2013

Three Essays in Non-market Valuation and Energy Economics

Yongjie Ji
Iowa State University

Three essays in non-market valuation and energy economics

by

Yongjie Ji

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Economics

Program of Study Committee:

Joseph A. Herriges, Co-Major Professor

James B. Bushnell, Co-Major Professor

Catherine L. Kling

Brent Kreider

John R. Schroeter

Iowa State University

Ames, Iowa

2013

DEDICATION

I would like to dedicate this thesis to my wife Xiaobo Xiong and to my daughter Jennifer without whose support I would not imagine I have been able to complete this work. I would also like to thank my family for the understanding and help in the last six years

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my thanks to those who helped me with various aspects of conducting research and the writing of this thesis.

First and foremost, thanks Dr. Herriges and Dr. Bushnell for their guidance, patience and support throughout my research and the whole writing of these essays. Regular meetings with them not only help me finish the study, but also learn the important things about how to be a solid and qualified researcher.

The same gratitude to Dr. Kling, thanks for the opportunity to be included in the non-market valuation research group. Without the support of the data and valuable group discussion, I can not finish my dissertation work. Also thanks for her funding opportunities which alleviate the common pressure faced by graduate students like me.

I want to take the chance to show my thanks for Dr. Kreider and Dr. Schroeter. Thanks for serving in my committee and give me valuable suggestions on the research work.

I would also like to thank my friends for their encouragement and useful discussion. Thank you, Jingbo Cui, Bo Xiong, Hailong Jin, Hang Qian, Alicia Rosburg, Adriana Valcu, Subhra Bhattacharjee and Babatunde Abidoye.

CHAPTER 1. General Introduction

1.1 Overview

The focus of my dissertation is in two areas: modeling recreation behavior with limited information and the interaction between two greenhouse mitigation instruments in the power market.

My first dissertation chapter, entitled "*Modeling Recreation Demand when the Access Point is Unknown*", seeks to use the aggregation technique to model Iowan's riverine recreation behavior without knowing their detailed access points. The task of modeling the recreation demand for geographically large sites, such as rivers and beaches or large parks with multiple entrances, is often challenged by incomplete information regarding the access point used by the individual. Traditionally, analysts have relied upon convenient approximations, defining travel time and travel distances on the basis of the midpoint of a river or beach segment or on the basis of the nearest access point to the site for each individual. In this paper, we instead treat the problem as one of aggregation, drawing upon and generalizing results from the aggregation literature. The resulting model yields a consistent framework for incorporating information on site characteristics and travel costs gathered at a finer level than that used to obtain trip counts. We use a series of Monte Carlo experiments to illustrate the performance of the traditional mid-point and nearest access point approximations. Our results suggest that, while the nearest access point approach provides a relatively good approximation to underlying preferences for a wide range of parameter specifications, use of the midpoint approach to calculating travel cost can lead to significant bias in the travel cost parameter and corresponding welfare calculations. Finally, we use our approach in modeling recreation demand for the major river systems in Iowa using data from the 2009 Iowa Rivers and River Corridors Survey.

The second paper in my dissertation, entitled "*Modeling Recreation with Partial Trip Information*", tries to use the same aggregation technique in another set of situations with partial information about residents' visitation patterns. Full information about visitation pattern to all the related recreational sites is unavailable with surveys yielding trip information to a subset of possible sites. Conventional methods tend to focus on the sites with trip information and discard the sites with partial trip information. In this paper, we treat the partial information as an aggregation choice for this group of sites. In doing so, a similar aggregation modeling technique is proposed then, under some circumstances, allows one to recover preference parameters and avoid the possible bias caused by the conventional methods. A series of Monte Carlo simulations are conducted to study the possible bias caused by conventional methods and the performance of the aggregation model when the application is possible. The results show that the aggregation model performs quite well in recovery of preference and subsequent welfare analysis. Both methods are applied to data from 2009 Iowa lake and river projects. The results show that both methods give qualitatively similar preference parameters but produce significant differences in terms of the welfare measures.

The third paper in my dissertation, entitled "*Carbon Tax, Wind Energy and GHG reduction-ERCOT as an Example*", seeks to evaluate the performance of two greenhouse gas intervention policies in the Texas ERCOT, power market.In the battle to control the greenhouse gas (GHG) emission, a prominent component contributing to the climate change, there are several schemes already taken by governments. Direct targeted policies, such as cap-and-trade program or a potential carbon tax, and indirect policies, such as promotion of renewable energies are receiving governments endorsements worldwide. With data from the Texas ERCOT power market, we develop a simple electricity generation dispatch model to analyze the relative performance in emission reduction when a carbon tax and significant amount of wind generation co-exist in the power grid. The simulation results show that during the research period, both policies have significant effects on reduction of carbon dioxide emission under hypothetical policy scenarios. The combination of a carbon tax policy and the promotion of wind energy seems more effective to achieve big reduction targets in the short run.

1.2 Dissertation Organization

The overall structure of the dissertation is ordered as follows. The next three chapters present the essays described above. Each of these chapters is considered as a stand-alone paper. All the supplement tables and figures are included in the appendices.

CHAPTER 2. Modeling Recreation Demand when the Access Point is Unknown

2.1 Introduction

Recreation demand (or travel cost) models provide one of the primary tools for valuing environmental amenities, inferring value by observing the full costs incurred by the individual or household in reaching sites in a choice set. While there are a myriad of conceptual issues in defining travel costs themselves (see, e.g., (9) and (23)), practitioners are typically content with computing these costs as the sum of out-pocket costs (usually a fixed mileage rate times round-trip travel distance) and an opportunity cost for the individual's travel time (often valued at a fixed fraction of the individual's wage rate times round-trip travel time). However, in applications where the specific access point used to visit the "site" is unknown, the appropriate way to compute the cost of access can be unclear. A prime example is river based recreation. Surveys can elicit information on the number of trips to one or more river segments during the course of a season, but typically do not acquire information on the precise access point used by the individual on each choice occasion. This makes the computation of travel costs problematic in that the analyst cannot precisely compute either the travel distance or travel time. At best, in these cases travel costs can be bounded by considering the nearest and furthest access points along the river segment. Similar problems emerge in the context of beaches (e.g., (14)) and wetlands (e.g., (24)) recreation, or, more generally, any large geographic regions.

A number of solutions to this problem are employed in the literature, including computing travel costs based upon the nearest access point for each individual or using the midpoint along the river segment as the assumed point of entry for everyone. The issue with these *ad hoc* approaches is that they implicitly make assumptions regarding the role of travel costs (the

marginal utility of income) in the individual's decision making that are inherently inconsistent with the broader models used to represent the choice *among* river segments. For example, using the nearest access point along a river segment to compute travel cost implicitly assumes that travel cost is *the* determining factor in choosing where along a river segment to recreate (essentially assuming that the marginal utility of income is infinite), whereas in the broader models of the choice among river segments (say a RUM model) travel cost is but one of the factors in site selection (implying a finite marginal utility of income) (e.g., (14) and (24)).

In this paper, we consider an alternative approach, treating segment level trip data as the aggregation of underlying access-point level trip information. A logit structure is used to construct the choice probabilities for this aggregated data and to consistently recover preference parameters. A series of Monte Carlo exercises are used to compare and contrast the performance of this approach versus both the mid-point and nearest access point (or shortest distance) approximations used in the literature. The simulation results suggest that our model successfully recovers the underlying preference parameters, while the two traditional approaches vary in their estimation of the key travel cost parameter and subsequent welfare estimates. The shortest distance model generally provides a good approximation over a wide range of model parameterizations. In contrast, the commonly used midpoint model generates bias in both the travel cost parameter and subsequent welfare estimates that increases substantially as the number of river segments increases and travel costs become a more important determinant of behavior. In addition, we apply our approach, along with the midpoint and nearest access point approaches, using data from the 2009 Iowa River Survey. The survey was conducted in late 2009, eliciting information on the visitation patterns of 10000 randomly chosen Iowans to 73 identified river segments in the state.

The remainder of this paper is organized as follows. Section 2 provides a brief review of the literature. Our modeling approach to handling the missing access point data is then described in section 3. Section 4 describes a Monte Carlo exercise used to illustrate the scope of the bias from using either the midpoint or nearest access point approximations to travel cost. Finally, section 5 describes the 2009 Iowa Rivers project application, including both a description of the data and the resulting parameter estimates. Section 6 concludes the paper.

2.2 Related Literature

The problem of missing access point data is directly related to the issue of site aggregation encountered in recreation demand analysis (See, e.g., (4), (23) and (19)). Whereas missing access point data essentially forces the aggregation of possible "sites," practitioners have often intentionally aggregated elementary sites for computational convenience. In this literature, a wide variety of aggregation schemes have been considered, including county level aggregation, activity based aggregation, aggregation of familiar or unfamiliar sites and distance-based aggregation (See, e.g., (14), (23), and (22)). The potential bias associated with aggregation is the major concern in this context. The magnitude of bias depends on the degree of the aggregation and the heterogeneity across the aggregated elementary sites. Unfortunately, the nonlinearity of the RUM model makes the direction of bias generally ambiguous with respect to both preference parameter estimates and in terms of the welfare change induced by site loss or a change in site characteristics.

Kaoru and Smith (15) were the first to analyze the effects of aggregation on preference parameter estimation and welfare measurement in the context of recreation demand. Their work suggested that models with only a mild degree of site aggregation (i.e., 35 sites aggregated to 23 or 11 composite sites) performed relatively well in characterizing recreation behavior. The results, however, were not as promising in terms of subsequent welfare calculations. For example, the welfare impact from the closure of an aggregate site was understated by more than a factor of two using either site aggregations. The estimated welfare gain from site quality improvements faired even worse, being understated by a factor of five when 11 composite sites were used (See Kaoru *et al.* (10)).

Parson and Needelman's (23) subsequent paper identified two distinct sources of bias stemming from site aggregation, one linked to the number of sites being aggregated (the so-called *size* effect) and the other tied to the degree of heterogeneity among the sites being combined. Specifically, drawing on earlier work in the transportation literature by Ben-Akiva and Lerman (2), Parsons and Needleman note that, if the utility received by individual i from choosing an

elementary site j is given by

$$U_{ij} = V_{ij} + \epsilon_{ij} \qquad j = 1, \ldots, J, \tag{2.1}$$

where the ϵ_{ij}'s are distributed *i.i.d.* Gumbel with mode 0 and scale parameter μ, then the utility associated with choosing the aggregate site s ($s = 1, \ldots, S$) is given by:[1]

$$\begin{aligned} U_{is} &= \max_{j \in A_s} U_{ij} && (2.2) \\ &= \bar{V}_{is} + \mu ln J_s + \mu ln B_{is} + \epsilon_{is}, && (2.3) \end{aligned}$$

where A_s denotes the set of elementary sites associated with the aggregate site s,

$$\bar{V}_{is} = \frac{1}{J_s} \sum_{j \in A_s} V_{ij}, \tag{2.4}$$

J_s denotes the number of sites associated with aggregate site s,

$$B_{is} = \frac{1}{J_s} \sum_{j \in A_s} exp[\mu^{-1}(V_{ij} - \bar{V}_{is})], \tag{2.5}$$

and the $\epsilon'_{is}s$ are again distributed *i.i.d.* Gumbel with mode 0 and scale parameter μ. Estimating a model of aggregate site choice using only average site characteristics (including travel cost) corresponds to specifying that the utility from visiting aggregate site s is given by:

$$U_{is} = \bar{V}_{is} + \epsilon_{is} \tag{2.6}$$

Comparing equations (2.3) and (2.6), it is clear that the latter specification suffers potential bias due to two omitted variables: (a) a *size* variable reflecting the number of sites in the aggregate alternative s (i.e., lnJ_s in (2.3)) and (b) a measure of the heterogeneity of the sites being combined (reflected by lnB_{is} in (2.3)).[2] The general nature of the problem does not change if, in lieu of $\bar{V}_{is}$, an alternative proxy (V^p_{is}) is used to characterize the aggregate site utility (e.g., by using the nearest access point or site midpoint to determine travel cost). In this case, equation (2.3) simply becomes

$$U_{is} = V^p_{is} + \mu ln J_s + \mu ln B^p_{is} + \epsilon_{is}, \tag{2.7}$$

[1] The standard deviation of the ϵ_{ij}'s is given by $\mu\pi/\sqrt{6}$. Note that the scale parameter referred to in Parsons and Needleman corresponds to our μ^{-1}.

[2] Note that while the impact of these omitted variables can be mitigate by making the aggregates similar in size and minimizing the degree of heterogeneity across sites in terms of site attributes, site heterogeneity will necessarily persist in the form of heterogeneous travel costs to the elemental sites.

where

$$B_{is}^{p} = \frac{1}{J_s} \sum_{j \in A_s} exp[\mu^{-1}(V_{ij} - V_{is}^{p})]. \tag{2.8}$$

Modeling aggregate site utility using $U_{is} = V_{is}^{p} + \epsilon_{is}$ would again be subject to omitted variables bias.

Parsons and Needleman provide empirical evidence as to the scope of aggregation bias. Specifically, using data on fishing trips to 1133 lakes in Wisconsin, they estimate three models: one using the full choice set as in (2.1), a second using site aggregates with only average site characteristics as in (2.6), and a third using site aggregates with both average site characteristics and a size correction (but no heterogeneity correction). Two levels of aggregation are considered (9 regions and 61 counties). The results suggest that ignoring both heterogeneity and size factors leads to significant bias in parameter estimates (except for the price coefficient) and that, while the size correction alone works well with limited aggregation, the size corrected model performs poorly when large numbers of sites are aggregated. The authors suggest minimizing heterogeneity of sites within aggregates and controlling for the number of sites in the aggregate groups. A series of subsequent papers have largely confirmed the findings in Parsons and Needleman (23) (including Kaoru, Smith, and Liu (10), Feather (4), Feather and Lupi (14) and Parson, Plantinga and Boyle (22)).

In a more recent paper, Haener *et al.* (7) suggest that, while analysts may choose not to model detailed site visitation data, they often have access to detailed site characteristics data, including travel costs. As such, they should be able to form both the *size* and *heterogeneity* correction terms in equation (2.3) and obtain consistent parameter estimates. Their empirical analysis, however, suggests that the size correction alone mitigates much of the aggregation bias. For their application, site heterogeneity appears to not play a significant role.

2.3 Methodology

The approach followed in the aggregation literature is based on the underlying structure of the logit model, yielding the specific *size* and *heterogeneity* correction terms identified in equation (2.3). The problem, however, is more general and the solution need not rely on the

logistic structure. Rather than observing the elemental site visitation data, we observe only whether one of a series of sites is visited. Specifically, in the case of a single choice, let y_{isj} equal 1 when individual i visits elemental site j in aggregate site s (and equals 0 otherwise). Observing visitation data for the aggregate site s corresponds to observing $y_{is\bullet}$, where

$$y_{is\bullet} = \sum_{j \in A_s} y_{isj}. \tag{2.9}$$

The corresponding choice probability for the aggregate site is then simply the sum of the individual choice probabilities; i.e.,

$$P_{is\bullet} = Pr[y_{is\bullet} = 1] \tag{2.10}$$

$$= \sum_{j \in A_s} Pr[y_{isj} = 1] = \sum_{j \in A_s} P_{isj}. \tag{2.11}$$

In this section, we begin by laying out the implications of this aggregation in the context of a simple repeated logit model of riverine recreation, linking it to the existing literature on site aggregation. We then extend to model to the nested and mixed logit settings and to the case in which site attributes are available at the elemental site level.

The Repeated Logit Model

Following Morey, Rowe and Watson (15), we model an individual's riverine recreation using a repeated random utility maximization (RUM) framework. In particular, we assume that there are T choice occasions in each year. On each choice occasion, the individual decides either to visit one of the river segments or to stay at home. There are S river segments and J_s access points along the segment. The conditional utility that household i receives from visiting river segment s ($s = 1, \ldots, S$) via access point j ($j \in A_s$) on choice occasion t ($t = 1, \ldots, T$) is assumed to take the form

$$U_{isjt} = \alpha_s + \beta C_{isj} + \epsilon_{isjt}, \tag{2.12}$$

where α_s is a segment specific constant reflecting all segment attributes and C_{isj} is the travel cost of reaching access point j along river segment s for household i.[3] The error term ϵ_{isjt}

[3]The assumption that there is single segment specific constant, rather than an alternative specific constant for each access point (i.e., an α_{sj}), implicitly assumes that there is no heterogeneity in site attributes along the river segment, an assumption that will be relaxed below. Given this assumption, the only source of heterogeneity across access points is in terms of the travel cost C_{isj}.

captures unobserved factors influencing the choice made by the household. Letting $s = j = 0$ denotes the option of choosing to stay at home on a given choice occasion, the relevant conditional utilities can be summarized as

$$U_{isjt} = \begin{cases} \epsilon_{isjt} & \text{if } s = j = 0 \\ V_{isj} + \epsilon_{isjt} & \text{otherwise} \end{cases} \tag{2.13}$$

where $V_{isj} = \alpha_s + \beta C_{isj}$, and V_{i00} has been normalized to zero for the stay-at-home option. Assuming that the ϵ_{isjt}'s are *i.i.d.* type I extreme value random variables, individual i will choose to visit the segment-access point combination sj, denoted by $y_{isjt} = 1$, with the probability of

$$P_{isjt} = Pr(y_{isjt} = 1) = \frac{exp(V_{isj})}{1 + \sum_{r=1}^{S} \sum_{k \in A_r} exp(V_{irk})} = P_{isj} \quad \forall t. \tag{2.14}$$

If the elementary choices made by households (i.e., the y_{isjt}'s) were observed, we could form the appropriate likelihood function on the basis of equation (2.14) and estimate the parameters of the model. Instead, information is only provided at the segment level; i.e.,

$$y_{is\bullet t} = \sum_{j=1}^{J_s} y_{isjt}. \tag{2.15}$$

However, we can still use equation (2.14) to construct the relevant choice probabilities. In particular, we have

$$P_{is\bullet t} = Pr(y_{is\bullet t} = 1) = \frac{\sum_{j \in A_s} exp(V_{isj})}{1 + \sum_{r=1}^{S} \sum_{k \in A_r} exp(V_{irk})} = P_{is\bullet} \quad \forall t \tag{2.16}$$

where $y_{is\bullet t}$ equals 1 if the individual chooses to visit the segment s at some unknown access point along this segment.[4] These aggregate probabilities provide the basis for estimating a repeated logit model using the aggregated data and maximum likelihood estimation. In particular, the contribution of individual i to the log-likelihood function is given by:

$$\mathcal{L}_i(\boldsymbol{n}_i) = \sum_{s=0}^{S} n_{is\bullet} ln \left(\sum_{j \in A_s} P_{isj} \right) \tag{2.17}$$

$$= \sum_{s=0}^{S} n_{is\bullet} ln(P_{is\bullet})$$

$$= \left\{ \sum_{s=1}^{S} n_{is\bullet} ln \left[\sum_{j \in A_s} exp(V_{isj}) \right] \right\} - T \cdot ln \left[1 + \sum_{r=1}^{S} \sum_{k \in A_r} exp(V_{irk}) \right], \tag{2.18}$$

[4] A similar approach was suggested and applied by Kurkalova and Rabotyagov (13) in a binary model when county level, rather than farm level, data was available in an agricultural technology adoption setting.

where $\boldsymbol{n}_i = (n_{i0\bullet}, \ldots, n_{iS\bullet})$ and $n_{is\bullet} = \sum_{t=1}^{T} y_{is\bullet t}$ denotes the total number of times aggregate alternative s is chosen across the T choice occasions. Note that the specification of the log-likelihood function in (A.3) holds in general when only aggregate data are available, but that (A.4) holds specifically for the logit formulation of the choice probabilities. From a programming point of view, (A.4) provides all that is needed in terms of estimation.[5] It is not necessary to reduce the expression further. However, it is instructive to do so. Specifically, note that we can rewrite equation (2.16) as

$$P_{is\bullet} = \frac{exp(V_{is\bullet})}{1 + \sum_{n=1}^{S} exp(V_{in\bullet})} \tag{2.19}$$

where

$$V_{is\bullet} = ln\left[\sum_{j \in A_s} exp(V_{isj})\right] \tag{2.20}$$

$$= ln\left[\sum_{j \in A_s} exp(\alpha_s) exp(\beta C_{isj})\right]$$

$$= \alpha_s + ln\left[\sum_{j \in A_s} exp(\beta C_{isj})\right] \tag{2.21}$$

$$= \alpha_s + \beta C_{is\bullet} \tag{2.22}$$

with

$$C_{is\bullet} \equiv \frac{1}{\beta} ln[\sum_{j \in A_s} exp(\beta C_{isj})]. \tag{2.23}$$

The term $C_{is\bullet}$ can be thought of as the aggregate price for segment s. Indeed, viewing (A.28) as a function of the access point travel costs (i.e., the C_{isj}'s), a first order Taylor-series approximation of C_{is} around the mean segment travel cost yields

$$C_{is\bullet} \approx \sum_{j \in A_s} P_{isj|s} C_{isj} \tag{2.24}$$

where $P_{isj|s} \equiv P_{isj}/P_{is\bullet}$ denotes the probability that access point j is chosen, given segment s has been selected. The segment level travel cost is just a probability weighted average of the access point travel costs. However, because C_{is} involves the unknown preference parameter

[5] This assumes, of course, that the model remains identified, an issue that is returned to below.

β, there is no promising way to construct it *ex ante* for use in estimation. Consequently, the conditional indirect utility function for the aggregate site s (i.e., $V_{is\bullet}$) is no longer linear in the parameter β, as can be seen in (2.21), making estimation potentially more difficult.[6]

The alternative approach typically used in the literature is to replace $C_{is\bullet}$ in equation (2.22) with a proxy (C_{is}^{p}), computing travel cost of the basis of either the nearest access point ($p = min$) or the midpoint ($p = mid$) of the segment. The advantage of doing so is that the corresponding conditional indirect utility function is once again linear in its parameters, with $V_{is\bullet}^{p} = \alpha_s + \beta C_{is}^{p}$. The problem, as noted above, is that the subsequent parameter estimates will be subject to omitted variable bias, since:

$$\begin{aligned} V_{is\bullet} &= \alpha_s + ln\left[\sum_{j\in A_s} exp(\beta C_{isj})\right] \\ &= \alpha_s + \beta C_{is}^{p} + ln\left[\sum_{j\in A_s} exp(\beta[C_{isj} - C_{is}^{p}])\right] \\ &= \alpha_s + \beta C_{is}^{p} + ln(J_s) + ln\left[\frac{1}{J_s}\sum_{j\in A_s} exp(\beta[C_{isj} - C_{is}^{p}])\right] \\ &= V_{is\bullet}^{p} + ln(J_s) + ln\left[\frac{1}{J_s}\sum_{j\in A_s} exp(\beta[C_{isj} - C_{is}^{p}])\right], \end{aligned} \tag{2.25}$$

where the last two terms are the size and heterogeneity corrections identified in the aggregation literature.[7] The minimum distance proxy, $C_{is}^{min} = \min_{j\in A_s}\{C_{isj}\}$, has the intuitively appealing property that the omitted variable bias disappears as the marginal utility of income increases (i.e., $V_{is\bullet}^{min} \to V_{is\bullet}$ as $\beta \to -\infty$).

Finally, the analysis above assumes that the conditional utilities (V_{isjt}) derived from the elemental access sites differ only terms of travel cost, sharing a common segment specific constant. As argued in the Appendix A below, a general structure allowing for access point specific constants (i.e., with $V_{isj} = \alpha_{sj} + \beta C_{isj}$) could be weakly identified.[8]

[6] It is, however, the case that $V_{is\bullet}$ is still linear in the segment specific constants α_s. As a result, the model will still be mean fitting (i.e., the actual segment shares will precisely equal the mean fitted shares) and the contraction mapping algorithm outlined in Murdock (1) can still be used in estimation.

[7] Here we have normalized the scale parameter $\mu = 1$.

[8] By weak identification, we mean the identification relies on the logistic choice structure we assumed in the first place. For example, if we assume a linear probability model for the access point level choice, the identification goes away.

The model can, however, be generalized to control for observable site qualities, with

$$V_{isj} = \alpha_s + \beta C_{isj} + \gamma X_{sj}, \tag{2.26}$$

where X_{sj} denotes an attribute of access point j along segment s. The segment level choice probabilities in (A.26) would apply, except that equation (2.21) would now be replaced with:

$$V_{is\bullet} = \alpha_s + ln\left[\sum_{j \in A_s} exp(\beta C_{isj} + \gamma X_{sj})\right] \tag{2.27}$$

There are, however, several problems with this approach. While such a model is identified, it is only structurally identified (See Appendix A), relying heavily on the logit structure to identify γ.[9] This weakness is potentially exacerbated by limited variation in site attributes within a segment. In the extreme, if there is no variation in access point attributes, their impact would be captured by the segment specific constant α_s.

The Nested Logit Model

While the focus of our Monte Carlo and empirical analysis below is on the mixed logit generalization of the logit model, in this section we touch briefly on the implications of aggregated data for the more traditional nested logit models. We consider two nested logit specifications.

Specification 1: Trip Nest

In the first specification, all of the segments (and their associated access points) are grouped together in a single nest. In this case, the choice probability for access point j becomes:

$$P_{isj} = exp(\tilde{V}_{isj})\left[\sum_{r=1}^{S}\sum_{k \in A_r} exp(\tilde{V}_{irk})\right]^{\theta-1}\left\{1 + \left[\sum_{r=1}^{S}\sum_{k \in A_r} exp(\tilde{V}_{irk})\right]^{\theta}\right\}. \tag{2.28}$$

where

$$\tilde{V}_{isj} = \frac{V_{isj}}{\theta} = \frac{\alpha_s}{\theta} + \frac{\beta C_{isj}}{\theta} = \tilde{\alpha_s} + \tilde{\beta} C_{isj}. \tag{2.29}$$

As shown in Appendix A, the choice probability for aggregate site s retain the general nested logit structure, with

$$P_{is\bullet} = exp(\tilde{V}_{is\bullet})\left[\sum_{r=1}^{S}\sum_{k \in A_r} exp(\tilde{V}_{ir\bullet})\right]^{\theta-1}\left\{1 + \left[\sum_{r=1}^{S}\sum_{k \in A_r} exp(\tilde{V}_{ir\bullet})\right]^{\theta}\right\} \tag{2.30}$$

[9] One need only consider the alternative linear probability model to see this. In the case of the linear probability model, the segment level probability becomes a function of the sum of the access point attributes, which is perfectly collinear with the segment level alternative specific constant.

where

$$\tilde{V}_{is\bullet} = \tilde{\alpha}_s + \tilde{\beta}\tilde{C}_{is} \tag{2.31}$$

with

$$\tilde{C}_{is} = \frac{1}{\tilde{\beta}} ln \left[\sum_{j \in A_s} exp(\tilde{\beta} C_{isj}) \right]. \tag{2.32}$$

All of the underlying parameters of the model (i.e., α_s, β, and θ) remain identified.

