preference heterogeneity, I calculate average MWTP as $\mathbb{E}[\widehat{MWTP}] = \hat{\beta}_d/\hat{\alpha}_c$. In the case of the random parameters logit models, I calculate individual-level MWTP for DO as $\widehat{MWTP}_i = \hat{\beta}_i^d/\hat{\alpha}_i^c$. Comparing MWTP estimates reported in Table 5, the upward bias in the travel cost parameter estimates when not accounting for the travel cost endogeneity results in upward biased MWTP estimates.

To understand the impact of accounting for travel cost endogeneity on partial equilibrium welfare estimates, I consider a simple policy counterfactual. In particular, I calculate the

Figure 5. Change in Expected Compensating Surplus

(a) ECS without control function **(b)** ECS with control function



Notes: Distribution of the change in expected compensating surplus from a 10% increase in DO levels at all recreation areas (a) not accounting for the endogeneity of travel cost and (b) accounting for the endogeneity of travel cost using a control function approach. (a) uses parameter estimates reported in column (2) of Table 5 and (b) uses parameter estimates reported in column (4) of Table 5.

change in consumer surplus resulting from a 10% increase in DO levels across all recreation sites in the data. I focus on comparing resulting welfare estimates based on the two choice models which allow for preference heterogeneity, reported in columns (2) and (4) of Table 5. Given the specification of utility in these choice models, expected utility per trip for individual i on choice occasion t is given by:

$$\mathbb{E}[\hat{v}_{it}] = \log\left(\sum_j \exp(\hat{v}_{ijt})\right) + C \tag{10}$$

where C is an unknown constant which indicates that the absolute level of utility is not identified in random utility models, only utility differences across alternatives in the choice set (Train, 2009). To get a monetary measure of compensating surplus per trip, I can divide Equation 10 by the individual-level travel cost parameter, $\hat{\alpha}_i^c$. Consider a change in DO from DO^0 to DO^1. The corresponding absolute change in money-metric utility is identified and, using Equation 10, can be written as:

$$\mathbb{E}\left[\Delta\widehat{CS}_{it}\right] = \frac{1}{\hat{\alpha}_i^c}\left[\log\left(\sum_j \exp(\hat{v}_{ijt}^1)\right) - \log\left(\sum_j \exp(\hat{v}_{ijt}^0)\right)\right] \tag{11}$$

Figure 5 plots distributions of $\mathbb{E}\left[\Delta\widehat{CS}_{it}\right]$ for an across-the-board increase in DO of 10% (i.e., a 10% improvement in water quality) using choice model parameter estimates reported in columns (2) and (4) of Table 5. Overall, both distributions are quite concentrated, with very moderate heterogeneity in the form of a long right tail. The mean and median of the distribution of welfare estimates which account for travel cost endogeneity are lower than those which do not account for this endogeneity. This is expected given the change in travel cost parameter estimates across these models.

While the differences in travel cost parameter estimates and the resulting welfare estimates across models which do and do not account for travel cost endogeneity indicate that failure to do so can result in meaningfully different results, little has been said about the direction of the bias uncovered in these results. The focus thus far in describing the particular sources of travel cost endogeneity of concern has been on individuals who highly value access to high quality recreational opportunities or high value environmental amenities might choose to live near high quality recreation sites, thereby selecting into lower travel costs. If this were the only sorting pattern of concern, then the travel cost parameter when not accounting for the endogeneity of travel cost would be downward biased (i.e., more negative) as the individuals who choose to live near recreation sites would appear more responsive to price than they actually are. Indeed, these are individuals who are willing to move to be closer to recreation sites. However, there are other factors leading to the non-uniform, non-random distribution of residences and recreation sites which are of concern. In particular, there is a concern that high value, more pristine recreation sites may be more prone to be located away from high density population centers. With both sorting patterns possible at once, the net sign of the bias when not accounting for travel cost endogeneity due to spatial sorting is ambiguous. Thus, it is not surprising that the results reported in Table 5 find that travel cost parameters in this setting are upward biased (i.e., less negative) when not accounting for this endogeneity concern.

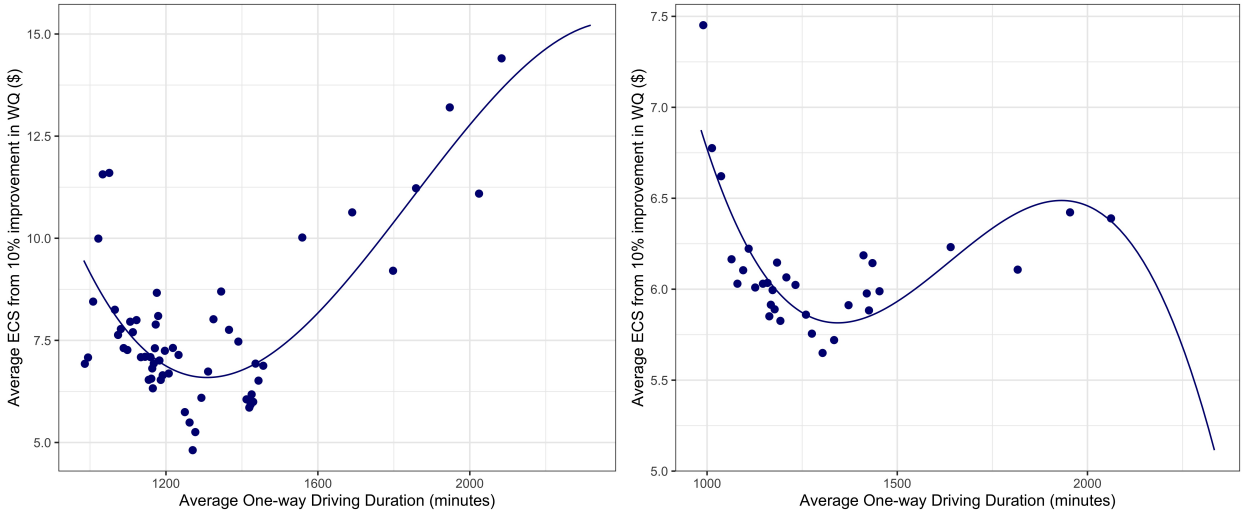
To further explore the spatial distribution of preferences for water quality, I calculate average change in welfare for the 10% overall improvement in DO counterfactual for all trips originating from each ZIP code in July 2018 as

$$\mathbb{E}\left[\Delta\widehat{CS}_{z}\right] = \frac{1}{N_z}\sum_{i=1}^{N_z}\mathbb{E}\left[\Delta\widehat{CS}_{it}\right] \tag{12}$$

where z indexes ZIP codes and N_z is the number of individual trips originating from ZIP code z in the period of the data used in estimation (i.e., July 2018). I then construct a ZIP code-level measure of proximity to recreation areas in the choice set as the average one-way driving

Figure 6. Welfare Changes and One-way Driving Distance

(a) Welfare change without control function **(b)** Welfare change with control function



Notes: Binned scatterplots of the average welfare change under the 10% DO improvement counterfactual (calculated using Equation 12) as a function of average one-way driving duration (a) not accounting for the endogeneity of travel cost and (b) accounting for the endogeneity of travel cost using a control function approach. (a) uses parameter estimates reported in column (2) of Table 5 and (b) uses parameter estimates reported in column (4) of Table 5.

duration from each ZIP code to each recreation area. Figure 6 shows binned scatterplots of the average welfare change under the 10% DO improvement counterfactual (calculated using Equation 12) as a function of average one-way driving duration. Figure 6 reveals that not accounting for travel cost endogeneity, welfare changes under the counterfactual are high for counties at the lower and higher ends of the average duration spectrum, which is consistent with both forms of non-uniform spatial sorting described above.

7 Conclusion

Recreation demand models inform decision-making across a wide range of applications, from regulatory impact analysis and resource management to public health and environmental litigation. Careful estimation of model parameters in these settings is critical to ensure unbiased inferences when making policy, regulatory, and legal decisions. I show that a common assumption in the literature—namely, that individuals' proximity to recreation sites is exogenous to their recreation decisions—can produce biased results and demonstrate a simple, feasible approach to relaxing this assumption in empirical applications. In numerical simulations, I find that accounting for travel cost endogeneity using a two-stage control

function approach yields greater performance than existing approaches to account for unobservables in recreation demand models, substantially reducing statistical bias. This can lead to dramatically different inferences in real world settings: estimating a nationwide model of demand for campsite reservations as a function of water quality, I find that not correcting for endogeneity via a control function nearly doubles estimates of consumers' willingness-to-pay for improvements in water quality. While these findings might be concerning to policymakers and practitioners who rely on the conclusions from recreation demand models, the relatively simple fix for which I advocate in this paper should restore faith in this crucial methodology moving forward.

References

Abidoye, Babatunde O., and Joseph A. Herriges. 2012. "Model Uncertainty in Characterizing Recreation Demand." *Environmental and Resource Economics*, 53(2): 251–277.

Abidoye, Babatunde O., Joseph A. Herriges, and Justin L. Tobias. 2012. "Controlling for Observed and Unobserved Site Characteristics in RUM Models of Recreation Demand." *American Journal of Agricultural Economics*, 94(5): 1070–1093. Publisher: [Agricultural & Applied Economics Association, Oxford University Press].