Specification 2: Segment Nests

In the second specification, the access points within each segment form distinct nests. In this case, (A.29) is replaced with

$$P_{isj} = exp(\breve{V}_{isj}) \left[\sum_{k \in A_s} exp(\breve{V}_{isk}) \right]^{\theta_s - 1} \left\{ 1 + \sum_{r=1}^{S} \left[\sum_{k \in A_r} exp(\breve{V}_{irk}) \right]^{\theta_r} \right\}, \tag{2.33}$$

where

$$\breve{V}_{isj} = \frac{V_{isj}}{\theta_s} = \frac{\alpha_s}{\theta_s} + \frac{\beta C_{isj}}{\theta_s} = \breve{\alpha}_s + \breve{\beta}_s C_{isj}. \tag{2.34}$$

Now (as shown in Appendix A), the choice probability for segment s becomes

$$P_{is\bullet} = \exp(\breve{V}_{is\bullet}) \left\{ 1 + \sum_{r=1}^{S} exp(\breve{V}_{ir\bullet}) \right\}^{-1} \tag{2.35}$$

where

$$\breve{V}_{is\bullet} = \alpha_s + \beta \breve{C}_{is} \tag{2.36}$$

with

$$\breve{C}_{is} = \frac{1}{\breve{\beta}_s} ln \left[\sum_{j \in A_s} exp(\breve{\beta}_s C_{isj}) \right]. \tag{2.37}$$

Again, all of the underlying parameters of the model (i.e., α_s, β, and θ_s) remain identified with aggregate data. However, unlike in the previous case, the segment choice probabilities look like a standard logit model. Identification of θ_s (which is equivalent to identification of $\breve{\beta}_s$) hinges on the structure of the nonlinear relationship in (A.38). One observation regarding (A.38) is that as $\theta_s \to 0$, the travel cost index $\breve{C}_{is} \to C_{is}^{min}$, which reduces the variability needed to identify θ_s.

Normal Error Component Logit Mixture Models

Normal Error Component Logit Mixture(NECLM) models have become a popular alternative to nested logit model as a means of inducing correlation patterns among alternatives in the choice set (See, e.g., Herriges and Phaneuf (8) and Walker *et al.* (25)). Unobservable factors, shared by one or more of the alternatives, are introduced into the conditional utility functions in the form of normally distributed error components. One of the advantages the approach provides is the ability to create more complex and overlapping nests, rather than relying on the usual tree structure assumed by nested logit. At the same time, as Walker *et al.* (25) note, model identification can be more difficult to establish and spurious results can be obtained for models that are not identified if care is not taken in simulating the requisite choice probabilities. In this section, we discuss the identification of several NECLM model in the context of aggregate choice data. Though aggregate choice probabilities analogous to (2.16) hold, the identification of parameters in the model becomes difficult to establish. The requisite conditions for identification developed in Walker *et al.* (25) are employed.

Specification 1: Trip Nest

An error component structure similar to specification 1 of the nested logit model in the previous section would replace (A.2) with

$$\tilde{U}_{isjt} = \begin{cases} \epsilon_{isjt} & \text{if } s = j = 0 \\ V_{isj} + \tau_{it} + \epsilon_{isjt} & \text{otherwise} \end{cases} \tag{2.38}$$

where $\tau_{it} \sim N(0, \sigma^2)$ and ϵ_{isjt} is distributed *i.i.d.* Gumbel with mode zero and scale parameter μ. In this case, the corresponding utility for the aggregate segment level alternative becomes:

$$\begin{aligned} \tilde{U}_{is\bullet t} &= \max_{j \in A_s} \tilde{U}_{isjt} \\ &= \max_{j \in A_s} (V_{isjt} + \tau_{it} + \epsilon_{isjt}) \\ &= \left(\max_{j \in A_s} V_{isjt} + \epsilon_{isjt} \right) + \tau_{it} \\ &= V_{is\bullet} + \tau_{it} + \epsilon_{is\bullet t} \end{aligned} \tag{2.39}$$

where ϵ_{isjt} is distributed *iid* Gumbel with mode zero and scale parameter μ. The unconditional

choice probabilities have the same form as those for the disaggregate data, *except* that the aggregate cost variable becomes C_{is} in equation (2.32); i.e.,

$$P_{is\bullet} = \int \frac{exp(V_{is\bullet} + \tau)}{1 + \sum_{n=1}^{S} exp(V_{in\bullet} + \tau)} f(\tau) d\tau \tag{2.40}$$

where $f(\tau)$ denotes the pdf for τ_{it}. Appendix A provides the proof that the parameters of this model (i.e., α_s, β and σ) are identified.[10]

Specification 2: Segment Nests

An error component structure similar to specification 2 of the nested logit model in the previous section would replace (A.2) with

$$\tilde{U}_{isjt} = \begin{cases} \epsilon_{isjt} & \text{if } s = j = 0 \\ V_{isj} + \tau_{ist} + \epsilon_{isjt} & \text{otherwise} \end{cases} \tag{2.41}$$

where $\tau_{ist} \sim N(0, \sigma_s^2)$ and ϵ_{isjt} is distributed *i.i.d.* Gumbel with mode zero and scale parameter μ. In this case, the corresponding utility for the aggregate segment level alternative becomes:

$$\begin{aligned} \tilde{U}_{is\bullet t} &= \max_{j \in A_s} \tilde{U}_{isjt} \\ &= \max_{j \in A_s} (V_{isjt} + \tau_{ist} + \epsilon_{isjt}) \\ &= V_{is\bullet} + \tau_{ist} + \epsilon_{is\bullet t} \\ &= \alpha_s + \beta C_{is} + \tau_{ist} + \epsilon_{is\bullet t} \end{aligned} \tag{2.42}$$

where $\epsilon_{is\bullet t}$ is distributed *i.i.d.* Gumbel with mode zero and scale parameter μ. The model in equation (2.42) is analogous to the *alternative-specific variance model* considered by Walker *et al.* (25), except that C_{is} is a nonlinear function of the model parameter β. Appendix A provides the proof that the parameters of this model (i.e., α_s, β and σ_s) are identified.

2.4 Monte Carlo Simulation

The goal of this section is to evaluate the performance of the two standard travel cost proxies (i.e., the midpoint and shortest distance measures), relative to using the aggregated travel cost

[10] Similar results apply if it is assumed that τ_{it} is constant over time with $\tau_{it} = \tau_i \sim N(0, \sigma^2)$.

index C_{is}, in recovering preference parameters and calculating welfare change associated with the loss of a river segment. The simulation scenarios vary the river and parameter configurations along four dimensions:

- Price responsiveness: As noted above, the shortest distance proxy becomes more appropriate as travel cost becomes the dominant consideration in site selection. Thus we would expect the use of this proxy to create less bias as β increases. We consider the three levels of β shown in Table 1.1.

- Number of river segments: We consider three levels for the number of aggregate river segments S (S=5,10 and 20).

- River/Population configurations: Ferguson and Kanaroglou (5) note that the shape of the spatial object (river segments in this paper) and the spatial distribution of households will affect the heterogeneity among the aggregated sites, though they did not examine it's specific impact on estimation results. We consider four possible configurations for river segments and population centers. In two of the configurations, the rivers are assumed to be straight segments, 50 miles in length, whereas in the other two configurations the rivers are 50 miles long, but kinked at the midpoint. Population is either uniformly distributed or centers around the first two segments. Thus we have four possibilities along this dimension:

  - B: The base scenario with straight river segments and no population centers
  - K: Kinked river segments, with no population centers
  - P: Straight river segments, with population centers
  - C: The combination of kinked river segments and population centers

- Water quality: We consider two types of conditional utility functions. The first consists of segment specific constants along with travel cost (i.e., as depicted in equation 2.12). The second includes an additional term, representing say water quality, along the lines of (2.26).

A total of 72 ($2 \times 4 \times 3 \times 3$) Monte Carlo scenarios were considered, with 100 replications for each scenario. In each case, a simple logit structure is employed, though similar results are obtained when nesting among sites is allowed. Details of the data generation process are provided in Appendix B.

For each scenario, three models are used to recover the underlying preference parameters: two based on standard proxies for the aggregate site travel cost (i.e., the midpoint and shortest distance proxies) and one based on the aggregated choice probabilities. In addition, we consider the performance of the models in estimating the welfare costs associated with the closure of river segment 1 using the standard log-sum formula. In the context of the simple logit structure, this reduces to

$$CV_1 = \frac{1}{\beta} ln(1 - P_{i1\bullet}) \approx \frac{-P_{i1\bullet}}{\beta} \tag{2.43}$$

Since all three models have the basic logit structure with segment level alternative specific constants, they are mean fitting (i.e., the fitted choice probabilities will equal the average observed choice shares). Any bias in CV_1 will be driven by bias in the travel cost parameter.

2.4.1 Simulation Results

Table 1.2 summarizes the results for the first 36 of the Monte Carlo experiments; i.e., those without access-point level attributes (labeled here as water quality). We focus our attention on the travel cost parameter β, as it is the main determinant of subsequent welfare measures, with the alternative specific constants (the α_j's) changing to insure that the model is mean-fitting. The first half of Table 2 provides the mean absolute percentage error in the estimated travel cost parameter, β. Several results emerge here. First, the aggregated choice probabilities approach successfully recovers the underlying travel cost parameter, with a mean absolute percentage error of 0.3 percent or less. This should not be surprising, since it represents the true data generating process in this case. The aggregation of choice data to the segment level represents a loss of information, and hence efficiency, but does not alter the underlying model. Second, the shortest distance proxy also performs quite well, with a mean absolute percentage error that is typically less than 5 percent. However, the midpoint proxy does not do as well, particularly

when both the number of segments and the travel cost parameters are large. When $S = 20$ and $\beta = -0.1$, the mean absolute percentage error ranges between 15.8 and 19.6 percent. As we would expect, bias in the travel cost parameter translates directly into bias for the corresponding welfare measures. Table 1.2b compares the mean absolute percentage error for CV_1 across the same 36 Monte Carlo experiments. Again, the aggregated choice probabilities approach successfully recovers the true compensating variation associated with the loss of site 1, and shortest distance proxy works reasonably well. However, the midpoint proxy results in mean absolute percentage errors exceeding twenty-five percent.

Table 1.3's reports similar summary statistics for the cases in which water quality is included in households' recreational utility function at the access point level. Before proceeding with describing the results in Table 1.3's, two facts should be noted. First, there are two versions of the shortest distance model. In version 1, the water quality associated with the nearest site for each individual is used as their water quality measure, whereas in version 2 the average water quality over the entire segment is used. Second, for two of the specifications (i.e., the midpoint model and the shortest distance model, version 2), the water quality does not vary by individual and, hence, β_w is not identified (being collinear with the alternative specific constant for that site).

Starting with the results in Table 1.3a for the travel cost coefficient, the findings are similar to those in Table 1.2a. The aggregated choice probability model does a good job in recovering the underlying price coefficient, with the nearest distance proxy performing reasonably well. The midpoint proxy again suffers from the largest bias, with the mean absolute percentage error being higher when the price coefficient is larger and when there are more segments in the choice set. In Table 1.3b, we see that the aggregated choice probability model recovers the water quality parameter reasonably well, with a mean absolute percentage error that is typically less than five percent, though it reaches as high as 9.8 percent. Interestingly, the error rate appears to be highest when the price coefficient is small, perhaps because in that situation the water quality factor becomes a more dominant determinant of the individual's choice. The shortest distance proxy, in contrast, does a poor job in recovering β_w, with the mean percentage error typically exceeding fifty percent. Finally, as Table 1.3c indicates, the inclusion

of access point water quality attributes does not change the basic conclusions in terms of the estimated compensating variation associated with losing a site (CV_1). This is not surprising, as CV_1 is largely driven by the travel cost coefficient. Again, we find that the aggregated choice probability approach has the lowest mean absolute percentage error, followed by the shortest distance proxy and then the midpoint proxy approach.

2.5 An Application to Iowa Rivers

This section of the paper provides an application of the aggregated choice probability approach to the study of recreational river usage in Iowa. The primary data source for our analysis is the 2009 Iowa River Survey, funded by the Iowa Department of Natural Resources and the USEPA. The purpose of the survey was to gather baseline information about riverine recreation along 73 key river and stream segments in the state, depicted in Figure 1. The survey, conducted by mail, elicited data from each respondent regarding their total number of trips in 2009 to each of the river segments, as well as information regarding the individual's socio-demographic characteristics. However, information is not available regarding the specific point used by recreationists to access a given river or stream segment. With the segments ranging in length from 26 to 121 miles, considerable uncertainty exists in the imputed travel costs to the segment. We use the aggregated choice probability model to implicitly construct a travel cost index for each segment. The results are compared to models estimated using both the shortest distance and midpoint specifications commonly employed in the literature.[11]

2.5.1 Models

A total of three models are estimated using the Iowa Rivers data. All of the models are based on repeated logit version of the normal error component logit mixture (NECLM). Specifically, we employ a structure similar to (2.38), but with $\tau_{it} = \tau_i \sim \mathcal{N}(0, \sigma_\tau^2)$. This creates a nesting of all trip alternatives and a common error term inducing correlation across choice occasions.

[11] A copy of the survey instrument is included in the appendix.

The models start from the same basic structure for access-point level utility, with

$$\tilde{U}_{isjt} = \begin{cases} \gamma S_i + \epsilon_{isjt} & \text{if } s = j = 0 \\ \alpha_s + \beta C_{isj} + \tau_i + \epsilon_{isjt} & \text{otherwise} \end{cases} \tag{2.44}$$

where S_i denotes socio-demographic characteristics of individual i, potentially influencing their propensity to take trips. The three models differ in terms of how they handle aggregation, with one employing aggregated choice probabilities, while the other two employ the shortest distance and midpoint travel cost proxies, respectively. Following Murdock (1), a second stage regression of the alternative specific constants on segment characteristics is used to examine the role of site characteristics on recreation demand. Specifically, we run the second stage regression:

$$\hat{\alpha}_s = \alpha_0 + \delta Z_s + \xi_s \tag{2.45}$$

where Z_s denotes a vector of observable site characteristics for segment s.

2.5.2 The Iowa Rivers Data

After focus groups and pre-testing of the survey instrument, the 2009 Iowa Rivers Survey was mailed to a total of 10,000 Iowa households, beginning in November of 2009. Multiple mailings of the survey, as well as a postcard reminder and an incentive of $12 for completing the survey, were used to increase survey response. Among all the surveys mailed, 4758 surveys were returned, for an overall response rate of 49% among deliverable surveys. Of the returned surveys, 4137 are employed in the analysis below.[12] Table 1.4 provides a summary of the demographic characteristics of the survey respondents.The demographic statistics show that, on the average, a respondent to the survey tends to be an older, female Iowan with college level or equivalent education.[13]

[12] A total of 176 returned surveys were unusable because the respondents did not provide their numbers of visits to the river segments depicted in Figure 1. An additional 445 respondents reported taking more than 52 trips to individual river segments and were excluded from the sample. The focus of our analysis is on day-trips to the river segments. Setting the maximum numbers of trips to 52 allows for one trip per week. While this specific cut-off is arbitrary, the goal here is to focus on day trips and to exclude individuals who report large numbers of trips simply because they live on or near a specific river segment. Similar cutoffs have been used in other recreational studies (e.g., Egan *et al.* (2)) and were not found to significantly impact the results of the analysis.

[13] According to US Census, approximately 32 percent of adult Iowans are over 60, whereas in the survey sample this figure is somewhat higher at 36 percent. Likewise, among respondents the percentages of females (70 versus

Table 1.5 provides an overall summary of the data on trips to each of the river segments, both in terms of the percentage of the respondents who report visiting a given river segment and in terms of total numbers of trips to the segments. As the data indicate, the segments vary considerably in terms of popularity. River segment 71 (the Mississippi River between Clinton and Muscatine, Iowa) is the most popular, visited a total of 1591 times by just under six percent of the sample. At the other end of the spectrum, segment 37 (Big Cedar Creek, in northwest Iowa) is the least popular, visited only 20 times by 0.24 percent of the sample. In total, forty-eight percent of the sample visit at least one river segment during the course of the year, with an average number of trips per year of over six.

The travel cost variables (C_{isj}, C_{is}^{min}, and C_{is}^{mid}) must be constructed for each access point and river segment. There are several issues in doing so. First, a complete set of access points are not available for the Iowa rivers and streams.[14] In the current application, we divide each of the river segments into approximately twenty-mile sub-segments, defining "access points" in terms of the mid-point of each of these sub-segment. This process results in a total of 300 access-points. The numbers of sub-segments per segment ranges from one for River segment 1 (Rock River in northwest Iowa) to eight for River segment 56 (for portions of the Wapsipinicon River in eastern Iowa). Second, given these access points, travel cost must be calculated. PCMiler is used to compute both the round-trip distance (d_{isj}) and travel time (t_{isj}) between the individual's home and the relevant access point. Travel costs are then computed as $C_{isj} = g \cdot d_{isj} + (w_i/3) \cdot t_{isj}$, where g is per mile vehicle cost and w_i denotes the individual wage rate.[15] As indicated at the top of Table 1.5, the average round trip travel cost is approximately \$162.85, ranging from close to zero to almost \$500.

The final data category consists of river site characteristics (i.e., the Z_s's) summarized in Table 1.6. The following river characteristics were constructed:

- *LENGTH* indicates the length of the river segment or sub-segment;

50.4) and those with college degrees (69 versus 24.9) are higher than in the general Iowa population. The average family size of 2.4 is virtually the same in the sample as in the Iowa population as a whole.

[14] Indeed, for a number of activities, such as hunting, bird watching, etc., it is not clear what criteria to use in defining access points.

[15] The value of g was set to 54 cents per mile based on the 2009 AAA annual driving cost for an average sedan with 15,000 miles per year driving. The wage rate w was set at household income divide by 2000 times the number of adults in the household.

- *CANOE* indicates the percentage of the river segment or sub-segment that is considered canoeable, as defined by the Iowa Department of Natural Resources (IDNR);
- *OUTCROPPING* is a count of the number of outcroppings along the river segment or sub-segment, thought to contribute to the scenic nature of the river;
- *WATERBODY, WETLAND FOREST, GRASS, CROP*, and *DEVELOPED* indicate the percentage of the river corridor (defined as 75 meters on either side of the center-line for the river) that is water, wetland, forest, grassland, cropland and developed (industrial, commercial or residential) land, respectively;
- *IWQI* denotes a water quality index developed by the IDNR;[16]
- *MIWQI* is a dummy variable equal to one for river segments or subsegments for which the *IWQI* is not available;
- *FISH* denotes the number of fish species found along the river segment;[17] and
- *MFISH* is a dummy variable that equals one for river segments or subsegments without fish species data.

As Table 1.6 indicates, just over sixty percent of the Iowa river segments are canoeable. Not surprisingly, cropland is the largest form of land cover along the river segments (close to forty percent), with forested land being the second most common at under thirty percent. In terms of our two primary water quality measures, there are several important factors to note. First, both measures are available for only a fraction of the segments or sub-segments. Second, water quality along river segments is particularly difficult to capture, as rivers quality levels are only measured at selective sites. Even short distances from the monitoring site, water quality can be substantially different, depending on the river currents. Finally, the trips included a wide range of activities, from bird-watching to swimming, with the water quality measure being more or less salient depending upon the specific activity engaged in. With these concerns in

[16]This water quality index differs from USEPA's national water quality index. Details regarding its construction are available at Iowa Department of Natural Resources's webstie. One drawback of the *IWQI* measure is that it is available for less than 70 percent of the river segments and less than 30 percent of the river sub-segments.

[17]This information was provided directly by the IDNR.

mind, we consider below the use of turbidity as an alternative proxy for overall water quality. Turbidity is a measure of the cloudiness of water and, as such, is readily visible to recreators. As with $IWQI$, turbidity is only available for a fraction of river segments, so MTURBIDITY, a dummy variable for segments missing turbidity data, is also included in the model.

2.5.3 Results

The models estimated using the Iowa Rivers data are summarized in Tables 1.7, 1.8a and 1.8b. As indicated in Section 2.5.1, all three of these models employ a normal error component logit mixture (NECLM) structure and are estimated in two stages. The first stage involves estimating the alternative specific constants (i.e., ASC's α_s), the travel cost coefficient (β), and the parameters associated with the sociodemographic factors thought to influence the individual's propensity to stay at home (γ), as well as the variance of the trip error component (i.e., σ_τ^2).

Given the large number of ASC's (73), we refrain from reporting all of them here.[18] However, the ASC's can be used as an indicator of the relative "appeal" of each site, controlling for travel cost. That is, all else equal, a site with a larger ASC is preferred to a site with a lower ASC. Table 1.7 illustrates the implied ranking of the river segments on the basis of the estimated ASC's from each model. The second column in Table 1.7 also provides the ranking of sites by total visitation. Not surprisingly, all of the top river segments based on visitation rates are near population centers. For example, segments 23 and 24 rank 6th and 3rd, respectively, in terms of total visitation and are located near the state's largest city, Des Moines. However, the rankings of these segments drop substantially once travel cost is controlled for. Both segment 23 and 24's rank fall out of the list of top ten sites altogether. For models 1, 2 and 3, the relative "appeal" of the river segments do differ, though they share some common features. Seven of the top ten segments and six of the bottom ten segments are the same across the three models.

Table 1.8a provides the other stage 1 parameter estimates. As expected, the travel cost coefficient is negative and statistically significant under all three specifications. Indeed, β varies

[18] These ASCs are reported in the appendix.

relatively little across the three models. The three models also generally agree as to the impact of age and boat ownership, with older individuals and those without a boat being more likely to stay at home. On the other hand, differences do emerge in terms of other factors. For example both females and college educated individuals are found to be more likely to take trips in models 2 and 3, whereas these factors are statistically insignificant in model 1. Also, larger households are significantly less likely to take trips according to model 1. Perhaps most importantly, the variance of the trip error component (i.e., σ_τ^2) is substantially larger in model 1 than in models 2 and 3, indicating a greater similarity across (and correlation among) the utilities received from trip options.[19] This is analogous to a greater degree of "nesting" among the trip options in a nested logit setting. As will be seen below, this has implications for estimated welfare effects.

The second stage associated with estimating models 1 through 3 involves regressing the estimated segment level ASC's on segment characteristics, as depicted in equation (2.45). The results are reported in Table 1.8b. The parameter estimates again suggest some consistency across the three specifications. In all three cases, river segments that are canoeable, relatively wide (i.e., with a larger value for $WATERBODY$), and contain a larger number of fish species are more appealing (i.e., have a larger ASC). Not surprisingly, the segments associated with the border rivers (i.e., the Mississippi and Missouri Rivers) are also more popular, as these rivers provide opportunities for activities (such as power boating) not available for most other river segments. Longer river segments are significantly less appealing according to Model 1, whereas they are significantly more appealing according to Model 3. Finally, while the parameter on Iowa's water quality index ($IWQI$) has the expected sign, it is not statistically significant in any of the models. This is not surprising. As suggested above, $IWQI$ is likely to be a poor measure of the perceived water quality along the river segments. First of all, it is measured for only a portion of the river segments and then only at specific monitoring sites. Moreover, given the wide range of activities associated with the trips being reported in the survey (from hiking and bird watching, to swimming and fishing), the $IWQI$ is likely to be salient only for

[19] The correlation coefficient could calculated through the formula, $\sigma_\tau^2/(\sigma_\tau^2 + \pi^2/6)$, where $\pi^2/6$ is the variance of a standard logistic distribution.

a fraction of the reported trips.

At the bottom of Table 1.8b, we provide the predicted welfare implications associated with closure of river segment 71 (the most visited site) and the closure of all 73 river segments. The three models provide significantly different welfare estimates. The model based on aggregate choice probabilities for the river segments (i.e., model 1) yields a substantially higher welfare estimate under both scenarios. The compensating variation associated with closing site 71 is over thirty percent higher in model 1 than in model 2 (which uses the shortest distance proxy for travel cost), and nearly twenty percent high than in model 3. The former result is not surprising, since model 2 is guaranteed to understate the travel cost associated with visiting a river segment and, hence, undervalue the resulting lost trips. The differences are even larger when closing all seventy-three river segments.

Given the limitations associated with the Iowa water quality index ($IWQI$), Table 1.9 provides results from an alternative specification for the Stage 2 model of the ASC's (i.e., equation 2.45), in which $IWQI$ is replaced with $TURBIDITY$. The qualitative findings for most variables are similar to those from Table 1.8b. However, in the case of the aggregate choice probability model, turbidity is found to be a negative and statistically significant factor. On the hand, while fish stocks still have a positive coefficient, they are no longer a statistically significant factor. These results highlight the importance of viewing all three water quality measures ($IWQI$, $TURBIDITY$, and $FISH$) as proxies for water quality in the river segments, rather than direct causal influences in determining the appeal of a given site.

2.6 Concluding Remarks

The task of modeling recreation demand is often complicated by incomplete information regarding specifically where an individual travels to in visiting a geographically large site. While midpoint and shortest distance travel cost measures are often used as proxies for the unobserved travel cost, the resulting parameter estimates are likely to suffer from omitted variables bias. In this paper, we suggest that the problem be viewed instead as one of implicit site aggregation. In general, the probability of any aggregate site being visited is simply the sum of the probabilities that its component sites would be chosen. Using an underlying logit structure,

the existing aggregation literature provides the specific functional form for the aggregated site choice probabilities, as well as an explicit characterization of the omitted variables that arise when using proxy variables for site attributes (including travel cost). We show that, in the context of the RUM model with a full set of alternative specific constants, the appropriate travel cost for an aggregate site is a probability weighted average of the travel cost to the component sites. We also generalize the existing aggregation models to include Normal Error Component Logit Mixture(NECLM) models that have become increasingly popular in the literature, paying particular attention to concerns regarding model identification in these settings.

A Monte Carlo exercise illustrates that the use of travel cost proxies can potentially lead to significant bias in characterizing recreation demand. In particular, while the nearest access point approach provides a relatively good approximation to underlying preferences for a wide range of parameter specifications, use of the midpoint approach to calculating travel cost can lead to significant bias in the travel cost parameter and corresponding welfare calculations. Finally, an application is provided drawing on data from the 2009 Iowa Rivers and Rivers Corridors Survey. We find that the use of either the midpoint or shortest distance travel cost proxies yields a substantially smaller estimate of the welfare impacts of site loss than what is obtained using a model based on aggregated site probabilities.

Bibliography

[2] Ben-Akiva, M., and S. Lerman (1985) "Discrete Choice Analysis: Theory and Application to Travel Demand", Cambridge: The MIT Press.

[2] Egan, K.J., Herriges, J.A., Kling, C.L. and Downing, J.A. (2009) "Valuing water quality as a Function of Water Quality Measures" *American Journal of Agricultural Economics*, Vol. 91, No. 1, pp. 106-123.