Albouy, David, Walter Graf, Ryan Kellogg, and Hendrik Wolff. 2016. "Climate Amenities, Climate Change, and American Quality of Life." *Journal of the Association of Environmental and Resource Economists*, 3(1): 205–246. Publisher: The University of Chicago Press.

Bajari, Patrick, Jane Cooley Fruehwirth, Kyoo il Kim, and Christopher Timmins. 2012. "A Rational Expectations Approach to Hedonic Price Regressions with Time-Varying Unobserved Product Attributes: The Price of Pollution." *The American Economic Review*, 102(5): 1898–1926.

Bayer, Patrick, and Christopher Timmins. 2007. "Estimating Equilibrium Models Of Sorting Across Locations*." *The Economic Journal*, 117(518): 353–374. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1468-0297.2007.02021.x.

Bayer, Patrick, Nathaniel Keohane, and Christopher Timmins. 2009. "Migration and hedonic valuation: The case of air quality." *Journal of Environmental Economics and Management*, 58(1): 1–14.

Bento, Antonio, Matthew Freedman, and Corey Lang. 2014. "Who Benefits from Environmental Regulation? Evidence from the Clean Air Act Amendments." *The Review of Economics and Statistics*, 97(3): 610–622. Publisher: MIT Press.

Bishop, Kelly C., Nicolai V. Kuminoff, H. Spencer Banzhaf, Kevin J. Boyle, Kathrine von Gravenitz, Jaren C. Pope, V. Kerry Smith, and Christopher D. Timmins. 2020. "Best Practices for Using Hedonic Property Value Models to Measure

Willingness to Pay for Environmental Quality." *Review of Environmental Economics and Policy*, 14(2): 260–281. Publisher: Oxford Academic.

Bureau of Economic Statistics. 2019. "Outdoor Recreation Satellite Account, U.S. and States, 2019." Bureau of Economic Statistics.

Egan, Kevin J., Joseph A. Herriges, Catherine L. Kling, and John A. Downing. 2009. "Valuing Water Quality as a Function of Water Quality Measures." *American Journal of Agricultural Economics*, 91(1): 106–123. Publisher: [Agricultural & Applied Economics Association, Oxford University Press].

English, Eric, Roger H. von Haefen, Joseph Herriges, Christopher Leggett, Frank Lupi, Kenneth McConnell, Michael Welsh, Adam Domanski, and Norman Meade. 2018. "Estimating the value of lost recreation days from the Deepwater Horizon oil spill." *Journal of Environmental Economics and Management*, 91: 26–45.

Glaeser, Edward L., ed. 2010. *Agglomeration Economics. National Bureau of Economic Research Conference Report*, Chicago, IL:University of Chicago Press.

Greenstone, Michael, and Justin Gallagher. 2008. "Does Hazardous Waste Matter? Evidence from the Housing Market and the Superfund Program." *The Quarterly Journal of Economics*, 123(3): 951–1003. Publisher: Oxford University Press.

Hallstrom, Daniel G., and V. Kerry Smith. 2005. "Market responses to hurricanes." *Journal of Environmental Economics and Management*, 50(3): 541–561.

Hanemann, W.M. 1978. "A Methodological and Empirical Study of the Recreation Benefits from Water Quality Improvement." *PhD. DIssertation (Economics), Harvard University.*

Heckman, James J. 1978. "Dummy Endogenous Variables in a Simultaneous Equation System." *Econometrica*, 46(4): 931–959.

Hicks, Robert L., and Ivar E. Strand. 2000. "The Extent of Information: Its Relevance for Random Utility Models." *Land Economics*, 76(3): 374–385. Publisher: [Board of Regents of the University of Wisconsin System, University of Wisconsin Press].

Hjerpe, Evan, Anwar Hussain, and Thomas Holmes. 2020. "Amenity Migration and Public Lands: Rise of the Protected Areas." *Environmental Management*, 66(1): 56–71.

Hotelling, Harold. 1947. "Letter to the National Parks Service." In *An economic study of the monetary evaluation of recreation in the National Parks*. US Department of the Interior, National Parks Service.

Keiser, David A, and Joseph S Shapiro. 2019. "Consequences of the Clean Water Act and the Demand for Water Quality." *The Quarterly Journal of Economics*, 134(1): 349–396.

Klaiber, Allen H., and Daniel J. Phaneuf. 2010. "Valuing open space in a residential sorting model of the Twin Cities." *Journal of Environmental Economics and Management*, 60(2): 57–77.