[4] Feather, P. (1994), "Sampling and Aggregation Issues in Random Utility Model Estimation," *American Journal of Agricultural Economics*, **76**, pp. 926-33.

[5] Ferguson, M. R., and P. S. Kanaroglou (1998). "Representing the Shape and Orientation of Destinations in Spatial Choice Models," *Geographical Analysis* **30(2)**, pp. 119-137.

[14] Feather, P. and Lupi, F. (1998) "Using partial aggregation to reduce bias in random utility travel cost models" *Water Resources Research* Vol. 34, No. 12, pp. 3595-3603.

[9] Haab, T., and K. E. McConnell (2002), *Valuing Environmental and Natural Resources: The Econometrics of Non-Market Valuation*. Northampton, MA:Edward Elgar.

[7] Haener, M. K., Boxall, P. C., and Adamowicz, W. L., (2004) "Aggregation bias in Recreation site choice models: resolving the resolution problem," *Land Economics*, **80(4)**, pp. 561-574.

[8] Herriges, J. A., and D. J. Phaneuf, (2002) "Inducing Patterns Correlation and Substitution in Repeated Logit Model of Recreation Demand," *American Journal of Agricultural Economics*, **84**(4): 1076-1090

[15] Kaoru Yoshiaki, and V. K. Smith. (1990), " 'Black Mayonnaise' and Marine Recre- ation: Methodological Issues in Valuing a Cleanup." Marine Policy Center, Woods Hole Oceanographic Institution, Woods Hole, MA.

[10] Kaoru, Y., V.K. Smith, and J. Liu (1995), "Using Random Utility Models to Estimate the Value of Estuarine Resources," *Americal Journal of Agricultural Economics*, Vol. 77, No. 2, pp 141-151.

[12] H. Allen Klaiber and Roger H. von Haefen. (2008) "Incorporating Random Coefficients and Alternative Specific Constants into Discrete Choice Models: Implications for InSample Fit and Welfare Estimates," Working Paper

[12] Herriges, J., and C. Kling. (1997) "The Performance of Nested Logit Models When Welfare Estimation is the Goal," The American Journal of Agricultural Economics, 79: 792-802.

[14] Kling, C., and C. Thomson (1996), "The Implications of Model Specification for Welfare Estimation in Nested Logit Models," *American Journal of Agricultural Economics*, **78**, pp. 103-14.

[13] Kurkalova, L.A. and S.S. Rabotyagov (2006), "Estimation of a Binary Choice Model with Grouped Choice Data," *Economics Letters* **90**(2): 170-175.

[15] Morey, Robert D. Rowe and Michael Watson (1993), "A Repeated Nested-Logit Model of Atlantic Salmon Fishing", *American Journal of Agricultrual Economics*, **75(3)**, pp. 578-592.

[1] Murdock, Jennifer. (2006) "Handling unobserved site characteristics in random utility models of recreation demand", *Journal of Environmental Economics and Management*,Vol. 51, No.1 pp. 1-25.

[23] Parsons, G., and M. Needelman (1992), "Site Aggregation in a Random Utility Model of Recreation," *Land Economics*, **68**: 418-33.

[22] Parson, G., Plantinga, A.J. and Boyle, K.J, (2000) "Narrow Choice Sets in a Random Utility Model of Recreation Demand", *Land Economics*, Vol. 76, No. 1, pp. 86 - 99.

[19] Parsons,G.R., and Hauber, A.B. (1998), "Spatial Boundaries and Choice Set Definition in a Random Utility Model of Recreation Demand" *Land Economics*, **74**:3248.

[24] Phaneuf, D. J., and J. A. Herriges (2000), "Choice Set Definition Issues in a Kuhn-Tucker Model of Recreation Demand," *Marine Resource Economics*, **14**, pp. 343-55.

[21] Phaneuf, D. J., and J. A. Herriges (2002), "Inducing Patterns Correlation and Substitution in Repeated Logit Model of Recreation Demand," *American Journal of Agricultural Economics*, **14**, No.4, pp. 1076-1090.

[25] Phaneuf, D. J,. and Smith, V. K. (2005) " Recreation Demand Models" *Handbook of Environmental Economics* Edited by K. G. Mäler and J. R. Vincent, Vol. 2, pp. 672 - 751.

[23] Smith, V. K., W. H. Desvousges and M. P. McGivney (1983), "The Opportunity Cost of Travel Time in Recreation Demand Models," *Land Economics*, Vol. 59, No. 3 , 259-278.

[27] Train, K.,(2003), "Discrete Choice Methods with Simulation," Massachusetts: Cambridge University Press

[25] Walker, J.L., M. Ben-Akiva, and D. Bolduc (2007) "Identification of Parameters in Normal Error Component Logit-Mixture (NECLM) Models" *Journal of Applied Econometrics* Vol. 22 pp 1095-1125.

Table 1.1 Description of Monte Carlo Designs

Experiment Design	Description	Number of Variations
Price responsiveness (i.e., value of β)	$\beta = \{-0.01, -0.05, -0.09\}$ w/o Water $\beta = \{-0.02, -0.05, -0.1\}$ w/ Water	3
Number of river segments	s=5,10,20	3
Population and river characteristics	Base (B) Population Center (P) Nonlinear (Kinked) Rivers (K) Combined Population centers + Kinked Rivers (C)	4
Water Quality	Included (w/ Water) Not included (w/o Water)	2
	Total	72

Table 1.2a Mean Absolute Percentage Error in Estimated β (w/o Water Quality)

Pop./River Config.	Number of Segments	Agg. Choice Prob.			Midpoint Proxy			Shortest Distance Proxy		
		β = -0.01	-0.05	-0.10	-0.01	-0.05	-0.10	-0.01	-0.05	-0.10
B	5	0.1	0.1	0.1	1.2	5.9	8.5	0.3	0.1	0.0
	10	0.1	0.1	0.1	1.2	6.5	13.7	1.2	1.0	1.0
	20	0.1	0.1	0.0	1.4	6.8	15.8	0.2	1.5	1.6
K	5	0.2	0.1	0.1	1.8	5.4	13	0.3	0.3	0.3
	10	0.3	0.1	0.1	1.6	8.2	17.9	0.6	0.5	1.8
	20	0.1	0.1	0.2	1.3	8.2	17.3	0.5	0.7	0.6
P	5	0.1	0.1	0.1	1.4	7.0	13.6	0.7	4.0	5.1
	10	0.2	0.1	0.2	1.9	7.4	14.9	0.5	2.4	4.1
	20	0.1	0.1	0.1	1.2	9.0	19.6	1.5	3.0	3.5
C	5	0.1	0.1	0.1	0.4	6.1	8.6	0.6	2.4	2.7
	10	0.1	0.1	0.2	0.9	8.1	15.4	1.0	0.5	2.6
	20	0.1	0.1	0.1	1.5	9.4	18.1	0.6	1.0	1.2

Table 1.2b Mean Absolute Percentage Error in Estimated CV_1 (w/o Water Quality)

Pop./River Config.	Number of Segments	Agg. Choice Prob.			Midpoint Proxy			Shortest Distance Proxy		
		β = -0.01	-0.05	-0.10	-0.01	-0.05	-0.10	-0.01	-0.05	-0.10
B	5	0.1	0.1	0.1	1.3	6.5	10.9	0.3	0.3	1.0
	10	0.1	0.1	0.1	1.5	7.2	16.1	1.5	1.1	1.0
	20	0.2	0.1	0.1	1.1	7.9	20.8	0.1	1.7	2.5
K	5	0.2	0.0	0.1	1.8	7.5	12.4	0.3	0.8	0.5
	10	0.1	0.3	0.4	1.4	9.1	16.9	0.4	0.1	1.1
	20	0.1	0.1	0.2	1.3	8.2	17.3	0.5	0.7	0.6
P	5	0.1	0.2	0.1	1.6	8.3	15.5	0.9	4.7	5.4
	10	0.2	0.2	0.3	2.0	8.5	17.7	0.5	3.0	5.0
	20	0.4	0.1	0.2	0.9	10.5	25.4	1.2	3.3	4.4
C	5	0.5	0.2	0.1	0.5	4.7	7.0	1.0	2.7	2.7
	10	0.2	0.1	0.1	1.0	10.0	19.6	0.9	0.6	3.2
	20	0.3	0.1	0.3	1.3	10.8	23.9	0.4	0.7	1.4

Table 1.3a Mean Absolute Percentage Error in Estimated β (w/ Water Quality)

Pop./River Config.	Number of Segments	Agg. Choice Prob.			Midpoint Proxy			Shortest Dist. Proxy ver.1			Shortest Dist. Proxy ver.2		
		β = -0.01	-0.05	-0.10	-0.01	-0.05	-0.10	-0.01	-0.05	-0.10	-0.01	-0.05	-0.10
B	5	0.1	0.1	0.1	1.6	4.2	9.3	1.0	1.7	1.8	0.7	0.4	0.5
	10	0.0	0.1	0.1	1.9	5.4	10.7	0.1	2.6	4.7	0.6	2.1	4.6
	20	0.1	0.1	0.1	2.5	6.8	15.1	1.3	2.4	3.9	0.9	1.8	3.2
K	5	0.1	0.1	0.1	1.3	4.7	13.3	0.6	0.6	1.9	0.4	0.3	1.8
	10	0.2	0.1	0.1	2.6	6.2	14.7	0.2	0.2	1.5	0.2	0.1	1.5
	20	0.1	0.1	0.1	3.1	9.0	18.9	0.2	0.5	1.2	0.2	0.3	0.9
P	5	0.2	0.2	0.1	1.7	6.0	13	0.3	0.8	3.1	1.8	0.2	3.8
	10	0.1	0.1	0.1	2.2	6.4	16.8	0.7	0.6	3.8	1.1	0.3	3.1
	20	0.1	0.1	0.1	2.3	8.0	19.4	1.9	3.3	4.8	1.4	2.5	4.0
C	5	0.0	0.3	0.2	2.5	7.0	14.1	0.1	2.5	3.5	0.3	1.8	3.6
	10	0.2	0.1	0.1	3.7	5.5	10.5	0.3	0.5	1.7	0.1	0.7	2.0
	20	0.1	0.1	0.1	3.9	9.3	22.6	0.4	0.8	2.3	0.4	0.8	2.3

Table 1.3b Mean Absolute Percentage Error in Estimated β_w (w/ Water Quality)

Pop./River Config.	Number of Segments	Agg. Choice Prob.			Midpoint Proxy			Shortest Dist. Proxy ver.1			Shortest Dist. Proxy ver.2		
		β = -0.01	-0.05	-0.10	-0.01	-0.05	-0.10	-0.01	-0.05	-0.10	-0.01	-0.05	-0.10
B	5	5.4	3.2	1	-	-	-	66	23.8	28.5	-	-	-
	10	0.2	0.5	0.1	-	-	-	90.4	81	82	-	-	-
	20	4	0.6	0.1	-	-	-	142	200.7	233.1	-	-	-
K	5	0.3	0.1	0.6	-	-	-	84.4	52.8	24.2	-	-	-
	10	1.2	1.1	0.6	-	-	-	98.7	91.3	80.8	-	-	-
	20	6.2	0.2	0.8	-	-	-	123.2	136.7	157.3	-	-	-
P	5	9.8	0.2	0.2	-	-	-	62.5	28.8	5.2	-	-	-
	10	0.8	0.7	0.2	-	-	-	89.5	63.6	38.2	-	-	-
	20	6.8	1.3	0.7	-	-	-	151.4	204.8	220.3	-	-	-
C	5	2.9	0.2	1	-	-	-	75.8	48.5	34.7	-	-	-
	10	3.2	0.3	0.2	-	-	-	95.2	82.4	76.1	-	-	-
	20	1.5	0.2	0.4	-	-	-	109.6	119.1	126.3	-	-	-

Table 1.3c Mean Absolute Percentage Error in Estimated CV_1 (w/ Water Quality)

Pop./River Config.	Number of Segments	Agg. Choice Prob.			Midpoint Proxy			Shortest Dist. Proxy ver.1			Shortest Dist. Proxy ver.2		
		β = -0.01	-0.05	-0.10	-0.01	-0.05	-0.10	-0.01	-0.05	-0.10	-0.01	-0.05	-0.10
B	5	0.2	0.1	0.1	1.9	6.4	9.2	1	0.9	0.2	0.5	1.8	0.7
	10	0.2	0.2	0.1	2.1	4.9	6.5	0.1	3.3	7.1	0.5	2.4	6.5
	20	0.4	0.7	0.4	3	7.7	16.9	1.7	2.9	3.4	1.2	2.5	2.5
K	5	0.2	0.2	0.1	2.8	2.1	5.1	0.7	0.4	0.7	0.2	0.6	1.8
	10	0.1	0.2	0.1	2.8	6.6	21.7	0.3	0.3	1.9	0.3	0.2	2.3
	20	0.1	0.3	0.3	3	9.5	24.3	0.2	0.2	1.5	0.1	0.5	0.7
P	5	0.3	0.3	0.2	2.1	7.1	14.2	0.3	1	3.1	2.2	0.5	2.5
	10	0.4	0.2	0.2	2.7	8.1	21.2	0.5	0.8	3.7	0.9	0.2	2.8
	20	0.1	0.2	0.2	2.4	9.9	26.5	1.9	3.8	6.8	1.3	2.5	4.3
C	5	0.1	0.2	0.1	2.2	7.8	12.9	0.1	2.6	2.8	0.2	2.1	3.1
	10	0.3	0.1	0.3	3.8	4.8	5.8	0.4	1.2	3.2	0.2	1.3	3.4
	20	0.1	0.3	0.1	4.1	10.7	29.3	0.3	0.8	2.5	0.4	0.8	2.5

Table 1.4 Summary Statistics Demographic Characteristics (N=4137)

Variable	Description	Mean	Std.Dev
Age1	Dummy Variable for Age Group: 18-25	0.01	0.09
Age2	Dummy Variable for Age Group: 26-34)	0.06	0.24
Age3	Dummy Variable for Age Group: 35-49)	0.21	0.41
Age4	Dummy Variable for Age Group: 50-59)	0.25	0.43
Age5	Dummy Variable for Age Group: 60-75)	0.31	0.46
Age6	Dummy Variable for Age Group: 76-)	0.15	0.36
Female	Dummy Variable for Females	0.70	0.46
College	Dummy Variable for College Degree	0.69	0.46
Size	Number of Adults	1.88	0.65
Kids	Number of Children	0.55	1.00
Employed	Dummy Variable for Employed	0.59	0.49
Student	Dummy Variable for Students	0.01	0.07
Retired	Dummy Variable for Retirees	0.36	0.48
Boat	Dummy Variable for owning a boat	0.23	0.42

Table 1.5 Trip Summary Statistics (N=4137)

Segment #	% Visitors	Total Trips	Segment #	% Visitors	Total Trips
1	1.93%	369	38	0.58%	116
2	1.21%	253	39	0.75%	68
3	0.94%	193	40	1.35%	293
4	0.29%	76	41	1.04%	232
5	1.57%	290	42	0.56%	113
6	1.16%	143	43	1.31%	303
7	0.31%	52	44	0.19%	31
8	0.12%	23	45	2.05%	533
9	0.29%	59	46	3.87%	800
10	0.34%	61	47	3.63%	737
11	0.56%	105	48	0.48%	76
12	0.22%	37	49	1.60%	462
13	0.75%	183	50	2.20%	381
14	0.65%	80	51	0.58%	116
15	0.58%	113	52	3.75%	1071
16	0.60%	154	53	3.92%	877
17	0.17%	25	54	3.05%	653
18	0.41%	60	55	1.76%	308
19	0.34%	59	56	3.60%	742
20	0.99%	171	57	0.56%	85
21	1.09%	314	58	3.41%	659
22	1.28%	194	59	0.73%	84
23	5.25%	979	60	1.55%	248
24	7.23%	1513	61	1.47%	142
25	3.02%	675	62	1.47%	232
26	1.69%	399	63	1.33%	111
27	0.46%	84	64	2.51%	314
28	0.60%	123	65	1.43%	208
29	0.77%	175	66	3.87%	884
30	0.58%	104	67	1.33%	245
31	0.24%	58	68	4.33%	676
32	0.58%	82	69	7.03%	1246
33	1.28%	193	70	6.24%	1578
34	2.30%	468	71	5.73%	1591
35	0.92%	234	72	2.34%	483
36	1.16%	272	73	2.34%	740
37	0.24%	20	Overall	47.76%	6.24

Table 1.6 Summary Statistics of River Attributes

Variable	Mean	Std.Dev	Min	Max
C_{isj}	162.85	79.15	0.08	485.14
73 river segments				
LENGTH	83.2	33.30	26.9	161.8
CANOE	62.82	27.13	0.00	99.00
OUTCROPPING	18.77	31.66	0.00	141.00
WATERBODY	0.22	0.30	0.00	1.00
WETLAND	0.01	0.01	0.00	0.05
FOREST	0.24	0.19	0.00	0.71
GRASS	0.11	0.09	0	0.40
CROP	0.39	0.29	0	0.92
DEVELOPED	0.02	0.04	0	0.24
IWQI	31.55	23.00	0	75
MIWQI	0.32	0.47	0	1
FISH	30.29	18.52	0	71
MFISH	0.12	0.33	0	1

Table 1.7 Ranking of River Segments

	Rank	Visitation	Model 1 Agg. Prob.	Model 2 Shortest Distance	Model 3 Midpoint Proxy
Top 10	1	71	68	68	70
	2	70	69	69	69
	3	24	73	70	68
	4	69	70	73	71
	5	52	71	71	73
	6	23	2	64	66
	7	66	65	52	52
	8	53	25	65	72
	9	46	64	66	65
	10	56	72	25	64
Bottom 10	64	18	29	31	31
	65	9	42	57	10
	66	19	7	19	48
	67	31	19	7	39
	68	7	12	30	30
	69	12	37	12	12
	70	44	17	17	17
	71	17	8	8	8
	72	8	44	44	44
	73	37	30	37	37

Note: 1 A list of river id and their names could be found in the survey sample.

Table 1.8a Estimation Result of Nested Logit Specifications (Stage 1)

Variable	Agg. Choice Prob.		Shortest Dist. Proxy		Midpoint Proxy	
	Est.	Std.Dev	Est.	Std.Dev	Est.	Std.Dev
Travel Cost Variable						
TC	-0.035***	0.001	-0.034***	0.001	-0.033***	0.001
Demographics						
AGE 2	2.526***	0.365	2.827***	0.598	4.534***	0.449
AGE 3	2.904***	0.300	2.581***	0.576	4.460***	0.425
AGE 4	3.795***	0.288	3.077***	0.581	4.897***	0.426
AGE 5	4.679***	0.332	3.674***	0.589	5.400***	0.436
AGE 6	6.110***	0.422	4.712***	0.615	6.672***	0.458
FEMALE	0.284	0.157	−0.561***	0.116	−0.495***	0.101
COLLEGE	−0.057	0.146	−0.514***	0.110	−0.625***	0.100
SIZE	0.589***	0.028	-0.055	0.090	0.050	0.078
KIDS	0.347***	0.080	-0.014	0.070	-0.057	0.063
EMPLOYED	6.941***	0.262	0.453	0.290	0.873***	0.275
STUDENT	0.880*	0.446	2.478***	0.835	3.902***	0.754
RETIRED	7.218***	0.337	0.439	0.321	0.745***	0.303
BOAT	−1.899***	0.179	−1.913***	0.144	−1.840***	0.136
Nest Variable						
σ_τ	3.758***	0.105	2.920***	0.054	3.058***	0.051

1. *,**,*** represent significant levels at 10%, 5% and 1% respectively

Table 1.8b Estimation Result of Nested Logit Specifications (Stage 2)

Variable	Agg. Choice Prob.		Shortest Dist. Proxy		Midpoint Proxy	
	Est.	Std.Dev	Est.	Std.Dev	Est.	Std.Dev
Site Attributes						
LENGTH	−0.008**	0.004	-0.001	0.004	0.009**	0.004
CANOE	0.010**	0.005	0.009*	0.005	0.008*	0.005
OUTCROPPING	0.003	0.004	0.004	0.004	0.001	0.004
WATERBODY	1.522**	0.782	1.688**	0.827	1.857**	0.746
WETLAND	13.338	12.324	13.887	12.643	19.386	11.902
GRASSLAND	1.257	1.572	1.619	1.549	0.974	1.387
CROPLAND	-0.210	0.729	-0.041	0.740	-0.024	0.689
DEVELOPED	-0.902	3.815	-0.289	3.877	-0.157	3.795
IWQI	0.006	0.014	0.007	0.013	0.005	0.013
MIWQI	0.078	0.677	0.180	0.641	0.231	0.677
FISH	0.014*	0.007	0.013*	0.008	0.012*	0.007
MFISH	0.578	0.360	0.562	0.362	0.356	0.354
BORDER	1.362***	0.479	1.253***	0.483	1.304***	0.452
CONSTANT	3.301***	1.130	−5.433***	1.218	−3.277***	1.119
Welfare Calculation ($/year/household)						
Loss of River 71	17.03		12.92		14.30	
Loss of All Rivers	668.94		361.37		420.23	

1.Bootstrap is used to get standard deviations for the second stage estimation of river attributes.
2. *,**,*** represent significant levels at 10%, 5% and 1% respectively
3. The CV figures are coming from 1000 numerical simulations using equation 2.43

Table 1.9 Estimation Results Using Turbidity as Water Quality Proxy

Variable	Agg. Choice Prob.		Shortest Dist. Proxy		Midpoint Proxy	
	Est.	Std.Dev	Est.	Std.Dev	Est.	Std.Dev
Site Attributes						
LENGTH	-0.008**	0.004	-0.001	0.004	0.009***	0.003
CANOE	0.010**	0.004	0.009*	0.005	0.009**	0.004
OUTCROP	0.002	0.004	0.003	0.004	0.001	0.004
WATERBODY	1.510**	0.693	1.685**	0.810	1.817**	0.778
WETLAND	11.734	11.968	12.319	12.060	17.085	10.669
GRASSLAND	1.125	1.402	1.541	1.407	0.873	1.314
CROPLAND	-0.105	0.644	0.013	0.712	0.097	0.690
DEVELOPED	-1.543	3.795	-0.872	3.635	-0.800	3.637
TURBIDITY	-0.004*	0.002	-0.003	0.002	-0.004	0.003
MTURBIDTY	-0.553*	0.321	-0.444	0.325	-0.398	0.324
FISH	0.010	0.007	0.010	0.008	0.008	0.007
MFISH	0.622*	0.343	0.590	0.373	0.408	0.326
BORDER	1.457***	0.424	1.324***	0.446	1.408***	0.447
CONSTANT	3.876***	0.666	-4.827***	0.762	-2.736***	0.739
Observations	73		73		73	
Adjusted R^2	0.653		0.600		0.679	

1.Bootstrap is used to get standard deviations for the second stage estimation of river attributes.
2. *,**,*** represent significant levels at 10%, 5% and 1% respectively
3. The CV figures are coming from 1000 numerical simulations using equation 2.43

CHAPTER 3. Modeling Recreation with Partial Trip Information

3.1 Introduction

Recreation demand (or travel cost) models provide one of the primary tools for valuing environmental amenities, inferring value by observing the full costs incurred by the individual or household in reaching the site or sites of interest (see, e.g., (9)). Analysts, however, often have information on the visitation patterns to only a subset of recreational sites that are available to the individual. By focusing on only a subset of the available choice alternatives, the un-modeled sites are implicitly imbedded into the outside option (i.e., the stay-at-home alternative). Doing so ignores the substitution possibilities to these alternatives, exposing the resulting preference parameters, and subsequent welfare estimates open to possible bias.

Relying on the site aggregation literature, we explore the possibility of consistently estimating the preference parameters under the random utility maximization framework. To employ any of the procedures described below, the recreational data used should at least include information on the characteristics for all the sites, specifically both for lakes and for rivers in this paper. With the extra information on one set of sites for which we do not have households' detailed visitation information, we can separate the *stay-at-home option* into two parts, one is the standard "stay-at-home option" and the other is the aggregated choice for sites without detailed trip information. Thus the probability of observing a household choosing *not* to visit the set of sites for which we have the full information will be the sum of the probability of staying at home and the probability of visiting any of the sites in the other set. By doing so, we use aggregation techniques in the literature (see, e.g., (4),(23) and (10)) to estimate the model even when we have only partial information. To mimic the possible nested structure among sites, the aggregation is conducted under the framework of the normal error-component

mixed logit models (NECLM).[1] Identification of parameters in the model is a key concern when we apply the aggregation technique. Walker *et al.* (2) gives the useful guidance to judge the identification of the NECLMs. Due to the complication brought about by the aggregation over several subsets of options, we are not able to formally prove the identification of the aggregation model. However, a series of Monte Carlo simulations illustrate the subset of the models that can be practically estimated.

We conduct two categories of simulations. In the first set of simulations, we specify the conditional utility to be a linear function of site attributes and the travel cost. It fits into situations where focused recreational sites are essentially the same, e.g. a recreational survey about people's visitation to local lakes, but only detailed trip information about famous lakes is explicitly asked. In the second set of simulations, we assume the conditional utility depends on the travel cost and alternative (site) specific constants (ASC). The key difference between these two settings is the way those site related values, such as the amenities brought about by better water quality, enter into conditional utilities. Technically, including ASCs in the maximum likelihood estimation will absorb the effects of other site attributes and the coefficients of these site attributes can only be recovered in the second stage regression. If there were concerns about omitted variables, the possible bias could be controlled with ASCs in the utility function. With the estimation strategy proposed in Murdock (1), the inclusion of ASCs in recreation models has become the norm in the literature. In addition to these technical concerns, there are some practical reasons for the second set of simulations. The recreation sites considered in the application to the Iowa Lake and River project have significant differences in terms of site attributes and data availability. For example, the site size, one of commonly used site attributes in the literature, is not comparable for lakes and rivers. The area of water body may be a good metric for lake sites, while the total length seems more appropriate for river segments. To avoid the difficulty in the construction of site attributes, we will include ASCs in the application work and recover the values of attributes in a second-stage analysis. To get a sense of applicability and performance of the aggregation model, it is useful to have some

[1]The mixed logit model could mimic the substitution pattern among alternatives represented by the nest structure by choosing appropriate group dummy variables (See, e.g., (27), (24)).

counterpart simulations.[2] For each simulated data set, we apply three modeling techniques to fit the data: the full information model in which all visitation information is available to researchers, the aggregation model in which we assume the visitation information to a subset of sites is only aggregately observed by researchers and the partial model in which we assume the unobserved sites are pooled with the outside *stay-at-home* option. The performance of each modeling techniques is evaluated along three dimensions: the recovery of preference parameters, the recovery of error structures and the welfare measure of site loss.

The simulation results confirm our speculation that the aggregation model can practically recover the preference and error structure parameters. In both sets of simulations, the mean absolute percentage error of preference parameters (MAPE), a metric used to evaluation the quality of estimation result, is almost negligible, while the partial model generally produces biased estimation of preference parameters. In terms of error structures, the aggregation model performs quite well in terms of estimation in a subset of scenarios.[3] Specifically, if the underlying structure is simple, the bias measure by MAPE is small. Once the error structure becomes more complex, the bias increases substantially.[4] The welfare measure of site loss is also a good metric to assess the performance of different recreational models. In the simulation, we also calculate the compensating variation (CV) of losing an individual site.[5] The comparison in the welfare measure produces mixed results. In the first set of simulation without ASCs, the bias in CV is significantly larger and in line with the finding in the literature. Once ASCs are added into the model, the bias from the partial model almost disappears and the scale of bias is of the same order of that from the aggregation model. The CV bias from the aggregation model is negligible in both sets of simulations regardless as to whether ASCs are included or not.

The aggregation technique is also applied to the unique data set from 2009 Iowa Lake and River Project. The Iowa River Project is a cross-sectional study of the 2009 visitation patterns of state residents to the 73 major river segments in the Iowa. The 2009 Iowa Lake Project is the

[2]The details about the simulation setup are given in the simulation section.

[3]Since the partial model implicitly rearranges the underlying error structure, we do not explicitly compare the performance of the partial model in this regard.

[4]The simple structure is referred to the case 1 and 2 in Figure 1, and the structures represented by case 3 and 4 in Figure 1 are more complex.

[5]This individual site is one of sites which have explicit visitation.

continuation of a four-years (2002-2005) Iowa Lake Project of visitation information of state residents to the 100+ instate lakes. Questionnaire including recreational visitation questions and social economic questions were sent to 10,000 randomly selected Iowa households via mail in 2010. Among the 10,000 households, there were 2,500 overlapping households who receives both survey questionnaires. In total, the survey yielded around 5,000 completed surveys in both projects. The randomness and the overlapping design features permit us to explicitly model the non-overlapping samples with the aggregation technique and insure identification given the existence of the overlapping samples.

The results show that the preference parameters estimated in different models agree in direction but differ in magnitudes and significance levels. Judging from the welfare evaluation results, the figures produced from the partial models are moderately different from the results in the aggregation models.

The remained of this paper is organized as follows: Section 2 provides a brief literature review. The modeling technique and identification issue is discussed in section 3. Section 4 contains the detailed data generation process and the results of Monte Carlo simulation. Data description and estimation results of the application is given in section 5. Section 6 concludes the paper.

3.2 Related Literature

3.2.1 RUM model in recreational literature

McFadden (16) established the random utility maximization (RUM) models to study the discrete choices made by individuals from among a set of alternatives. In RUM framework, individuals are assumed to have full information about the factors that affect their utilities and make decisions based on that. Researchers do not have the full information about individuals preference and treat the unobservable parts as random variables following an assumed distribution. Combined with the observed choices, researchers can derive the probabilities associated with these choices and estimate preference parameters and conduct statistical inferences thereafter. Two popular models are logit models if the distribution of the unobserved factors is

assumed to belong to the family of extreme value random variables and probit models if the distribution belongs to normal distributions (See, e.g., (27)).

3.2.1.1 Multinomial logit and IIA

The central building block in RUM models is the conditional indirect utilities, U_{ij}, that individual i receives from choosing alternative j (sites in recreation literature). U_{ij} is usually assumed to be a sum of two additive parts, V_{ij} and ϵ_{ij}. V_{ij} is the deterministic part of U_{ij} and usually is a linear function of factors (X_{ij}) observed by researchers, including site attributes, demographic information about individuals and the interaction of individual variables and site variables. ϵ_{ij} is known to individuals, but unobserved by researcher. The individual i will choose alternative j denoted by $y_{ij} = 1$ if alternative j gives him the highest utility. Mathematically,

$$y_{ij} = \begin{cases} 1 & \text{if } V_{ij} + \epsilon_{ij} \geq V_{ik} + \epsilon_{ik} \text{ for } k = 1, \ldots, J \\ 0 & \text{otherwise} \end{cases}$$

If ϵ_{ij} is independent and identical (IID) distributed as type one extreme value random variables with scale parameter μ, the model is a multinomial logit model.[6] The probability of choosing alternative j has the close form (See, e.g., (16)).

$$P_{ij} = \frac{exp(X_{ij}\beta/\mu)}{\sum_{k=1}^{J} exp(X_{ik}\beta/\mu)} \tag{3.1}$$

A well known limitation of the multinomial logit model is that it imposes the independent of irrelevant alternative (IIA) property with the probability ratio of any two alternatives is independent of other alternatives. IIA property limits the possible substitution pattern among alternatives and is often rejected as a restriction on preferences in practice. For this reason, it is more common in the literature to employ more flexible distributional assumptions, such as nested logit or mixed logit (See, e.g., (27), (16)).

3.2.1.2 Nested Logit and Mixed logit

Based on the general extreme value distributions, nested logit models are proposed in the literature to allow for more flexible substitution pattern among alternatives (See, e.g., (28)).

[6] μ is usually normalized to 1 to achieve the point identification of model parameters.

Alternatives are grouped into nests according to some common attributes, which allows for correlation (and great substitutability) among alternatives in the same nest while still yielding relatively simple choice probability formulas (See, e.g., (14), (24) and (12) among others).

Although introduction of nested logit allows for more flexible substitution patterns, nested logit suffers some shortcomings. One of them is that the preference parameters (β) and welfare estimate can be very sensitive to the assumed nest structure (See, e.g., (14)). Although nested logit models allow the induced substitution pattern to relax the IIA property, the ability to allow heterogenous preference on site attributes is limited.

McFadden and Train ((20)) develops the mixed logit model by introducing another layer of heterogenous and unobservable uncertainty on the preference parameters in the standard multinomial logit model. The preference parameter vector β is assumed to be a random vector from researchers' perspective. Mathematically,

$$\beta \sim f(\beta|\theta) \text{ where } \theta \text{ is the set of distribution parameters to be estimated.}$$

The conditional choice probability of choosing alternative j is then given by

$$P(y_{ij} = 1|\beta = \beta_r) = \frac{exp(X_{ij}\beta_r)}{\sum_{k=1}^{J} exp(X_{ij}\beta_r)}$$

and the unconditional probability then is

$$P(y_{ij} = 1|\theta) = \int P(y_{ij} = 1|\beta_r) f(\beta_r|\theta) d\beta.$$

Mixed logit model can mimic any random utility models (See, e.g., (20),(27)). As the computation technologies advances and cost dramatically decreases, mixed logit model has become more popular and a standard framework to study the discrete choice models.

Though the assumptions used in the nested logit model is different from the ones used in the mixed logit model, the substitution pattern captured in a nested structure can be mimicked by an error component mixed logit model by introducing a dummy to alternatives in the same nest and imposing a random coefficient for this dummy (See, e.g., (27), (24)). The advantages of using a mixed logit model instead of analog nested logit model based on certain general extreme value (GEV) structure are two fold. First, statistical test on the possible correlation among

alternatives is much easier in a mixed logit model than in nested models. In a mixed logit model, identifying a possible nest among some alternatives can be achieved by simply adding a dummy variable for those alternatives and testing whether the corresponding coefficient is significantly different from zero. While in nested models, we must estimate several competing nested structures and decide which one is the best one based on certain criteria (*e.g.,* likelihood dominance criteria suggested by Pollak and Wales (26)). The limitation of this approach is the number of models need to be estimated increases as the number of possible structure supposed. Another advantage is that mixed logit model is easy to be extended to incorporate more flexible substitution patterns across alternatives and choice occasions (such as models with overlapping nests).

A problem with using error component mixed logit model to replace a GEV based nested model is that usually there is no closed form for unconditional probabilities and the identification issue may become more complicated than in nested models. The identification in RUM models is different from the ones caused by omitted variables or endogenous variables in OLS. In RUM models, the preference will not change as we scale the utility levels of all the alternatives. There are essentially infinite vectors of parameters could represent the same preference, thus we usually normalize some parameters in the distribution of error term ϵ to a certain value to pin down the unique combination of parameters, like assuming a standard type I extreme value distribution for logit models. Although there are numerous applications of mixed logit models in the literature, only Walker *et al.* (2) to our knowledge proposes conditions to check the identification issue in a subset of mixed logit models called normal error component logit mixture. Three conditions are proposed in the paper to check the identification issue before estimation.[7] We will follow the guidelines in that paper to check identification of our proposed models.

[7]Checking the singularity of the hessian matrix of the log-likelihood functions after estimation is an empirical method to check the identification. However the mixed logit model are usually estimated with numerical simulations which makes this judgement less reliable. See (3)

3.2.2 Aggregation models

Using aggregate probability model to conquer the partial information issue discussed in this paper is closely related the aggregation problem addressed elsewhere in the literature. Historically, aggregation was used in the literature to alleviate the computation burden when researchers were faced with the problem of modeling model a large number of alternatives in the choice set. Consider the standard RUM model in which the utility that individual i receives from visiting site j is given by

$$U_{ij} = V_{ij} + \eta_{ij}$$

where η_{ij} is a *IID* type one extreme value random variable with scale parameter of μ. Grouping together a subset of sites as group a, say sites $j = 1, \ldots, J_a$, the utility for individual i to choose this group will be

$$U_{ia} = \mu ln[\sum_{j=1}^{J_a} exp(\frac{V_{ij}}{\mu})] + \eta_{ia}$$

where η_{ia} is distributed as a type one extreme value random variable with scale μ. The utility can then be written as

$$\begin{aligned} U_{ia} &= V_{ia} + \eta_{ia} \\ \text{where } V_{ia} &= \bar{V}_{ia} + \underbrace{\mu ln[\frac{1}{J_a} exp(\frac{(V_{ij} - \bar{V}_{ia})}{\mu})]}_{\text{heterogeneity term}} + \underbrace{\mu ln J_a}_{\text{size term}} + \eta_{ia} \end{aligned}$$

where $\bar{V}_{ia}$ is the average utility of alternatives in the aggregate group a. A RUM model can be used with the redefined aggregate site a.

Ben-Akiva and Lerman (2) and McFadden (17) both identified that there will be bias in the aggregation model when the heterogeneous term and size term are not included. The size term can be easily incorporated into the model, while the heterogeneity term is more difficult to control for because it contains information about unknown preference parameters in V_{ij} and $\bar{V}_{ia}$. In addition, the heterogeneity term is usually nonlinear in parameters, the possible direction and size of the bias cannot be determined a priori. There have been a number of studies in the literature examining the possible ways to mitigate the bias or evaluate the bias caused by

different aggregating schemes for specific data sets. The general conclusion is that aggregation biases increases as heterogeneity within aggregate groups increases and tends to overstate the welfare changes. Some form of heterogeneity control is suggested to be used whenever is possible (See, e.g., (25)).

Aggregation is common in random utility models (RUM) of recreation in the literature for the purpose of saving computational resources. Kaoru and Smith (15) were the first one to analyze the effects of aggregation on preference parameter estimation and welfare measure. Their pioneering work suggested the aggregation model performed relatively well in capturing recreation behavior, though it was not promising in welfare calculation in an experiment with mild aggregation, from 35 sites to 23 sites and 11 sites. Parsons and Needelman (23) conducted more experiments in a large scale of aggregation with a fishing recreation data set in Wisconsin. In this paper, the authors identify two sources of bias in the aggregation model, the degree of aggregation measured by the number of elementary sites in one aggregate site and the heterogeneity of elementary sites in the aggregation. Their results partly confirm the finding in Kaoru and Smith (15) that some level of aggregation still did reasonably well in modeling behavior, but extensive aggregation leads to inconsistent and unexplainable recreation behavior, as well as the wrong sign of some coefficients of important variables in the econometric model. In terms of welfare measure, even mild aggregation results in significant bias.

Feather and Lupi (14) proposes the partial aggregation approach in which the popular sites and policy-related sites are treated as individual sites, while the remaining sites are aggregated to some level. With an application to a data set on sport fishing at lakes in Minnesota, the authors find their partial aggregation model outperform the full aggregation models and the difference on preference parameter estimation between the partially aggregated model and the full information model became smaller when the degree of aggregation decreased. Parson, Plantinga and Boyle (22) suggests a similar partial aggregation approach by treating policy interested sites and their close substitutes as individual sites and aggregating other sites to some degree. Our aggregation model is similar to these partial aggregation models in the sense that some sites are modeled individually and other sites are modeled as a group.

With improvements in computing technology, Ferguson and Kanaroglou (5) suggests to esti-

mate the full version of the aggregation model explicitly including heterogenous terms and size terms, which are usually estimated without the heterogeneity term in a simple nest structure. Haener *et al.* (7) follows the suggestion and applies the method to model hunting behavior in Canada and suggests to use area of the aggregate hunting zone to control the heterogeneity in order to alleviate aggregation bias. Similar to these full version aggregation model, our aggregate (probability) models implicitly includes heterogenous terms and size terms, and thus will avoid the bias whenever the aggregation model is effective in the sense that it is identified with the standard normalization.

The hypothesized fundamental models on which aggregation work are based in these papers are either multinomial logit model or nested model. To our knowledge, aggregation models have not been explored in the normal error component framework. In this paper, we will propose an aggregate (probability) model to incorporate data sets with partial information to model the individuals' recreational choices whenever possible and practical.

3.3 Model Setup and Identification Issue

Walker *et al.* (2) points out the identification issue in the prevailing mixed logit models. We follow the conditions provided in that paper to discuss the identification issues that arise in our error-component mixed logit models.

We model individuals's recreational sites choice problem in the RUM framework. Suppose the utility individual i receives from visiting recreational site j is a linear function of explanatory variables describing individual i and site j's attributes (X_{ij}) and the error terms (ε_{ij}). That is

$$U_{ij} = \begin{cases} \underbrace{X_{ij}\beta}_{V_{ij}} + \varepsilon_{ij} & \text{if } j = 1, \ldots, J \\ \varepsilon_{i0} & \text{if } j = 0 \end{cases}$$

The error term (ε) here are not the independently and identically distributed (*i.i.d.*) type one extreme random variables, but instead are the sum of a series of random variables representing alternative nesting structures. In each choice occasion, individual i will choose to visit the site

j if it gives the highest utility among all the sites. That is

$$y_{ij} = \begin{cases} 1 & \text{if } U_{ij} \geq U_{ik} \text{ for } k = 1, \ldots, J \\ 0 & \text{otherwise} \end{cases}$$

Where $y_{ij} = 1$ means individual i chooses to visit site j.

In order to ease the discussion of identification, we employ a compact vector form:

$$\begin{aligned} U_i &= X_i\beta + \varepsilon_i \\ \varepsilon_i &= F\xi_i + \eta_i \\ Y_i &= [y_{i0}, \ldots, y_{iJ}]' \end{aligned} \tag{3.2}$$

Where X_i is a $((J+1) \times K)$ matrix, U_i, Y_i and ε_i are $((J+1) \times 1)$ vectors, ξ_i is an $(M \times 1)$ vector of M independent standard normally distributed variables, F is a $((J+1) \times M)$ matrix, $F\xi_i$ can be used to represent a wide variety of nesting structures. η_i is a $((J+1) \times 1)$ vector of *i.i.d.* extreme type one random variables. ξ_i and η_i are independent from each other. The followings are four nesting structures:

- Case 1: $F = 0_{(J+1)\times M}$, standard multinomial logit model (case 1 in Figure 1).
- Case 2: $F = [0 \quad \sigma_t \ldots \sigma_t]'_{(J+1)\times 1}$ and $\xi_i = \tau_t$, nested structure one (case 2 in Figure 1). This case nests together our two types of sites, lakes and rivers, into a broad "trip" nest.
- Case 3: $F = \begin{pmatrix} 0 & (\sigma_l)_{(1\times J_l)} & 0_{1\times J_r} \\ 0 & 0_{1\times J_l} & (\sigma_r)_{(1\times J_r)} \end{pmatrix}'$ and $\xi_i = [\tau_l \quad \tau_r]$, nested structure two (case 3 in Figure 1). This case groups lake sites together in a single nest and groups river sites together in a separate nest.
- Case 4: $F = \begin{pmatrix} 0 & (\sigma_t)_{(1\times J_l)} & (\sigma_t)_{1\times J_r} \\ 0 & 0_{1\times J_l} & (\sigma_r)_{(1\times J_r)} \\ 0 & (\sigma_l)_{(1\times J_l)} & 0_{1\times J_r} \end{pmatrix}'$ and $\xi_i = [\tau_t \quad \tau_l \quad \tau_r]$, nested structure three (case 4 in Figure 1). This case is combination of cases 2 and 3, with an overall "trips" nest and two sub-nests distinguishing lakes and rivers.

The parameters $\sigma_t, \sigma_l, \sigma_r$ are all positive real numbers, τ_t, τ_l, τ_r are independent standard normal random variables and $J = J_l + J_r$. Written in the forms of conditional utilities, the four cases will be

- Case 1

$$U_{ij} = \beta X_{ij} + \eta_{ij} \quad \forall j \in \text{ group L}$$
$$U_{ij} = \beta X_{ij} + \eta_{ij} \quad \forall j \in \text{ group R}$$
$$U_{i0} = \eta_{i0}$$

- Case 2

$$U_{ij} = \beta X_{ij} + \sigma_t \tau_t + \eta_{ij} \quad \forall j \in \text{ group L}$$
$$U_{ij} = \beta X_{ij} + \sigma_t \tau_t + \eta_{ij} \quad \forall j \in \text{ group R}$$
$$U_{i0} = \eta_{i0}$$

- Case 3

$$U_{ij} = \beta X_{ij} + \sigma_l \tau_l + \eta_{ij} \quad \forall j \in \text{ group L}$$
$$U_{ij} = \beta X_{ij} + \sigma_r \tau_r + \eta_{ij} \quad \forall j \in \text{ group R}$$
$$U_{i0} = \eta_{i0}$$

- Case 4

$$U_{ij} = \beta X_{ij} + \sigma_t \tau_t + \sigma_l \tau_l + \eta_{ij} \quad \forall j \in \text{ group L}$$
$$U_{ij} = \beta X_{ij} + \sigma_t \tau_t + \sigma_r \tau_r + \eta_{ij} \quad \forall j \in \text{ group R}$$
$$U_{i0} = \eta_{i0}$$

In these above four cases we assume there are three possible groups of sites in the model, stay-at-home option, a group (L) consisting of J_l sites and a group (R) consisting of J_r sites. This division is used in our Monto Carlo simulations reflects the two groups of sites, *Lakes*

and *Rivers*, in our application to the 2009 Iowa Lake and River Data. However, the following discussion of identification is not limited to the two groups setup and the argument for identifiability of the model can be easily extended into other setups.

When the site level trip information is available, a full information likelihood function can be constructed based on the probability of visiting site j by individual i,

$$Pr(y_{ij} = 1|\xi_i) = \frac{exp((X_{ij}\beta + F_j\xi_i)/\mu)}{\sum_{k=0}^{J} exp((X_{ik}\beta + F_k\xi_i)/\mu)}.$$

Because the error term of ξ_i is not known to researchers, the unconditional probability will be

$$Pr(y_{ij} = 1) = \int_{\xi} Pr(y_{ij} = 1|\xi_i)d\xi_i$$

and the log-likelihood function will be

$$llk(i) = \sum_{j=0}^{J} Pr(y_{ij} = 1)^{y_{ij}}$$

This likelihood function can be estimated with an unbiased, smooth simulator with a sequence of R random draws or Halton sequences for the unknown variable, ξ (See, e.g., (18)).

When researchers only have the group trip information for a subset of sites, the full information model cannot be used. In such situations, the aggregation model may allow us recover the parameters in the model. For example, consider the case in which we do not have the individuals' trip information to sites in group R and instead we have only the group trip information y^a, defined as

$$y^a = y_{i0} + \sum_{j=1}^{J_r} y_{ij} \quad \forall j \in \text{ group R}$$

i.e., we only have the aggregated choice information for sites in group R and staying-at-home option. This partial trip information allows us to construct the probability of observation $y^a = 1$ which is the sum the probabilities of visiting each site j in this group.

$$\begin{aligned} Pr(y_i^a = 1|\xi_i) &= \sum_{\forall j \in (0,\text{ group R})} Pr(y_{ij} = 1|\xi_i) \\ &= \sum_{\forall j \in (0,\text{ group R})} \frac{exp((X_{ij}\beta + F_j\xi_i)/\mu)}{\sum_{k=0}^{J} exp((X_{ik}\beta + F_k\xi_i)/\mu)} \\ &= \frac{1}{1 + \sum_{\forall k \in \text{ group L}} exp(\frac{(X_{ij}\beta + F_j\xi_i)}{\mu V_{ia}})} \end{aligned} \tag{3.3}$$

where

$$V_{ia} = \sum_{\forall j \in (0,\text{ group R})} exp((X_{ij}\beta + F_j\xi_i)/\mu)$$

We could think this probability is the one associated with a redefined site a with the utility

$$U_{ia} = \max \quad (U_{ij} = X_{ij}\beta + F_j\xi_j + \eta_{ij}, \quad j \in \{0, \text{group R}\}) \tag{3.4}$$

Modifying the log-likelihood function accordingly seems profitable in the sense we still have a well defined likelihood function. However, the aggregated nature of this model will limit its potential applicability because of potential identification concerns. Some parameters originally identified in the full information model may no longer be identified in the aggregation models based on this partial trip information.

The conditional utility of visiting the new aggregate site a is complicated and the aggregation model is no longer a NECLM discussed in Walker *et al.*(2) even if the underlying model is a NECLM. Though general conclusions about the identification is out of this paper's scope, we speculate that the identification of the aggregation model relies on the richness of the variation embedded in the travel cost variable. The spirit of the conditions of Walker *et al.* (2) is to make sure the variation in the covariance matrix of utility differences is rich enough to identify the parameters in the error structures. In the examples used by Walker *et al.*, the covariance matrix solely depends on the structure of the error terms and there is no interaction with preferences. While here in the aggregation model, the preference information directly enters into the covariance matrix. The potential richness of the elements in the covariance matrix can be larger, thus we speculate that the model could be practically identified if the variation of preference variables is large.

The conventional modeling technique in this situation is to treat y^a as the choice of staying-at-home option, and to model the data accordingly. Mathematically,

$$Pr(y^a = 1|\xi_i) = Pr(y_{i0} = 1|\xi_i) = \frac{1}{1 + \sum_{\forall k \in \text{ group L}} exp((X_{ik}\beta + F_k\xi_i)/\mu)} \tag{3.5}$$

the unconditional likelihood of individual i changes accordingly. Comparing this probability with the one in the aggregation model (3.3), these two will be identical to each other if the

conditional utility of the aggregate site a, V_{ia}, is constant for all the individuals. This condition will not hold in general, thus the conventional model is not correctly specified, the estimation is open to the possible bias.

3.4 Monte Carlo Simulation

In this section, we will conduct two sets of simulations which differ in terms of the structure of the conditional utility function. In the first set of simulations called *simulation without ASCs*, the conditional utility is a function of travel cost and the site attribute(s). This specification fits to the situation in which sites with trip information and sites with partial trip information are essentially the same set of alternatives whose value can be captured by the same set of site attributes. In the second set of simulation, *simulation with ASCs*, the conditional utility function depends on the travel cost and alternative specific constant (ASC). This setting mimics the situation in which sites with trip information and sites with partial trip information belong to two closely related choice alternatives, such as river segments and lakes. ASCs are used to capture site specific values which could be a function of site attributes. A second stage regression can be used to fully recover the contribution of these site attributes to the conditional utility. In the next section, a similar structure of conditional utility is used.[8]

3.4.1 Data Generation Process

Data generation process are generally following the RUM structure discuss in the above section. That is,

Step 1: Generation of the part of conditional utility denoted by V_{ij}. V_{ij} in each set of simulations is generated as follows.

a. Simulation without ASCs

$$V_{ij} = \beta TC_{ij} + \beta_w W_j$$

[8]Strictly speaking, there is a slight difference in the application that ASCs are also applied to sites with partial trip information while in the simulation ASCs are only used for sites with trip information. The unique structure of the application data set allows use to do this without any concern on the identification.

b. Simulation with ASCs

$$V_{ij} = \begin{cases} \alpha_j + \beta TC_{ij} & \text{if site } j \text{ has individual visitation information.} \\ \beta TC_{ij} & \text{otherwise.} \end{cases}$$

The values of parameters used in the simulation are $\beta = -0.05$, $\beta_w = 1$, TC is uniformly distributed between 10 and 30, W is uniformly distributed within 0 and 1 and αs are equally spaced between -2.5 and -0.5.

Step 2: Generation of utility part defined by ε_i. $\epsilon_{ij} = \sigma_t\tau_t + \sigma_l\tau_l + \sigma_r\tau_r + \eta_{ij}$, where $\sigma_t, \sigma_l, \sigma_r$ are positive numbers, τ_t, τ_l, τ_r are *iid* standard normal random variables and η_{ij} is a type one extreme value random variable. Group L and R have equal number of sites.

Step 3: Generation of the site choice for T = 52 choice occasions. For each individual at a given occasion, $y_{ij} = 1, \quad j = 0, 1, \ldots, J$ iff $U_{ij} = V_{ij} + \epsilon_{ij} \geq U_{ik}, \quad \forall k = 0, 1, \ldots, J$. This process is repeated cross individuals and 52 times for each individual.

We anticipate that the number of sites (J) and substitution pattern implied by the combination of σ_t, σ_l and σ_r will affect the performance of the conventional models. Thus we estimate a total of 48 scenarios in the simulation by varying the value of $J, J = 10, 20, 40$ and $\sigma_s (s = t, l, r)$. In each scenario, three models are estimated, the full information model (*Full*), the aggregation model (*Aggregation*) and the conventional partial model (*Partial*). Each scenario will be replicated 100 times. Table 2.1 lists the value of the parameters in the error structures. Each set of parameters defines a different nesting structure with different within and cross nest correlation. For example, the scenario labeled as $S0$ implies that there are not any correlations in the error structure. The implied choice model is a standard multinomial logit model. The scenarios from $S1$ to $S3$ correspond to the nesting structure shown in case 2 of Figure 1 with equal within- and cross-nests correlation.

Since the increase of number of sites will increase the number of ASCs in the model and thus greatly increase the computation time needed to get the maximum likelihood estimation, we only consider a subset of scenarios, namely for the scenarios with $J = 10$ sites, in the second set of simulations for four nesting specifications (labeled as S0, S1, S5 and S8).