Kling, Catherine L., and Cynthia J. Thomson. 1996. "The Implications of Model Specification for Welfare Estimation in Nested Logit Models." *American Journal of Agricultural Economics*, 78(1): 103–114. Publisher: [Agricultural & Applied Economics Association, Oxford University Press].

Kuwayama, Yusuke, Sheila Olmstead, and Jiameng Zheng. 2020. "A More Comprehensive Estimate of the Value of Water Quality." *Working Paper*, 1–57.

Lupi, Frank, Daniel J. Phaneuf, and Roger H. von Haefen. 2020. "Best Practices for Implementing Recreation Demand Models." *Review of Environmental Economics and Policy*, 14(2): 302–323. Publisher: Oxford Academic.

McFadden, Daniel. 1974. "Conditional logit analysis of qualitative choice behavior." In *Frontiers in econometrics*.

Muehlenbachs, Lucija, Elisheba Spiller, and Christopher Timmins. 2015. "The Housing Market Impacts of Shale Gas Development." *The American Economic Review*, 105(12): 3633–3659. Publisher: American Economic Association.

Murdock, Jennifer. 2006. "Handling unobserved site characteristics in random utility models of recreation demand." *Journal of Environmental Economics and Management*, 51(1): 1–25.

National Park Service. 2020. "Visitation Numbers (U.S. National Park Service)." National Park Service.

Parsons, George R., Andrew J. Plantinga, and Kevin J. Boyle. 2000. "Narrow Choice Sets in a Random Utility Model of Recreation Demand." *Land Economics*, 76(1): 86–99. Publisher: [Board of Regents of the University of Wisconsin System, University of Wisconsin Press].

Parsons, G.R., D. Matthew Massey, and Ted Tomasi. 1999. "Familiar and Favorite Sites in a Random Utility Model of Beach Recreation." *Marine Resource Economics*, 14(4): 299–315. Publisher: [MRE Foundation, Inc., University of Chicago Press].

Petrin, Amil, and Kenneth Train. 2010. "A Control Function Approach to Endogeneity in Consumer Choice Models." *Journal of Marketing Research*, 47(1): 3–13.

Phaneuf, Daniel J., and V. Kerry Smith. 2005. "Recreation Demand Models." In *Handbook of Environmental Economics*. Vol. 2, 671–761. Elsevier.

Phaneuf, Daniel J., V. Kerry Smith, Raymond B. Palmquist, and Jaren C. Pope. 2008. "Integrating Property Value and Local Recreation Models to Value Ecosystem Services in Urban Watersheds." *Land Economics*, 84(3): 361–381. Publisher: [Board of Regents of the University of Wisconsin System, University of Wisconsin Press].

Rickman, Dan S., and Shane D. Rickman. 2011. "Population Growth in High-Amenity Nonmetropolitan Areas: What's the Prognosis?*." *Journal of Regional Science*, 51(5): 863–879. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9787.2011.00734.x.

Rosen, Sherwin. 1974. "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition." *Journal of Political Economy*, 82(1): 34–55. Publisher: University of Chicago Press.

Shaw, W. Douglass, and Michael T. Ozog. 1999. "Modeling Overnight Recreation Trip Choice: Application of a Repeated Nested Multinomial Logit Model." *Environmental and Resource Economics*, 13(4): 397–414.

Smith, V. Kerry, William H. Desvousges, and Ann Fisher. 1986. "A Comparison of Direct and Indirect Methods for Estimating Environmental Benefits." *American Journal of Agricultural Economics*, 68(2): 280–290. Publisher: [Agricultural & Applied Economics Association, Oxford University Press].

Tiebout, Charles M. 1956. "A Pure Theory of Local Expenditures." *Journal of Political Economy*, 64(5): 416–424.

Timmins, Christopher, and Jennifer Murdock. 2007. "A revealed preference approach to the measurement of congestion in travel cost models." *Journal of Environmental Economics and Management*, 53(2): 230–249.

Train, Kenneth E. 2009. *Discrete Choice Methods with Simulation.* . Second ed., New York, NY, US:Cambridge University Press.

Villas-Boas, J. Miguel, and Russell S. Winer. 1999. "Endogeneity in Brand Choice Models." *Management Science*, 45(10): 1324–1338.

von Haefen, Roger H., and Daniel J. Phaneuf. 2008. "Identifying demand parameters in the presence of unobservables: A combined revealed and stated preference approach." *Journal of Environmental Economics and Management*, 56(1): 19–32.

Ward, Frank A., and John B. Loomis. 1986. "The Travel Cost Demand Model as an Environmental Policy Assessment Tool: A Review of Literature." *Western Journal of Agricultural Economics*, 11(2): 164–178. Publisher: Western Agricultural Economics Association.