3.4.2 Simulation Results

Simulation without ASCs

Table 2.2a - 2.2d list the estimation results of preference parameters and nest parameters. The preference parameters considered there are the estimated value of β and β_w, the estimated value of marginal willingness to pay (MWTP) for water quality, which is defined as $MWTP = \frac{\beta_w}{\beta}$. MWTP is a good indicator for the value of water quality change given other conditions are fixed. A large deviation of estimated MWTP from the true value implies the CV calculation will be largely different from the true CV. The metric used in the comparison is the mean absolute percentage error (MAPE) between estimated parameters and the true value of these parameters. For example, the MAPE of α in R simulations is defined as

$$MAPE(\alpha) = \frac{1}{R}\sum_{r=1}^{R} abs(\frac{\hat{\alpha}_r - \alpha}{\alpha}) \tag{3.6}$$

where α is the true value and $\hat{\alpha}_r$ is the estimated value of α in r_{th} simulation.

There are several results that are worthy regarding the estimated parameters, β, β_w and MWTP(β_w/β) in Tables 2.2a - 2.2c. Firstly, the estimated value of β and β_w in the full information model are consistently close to the true value cross all the simulated scenarios, which is expected since the full information model is identified if we normalize $\mu = 1$, the value we used in the data generation process. Secondly, in the aggregation model, these two parameters are also recovered well with only slightly larger percentage errors. The majority of MAPEs are below 5% and the maximum percentage error is 13%. Also, as the number of sites increase, the difference between estimated value and true value becomes smaller. The correlation between groups of sites does appear not affect those findings substantially. Thirdly, the estimated values from the partial model are systematically and substantially different from the true values. As the correlations and the number of sites increase, the difference tends to become smaller. The lowest difference can reach 9% in terms of MAPE in our simulations.

The comparisons of nest parameters between the full information model and the aggregation model, $\sigma_t, \sigma_l, \sigma_r$, are showed in Table 2.2d. In general, the full model has the estimated values close to the true values, though the inaccuracy is larger compared with the estimation of preference parameters, such as β and β_w. The aggregation model does not perform as well

in recovering the nest parameters. The results on nest parameters are diverse. In the set of scenarios in which the underlying nest structure is the case 2 in Figure 1, it seems the model does reasonably well. However, in the most complicated structures (Case 3 and 4 in Figure 1), the estimation is relatively poor.

Table 2.2e shows the welfare comparison between models, using the full information model as the baseline. Recreational activities are modeled not only to recover the underlying preference, but also to analyze the welfare change due to some changes in recreational sites' accessability and in sites' attributes, such as water quality. In our simulation, we calculate the welfare change due to site closure, *i.e.*, we assume the first site is closed for residents in each scenario. The first observation is that the aggregation model produces almost identical welfare estimation despite its poor performance in recovering of the nest parameters, i.e., σ_t, σ_l and σ_r. The second observation is that there are substantial bias found in CV from the partial model. The direction of the bias varies depending on the number of sites in the simulation and correlation among them. Generally, as the number of sites and the correlation increases the bias also increases. The overestimation of welfare change partially confirms the finding in the aggregation literature if we think the partial model poorly controls for aggregation.[9]

Simulation with ASCs

A model without ASCs implies a strong assumption that researcher observes all the site specific information that affects the conditional utility of visiting any site. If we acknowledge that there may be some site specific attributes not observed, a model with ASCs is preferred to control the possible bias (i.e., Murdock (1)). We add the site specific values to individual sites with observed trip information in the simulation and estimate the simulated data with maximum likelihood estimation technique.

Table 2.3 shows the mean absolute percentage errors in estimated parameters. It is clear that ASCs in the partial model were poorly estimated with quite large bias. The bias positively correlates with site specific values. As in the simulations without ASCs, all three models successfully recover the parameter of travel cost (β). We argued early that the variation

[9]The partial model can be thought as an aggregation model in which the observed part of utility is normalized to zero.

pattern of site attributes could affect the identification of the aggregation model. As shown in the appendix, the aggregation substantially change the structure of the covariance matrix of utility differences by introducing the part of conditional utility determined by site attributes into the covariance matrix. Thus the variation needed to identify the model depends on the richness of site attributes. By replacing water quality with ASCs, we lower the possible variation in the covariance matrix brought about by the variation of water quality. Consequently, the percentage errors of parameters of normal errors increases when compared with the counterpart scenarios in simulations without ASCs. The large bias of ASCs in partial models implies the possible large bias in welfare measure at the first glance, while the results show almost no bias in welfare measure. The welfare change of the site loss depends on two things, the value of parameter (β), and the estimated probability of visiting the site. There is no significant bias in the parameter of β and the inclusion of ASCs in the model ensures that the estimated probabilities will be close to the observed probability which is a fixed number, thus, it is not so surprising to find no significant bias for both the aggregation model and the partial model.[10]

The results from both sets of simulation suggest that the aggregation model performs relatively well in terms of recovering the preference parameters and welfare analysis. The partial model performs substantially better in terms of welfare measure when ASCs were included in the regression. The findings from the application in the next section are generally consistent with these simulations.

3.5 Application to 2009 Iowa Lake and River Project

3.5.1 Iowa Lake and River Projects

The application provided below draws on a unique pair of data sets, providing information on lake and river visitation pattern of Iowa residents in 2009.[11] The unique feature of these two data sets is that 2,500 households who receive both river survey and lake survey give full

[10]The inclusion of ASCs in a standard multinomial logit model guarantees the predicted probability of visiting a site equals the observed probability. In mixed logit models, generally, the predicted probability will not *equal* the observed probability when ASCs are included in the model, though Babatunde *et al.* (1) find they typically do not differ substantially.

[11]The survey samples could be downloaded from Center for Agricultural and Rural Development at Iowa State University's websites: Lake Survey's Link and River Survey's Link

information about their recreational trip information to all river sites and lake sites. Among these overlap respondents, there are 1160 (46.4 %) respondents who returned both surveys.[12] The co-existence feature allows us to identify the aggregation model constructed for samples with only partial information by relying on the identification of the full model applied to these overlap samples.

3.5.1.1 2009 Iowa River Project

The data set has two primary sources: the Iowa 2009 River Survey conducted by the Department of Economics, Iowa State University and the location and attributes information about 73 identified river segments and streams in Iowa provided by the Iowa Department of Natural Resources (IDNR). The focus of the survey was on collecting the baseline information on Iowa households' riverine recreation activities in the year of 2009, along with demographic information and attitudes regarding the factors affecting respondents' recreation decisions. The final survey was conducted by mail in November, 2009 sent to 10000 randomly selected Iowa residents. The sampled residents were divided into two groups, 7500 households were selected to receive only the river survey and another 2500 households were selected to receive both the river survey and the lake survey. Among all the surveys mailed, 4758 surveys were returned, for a raw return rate of around 48%.[13] The IDNR provides us a series of shape files, a type of files used in ArcGis software, on geographic information and site attributes. Based on these shape files, we were able to calculate the distance and travel time for each household to each river access point by PCMiler street version 24 . In the following model, we simplify the travel distance calculation by setting the midpoint of each river as the access point for that river segment.[14]

[12] Only 1084 respondents' surveys are used in the estimation because of the data incompleteness issue.

[13] There are only 4084 qualified surveys for the analysis. Among those excluded surveys, some are uncompleted due to missing some information on trip and demographic information.

[14] Once we have the one way travel distance d, the travel cost is calculated by the formula: $C = 2gd + 2/3w * t$, where C is the travel cost, g is the per mile vehicle cost, w is the hourly wage rate and t is the travel time. Both d and t are calculated by PCMiler. The value of g is set to be 54 cents per mile according to the 2009 AAA annual driving cost for an average sedan with 15,000 miles per year driving. The wage rate w is calculated by dividing household annual income imputing from the survey by the number of adults in the households and 2000, the total working hours in a year. For the respondents who do not report income information, we assume the wage rate is 19 dollars per hour calculated from 2009 Iowa wage survey conducted by Iowa Workforce Development, Labor Force and Occupational Analysis Bureau.

3.5.1.2 2009 Iowa Lake Project

2009 Iowa lake project is a continuation of the four-years long study (2002-2005) about Iowans' visitation patterns and preferences to 132 important lakes in the state. The goal of the project is to combine the lake attributes, such as water quality, with Iowans' visitation information to assess the value of water quality improvement in those lakes. The 2009 surveys were sent out to 10,000 households, among them 5400 households are selected from respondents to the former 2005 lake survey and 4,600 new households are randomly selected from Iowa households. There are also 2,500 households receiving both the lake survey and the river survey. A total of 6,043 respondents returned their surveys, yielding a 61.43% response rate among deliverable surveys.[15]

3.5.2 Data Description

Table 2.4 provides summary statistics for trip information of surveyed households. For the pooled sample, the average number of trip to all the lakes is 6.12 per year and the number of trip to river segments is 6.07 per year, varying with some respondents taking zero trips while others took more than 52 per year.[16] In total, 59 % of the lake sample of respondents receiving the lake survey reports positive lake recreation trips with the average visitation at 10.36 per year. Among the river respondents who receive the river survey, 47.6 % of the sample report they had positive river trips in 2009 with the average visitation rate of 12.76 per year.

Table 2.5 shows summary statistics of demographics of different samples. For the pooled sample, the returned surveys show that the sample covers a population leaning toward middle age and senior people. On average, the respondent is more likely to be a female with college degree and comes from a family with two adults and one kid under age 6. The respondent have a 60% probability to be an employed worker and 36% chance to be a retired person. With small chances, around 1% and 3%, respectively, the respondent will be a student and unemployed person. There are around 36% respondents coming from a household owning a boat, such as

[15] There are some household not delivered due to reasons such as address change, deceased or refusal of receiving the survey, which makes the total number of deliverable survey less than 10,000.

[16] As a conventional assumption in the literature, the maximum allowed visits per year is set at 52. Those respondents who report more than 52 trips per year are excluded from the analysis.

fishing boats, canoes and so on. The summary statistics of demographics of lake sample, river sample and overlap sample are also shown in Table 2.5. There is no substantial difference among these samples since the respondents are randomly selected to represent typical Iowa residents by design.

The summary statistics of lake attributes are shown in Table 2.6. We include six lake attributes. *Secchi* depth is used in this paper to represent water quality, which is a measure of water transparency. Iowa's lakes presents great variation in this measure, with Secchi depth ranging from 0.2m in Lake Darling, Washington county to 7.8m in West Okoboji Lake, Dickinson County. The average Secchi depth in the sample lakes is 1.25m.[17] Iowa lakes varies substantially in terms of area (*Size*), with the smallest lake, Moorhead Lake in Ida county, covering only 10 acres, while the biggest lake, Red Rock lake in Marion county, covers 19,000 acres. There are also four dummy variables to measure the attractiveness of a lake. The first one is *ramp* which equals one if there is a cement boat-ramp at that site and equals zero otherwise. The second one is *state park* which equals one if there is a state park adjacent to the site. The third one is *wake* which equals one if motorized vessels are allowed to travel fast enough to create wakes in the lake and equals zero otherwise. The last one is *Handicap facility* which equals one if there are handicap facilities for disabled people at the site and equals zero otherwise.

We list the summary statistics of river attributes in Table 2.7. From the ArcGis shape files provided by Iowa DNR, we construct 11 river attributes. The length of 73 river segments, *Length*, varies from 26.9 miles to 161.8 miles. *Canoe* is a percentage variable measuring how much of the river segment is canoeable as identified by the IDNR. On average, more than 60% of the river segment is canoeable. The specific percentage of each river segment varies from zero percent to almost 100%. *Outcrop* measures the number of outcrops along the river banks, and varies from 0 outcrops along the bank to more than 100 outcrops. Using the built-in functions in Arcgis 10.1 and the shape file of Iowa Land Cover 2002, we also construct several

[17] Egan *et al.* ((11)) use more water quality measures based on a similar data set. Since the major purpose of this paper is not focused on how to choose water quality measures to best estimate household recreational behavior, only Secchi depth is included in the model at this stage and more water quality measures could be included in the future working if necessary.

surface-type variables: *Waterbody* for the share of water body within the corridor of 75 meters width on each side from the middle line of the river segments, *Wetland* for the share of wetland, *Forest* for the share of forest land, *Grass* for the share of grass land, *Crop* for the share of crop land, and *Developed* for the share of developed land. Within this wide rive corridor, the top three land cover types are crop land (39%), forest land (24%) and grass land (11%). Iowa water quality index (*IWQI*), adjusted from the national water quality index to incorporate Iowa specific situations, is used to represent the water quality status of river segments. The scale of this index is from 0 to 100 with higher values indicating better water quality. The river water quality is not monitored as frequently as lake sites. Also the coverage of monitoring sites is very poor in the sense that a significant number of river segments are not monitored, 23 river segments or 31.5% of 73 river segments in this sample. To capture this missing variable issue, *MIWQI*, a dummy variable, which equals one if the river segment is not monitored, is used. Within the monitored river segments, the mean value of IWQI is 31.55. This means in general the water quality of Iowa rivers is poor.[18] We also construct a variable, *FISH*, to measure the abundance of fish species. The Iowa DNR identifies more than 100 fish species in instate rivers and compiles the presence information at the segment levels delimited by dams on rivers. The average number of fish species is a little more than 30 and there are 3 river segments which do not have identified fish species and 9 river segments for which we do not have information regarding fish species. The dummy, *MFISH*, is used to indicate segments with missing fish species.

3.5.3 Model Setup and Results

The unique design of the Iowa lake project and river project distinguishes the respondents into three categories, the lake-only group (respondents who only receive the lake survey questionnaire), the river-only group (respondents who only receive the river survey questionnaire) and the overlap group (respondents who receive both river survey questionnaire and lake survey questionnaire). Based on these three groups, we estimate four models:

[18] Iowa DNR classifies five water quality categories based on the value of IWQI: very poor (0-25), poor(25-50), fair(50-70), good(70-90) and excellent(90-100). Other issues with river water quality are the poor coverage for the monitored river segment and unfrequent monitoring activities.

- Model 1 - pooled model
- Model 2 - overlap model
- Model 3 - lake partial model
- Model 4 - river partial model

In *pooled model*, the inclusion of the overlap group allows us to build an aggregate probability for the outside option, specifically separate the aggregate probability into the one for visiting a river (or lake) site and the one for staying-at-home option. In *overlap model*, the probability of visiting each site is specifically modeled for the overlap group. While in *lake partial model* and *river partial model*, we treat the outside options as if it is really a staying-at-home option and neglect it is an aggregated option for the lake-only group and the river-only group, respectively.

Nest structures are all assumed in all the four models and estimated by the error component mixed logit models. In *pooled model*, three normal distributed random variables are assumed to mimic the nest structure. One is used to introduce the cross river sites and lake sites correlation. The other two are introduced to allow for the correlation within lake sites or river sites. These three random variables are also assumed in *overlap model*. In *lake partial model* and *river partial model*, only one random variable is used to mimic the correlation among lake (river) sites. Test on the presence of nest structures can be performed after estimation by testing whether the parameters of the standard deviation for these normal distributions are significantly different from zero.

The estimation results of these four models are listed in Table 2.8. The travel cost coefficient in the three models are quiet similar and stable, ranging from 0.0305 to 0.0334. Comparing with young people, elder people tend to stay at home more often. Female respondents choose more visits when compared with male respondents.[19] A person with college degree is more likely to stay at home, though the effect is not significant in the *lake partial model*. Persons coming from big families are also more likely to choose water based recreational activities, though the effect is not significant in the *river partial model*. The number of kids in the household seems not

[19]In the questionnaire, we specifically ask the individual visits to lake sites or river sites, while whether the respondent will answer from the family perspective is out of our control. Thus we suggest readers do not take too much emphasis on this gender difference.

a significant factor when making the visitation decision, although the effects in *pooled model* and *overlap model* are marginally significant but opposite directions. For employment status variables, there are no significant effects found in these models. For students, the effects are all significantly negative on visits to river or lake sites except in *overlap model*. For retired people, we find a significant positive effect on visits only in *pooled model*. In other models, this effect is not significant. The ownership of a boat is found universally positively affect residents' either lake recreation or river recreation. Purchasing a boat has already show the strong preference on the water-based recreation activities, it is very natural to find the positive effects.

One of the advantages to estimate an error-component version of nested logit model is that it is straightforward and easy to test the possible structures. In the *overlap model* and *pooled model*, we estimate three error terms associating with trip error, lake-related error and river-related error. A joint restriction of the standard deviation of these three normal error terms to being zero is equivalent to reducing the model to the multinomial logit specification. The test is carried out with the Wald test ((8)). The joint test and individual tests all suggest that we rejected restricting the model from the two-level nesting structure depicted in Figure 1. The implied correlation cross and within lake (river) sites are showed in Table 2.9.[20] In the *lake partial model* and *river partial model*, the tests also suggest there is a nest structure among lake sites or river sites, respectively. If the *pooled model* is the true model, the two partial models capture the within correlation quite well and the *river partial model* performs relatively better.

The coefficients of certain lake and river attributes are reported in Table 2.10 respectively, these results are from the second stage regression suggested in Murdock ((1)). For lake attributes, most of the coefficients agree in terms of sign and significance except the Wake dummy. We found the Secchi is an appealing water quality measure.[21] The results also show that Iowan like to visit lakes with a larger water area. For other attributes, only the coefficient of wake

[20] If the three normal errors are characterized by zero mean and σ_t, σ_l and σ_r as standard deviations respectively, the correlation cross lake (river) sites is calculated as $\sqrt{\frac{\sigma_t^2}{\sigma_t^2+\pi^2/6}}$, the correlation within lake sites is $\sqrt{\frac{\sigma_t^2+\sigma_l^2}{\sigma_t^2+\sigma_l^2+\pi^2/6}}$ and the correlation within river sites is $\sqrt{\frac{\sigma_t^2+\sigma_r^2}{\sigma_t^2+\sigma_r^2+\pi^2/6}}$

[21] There are some other water quality for lakes, such as total suspended particles and water quality index. We only include Secchi here is because it has been used in other studies and found to be a good proxy for water quality.

dummy is found to be positively significant in the lake partial model. This coefficient is positive in the pooled model and negative in the overlap model, but not significant. We found all the coefficient of other attributes are significant in both models. Slight difference in the magnitude of values are found between the full model and lake partial models. In both models, water quality represented by the Secchi depth, is found to positively and significantly affect people's recreation decisions. The lakes with more recreational amenities, indicated by the dummy variables, Ramp, Handicap facilities and State Park, attract more visitation.

The estimation results of river attributes are also in Table 2.10. The coefficients in all the three models mostly agree with each other although sometimes they differ in the statistical significance. For example, a longer river segment attracts more visits implied in all three model and the effect is found not to be significant only in the overlap model. Using the forest land share along the river bank as the default land cover type, the more share of water body area and wetland area usually means the river is more attractive, though this appealing effect is only significant in the overlap model. Neither of these three model finds the water quality measure, Iowa water quality index, imply Iowan prefer to visit the river segment with a high IWQI. We think the lack of significance in our models reflects the current situation about water quality monitoring work for rivers. Taking Iowa as example, there are roughly 100 monitoring sites for all the rivers in Iowa. Based on our sample, the total length of 73 identified river segments is around 6000 miles. That means there are less than 2 monitoring sites for every 100 miles of river. At the same time, the frequency of monitoring is very low. The water quality file passed by the Iowa DNR shows that for most of the monitor sites, there are only 2 observations each year. Thus it is very likely the water quality measure do not reflect the relevant ones which really affect Iowa households' decisions.

Recreation models are often used to calculate welfare effect due to changes of site closing or programs targeting to improve water qualities of certain sites. We consider two type of welfare measure, the loss of site and changed site attributes. In the loss of site scenario, one lake site (Lake Saylorville) and one rive site (Mississippi-the part from Clinton County to Muscatine County) are assumed to be closed. The reason of choosing these two sites is they receive the most visits in the survey year. In terms of water quality change, we only consider the change

of Secchi depth for lake sites because the coefficient of water quality variable, *IWQI*, is not significant in our models. The specific scenario is that the Secchi depth of all the lakes are improved to the level of lake West Okoboji which has the Secchi depth of 7.8 meters. The welfare results are showed in Table 2.11.

The results show the welfare measure (CV) calculated from the pooled model and the overlap model for the site loss are similarly, and there is a highly overlap between their 90% quantiles. While the partial models produce significant smaller values. In terms of Secchi depth change, the results from three models do not agree with each other. The overlap model gives the highest value and is followed by the pooled model and lake partial model. If the pooled model was the model we can trust, the partial models do produce the statistically significant but quite moderate deviation in welfare analysis.

3.6 Conclusion

Households' recreational data are extensively used to connect individuals' visitation behavior with recreational facilities and attributes and to evaluate their economic values. Many of these studies are built on survey data which may only have partial information about individuals' recreational site choices. In this paper we propose a modeling technique based on the site aggregation literature to tackle this problem. Though the complex structure induced by the aggregation technique sheds shadows in the identification of the model, two sets of Monte Carlo simulations shows the aggregation model works quite well in some circumstances. Compared with the conventional modeling techniques, the aggregation model have relatively better performance in recovering preference parameters and welfare calculation. With alternative specific constants in the model, the conventional model can produce reasonable welfare measure as good as the aggregation model.

The application to the unique data from Iowa 2009 Lake and River projects shows the difference between a variety of modeling techniques may not be as significant as the ones found in the simulations. We applied the pooled model with the aggregation technique along with other two models, the conventional partial model on partial samples and the full information model based on the overlap samples. Though the partial model is open to bias in theory, we

find slight differences in preference parameters in terms of magnitude of values of coefficients and significance levels. With regards to welfare evaluation, the full model and partial models do produce statistically significant but moderate difference.

The underlying structures could make it difficult, if not possible, to find the identification conditions for the aggregation model, thus in some situations, the conventional partial model will be the one possible way to model the recreation data available to researches. On the other hand, our simulation work shows that the potential bias in doing so could be quite large. The nonlinearity of RUM models make it difficult to tell the possible direction and the extent of this bias. How to mitigate this bias deserves more future attentions.

Bibliography

[1] Babatunde, Abidoye O., Herriges, Joseph A. and Tobias, Justin L., (2012) "Controlling for Observed and Unobserved Site Characteristics in RUM Models of Recreation Demand," *American Journal of Agricultural Economics*, **94**(5) pp. 1070 - 1093.

[2] Ben-Akiva, M., and Lerman, S.R. (1985), "Discrete Choice Analysis: Theory and Application to Travel Demand." Boston, MA: MIT Press.

[3] Chiou L., Walker J.L., (2007) "Masking identification of discrete choice models under simulation methods," *Journal of Econometrics* **141** pp. 683-703.

[4] Feather, P. (1994), "Sampling and Aggregation Issues in Random Utility Model Estimation," *American Journal of Agricultural Economics*, **76**, pp. 926-33.

[5] Ferguson, M. R., and P. S. Kanaroglou (1998). "Representing the Shape and Orientation of Destinations in Spatial Choice Models," *Geographical Analysis* **30(2)**, pp. 119-137.

[14] Feather, P. and Lupi, F. (1998) "Using partial aggregation to reduce bias in random utility travel cost models" *Water Resources Research* Vol. 34, No. 12, pp. 3595-3603.

[7] Haener, M. K., Boxall, P. C., and Adamowicz, W. L., (2004) "Aggregation bias in Recreation site choice models: resolving the resolution problem," *Land Economics*, **80(4)**, pp. 561-574.

[8] William H. Greene "Econometric Analysis (6th)",*Prentice Hall*, pp

[9] Haab, T., and K. E. McConnell (2002), *Valuing Environmental and Natural Resources: The Econometrics of Non-Market Valuation.* Northampton, MA:Edward Elgar.

[10] M. K. Haener, P. C. Boxall, W. L. Adamowicz, and D. H. Kuhnke (2004), "Aggregation Bias in Recreation Site Choice Models:Resolving the Resolution Problem," *Land Economics*, **80**, No. 4, pp. 561-574.

[11] K., Egan, C., Kling, J. A. Herriges and J., Downing (2009), "Valuing Water Quality as a Function of Water Quality Measures," *American Journal of Agricultural Economics*, **91**, No. 1, pp. 106-123.

[12] Kling, C., and Herriges, J. (1997), "The Model Performance of Nested Logit Models when Welfare Estimation is the Goal," *American Journal of Agricultural Economics,* Vol. 79, 792-802.

[13] Kurkalova, L.A. and S.S. Rabotyagov (2006), "Estimation of a Binary Choice Model with Grouped Choice Data," *Economics Letters* **90**(2): 170-175.

[14] Kling, C., and C. Thomson (1996), "The Implications of Model Specification for Welfare Estimation in Nested Logit Models," *American Journal of Agricultural Economics*, **78**, pp. 103-14.

[15] Kaoru Yoshiaki, and V. K. Smith. (1990). " 'Black Mayonnaise' and Marine Recre- ation: Methodological Issues in Valuing a Cleanup." Marine Policy Center, Woods Hole Oceanographic Institution, Woods Hole, MA.

[16] McFadden, D. (1974), "Conditional Logit Analysis of Qualitative Choice Behavior," In P. Zarembka, ed., *Frontiers in Econometrics*, New York: Academic Press.

[17] McFadden, D. (1978), "Modeling the Choice of Residential Location," in *Spatial Interaction Theory and Planning Models,* ed. by A. Karlgvist, et al. Amsterdam: North-Holland.

[18] McFadden, D. (1989), "A method of simulated moments for estimation of discrete response models without numerical integration", *Econometrica*, Vol. 57, 995C1026.

[19] McCulloch R. and Rossi P. E., (1994) "An exact likelihood analysis of the multinomial probit model", *Journal of Econometrics,* Vol. 64, 207-240.

[20] McFadden, D. and K. Train. (2000), "Mixed MNL Models for Discrete Response," *Journal of Applied Econometrics.*" Vol 15, issue 5, 447-470.

[1] Jennifer Murdock (2006) "Handling unobserved site characteristics in random utility models of recreation demand", *Journal of Environmental Economics and Management*,Vol. 51, No.1 pp. 1-25.

[22] Parson, G., Plantinga, A.J. and Boyle, K.J, (2000) "Narrow Choice Sets in a Random Utility Model of Recreation Demand", *Land Economics*, Vol. 76, No. 1, pp. 86 - 99.

[23] Parsons, G., and M. Needelman (1992), "Site Aggregation in a Random Utility Model of Recreation," *Land Economics*, **68**: 418-33.

[24] Phaneuf, D. J., and J. A. Herriges (2002), "Inducing Patterns Correlation and Substitution in Repeated Logit Model of Recreation Demand," *American Journal of Agricultural Economics*, **14**, No.4, pp. 1076-1090.

[25] Phaneuf, D. J,. and Smith, V. K. (2005) " Recreation Demand Models" *Handbook of Environmental Economics* Edited by K. G. Mäler and J. R. Vincent, Vol. 2, pp. 672 - 751.

[26] Pollak, R., and T. Wales (1991), "The Likelihood Domi- nance Criterion," *Journal of Econometrics,* Vol. 47, 227-42.

[27] Train, K.,(2003), "Discrete Choice Methods with Simulation," Massachusetts: Cambridge University Press

[28] Train, K., D. McFadden, and M. Ben-Akiva, (1987), "The demand for local telephone service: A fully discrete model of residential calling patterns and service choice," *Rand Journal of Economics*, Vol. 18, 109 - 123.

[2] Walker, J.L., M. Ben-Akiva and D. Bolduc, (2007), "Identification of Parameters in Normal Error Component Logit-Mixture (NECLM) Models," *Journal of Applied Econometrics*, Vol. 22, 1095-125.

Figure 2.1 Nest Structures

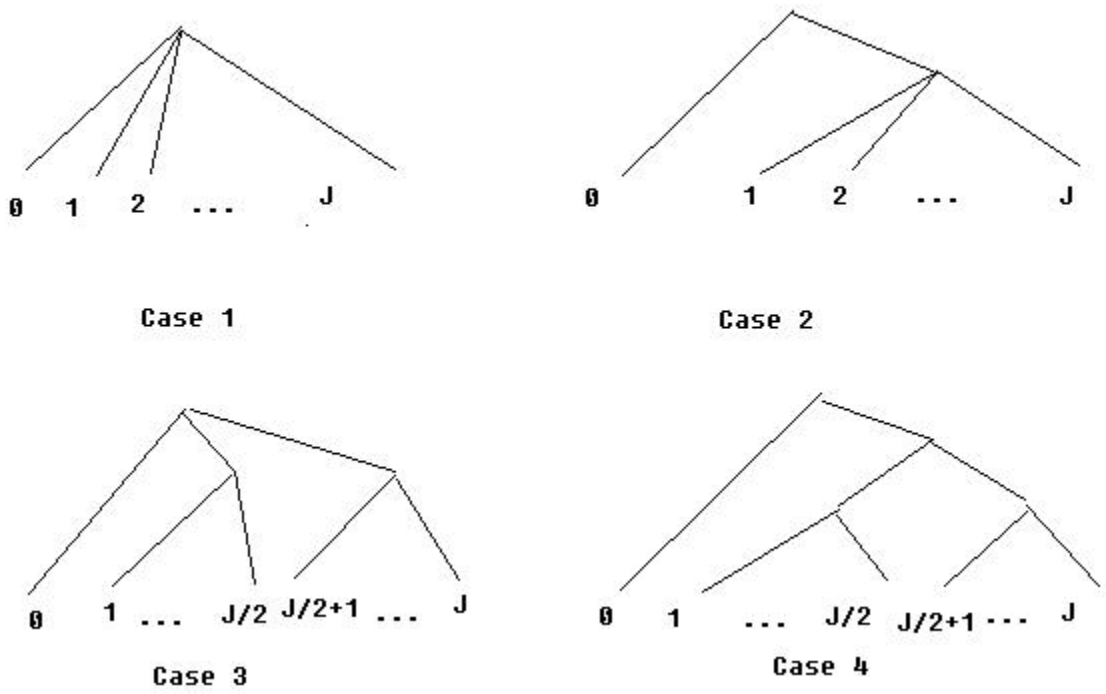


Table 2.1 Specification of Error Terms in Simulation

		σ_t	σ_l	σ_r	Correlation in Lake	Correlation in River	Cross Correlation
Case 1	S0*	0	0	0	0.00	0.00	0.00
Case 2	S1*	0.5	0	0	0.13	0.13	0.13
	S2	1	0	0	0.38	0.38	0.38
	S3	1.5	0	0	0.58	0.58	0.58
Case 3	S4	0	0.1	0.1	0.01	0.01	0.00
	S5*	0	0.5	0.5	0.13	0.13	0.00
	S6	0	1	1	0.38	0.38	0.00
Case 4	S7	0.5	0.1	0.1	0.14	0.14	0.13
	S8*	0.5	0.5	0.5	0.23	0.23	0.12
	S9	0.5	1	1	0.43	0.43	0.09
	S10	1	0.1	0.1	0.38	0.38	0.38
	S11	1	0.5	0.5	0.43	0.43	0.35
	S12	1	1	1	0.55	0.55	0.27
	S13	1.5	0.1	0.1	0.58	0.58	0.58
	S14	1.5	0.5	0.5	0.60	0.60	0.54
	S15	1.5	1	1	0.66	0.66	0.46

S0, ..., S15 are used to label different scenarios.

The scenarios with "*" appear in the simulation with ASCs.

Table 2.2a Mean Absolute Percentage Error in Estimated Marginal Utility of Income ($\beta = 0.05$)

		σ_t	σ_l	σ_r	**Full Model**			**Aggregation Model**			**Partial Model**		
					$J = 10$	$J = 20$	$J = 40$	$J = 10$	$J = 20$	$J = 40$	$J = 10$	$J = 20$	$J = 40$
Case 1	S0	0	0	0	1	1	1	1	1	1	103	148	201
Case 2	S1	0.5	0	0	1	1	1	1	1	1	85	32	20
	S2	1	0	0	1	1	1	2	1	1	70	31	20
	S3	1.5	0	0	1	1	1	1	1	1	52	30	20
Case 3	S4	0	0.1	0.1	1	1	1	1	1	1	82	32	20
	S5	0	0.5	0.5	1	1	1	1	1	2	44	26	18
	S6	0	1	1	1	1	1	1	1	1	25	18	14
Case 4	S7	0.5	0.1	0.1	1	1	1	1	1	1	80	32	20
	S8	0.5	0.5	0.5	1	1	1	2	1	1	43	26	18
	S9	0.5	1	1	1	1	1	2	1	1	24	18	14
	S10	1	0.1	0.1	1	1	1	3	1	1	64	31	20
	S11	1	0.5	0.5	1	1	1	1	1	1	42	26	18
	S12	1	1	1	1	1	1	2	1	1	25	18	14
	S13	1.5	0.1	0.1	1	1	1	8	1	1	51	30	20
	S14	1.5	0.5	0.5	1	1	1	2	1	1	38	25	18
	S15	1.5	1	1	1	1	1	2	1	1	24	19	14

Table 2.2b Mean Absolute Percentage Error in Estimated Coefficient of Water Quality ($\beta_w = 1$)

		σ_t	σ_l	σ_r	**Full Model**			**Aggregation Model**			**Partial Model**		
					$J = 10$	$J = 20$	$J = 40$	$J = 10$	$J = 20$	$J = 40$	$J = 10$	$J = 20$	$J = 40$
Case 1	S0	0	0	0	1	1	1	1	2	1	74	92	113
Case 2	S1	0.5	0	0	1	1	1	2	1	1	60	19	12
	S2	1	0	0	1	1	1	2	2	1	51	19	12
	S3	1.5	0	0	1	1	1	2	1	1	42	19	13
Case 3	S4	0	0.1	0.1	1	1	1	2	1	1	60	23	13
	S5	0	0.5	0.5	1	1	1	2	1	1	36	18	12
	S6	0	1	1	1	1	1	2	2	1	22	12	10
Case 4	S7	0.5	0.1	0.1	1	1	1	2	1	1	57	22	12
	S8	0.5	0.5	0.5	1	1	1	2	1	1	33	17	12
	S9	0.5	1	1	1	1	1	2	1	1	20	12	9
	S10	1	0.1	0.1	1	1	1	4	1	1	51	23	13
	S11	1	0.5	0.5	1	1	1	2	1	1	35	18	12
	S12	1	1	1	1	1	1	2	1	1	19	12	9
	S13	1.5	0.1	0.1	1	1	1	7	2	1	42	22	12
	S14	1.5	0.5	0.5	1	1	1	2	2	1	32	18	12
	S15	1.5	1	1	1	1	1	2	1	2	21	13	10

Table 2.2c Mean Absolute Percentage Error in Estimated MWTP ($\frac{\beta_w}{\beta} = 20$)

		σ_t	σ_l	σ_r	**Full Model**			**Aggregation Model**			**Partial Model**		
					$J = 10$	$J = 20$	$J = 40$	$J = 10$	$J = 20$	$J = 40$	$J = 10$	$J = 20$	$J = 40$
Case 1	S0	0	0	0	1	1	1	2	2	2	88	97	105
Case 2	S1	0	0.1	0.1	1	1	1	2	2	2	79	39	27
	S2	0	0.5	0.5	1	1	1	2	2	2	71	38	27
	S3	0	1	1	1	1	1	2	2	2	62	38	27
Case 3	S4	0	0.1	0.1	1	1	1	2	2	2	78	41	27
	S5	0	0.5	0.5	1	1	1	2	2	2	56	35	25
	S6	0	1	1	1	1	1	2	2	2	38	26	21
Case 4	S7	0.5	0.1	0.1	1	1	1	2	2	2	77	40	27
	S8	0.5	0.5	0.5	2	1	1	3	2	2	54	35	25
	S9	0.5	1	1	2	1	1	2	2	2	36	26	20
	S10	1	0.1	0.1	1	1	1	6	2	2	71	41	27
	S11	1	0.5	0.5	2	1	1	2	2	2	55	35	26
	S12	1	1	1	1	1	1	2	2	2	35	26	20
	S13	1.5	0.1	0.1	1	1	1	13	2	2	62	39	27
	S14	1.5	0.5	0.5	2	2	1	3	2	2	51	35	26
	S15	1.5	1	1	2	2	1	3	2	2	36	27	21

Table 2.2d Estimation Results of Nest Parameters

	Scenario	σ_t	σ_l	σ_r	**Full Model**								**Aggregation Model**							
					J=10			J=20			J=40		J=10			J=20			J=40	
					σ_t	σ_l	σ_r	σ_t	σ_l	σ_r	σ_t	σ_l	σ_r	σ_t	σ_l	σ_r	σ_t	σ_l	σ_r	σ_t	σ_l	σ_r
Case 1	S0	0	0	0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Case 2	S1	0.5	0	0	1	-	-	1	-	-	1	-	-	1	-	-	1	-	-	7	-	-
	S2	1	0	0	1	-	-	1	-	-	1	-	-	1	-	-	1	-	-	1	-	-
	S3	1.5	0	0	1	-	-	1	-	-	1	-	-	1	-	-	1	-	-	1	-	-
Case 3	S4	0	0.1	0.1	-	1	3	-	2	6	-	4	7	-	17	2	-	16	2	-	10	9
	S5	0	0.5	0.5	-	1	1	-	1	1	-	1	1	-	8	26	-	3	20	-	3	26
	S6	0	1	1	-	1	1	-	1	1	-	1	1	-	3	11	-	8	20	-	8	21
Case 4	S7	0.5	0.1	0.1	1	6	5	1	2	4	1	7	4	9	19	8	3	23	5	14	11	7
	S8	0.5	0.5	0.5	1	1	1	1	1	1	1	1	1	5	5	27	4	5	16	6	1	28
	S9	0.5	1	1	1	1	1	2	1	1	3	1	1	26	4	18	27	6	22	30	7	23
	S10	1	0.1	0.1	1	3	7	1	3	9	1	1	13	57	59	214	1	3	20	1	11	10
	S11	1	0.5	0.5	1	1	1	1	1	2	2	1	1	12	30	35	7	19	8	4	19	2
	S12	1	1	1	2	1	1	2	1	1	4	1	1	1	3	14	3	1	14	3	2	15
	S13	1.5	0.1	0.1	1	8	9	1	9	4	1	7	6	91	91	569	3	25	22	1	13	5
	S14	1.5	0.5	0.5	1	2	1	1	1	1	1	2	1	30	69	85	10	44	34	6	43	24
	S15	1.5	1	1	1	1	1	2	1	1	3	1	1	12	8	14	8	8	4	2	6	6

1. "-" represents the parameter does not appear in the model.

Table 2.2e Welfare Change (CV) in Simulation

		Aggregation Model				**Partial Model**			
	$\sigma_t =$	0	0.5	1.0	1.5	0	0.5	1.0	1.5
				J = 10					
$\sigma_l = \sigma_r$	0	1.00	1.00	1.00	1.00	0.56	0.70	0.85	1.02
	0.1	1.00	1.00	0.99	0.98	0.72	0.73	0.87	1.03
	0.5	1.00	1.00	1.00	1.01	1.03	1.07	1.10	1.16
	1.0	1.01	1.00	1.01	1.01	1.22	1.23	1.27	1.30
				J = 20					
$\sigma_l = \sigma_r$	0	1.00	1.00	1.00	1.00	0.48	1.25	1.29	1.34
	0.1	1.00	1.00	1.00	1.00	1.24	1.25	1.28	1.34
	0.5	1.00	1.00	1.00	1.00	1.28	1.30	1.32	1.36
	1.0	1.01	1.00	1.00	1.01	1.34	1.34	1.38	1.40
				J = 40					
$\sigma_l = \sigma_r$	0	1.00	1.00	1.00	1.00	0.41	1.39	1.41	1.44
	0.1	1.00	1.00	1.00	1.00	1.39	1.40	1.42	1.43
	0.5	1.00	1.00	1.00	1.00	1.39	1.40	1.42	1.45
	1.0	1.00	1.00	1.00	1.00	1.41	1.41	1.43	1.45

1. The figures in this tables are average ratios of CV from each of these two models over CV from the corresponding Full model.

Table 2.3 Mean Absolute Percentage Error in Estimated Parameters with ASCs (J=10)

	σ_t	σ_l	σ_r	α_1	α_2	α_3	α_4	α_5	β	σ_t	σ_l	σ_r	CV
Full Model													
S0	0	0	0	1	1	1	1	2	1	-	-	-	0
S1	0.5	0	0	1	1	1	1	2	1	2	-	-	0
S5	0	0.5	0.5	1	1	1	2	3	1	-	3	2	0
S8	0.5	0.5	0.5	1	1	1	2	4	1	4	4	3	0
Aggregation Model													
S0	0	0	0	1	1	1	2	2	2	-	-	-	2
S1	0.5	0	0	1	1	1	1	2	2	7	-	-	2
S5	0	0.5	0.5	1	2	2	3	5	3	-	18	80	2
S8	0.5	0.5	0.5	2	2	3	5	9	2	66	23	82	2
Partial Model													
S0	0	0	0	43	53	71	107	213	2	-	-	-	2
S1	0.5	0	0	44	55	73	109	219	2	-	-	-	2
S5	0	0.5	0.5	44	55	73	109	218	3	-	-	-	2
S8	0.5	0.5	0.5	45	56	75	112	224	2	-	-	-	2

Table 2.4 2009 Iowa Lakes and Rivers Survey Statistics

Survey Type	Variable	Mean	Std. Dev.	Min	Max
Pooled Sample	Lake trips	6.12	9.27	0	52
	River trips	6.07	10.01	0	52
Overlap Sample	Single Day Trips	9.97	12.42	0	52
	Lake Trips	4.53	7.52	0	52
	River Trips	5.44	8.44	0	52
Lake Sample (+)	Lake Trips	10.36	10.08	1	52
River Sample (+)	River Trips	12.76	11.19	1	52

+:Respondents who report positive recreational trips in 2009.

Table 2.5 Summary Statistics of Demographics

Variable	Description	Pooled Sample	Overlap Sample	Lake Sample	River Sample
		Mean	Mean	Mean	Mean
Age1	1(Age: 18-25)	0.01	0.01	0.01	0.01
Age2	1(Age: 26-34)	0.05	0.08	0.05	0.06
Age3	1(Age: 35-49)	0.25	0.23	0.23	0.21
Age4	1(Age: 50-59)	0.26	0.25	0.27	0.25
Age5	1(Age: 60-75)	0.31	0.30	0.31	0.31
Age6	1(Age: 76-)	0.15	0.13	0.14	0.15
Female	1, 0	0.69	0.71	0.68	0.70
College	1, 0	0.70	0.67	0.70	0.69
Size	# of adults	1.89	1.91	1.91	1.88
Kids	# of kids	0.55	0.65	0.56	0.55
Employed	1, 0	0.60	0.62	0.61	0.59
Student	1, 0	0.01	0.01	0.01	0.01
Retired	1, 0	0.36	0.35	0.35	0.37
Boat	1(Owning a boat)	0.16	0.22	0.13	0.22

1. The corresponding observations are 8316, 1084, 5352 and 4084, respectively.

Table 2.6 Summary Statistics for Lake Attributes

Variable	Unit	Mean	Std. Dev.	Min	Max
SECCHI	Meters	1.25	1.07	0.2	7.8
SIZE	Acres	662.41	2105.41	10	19000
RAMP	0, 1	0.85	0.36	0	1
WAKE	0, 1	0.66	0.48	0	1
HANDICAP FACILITY	0, 1	0.38	0.49	0	1
STATE PARK	0, 1	0.39	0.49	0	1

Table 2.7 Summary Statistics of River Attributes

Variable	Unit	Mean	Std.Dev	Min	Max
LENGTH	Miles	83.20	33.30	26.90	161.8
CANOE	%	62.82	27.13	0.00	99.00
OUTCROPPING	counts	18.77	31.66	0.00	141.00
WATERBODY	%	0.22	0.30	0.00	1.00
WETLAND	%	0.01	0.01	0.00	0.05
FOREST	%	0.24	0.19	0.00	0.71
GRASS	%	0.11	0.09	0	0.40
CROP	%	0.39	0.29	0	0.92
DEVELOPED	%	0.02	0.04	0	0.24
IWQI	(0-100)	31.55	23.00	0	75
MIWQI	0, 1	0.32	0.47	0	1
FISH	counts	30.29	18.52	0	71
MFISH	0, 1	0.12	0.33	0	1

Table 2.8 Estimation Results of Iowa Lake and River Project Data

	Pooled Model		**Overlap Model**		**Lake Partial Model**		**River Partial Model**	
	Est.	Std.Dev	Est.	Std.Dev	Est.	Std.Dev	Est.	Std.Dev
Travel Cost Variables								
TC	0.0316***	0.0001	0.0335***	0.0003	0.0325***	0.0002	0.0305***	0.0002
Demographics								
Age2	1.5313***	0.2355	1.8464***	0.3745	1.5775***	0.4374	1.9472***	0.4816
Age3	1.7391***	0.2258	2.1564***	0.3764	1.2755***	0.4210	1.7743***	0.5114
Age4	1.9749***	0.2138	2.1225***	0.3861	1.3919***	0.4213	2.1924***	0.5033
Age5	2.3142***	0.2279	2.7223***	0.4272	1.6494***	0.4262	2.8404***	0.4956
Age6	3.0515***	0.2673	2.8843***	0.4321	2.5493***	0.4403	3.8351***	0.5246
Female	-0.7814***	0.0754	-0.6312***	0.0969	-0.4547***	0.0749	-0.4042***	0.1387
College	-0.5792***	0.0860	-0.2422***	0.0919	-0.0364	0.0759	-0.5869***	0.1218
Size	-0.1309***	0.0445	-0.4276***	0.0674	-0.1843***	0.0560	-0.1227	0.1013
Kids	-0.1009**	0.0514	-0.0406	0.0562	-0.0104	0.0418	-0.0558	0.0567
Worker	-0.1409	0.1400	0.1573	0.2065	0.0723	0.1723	0.1214	0.3881
Student	1.5572***	0.3351	-1.1944***	0.3600	0.8660*	0.5121	1.7610***	0.5050
Retired	-0.4794**	0.2124	0.3052	0.2820	-0.0846	0.1952	-0.0869	0.3780
Boat	-1.4682***	0.0804	-1.5613***	0.1097	-0.9248***	0.1000	-1.8401***	0.1367
Nest Parameters								
Trip	2.7563***	0.0450	2.5782***	0.0647	2.2360***	0.0315	2.9631***	0.0528
Lake	0.9417***	0.0611	1.1318***	0.0426	-		-	
River	1.8848***	0.0422	1.6734***	0.0560	-		-	

Table 2.9 Implied Nest Correlations

Model	Cross Correlation	Within Lakes	Within Rivers
Pooled Model	0.91	0.92	0.93
Overlap Model	0.91	0.91	0.93
Lake Partial Model		0.87	
River Partial Model			0.92

Table 2.10 Estimation of Site Attributes (Stage 2)

Variable	Pooled Model	Overlap Model	Partial Model
Lake Attributes			
Secchi	0.21***	0.25***	0.21***
	(4.37)	(2.87)	(4.15)
ln(Size)	0.49***	0.48***	0.50***
	(11.70)	(8.29)	(11.57)
Ramp	-0.13	0.04	-0.14
	(-0.70)	(0.13)	(-0.75)
Wake	0.19	-0.09	0.22*
	(1.57)	(-0.44)	(1.80)
Handicap Facilities	0.02	0.18	0.01
	(0.18)	(0.87)	(0.09)
State Park	0.17	0.27	0.16
	(1.29)	(1.35)	(1.23)
Constant	-6.62***	-6.94***	-6.80***
	(-28.48)	(-18.11)	(-28.87)
Observations	130.0	130.0	130.0
Adj. R-sq	0.64	0.47	0.64
River Attributes			
Length	0.0073*	0.0069	0.0073*
	(1.80)	(1.37)	(1.88)
Canoe	0.0089*	-0.0010	0.0083
	(1.72)	(-0.12)	(1.59)
Outcropping	0.0016	0.0059	0.0017
	(0.37)	(1.20)	(0.37)
Waterbody	1.4379	2.0981**	1.4804
	(1.60)	(2.03)	(1.63)
Wetland	18.4493	29.3198*	19.3430
	(1.49)	(1.84)	(1.44)
Grass	0.6152	1.5130	0.5567
	(0.36)	(0.63)	(0.32)
Crop	-0.3843	-0.5286	-0.4043
	(-0.47)	(-0.52)	(-0.49)
Developed	0.2113	0.7292	0.3161
	(0.05)	(0.13)	(0.07)
IWQI	0.0061	-0.0186	0.0061
	(0.43)	(-0.87)	(0.42)
MIWQI	0.2996	-0.7824	0.3067
	(0.42)	(-0.77)	(0.42)
Fish	0.0098	0.0136	0.0107
	(1.20)	(1.35)	(1.33)
MFISH	0.4738	0.1461	0.4692
	(1.27)	(0.25)	(1.23)
Border	1.2020**	1.6624***	1.2014**
	(2.57)	(2.80)	(2.55)
Constant	-6.7099***	-5.1286***	-6.6469***
	(-5.26)	(-3.37)	(-5.39)
Observations	73.	73.	73.
R^2	0.668	0.589	0.681

t statistics in parentheses

* $p \leq 0.1$, ** $p \leq 0.05$, *** $p \leq 0.01$

Table 2.11 Welfare Measures

Model	Mean	5%	95%
Loss of Lake 103 (Lake Saylorville)			
Pooled Model	19.31	18.37	20.23
Overlap Model	19.42	17.41	21.50
Lake Partial Model	16.98	16.09	17.80
River Partial Model		-	
Loss of River 71 (Mississippi(Clinton to Muscatine))			
Pooled Model	17.42	16.50	18.33
Overlap Model	17.35	15.82	19.05
Lake Partial Model		-	
River Partial Model	14.73	13.79	15.68
Secchi Depth Change			
Pooled Model	235.76	206.94	264.95
Overlap Model	287.31	149.43	444.35
Lake Partial Model	218.08	190.54	244.39

1. The unit here is dollars per year per household.
2. The quantiles are from 1,000 simulations.
3. The figures in Secchi depth scenario is the absolute value of original CV.

CHAPTER 4. Carbon Tax, Wind Energy and GHG reduction- ERCOT as an Example

4.1 Introduction

The promotion of renewable energy and imposition of carbon tax are the two favored government policies for combating the climate change induced by the more anthropic GHG emission. Unlike the EU, the United States (US) does not have a national level cap-and-trade program for GHG emissions. The California Air Resources Board launched a regional program in 2013, and the northeastern RGGI market has operated since 2008. Although there is no national climate policy, the pressure faced by fossil fuel generators in the US is increasing in recent years. For example, the US Environmental Protection Agency (EPA) announced new point-source requirements for future fossil generators in a proposal of carbon standard for new power plants. New coal generators must have emission rate of carbon dioxide (CO2) of less than 1,000 pounds of CO2 per megawatt-hour (MWH), almost less than half of current levels.[1] Under these programs, the relative cost of fossil fuel generators, i.e., coal versus. gas, would be affected. As more pressures or high charges are imposed on carbon dioxide emission, the less generation from coal units would be predicted and thus lower CO2 emissions.

Using the data from Electric Reliability Council of Texas (ERCOT), we construct the cumulative marginal cost curves of fossil fuel generators, i.e. the supply curve of fossil fuel generation, at different carbon tax scenarios in Figure **??**. It is clear that when a carbon tax first imposed at the hypothetical level of 25 dollars per ton of carbon dioxide, the shape of this curve changes significantly (From Blue Curve to Brown Curve). Though we have difficulties

[1] On March 27^{th}, US EPA announced the proposal of "Standards of Performance for Greenhouse Gas Emission for New Stationary Sources: Electric Utility Generation Units". http://www.regulations.gov/#!documentDetail;D=EPA-HQ-OAR-2011-0660-0001

to show the composition of the generation by highlighting which part of the generation is from gas units or coal units, the structural change of this marginal cost curve obviously implies there must be significant changes in the generation composition, such as some gas units will move forward in the supply curve due to their lower emission of CO2. Given the specifications of generators, the resulted reshuffle of supply curve is becoming much smaller as carbon tax increases further from 50 dollars per ton of CO2 to 75 dollars per ton of CO2 (from green curve to red curve).

Table 3.1-1 lists the generation composition in the 4-quartiles under different carbon taxes scenarios. Confirmed with the findings in Figure **??**, the carbon tax on carbon dioxide changes the structure of the marginal cost curve. The coal units were pushed backward in the ladder in serving the system demand according to their marginal cost. In two extreme cases, no carbon tax versus the tax rate at 75$/CO2 ton, the coal units are almost eliminated from the generation fleet when the system demand was below 30kMWH (within the first 2 quartiles).[2]

At the same time, the fast expansion of renewable energy, such as wind farms and solar panels, increases the share in the energy production by substituting the fossil fuel energy in the electricity market. The expansion eventually lost momentum due to several factors. The current financial crisis constrained government budgets and thus the expenditure on subsidies to renewable energies. The incentive created by the energy act and state level renewable portfolio standards (RPS) have induced the expansion of wind farms faster than the anticipation. For example, the installed capacity in ERCOT has passed the target of 2025 set by the state RPS in 2009.[3] It has been a non-negligible generation source in ERCOT for several years due to this fast increase.

The intermittency of the wind power has brought about researchers' interests on the actual environment benefits of wind power and the affecting factors. Unlike fossil fuel generators, wind farms only can produce electricity as wind blows and cannot increase the production when the

[2] The total installed capacity of coal and gas units in ERCOT is larger than 60,000MWHs. In Figure **??** and Table 3.1-1, we limit the maximum system demand at 60,000MWHs. Analysis of the full spectrum of fossil fuel generation fleet with almost 70,000 MWHs capacity could be constructed similarly.

[3] The Texas legislature set a goal of 10,000 MW of renewable capacity in 2025, the capacity has been reached in 2009 with a total of 10,069 MW installed renewable capacity. http://www.ercot.com/news/press_releases/show/517.

demand increases as those fossil fuel generator do. The induced replacement of "dirtier" energy by wind power therefore also depends on the correlation between wind and the demand. If in the peak demand hours there is little wind, as is common in several markets, the ability of wind to replace coal or gas generation will be greatly limited.

The penetration of renewable energy, such as wind or solar panels, complicates the effectiveness of carbon tax or cap-and-trade programs. The introduction of volatile renewable energy adds another layer of uncertainties into the system in addition to the volatility of system demand. As Table 3.1-2 showed below, the extensive wind power in ERCOT significantly change the order of coal and gas generators in the aggregate supply curve in a typical day in the summer/winter.[4]

Though the above calculation is very primitive, it is still quite clear that the penetration of wind power in the system could affect the final generation allocation between coal generators and gas generators. Given the level of wind generation, the impacts are likely to correlate with the carbon tax rates. In this paper, we assess the impacts of the interaction between the penetration of wind power and the carbon tax in a relatively more realistic way, including consideration of transmission constraints, with the data from ERCOT in the period of Jun, 2009 to Jun, 2010. Specifically, we solve a supply "dispatch" model for a target research period under a variety of scenarios defined by the hypothetical carbon tax rate (dollars per ton of CO2) and the penetration of wind powers (the installed capacity of wind turbines).

The simulation shows that both policies are effective in terms of adjusting the generation portfolio and reducing the CO2 emission. Given the technical configuration of the current fossil fuel generator fleet in ERCOT, it is not surprised to find out that the effects of carbon tax on generation reallocation are nonlinear with the tax rates. In the simulation, a carbon rate of 25 dollars per ton of CO2 would lead the share of gas generation to increase by more than 20% from currently 55% to almost 80% given the state quo of wind capacity. The percentage of gas generation in the total fossil fuel generation will eventually increase to more than 95% at a slower

[4]We set the input price at the average level in the research period. Namely, the price of coal is 1.86$/mmBTU and the price of natural gas is 4.09$/mmBTU. We pool together all the hourly observations from Jun – Aug and Dec – Feb respectively. Then the average demand and average wind generation are used to find out the generation mix from the marginal curves shown in Figure **??**. Clearly, these results will differ from the simulation results in section 5.

pace if the carbon tax further increases to as high as 70 dollars per ton of CO2. Consequently, the pattern of induced reduction of CO2 emission also shows this similar nonlinearity in the carbon tax. Compared with the current emission level, the CO2 will be reduced by roughly 23.4% at a carbon tax rate of 25$ per ton of CO2. The pace of emission reduction has not been slowed too much until a tax rate of 40$ per ton of CO2. This nonlinear relation suggests that there are some ''sweet" range in which the reduction role of a carbon tax will not be curtailed and beyond this range we may see almost negligible effects of the tax increase. Depending on the composition of the generation fleet, there are slightly zonal variations in terms of the adjustment pace of gas generation. Eventually, the gas generation will dominate the fossil fuel generation in each zone when the carbon tax is relative high.

If more wind turbines were installed, say doubling the current capacity, the environmental benefit in terms of CO2 emission reduction, suggested by our model, will be around 4% in a no carbon tax world. This reduction effect is slightly smaller than the reduction achieved in the change from a zero wind world to status quo, which is roughly 5.4% implied in our model. When the wind capacity was doubled, the reduction effect of wind energy varies from 4.1% to 4.9% depending on the carbon tax imposed. The results also show some spatial differences. Relative to the current zonal emission, the effect in west zone is no surprisingly largest since the majority of wind turbines were installed in this zone. However, if measured in physical terms, the largest emission reduction happens in the south zone.

We also construct a contour graph of emission reduction in the carbon-wind plane. [5] The isocline in the graph shows the trade-off between carbon taxes and wind energy for a certain goal of emission reduction. Based on the isoclines, if we want to achieve a 10% reduction of CO2 in ERCOT, we could use a single carbon tax at 15$ per ton of CO2 or use a combination of carbon tax at around 10$ per ton of CO2 and doubling the current wind capacity. If the targeted reduction is 35%, we need a single carbon tax beyond the highest rate of 70$ per ton of CO2 considered in our simulation or we could achieve the goal at a lower carbon tax rate

[5] Since our simulation only can generation scattered points in the wind expansion and carbon tax space, we use linear interpolating methods to fill other cells of wind-carbon pairs. In doing so, we could have a reduction matrix for all the pairs of carbon tax and wind expansion. The increments are 1$/ton of CO2 and 1% wind expansion.

(45$ per ton of CO2) with the doubled wind capacity.

The remaining of this paper is organized as follows: a brief summary of related literature is provided in the section 2. In section 5, we introduce the dispatch model used in this paper. We discuss the data sources in the section 4. Calibration and simulation results are presented in the section 5. We conclude this paper in the section 6.

4.2 Related Literature

Unlike fossil fuel generators, the ability of wind turbines to generate electricity is limited by the local weather conditions, such as wind speed. The availability of wind power is therefore stochastic and tends to show a different pattern from the demand of power. In many markets, the wind power is usually negatively correlated with the demand. When the electricity demand is highest in the daytime, wind power is at its lowest. Conversely wind power is often most abundant, when demand tends to be lowest. The implications of the intermittency of wind power and the mismatch between the demand and wind power have been received extensive interest from researchers.

Since the generation of electricity by wind farms is almost carbonless when compared with traditional fossil fuel units based on coal and natural gas, wind power is promoted as a means to mitigate carbon dioxide emissions in the power industry. The intermittency of wind power puts some restrictions on this substitution. Campbell (2008) builds up a theoretical model to show that in some circumstances, the extra intermittent renewable resources, such as wind, introduced to the power market may not mitigate the carbon dioxide emissions if much dirtier generation resources are used to replace the substituted, relative cleaner base load generation resources. Though this paper points out a theoretical possibility of a scenario in which more wind brings about more carbon dioxide emissions, the empirical papers based on real generation data do not support its realization in practice.

The lack of empirical examples in which more intermittent renewable energy can lead to higher emissions is partly because coal-fired generators are the usual "base-load" resources – resources that run whenever available - and it is unlikely to see a much dirtier resource to replace the coal generation. Bushnell (2010) uses several power markets in west United States

to analyze the induced equilibrium generation mix caused by increasing wind penetration. With more intermittent wind power introduced into the grid, the equilibrium coal-fired generation capacity declines and the more flexible combustion turbine generators, usually using natural gas as heat input, increases to serve as quickly adjustable marginal generation that compliments the intermittent wind supply. Fell and Linn (2012) uses a long-run investment model to study the investment decision on different energy resources, such as coal, natural gas, wind and solar panel and market outcomes under different policy scenarios. Though the correlation between wind and load will have some effects on final investment decisions, the changes happen mostly on the investment of gas generators. When wind is positively correlated with the load, less investment on gas generators is needed. Otherwise, more gas generators are needed. The key assumption in Campbell (2008) to have dirtier generation resources to replace the cleaner base load generation may not be realistic in the real world.

Another possibility of increasing pollution through more wind power is related to operations of specific plants. Volatile wind power will increase the uncertainty of the load if wind and system load are negatively correlated. In turn, more cycling and start-up of fossil fuel generators is needed to maintain the reliability of the system. Noting this fact, Valentino et al. (2012) uses Illinois data to model the impacts on the emission from fossil fuel generators under different wind penetration levels with unit commitment assumption. The results show that by replacing fossil fuel generation with wind power, the system wide emission from fossil fuel generators decrease as wind penetration increases. While the more wind power is incorporated into the system, the average emission rate of fossil fuel may increase as the result of cycling and start-ups. The study finds that the latter emission-increasing effect is minimal and dominated by the direct effect of replacing more fossil fuel generation.

Alongside the research based on economic dispatch models mentioned above, there is a line of empirical studies using reduced-form econometric models to analyze the environmental effects of renewable resources, mostly of wind power. These studies, such as (Cullen 2010; Kaffine, McBee, and Lieskovsky 2011; Novan 2010) among others, focus on the marginal effects of wind power on the emission reduction.

Cullen (2010) is among the first to use reduced-from regressions to study the emission

reduction effect of wind power in the power grids. He uses the generation data at individual fossil fuel unit level in ERCOT, in which there is the highest installed capacity of wind farms, to construct a unit level replacement of generation by wind power in the system. The relevant variables in the unit's information set are controlled in the multivariate regression models by assuming the wind power is exogenous. With the replacement of generation at unit level, the author then aggregate to the system emission reduction using the average observed emission rates of units. The results show that one extra MWH of wind power will, on average, reduce CO2 emissions by more than half ton. The average emission rates used by the author may not be the best one to evaluate the emission reduction since the emission rates depend on the efficiency of heat input. J. B. Bushnell and Wolfram (2005) shows that there usually exists a maximally efficient "sweet spot" in the generator's heat rate profile and the correlation between generation and heat rate is generally nonlinear. If the generator were forced to adjust its production to meet the power demand or to incorporate the wind power in the grid as found in (Cullen 2010), the average emission rates may misrepresent the true reduction of pollutants.

The EPA continuous emissions monitoring system (CEMS) reports the gross generation and pollutants emission data at hourly base for all the fossil fuel units with the installed capacity bigger than 25MW. These reports give researchers abundant high frequency data to directly analyze the effects on emission reduction. Kaffine, McBee, and Lieskovsky (2011) use similar reduced-form econometrical models based on CEMS data in several regional power markets, California, Midwest and ERCOT, to evaluate the marginal effects on emission of wind power. Another difference from Cullen (2010) is that the authors aggregate the emission to the system level instead of looking at the emission reduction from individual generation units. Their results show that the marginal reduction in emission varies among three markets studied and the variation depends on the existing generation mix in the system. Specifically, if the generation mix is much dirtier, the resulted emission reduction of extra one MWH wind power is lower.

Though the wind is exogenous, the direct use of wind generation may suffer endogenous problems due to the fact that the existing transmission capacity between wind-abundant regions and high demand regions is limited. The realized wind power in the system may have some

system information which directly correlates with the generation arrangement of individual generators and thus emissions. Novan (2010) uses the wind speed data in west zone of ERCOT, a region with the majority of installed wind capacity in Texas, to instrument the troubled wind generation variables in the unit generation regression models.[6] The instrumented models show that the marginal effect of emission reduction depends nonlinearly on the system load, the demand of electricity. And the location of wind farm is also important in determining the emission reduction. It is not surprising since the replaced marginal generation varies depending on the system load. If the load is high, the most likely marginal resource will be gas units. The avoided emission is lower than the case when the coal units are the marginal resources in low load scenario.

The marginal analysis based on reduced-from approach is useful when the major interest is on the marginal effects of wind power. While if we want to evaluate the changes caused by the expansion of wind power in large scale, an economic dispatch model of electricity generation, similar to the one used in Bushnell and Chen (2009), is more appropriate. In this study, the possible carbon tax and wind expansion scenarios will be evaluated with a simple dispatch model.

[6] The instruments test does suggest there are substantial problems caused by the endogenous variable.

4.3 Methodology

The goal of this analysis is to simulate outcomes under different scenarios in which a mixture of policies are introduced into the power market to reduce emissions from fossil fuel generation and to compare their relative performance using a simple dispatch model. The two specific policies considered in the simulation are the carbon tax and the continuing support for the expansion of wind capacities.

The simple dispatch model is set up as a cost minimization problem. The system operator needs to arrange the generation portfolio of all the generation units in its territory to meet the exogenous demand and at the same to maintain the reliability of the system with the transmission limits obeyed at the lowest possible cost.

Mathematicallythe cost minimization problem could be described as

$$\min \sum_{i,j,t}^{4,J_i,T} C(q_{ijt})$$

$$s.t. \sum_{j=1}^{J_i} q_{ijt} + wrate * wind_{it} + y_{it} \geq d_{it} \quad \forall i = 1,2,3,4; t = 1,\ldots,T \tag{4.1}$$

$$-\bar{T}_{kt} \leq \sum_{i=1}^{4} PTDF_{ikt} y_{it} \leq \bar{T}_{kt} \tag{4.2}$$

$$C(q_{ijt}) = (a_{ij} + p_{ijt} h_{ij}) q_{ijt} + tax_c e_{ijt} q_{ijt} \quad \forall i = 1,2,3,4; j = 1,\ldots,J_i; t = 1,\ldots,T \tag{4.3}$$

Where

- d_{ijt} and $wind_{it}$ are exogenous $\forall i, jandt$.
- $C(q_{ijt})$ is the cost function of unit j in zone i at time t.
- q_{ijt} is the generation of unit j in zone i at time t.
- d_{it} is the hourly demand of electricity at zone i at time t.
- $wrate$ is the expansion rate of wind capacity in zone i at time t. $wrate = 1$ stands for the status quo.

- $wind_{it}$ is the wind generation in zone i at time t, we do not have direct measure of this variable and the construction of this variable is described in the appendix.
- y_{it} is the net injection of power in zone i at time t. A positive value means the energy export and the negative value means the import of energy.
- $\bar{T}_{kt}$ is the transmission limit on transmission interface k at time t.
- $PTDF_{ikt}$ is the power transfer distributing factor. It measures the flow on transmission k at time t when you inject the energy at zone i.
- a_{ij} is the non-fuel cost for producing 1 MWH from unit j in zone i.
- p_{ijt} is the fuel price used by unit j in zone i at time t.
- h_{ij} is the heat rate of unit j in zone i, defined as mmBTU per MWH.
- tax_c is the assumed carbon tax measured at dollars per metric ton of CO2.
- e_{ij} is the emission rate of CO2 of unit j in zone i, defined as metric ton of CO2 emitted per MWH power.

The details about how to construct these variables is provided in the appendix[7]. The numerical model is written with AMPL software, and solved with the *Minos* solver.

The different scenarios are defined by the combination of carbon tax rate (tax_c) and the expansion rate the wind capacity ($wrate$). For example, the pair of $(tax_c = 0, wrate = 1)$ is the status quo and the pair of $(tax_c = 25, wrate = 1.2)$ represents a hypothetical situation in which a carbon rate of 25$ per ton of CO2 is imposed and the wind capacity in the network increases by 20 percent. The carbon tax rates are set from 20 dollars per ton of CO2 to 70 dollars per ton of CO2 at the increment rate of 5 dollars, in addition to the status quo without any carbon tax.[8] The wind scenarios are set from 0 percent of current capacity to 200 percent

[7] mmBTU is a unit of thermal energy, read as million British Thermal Units. 1 BTU equals 1055 joules. MWH is a unit of electricity energy, 1 MWH = 1,000 KWH, 1 KWH = 3.6 million joules.

[8] Nordhaus (2010) estimated the carbon price associated with the fulfillment of a 2 Celsius increase by the end of this century is 59 dollar per ton (at 2005 price). The Department of Energy newly released update on the energy-efficiency of microwave ovens implied the potential social cost of carbon (SCC) was around 33 dollars per ton at 2010 under a moderate assumption about the social discount rate (3%). The SCC could be larger with smaller discount rates.

of current capacity at the increment rate of 20% each.

In each scenario, we hold other conditions fixed. That is to say the cost minimization algorithm will find the optional zonal generation arrangements with the hypothetical carbon tax and wind penetration. The other conditions, such as the zonal demand, zonal generation from non-thermal energy sources (such as nuclear), fuel prices and transmission capacity, will be unchanged across different scenarios.[9] Since the demand is uncertain, this method implicitly makes very a strong assumption about the realization of these important variables.[10]By taking this assumption, we implicitly limit our research question to ask what the performance of ERCOT market will be in the hypothetical scenarios during the research period.

Once those cost minimization problems are solved. The model reports the zonal generation arrangements at generator level along with the corresponding CO2 emission. With this information in hand, we could construct outcome matrices to compare the relative performance in terms of total emission reduction of CO2 and the distribution of generation among coal plants and gas plants at ERCO-wide or each zone.

The availability of data forces us to focus on the research period from Jun, 2009 to Jun, 2010. The limiting factor for data is the information about the monthly average weighted shift factors in ERCOT. In the research period, ERCOT runs a simplified four-zone electricity network, which facilitates our modeling work to mimic the market of ERCOT. With this framework, it is important to get the average zonal shift factor information that is used to capture the energy flow between transmission interfaces and plays key roles in determining the congestion status on transmission interfaces. Unfortunately, these shift factors have been collected from public sources from Jun, 2009 to Jun, 2010, which limits our ability to expend the model to other periods.[11]

With significant penetration of renewable energy, especially wind energy, ERCOT is an

[9] The annual growth rate of ERCOT, the example power market, is 3.75% from 2009 to 2010. If necessary, the growth rate could be added into the model.

[10] The loads and wind generation are random such that if we want to see the robustness of results under uncertainty, some simulation work showing the randomness of these variables is needed. For example, we could draw from the empirical distribution of load to construct the possible realization of zonal loads. And using the wind forecast models to simulate the possible wind generation in the near future. GE Energy (2008) has the details to construct such a scenario.

[11] The extension of our model to a later period will faces another challenge. ERCOT zonal market ends at Dec 1^{st}, 2010 and switched to a nodal market after that.

ideal power market to serve our purpose to evaluate the interaction impact of carbon tax and substantial renewable energy. On their website, ERCOT provides several sets of useful information about the power market. However, we still have some data issues in building our model, which will be discussed with more detail in next sections.

4.4 Data Sources

We have several primary data sources for this analysis. Electric Reliability Council of Texas (ERCOT) has the information on zonal demand, generation and inter-zonal transmission. The U.S. EPA Continuous Emission Monitoring System (CEMS) provides hourly output for all fossil fuel power plants with a capacity bigger than 25 megawatt per hour (MWH). The U.S. Energy Information Administration (EIA) has the monthly and yearly information on net generation of power plants aggregated by fuel type and types of prime movers, such as stream turbine, combined cycle and combustion turbine, etc. It also has the monthly price of coal delivered to power generation facilities. The daily natural price is from Natural Gas Intelligence's daily gas price index.

4.4.1 Market Demand

ERCOT is one of the eight independent system operators in US and serves 85 percent of electricity demand in Texas. Unlike other power networks, ERCOT is quite isolated from neighboring power grids. There is limited power exchanged between ERCOT and surrounding grids, e.g., the exchanged load account for less 1 percent of daily load Cullen (2010). During our research period, the inter-grid net import energy accounts for 0.6 percent of the total load in 2009, 0.7 percent of the total load in 2010 (FERC (2009, 2010)).

Before December 1^{st}, 2010, ERCOT operated a power network market based upon four price zones (Zonal market). The whole of ERCOT territory was divided into four congestion zones, North, South, West, and Houston. Five commercial significant controls (CSC) are used to represent the inter-zonal power flow.[12] On the wholesale market, most of the transactions are

[12] After December 1^{st}, 2010, ERCOT moved to a more complex nodal market to increase the system dispatch efficiency.

arranged through long term bilateral contracts. Any residual electricity, capturing deviations from the bilateral positions, is settled in the real-time energy balancing market. The balancing market price is determined by ERCOT by choosing the lowest price to clear the market when there is no zonal congestion. If there is congestion, the different zones will settle at different prices reflecting the congestion costs for each zone. These congestion costs are based upon the "re-dispatch" costs necessary to adjust generation production between zones to solve the congestion.[13]

A summary of zonal hourly electricity demand is presented in Table 3.4-1. The average hourly demand of the whole ERCOT area is around 36 GWH. It ranges from around 21 GWH to 63 GWH during the sample period. There is significant spatial variation among four zones. The North zone has the highest hourly demand of electricity and accounts for 38% of the total demand. Hourly demand of electricity is lowest in the West zone and the share is less than 10% of the total demand. The demand of the South Zone is in pair with that of the Houston Zone, both accounts for roughly 25% of the total load.

4.4.2 Zonal Production

EPA CEMS reports include the hourly generation information on all the fossil fuel power units with at least 25 MW capacities in Texas. To conduct a zonal analysis, the units first should be assigned to one of the four zones. Though CEMS reports have the location and other attributes of the unit, there are no fields in the reports directly related to ERCOT zones. To match the units to four zones, we combine the information from several sources and papers to completely match units to each zone.[14]

Table 3.4-2 summaries the generation portfolio in ERCOT. Among the units covered in the CEMS reports, ERCOT have a fleet of fossil fuel power generation with a total capacity of 77 GW. Gas units dominate coal units by a margin of 46% (73% of gas and 27% of coal generation capacity). In line with the zonal demands, North zone has the biggest fossil fuel

[13] In the zonal market design, local (or intra-zonal) congestion is inevitable and ERCOT resolves these congestions by using special generation sources. Local congestion will not be reflected in the zonal price if there are no co-existed inter-zonal congestions.

[14] Details about the matching procedure are in the appendix.

power generation fleet, followed by the Houston zone and the South zone. The West zone has the smallest generation fleet. Among the gas units, more than half of the capacities are from the combined cycle units, followed by the steam turbine gas units.

Table 3.4-3 gives average hourly generation by coal and gas units. On average, coal plants produce 46% of the electricity among the total fossil generation while the capacity share of coal units is only 27%. This is because the cost of producing 1MWH power is lower for coal units and thus coal units are served as base load generation source. At the zonal level, the coal generation shares vary from 29% in Houston zone to 55% in North zone. The spatial variation of the share of coal generation suggest that if the imposed carbon tax is high enough, the switch trigged by the carbon tax will also present spatial variations.

Before using this generation information in the model, the capacity of the each unit needs to be adjusted since the generation information from EPA CEMS is the *gross* generation from each unit. In other words, EPA gross load includes the electricity consumed at the facility, such as the electricity consumed by the pollutant control devices, and the amount of load delivered to the grid network is smaller. In addition, CEMS seems not having a universal standard on how to report the gross generation for combined circle units that uses the residual heat from the first stage of combustion turbine to generate more electricity. For some units, EIA reports more net generation than gross generation reported by CEMS.[15] Another issue with CEMS gross generation is that it does not separate the generation from combined heat and power units. In Houston zone, the majority of gas generation comes from this type of power plants, which imposes difficulties on how to incorporate these units in the model. We will discuss the impacts of these units on outputs of baseline simulations in next section. At the same time, the zonal demand information from ERCOT does not contain this part of in-facility consumption and we need to account for this issue in the production model.

In terms of net generation, EIA has reports on both monthly and yearly net generation from each power plant. The information is summarized and grouped by fuel and "prime-mover"

[15] There are several power plants with gross generation reported in CEMS much smaller than the net generation reported by EIA. For example, CEMS reports the power plant with ORISPL code of 55215 has a gross generation of around 1.9 million MWHs in 2010. However, the net generation reported by EIA is much larger and around 2.9 million MWHs. The plant has combined cycle units and the net generation from the first stage of a combined cycle, combustion turbine part, is around 1.8 million MWHs.

(technology) type. Compared with the gross load from CEMS reports, a rough adjustment gross load to net generation can be done. Specifically, we construct a yearly gross-to-net ratio for each fuel type and each prime mover at the plant level. We assume this ratio is the same for all units within a plant if their fuel sources are the same and the prime mover types are the same.[16] We deflate the capacity of each unit according to this ratio in the simulation models. The emission ratios, defined as the amount of emissions (SO2, CO2 and NOX) of each generation unit, are also adjusted accordingly.

The capacity of each unit used in the model is also discounted to reflect the probability of forced outage. The available capacity of each unit is calculated to be $(1 - fof_i) * Cap_i$, where $fof_i$is the factor of forced outage of unit i and Cap_i is the already deflated capacity of unit i. This formula is similar to the ones used in (Fowlie (2009); Bushnell (2010)). The difference is that we specifically consider the net-to-gross generation adjustment. The fof_i information comes from the generator availability data system (GADS) data collected and maintained by the North American Electricity Reliability Council.

The heat efficiency (or heat rate), which represents the generator's efficiency to turn fuel input into electricity, is very critical in our production model. With the fuel prices held constant in our model, the heat rate solely determines the order of the generator in the supply curve. With the adjusted gross generation and the heat input reported in CEMS, we could construct the annual heat rates for all the units in CEMS. The heat rates calculated this way represent the efficiency of each unit to turn the heat content in coal or gas to net generation. There are also other methods to construct heat rates with similar meanings. One possibility is to use the net generation and fuel consumption information reported by EIA. The problem with this method is that the accuracy of fuel consumption for combined heat and power plants. Plant operators have flexibilities to decide the share of their total fuel consumption used by heating purpose or power-generation purpose. If the fuel consumption for power-generation is reported to be lower than the actual use, the calculated heat rate of net generation will be lower than the actual heat rate. Thus, the generation unit will be more favored in the model to be deployed

[16] These ratios are not always within the 0 to 1 interval because of underestimation of some combined cycle gas units

more frequently.

We calculate heat rates in both ways: combining information from CEMS and EIA or using EIA information alone. They are labeled as "EPA-EIA" and "EIA", respectively. Table 3.4-4 lists summary statistics about capacity, heat rates and emission rates of fossil fuel generators in ERCOT. On average, coal units are larger than natural gas units. The heat input needed to produce one MWH of electricity is almost the same for two types of units, while coal units usually are dirtier than natural gas units. Generally, the average coal unit emits around 1 metric ton of CO2 for every 1 MWH electricity generated, and the Co2 emission from an average gas unit is around 30% lower. The average gas unit produces negligible SO2 when compared with coal units. The emission rate of Nox from both types of units is almost the same. On average, more electricity is used in the facility by coal units than by gas units. This could be seen by comparing the adjusted capacity for both types of units. A drop from around 600 MW to around 500 MW happens for coal units, while the change for gas units is moderate.

4.4.3 Transmission Network

In the zonal market design, ERCOT's power network is modeled as four zones connected by five commercially significant constraints (CSC): West-to-North, North-to-West, South-to-North, North-to-South and North-to-Houston. ERCOT utilizes the dispatch software to arrange the production of the generation units in the network to meet the demand and manage the congestion on those five interfaces.

Table 3.4-5 shows the hourly power flow in these five CSCs during the research period. The largest energy flow happens between the Houston zone and the North zone. On average, the Houston zone receives an amount of 1.6 GWH of net flow of power from the North zone. The electricity exchange between West zone and North zone varies between the first half of year and the second half of year. In the period from Jan 2010 to Jun 2010, there are net flows of 561 MWHs hourly from West zone to North Zone, while from Jun 2009 to Dec 2009, the hourly exchange of power is only around 40 MWH on average. One possible reason may be the hourly wind generation is larger in the first half of the year. The distribution of monthly wind generation is summarized in the next subsection.

As more wind farms, mostly in the West Zone, were added to the system, the transfer of wind generation from the West to North Zones put great pressure on the congestion management for the two CSCs connecting these two zones. In 2010, there were a large number of involuntary curtailments of wind generation in the West Zone due to the system reliability and congestion requirements.

Figure 3.4-1 shows the wind generation curtailment in West zone in 2009 and 2010 (Figure 28 in Potomac Inc., 2010). The frequent curtailment of wind generation puts a restriction on our analysis. In our analysis, we simulation the expansion of wind capacity by inflating the implied wind generation we observed from the ERCOT reports. If the wind generation from the West Zone was frequently curtailed, the observed wind generation understates the available generation in that timeframe and thus all the related variables based on this variable will also be affected.[17]

4.4.4 Wind Expansion

Wind capacity in Texas has grown rapidly in recent years, along with several other states in US. Currently, Texas has the largest wind fleet among US states with total capacity of 12,212 MW at the end of 2012, followed by California (5,549 MW) and Iowa (5,137 MWH) (AWEA, 2013). The extent of wind penetration, measured by energy, in the Texas grid is also the highest among all the power grids in US. The annual share of wind generation in ERCOT is 2.9% in 2007, 4.9% in 2008, 6.2% in 2009, 7.8% in 2010 and 8.5% in 2011 (ERCOT, 2007-2011). Though the annual wind generation continues to increase in US, the share of wind among all the electricity generated was still less than 3% in 2011 (EIA, 2012). Clearly, ERCOT leads the nation both in terms of total installed capacity and the penetration rate in the power grids.

Figure 3.4-2 shows the expansion of wind capacity in ERCOT in recent years. ECROT started with a negligible amount of wind capacity at the beginning of last decades and has experienced a rapid expansion of wind capacity even since, especially after 2005. The newly installed wind capacity topped at 2008 with 2,760 MWH capacity installed in the single year. The newly installed capacity number flatted since 2010 and only 686 MWH (around one quarter

[17] When we read the simulation results, this issue should be kept in mind.

of the peaked installment) capacity was added in that year.

With the wind resources concentrating on the west and coastal area of the state, the spatial distribution of wind farms, at 2012, is that 7,531 MW has been installed in the west zone, 2,075 MW is located in the south zone and a smaller establishment of 232MW exists in the north zone.

The change in the pace of expansion is most likely driven by the evolution of government policies and programs, such as Energy Policy Act of 2002, 2005 and state level programs, such as Renewable Portfolio Standards (RPS). These programs provide several kinds of production based subsidy to the wind farms, i.e. the equivalent subsidies induced by RPS are estimated to range from $5 per MWH to $50 per MWH in US (Wiser 2008). And the federal Production Tax Credit initiated by Energy Policy Act is to grant 2.2 cents per KWH ($22 per MWH) as tax credit to renewable energy producers for the first 10 years of operation. Cullen estimates that the PTC alone accounts for almost 40% of the wholesale price of electricity in ERCOT during his research period (Cullen,2010). In more recent years, the market price of electricity in ERCOT has been brought down further by the lower price of natural gas. PTC alone accounts for an even share of the electricity price.[18] The other factors include the construction cost for the wind farm decreases dramatically due to the technology advances (Bolinger and Wiser, 2009).

As a type of intermittent energy, the substitution ability of wind energy is limited by the wind flow patterns, especially the correlation between electricity demand and wind flow. Figure 3.4-3 shows the pattern of hourly electricity demand and wind generation in ERCOT. Clearly, the correlation is negative. When the demand peaks around noon, the wind generation is the lowest of the day. This negative correlation between demand and wind generation extensively exists in other power market as well. This pattern suggests that wind generation could substitute the generation from different generators depending on what type of generators are the marginal units in a given hour. If coal generators are the marginal suppliers in most of the time, one unit of wind generation will have a larger emission reduction effect than when relatively

[18] The average electricity price in ERCOT reported in Potomac (2010) is $34.03 per MWH in 2009 and $39.40 per MWH in 2010.

cleaner natural gas generators serve as the marginal suppliers in the most of the day.

4.5 Results and Discussion

4.5.1 Baseline Comparison

Since we use a simplified dispatch model to model the possible outcomes in different scenarios, we first need to compare the simulation results with the observed data. As discussed before, we have difficulties to construct heat rates for each unit, especially for combined cycle units and combined heat and power units. CEMS does not separate the fuel input used for the heat purpose or electricity-generation purpose, thus the heat rate constructed by dividing the heat input by the gross generation (net generation) will overestimate the heat rate for combined heat and power input. At the same time, the fuel consumption reported also its own problem due to possible under reporting the fuel input. The direct consequences of using these two heat rates are that the production of gas generation will be overestimated in the region with substantial existence of combined heat and power units if the lower "EIA" heat rates are used. On the other hand, the gas generation in those areas will be underestimated if "EPA-EIA" heat rates are used in the model.

Before we discuss the simulation results from the baseline in which there is no carbon tax and the wind penetration is at current level, several challenges should be pointed out. First, we do not have a real time net load for each unit in CEMS reports. Though we have supplemented CEMS observations with information from EIA, the potential problems still could be large with the reasons discussed above. Secondly, the detailed, hourly data on wind generation at zonal level during our research period are not available in this research. The way we used here, disaggregating ERCOT wide hourly wind generation to zonal generation according to the share of installed capacity, is a crude estimation of the real hourly zone wind generation. Third, the transmission flows simulated in this simple model may significantly differ from the real inter-zonal energy flows. One source of bias is the unavoidable difference between the dispatch model used here and the real dispatch model used by ERCOT. The other source of bias is how to deal with the zonal congestion. ERCOT will use the dispatch model to solve the allocation of generation before the generation resources are deployed. If any congestion detected by the model *a priori*, ERCOT will choose different zonal clear prices to solve the congestion before

it is realized. In this way, the occurrence of zonal congestion could be maximally avoided. However, the real electricity flow in the transmission interface may be significantly lower than the physical transmission limit. Thus no congestion happens in the pre-defined congestion intervals.[19] In our model, there are no pre-designed zonal congestion costs to eliminate the possible congestion in any hour. If a region imports more electricity from other regions, the congestion management will increase the receiving price of generators within in this region to avoid the transmission congestion and thus increase the production of electricity in this region. Without these measures, the production would be less in this rejoin. As a result, our model tends to have more generation from the imported zone if everything else is equal.

These factors could interact with each other. For example, it is more likely for our model to have more generation from an imported region. This problem becomes more troublesome if the imported zone, Houston zone in ERCOT, happens to have dozens of combined heat and power units whose heat rates are constructed with limitations. If heat rates of these units are overestimated, the consequence will be less generation from Houston zone. Compounding with the congestion issue, the overall impact could still be less generation from Houston zone if the effect of overestimated heat rates dominates the impact of congestion management. On the other hand, if heat rates for combined heat and power units are underestimated, the combining effect will cause more generation from Houston Zone.

Table 3.5-1 reports the output from baseline simulations with different heat rate calculations. The results confirm our speculation about the potential effect of poorly measured heat rates. In the scenario with "EPA-EIA" heat rates, our simulation model produces a profile of zonal generation different from the observed profile. The most obvious difference is that the gas generation from Houston zone is much less than the observed gas generation in Houston zone. As argued before, "EPA-EIA" heat rates tend to overestimate the heat rates for combined heat and power units. The resulting lower gas generation forces the other two surrounding zones, north zone and south zone, to produce more gas generation to fill the gap. With "EIA" heat rates, the effects are completely opposite. The gas generation in Houston zone is more than the observed generation due to the underestimation of the heat rates of combined heat and power

[19] ERCOT dispatches generation resources at 15-minutes interval.

units. As a result, the gas generation in North zone and West zone decreases.

There are no significant ERCOT wide differences in terms of the allocation between coal generation and gas generation in both cases. The coal generation is generally in line with the observed coal generation at both zonal level and ERCOT wide. The possible reason is that the coal units are more likely to be the base load generators. Even if the heat rates of some gas units are underestimated, coal units will still be base load producer without carbon tax. Though the zonal profile of gas generation differs from the observed profile, this difference is not large from the ERCOT perspective.

From the perspective of CO2 emission, the story is slightly different. In general, the profile of generation determines the profile of CO2 emission. While in some case, it is difficult to rationalize the results. For example, the gas generation from South zone in "EIA" heat rate scenario is close to the observed generation in South zone. While CO2 emission from the model is significantly less than the observed emission. It suggests that the CO2 emission reported in CEMS may be a poor measure about the real emission. At this stage, we do not have a better way to measure CO2 emission. Fortunately, at the system level, the bias of CO2 is not as severe as the zonal emission.

The results from baseline simulations reveal the impacts caused by the data limitations. Both methods fail in giving us a reasonable approximation to the real power market of ERCOT in the research period. If we accept these limitations associated with the availability of the public data set, the simulation with "EIA" heat rates has relative merits over the simulation with "EPA-EIA" heat rates because they both do well to capture the coal generation and the former model has the overall better estimation about CO2 emission, which is the key focus of this study. The other reason is that the focus of this paper is how the carbon tax and wind penetration affect the allocation of generation and resulted CO2 emission. The most important re-allocation of generation will happen between coal units and gas units, not within gas units. The statistics of CO2 emission rate reported in Table 3.4-4 show that the gas units, on average, has less CO2 emission and the standard deviation of CO2 emission rate is also smaller than that of coal units. With the current model, we may get the zonal allocation of gas generation different from the real allocation in the similar scenarios. The likelihood to a

dramatic difference at the system level may not be as large as that.

4.5.2 Simulation Results

The baseline comparison shows some difference between model outcomes and the observed data. At this stage, we do not have more appealing ways to eliminate the difference and proceed to apply the models to different policy scenarios with the existing limitation.[20] We separate all the simulations into two sets. In one set of simulations, we use heat rates calculated from the combined information from CEMS and EIA, so called "EPA-EIA" simulation. In other set of simulations, we use heat rates calculated with the information only from EIA, so called "EIA" simulation. We will focus more on the results from the "EIA" simulations, and leave some results from "EPA-EIA" simulations in the appendix.

Figure 3.5-1 shows the share of gas generation among the total fossil fuel generation. The x axis shows the simulated percentage of wind power serving the system relative to the current actual level. For example, the 20% label in the axis does not mean the wind penetration rate is 20% of the total generation capacity in ERCOT. Instead, it means the wind capacity was at 20% level of current wind capacity. Thus, 200% means in the simulation we assume the total capacity of wind turbines was doubled. As discussed in the introduction, the carbon dioxide tax decreases the relative gap of the marginal cost between the coal units and gas units. Although the fuel cost of gas generators are much higher than that of coal generators, gas units produce only around 70% of carbon dioxide emission relative to coal units. Also, the fuel cost of both types of generators varies substantially within each category, depending on the vintage of the generators and the turbine technologies. As the carbon tax increases, the dirtiest coal units will be replaced, first with the cleanest gas units. Then the cleaner gas units replace the average coal units and eventually all the coal units will be pushed backward in the ladder of supply curve.

Given the fossil fuel units covered in the dispatch model, the share of electricity from gas units will decrease from around 55% to around 52% when wind capacity increases to 2 times of

[20] We also tried to use gross load instead of net load in the dispatch model. In both cases, we still find significant difference between the model and reality.

current level if there is no carbon tax. Once the carbon tax was introduced, the baseline share of gas generation increases about another 25% to around 80% given the current wind capacity. The expansion of wind capacity does not cause the same significant change as the case of no carbon tax. The share of gas generation is quite stable and even slightly increases when the carbon tax is high enough. This suggests that in present the wind power tends to replace the gas generation even the mismatched pace between system load and wind power. As carbon tax increases, the structure of marginal cost curve as showed in the introduction changes as well in the direction to move gas generation forward and push the coal generation backward in the supply curve. This shift tends to allow wind power to replace more coal generation than it did currently and thus it tends to raise the share of gas generation. The rough calculation in Table 3.1-2 also confirms this reasoning.

Since the structure of the fossil fuel generation fleet varies among four zones, we also show the zonal share of gas generation in Figure 3.5-2.

Since the majority of wind farms were added in the west zone, the increase to current wind capacity leads to the biggest change in the share of gas generation in the west zone. An increase of wind capacity beyond the current level has less effect on the gas share. Outside the western zone, zonal changes patterns generally follow the system-wide pattern. The carbon tax leads to significantly higher shares of gas generation, and that increase eventually diminishes due to the fact that when the carbon tax is high, enough (such as 70$ per ton of CO2), coal units have been already pushed to the far end of the supply order. At that point there is only a small opportunity for extra wind to substitute for the marginal coal units. Currently, gas units are more likely to serve as the marginal generation resource. The impact of wind capacity expansion is not as unambiguous as the role of carbon tax. The increase of wind power decrease the demand for fossil fuel generation, while the relative replacement of coal generation versus gas generation depends on the marginal cost structure and the pattern of wind generation and system load. If the wind tends to blow when coal units are the marginal generation source, then the wind power will replace more coal generation, and vice versa. As shown in the table 3.1-1, the majority of generation capacity in the first quartile of 60 GWH changes from coal units to gas units as carbon tax increases.

The results in terms of reduction of carbon dioxide emissions in different scenarios are shown in Figure 3.5-3. The Y axis is the percentage *reduction* relative to the baseline emission, e.g. the current level of carbon dioxide emissions. As shown in the Figure 3.5-1, the carbon tax causes production to migrate to cleaner gas units. Consequently, emissions of carbon dioxide decrease as the carbon tax increases. Given the structure of present generation fleet, emissions reduce nonlinearly as the carbon tax rate increases. The first two increases of 25$ per ton of Co2 lead to an almost 15% of reduction in emission. The further increase of 20$ to 70$ per ton of Co2 only leads the emission to reduce by less than 5%. This nonlinear relationship between a carbon tax and emissions is likely shared by other power markets given the heterogeneous marginal costs of fossil fuel units, and the key opportunity to switch from higher costs coal generation from lower cost, but dirtier coal production. Once a threshold carbon cost is reached, large scale substitution between coal and gas becomes economic. Beyond that threshold, additional carbon charges produce less dramatic reductions as most of the benefit from fuel switching has already been realized at that point. The specific threshold of this nonlinearity will depend upon the specific composition of the fossil fuel generation fleet.

The central purpose of this Chapter is to explore the interaction of these two policy instruments when applied simultaneously. Table 3.1-2 shows the heterogeneous impacts of wind generation at several carbon tax rates for two typical days. Similarly, Table 3.5-2 shows the reduction matrix of CO2 emission among a subset of simulated scenarios. There are several findings to highlight. First, the expansion of wind capacity always leads to a lower emission of CO2 at all carbon tax rates. We only present three possible wind capacity scenarios here (No wind, Status Quo wind capacity and Double Current wind capacity).[21] The increase from zero wind to status quo and the increase from status quo to double wind capacity both cause the carbon dioxide emission decreases at all simulated carbon tax rates. Second, the emission reduction effects of wind capacity expansion are heterogeneous. At the lower carbon tax rates, the wind expansion enhances the emission reduction effect of carbon tax. Comparing the consequence of carbon tax at the no wind and status quo scenarios, the simulated CO2 emission will decrease by 11.2% at the carbon tax rate of 25$ per ton of CO2 relative to the no carbon

[21] The full matrix is listed in the appendix.

tax case. The corresponding reduction is around 18.4% in the scenario with status quo wind capacity, the current wind farms help to reduce 7.2% more of carbon dioxide from the fossil fuel generators. This enhancing effect tops out at 7.4% when the carbon tax rates are in the range of 30$ - 40$ per ton of CO2 at the status quo wind capacity. If the wind capacity doubles, similarly, this enhancing effect is maximized at 4.9%, using status quo as the baseline, with the tax rates in the range of 20$ - 25$ per ton of CO2.

These findings reflect the change in the utilization of fossil fuel generators, i.e. the share of gas generation in different scenarios discussed earlier. When the carbon tax is moderate, only a subset of gas generators can serve as the base load units and some of coal units would be pushed to work as marginal generating units and not be completely pushed to the very end of the supply curve and almost have no change to be used in any demand levels. Thus when wind blows, more coal generation will be substituted and more reduction will be seen. When the carbon tax is very high, those coal units might be completely useless at almost all system demands. More wind generation solely substitutes gas generation and the emission reduction becomes smaller consequently.

Because of the transmission constraints, it is useful to look at the emission reduction effects in different zones. Table 3.5-3 shows the similar comparison as Table 3.5-2 for the four zones in ERCOT. Since the majority of installed wind capacity is located in west zone and the fossil generation fleet there is also the smallest among four zones in ERCOT, it is not surprising to see the impact of wind energy is felt most in the west zone. The relative reduction of carbon dioxide emission between the status quo wind installation and the no wind scenario is on the order of 25% at all hypothetical carbon tax rates. The impact on emission reduction becomes slightly larger when the wind capacity was further doubled. Unlike the system-wide pattern, the effect fluctuates as the carbon tax increases without a maximum. The most significant impacts of carbon tax are found in the north zone. About half of the coal generators (measured by the installed capacity) are located in the north zone, thus if the structure of the supply curve is reshuffled by the imposed carbon tax, it is natural to see the biggest reconstruction in the north zone. It is somewhat surprising to see the relative reduction of carbon dioxide emission in the Houston zone is as similar as that in the south zone, though the coal generation

resource in the Houston zone is almost three times of that in the south zone. Recall that in the calibration section, our model produces the significantly different generation portfolio from the actual production plan because of the combination effect of transmission constraints and the measurement of heat rates. In our model, the fossil fuel generation, especially gas generation, in the Houston zone has been over-used. This limits the possibility that in scenarios with a carbon tax the over-used gas units in the Houston zone cannot grasp the chance to substitute more generation from the replacement of coal generation.

The extra emissions reduction brought about by wind energy in the three non-western zones are relatively moderate in terms of percentage reduction. These effects peak at the south zone and north zone within the hypothetical carbon tax scenarios. In the Houston zone, it seems that doubling wind capacity could bring more reduction if higher carbon tax rates were imposed. Table 3.5-4 shows the carbon emission matrix in physical units (10ˆ6 tons of CO2). The majority of reduction of carbon dioxide comes in the north zone and south zone if carbon tax were imposed. In our hypothetical scenario with a 70$/ton of CO2 tax, our model predicts there will be 41.7 million tons of carbon dioxide avoided in the north zone alone, followed by the reduction of 19.4 million tons in the south zone and 16.9 million tons in the Houston zone, contrasted with just over 1 million tons in the west. This variation of emission reduction could be partly explained by the portfolio of generation resources. Remember that the north zone has the largest fossil fuel generation fleet and the highest power demand among all four zones. The Houston zone and south zone have smaller generation fleets with the capacity only slightly more than half of that in the north zone. The west zone has the smallest generation fleet with less than one tenth of that of the north zone.

Although other pollutants, such as SO2 and Nox, have not been explicitly modeled in the simulation, the output from the dispatch model still allows us to calculate their emission in different scenarios. Table 3.5-5 and Table 3.5-6 show the counterpart results for SO2 and Nox emission reduction in physical units, respectively. Since the sulfur content in natural gas is far less than that in the coal, SO2 emission is closely correlated with the coal generation. North zone has the largest coal generation fleet within ERCOT, thus it is natural to see the reduction of SO2 is biggest in the north zone. The additional reduction brought about by the expansion

of wind energy does not follow the same pattern. At any given carbon tax rate, the effect of wind expansion is biggest in the south zone though the generation fleet in the south zone is not the largest.

The relative reduction of Nox emission is showed in the Table 3.5-6. The Nox emission rates of both types of generator are similar with gas units having slightly smaller weighted average emission rates. Thus, the scale of reduction in Nox emission is significantly smaller than the reduction of SO2 emission. North zone still have been the most affected zone since it has the biggest coal generation fleet. Wind expansion will lead to the largest extra reduction of Nox in the south zone.

For these three typical pollutants, we find that the extra reduction of emission with wind expansion is largest in the south zone. This is not a coincidence because of two reasons. Firstly, there is a substantial existence of wind farms, roughly 20% of the whole wind capacity in ERCOT. When we hypothetically increase the wind penetration in the market, we will see an equivalent amount of fossil fuel generation will be replaced in the south zone. Secondly, though the extra wind from the west zone could be imported to the north zone, the substitution potential is limited by the transmission capacity between these two zones.

Figure 3.5-4 shows a contour graph of CO2 emission on the plane of carbon tax and wind expansion. To construct this graph, we need interpolate CO2 emission reduction at higher resolution beyond the hundred plus simulations in our model. The method we use to interpolate using a linear weighting scheme.[22] With the interpolation, we could draw several iso-reduction lines in terms of the extent of emission reduction as in Figure 3.5-4. There are several interesting results to highlight. First, there is obvious trade-off between carbon tax and wind expansion, reflected in the negative slope of the iso-reduction lines. For example, if the policy target of CO2 emission reduction is to cut a 10% from the current level, the iso-reduction line of 10% reduction implies that a sole carbon tax of 15$ per ton of CO2 could be imposed or a combination of carbon tax at around 7.5$ per ton of CO2 and doubling the fleet of wind turbines is also effective. Second, the iso-reduction line at a low level of emission reduction is

[22] For a policy scenario defined by the pair of $(tax_c, wrate)$, we first find out the cell defined by four pairs in our simulations, which covers this specific scenario. Then we use a linear weighting scheme to find out the possible reduction of CO2 emission in this scenario.

almost linear, but eventually becomes highly nonlinear at higher reduction levels. This can be seen in this graph with the isocline of 35% emission reduction. In contrast with the target of 10% emission reduction, if a carbon tax would be used to achieve a 35% reduction, the possible tax rate is beyond the range of carbon tax considered in our simulation, which is 70$ per ton of CO2. However, if the wind capacity was doubled, the carbon tax needed to imposed is less than 45$ per ton of CO2. This nonlinearity can also be explained by the supply curve of fossil fuel generators in the introduction section. At a higher carbon tax, the dirty units have been completely pushed backward such that dirty generators seldom have the chance to serve the load. In those situations, a direct substitution away from fossil fuel generation, which is caused by more wind energy, is more effective.

Though the contour graph shows some combinational values of both instruments, we need to point out here that our study is a short-run analysis. In reality, if a cap-and-trade program or a carbon tax is launched, it will change incentives for investment. Cleaner gas power plants will be added into the generation fleet, thus the replacement of gas generation to wind generation will not have the same effect as that of substituting dirty coal generation. At the same time, the carbon tax could push up the market price received by wind producers. That will increase the economic prospects of wind farms independent of the benefits offered by government programs, such as PTC and RSP.

Figure 3.5-5 shows the distribution of hourly market price in four scenarios, no carbon tax, 25$/ton of CO2, 50$/ton of CO2 and 70$/ton of CO2. If a PTC and RSP payment are not included, the market price could be thought as the price received by owners of wind farms absent these subsidies. The EIA (2013) estimates that the levelized cost of new wind turbine entering into service at 2018 is at 86.6$ (8600 Cents) per MWH, which does not include possible tax credits provided by various government programs.[23] Figure 3.5-5 shows that at the status quo, the hypothetical carbon tax should be high enough, such as a rate around 70$/ton of CO2, to free wind investment from government subsidiary programs. The similar distribution of market price when the wind capacity is doubled is shown in Figure 3.5-6. Compared with the distributions in Figure 3.5-5, it is clear that the larger the existing installed wind capacity

[23] The link to the summary of the report is http://www.eia.gov/forecasts/aeo/electricity_generation.cfm.

is, the higher carbon tax is needed to support the independent investment on wind turbines.

Several summary statistics about the operation of representative coal and gas unit are listed in Table 3.5-7.[24] The statistics we considered are the total working hours of the unit in each scenario and the corresponding share of hours in which the unit is operated. Two economic statistics are also summarized: the average market price when the unit is running and the average operational profit calculated by subtracting the fuel cost and CO2 emission cost from the received the market price.[25] It is clear that as carbon tax increases, the representative coal unit eventually becomes the marginal generator, which is suggested by the share of running time decreases and the average market price required to be profitable increases. In this process, the operational profit in running hours also decreases despite the price received by the unit increases dramatically due to the increasing CO2 emission cost. The representative combined cycle gas unit moves in the opposite direction. As carbon tax increases, the unit is almost running at full time, while it is running at a little more than 30% of the time in the base line without any carbon tax. The average operational profit also increases as carbon tax increases.

Though the correlation between wind generation and electricity demand in ERCOT is negative, the extra wind generation introduced into the market always has a pressure on the market clear price. However, the likely decrease in the price is moderate as showed in Table 3.5-7.[26] For both types of generation units, the impacts of doubling the current wind capacity seems quite mild for the representative units. These findings suggest that a program like a carbon tax or cap-and-trade program will have more impacts on determining the economic prospects of fossil fuel generation technologies than the expansion of wind energy in ERCOT.

[24] We define a representative generation in each fuel type as the unit with median "EIA" heat rate. For the gas unit, we limit it to be a combined cycle gas unit.

[25] If the gross revenue is negative in a given hour, it means this unit is not running. This hour will not be included in the calculation of average gross revenue.

[26] CEMS units may not cover the price – setting small gas generators because only relatively big generators are included in the program.

4.6 Conclusion

As the atmospheric concentrations of carbon dioxide passed 400ppm in May 9^{th}, 2013, it becomes more urgent to tackle the climate change issue. Though there is no universal consensus on the best policy measures to curb the greenhouse gas emission, the promotion of renewable energy and the application of market-based instruments, such as carbon tax or cap-and-trade program, have been adopted in several countries and regions to achieve individually-set reduction goals. Both types of policies reduce greenhouse gas emission via different channels. In this paper, we use an economic dispatch model to evaluate the interaction of these two policies, a carbon tax and wind energy expansion, in terms of CO2 emission reduction with the observational data from Texas ERCOT, a power market with sufficient renewable energy penetration, especially wind energy.

After the careful calibration of the model, the simulations carried out show that both policies lead to significant changes in the composition of fossil fuel generation. The imposed carbon tax will greatly change the favorability of gas generators and increase the share of gas generation, thus reduce CO2 emission accordingly. Depending on the composition and technical specification of the current fleet of fossil fuel generators in each zone, there are zonal spatial variations. As the carbon tax increases, the marginal effect on emission reductions is eventually diminished. Expansion of the fleet of wind turbines in ERCOT also has the similar effect on emission reductions. Since the wind energy will still account for a small share in the generation mix, even doubling the capacity will only achieve 5% or so reduction of CO2 emission. Depending on the rate of carbon tax imposed, the effect of wind capacity expansion ranges from 4.1% to around 4.9%, a 20% variation. A closer look at the contour graph implied by the results from our simulation reveals that the single carbon tax may be more effective when the reduction target is moderate and the combination of carbon tax and renewable energy promotion may achieve an ambitious reduction goal with lower costs.

There are several caveats about this paper. First, the gross-to-net generation adjustment in this paper is not ideal. The calculated residual demand (net of fossil generation) in this paper suggests our approximations may represent the real adjustment extremely poorly under some

circumstances. Poor adjustment could potentially put the certain generators in wrong places in the supply order and consequently bias the results. Secondly, this simple cost-minimization dispatch model without demand response poorly represents the real market. The implicit assumption behind the cost minimization is that the market is perfectly competitive and this may not be true in the real world Borenstein et al. (2002) found that the market power of the large producers substantially causes the equilibrium price deviates from the prices in the perfect competition case in California power market. The startup cost of wind turbine is not significant compared with the startup cost of a coal generation, if some producer has a generation fleet consisting of fossil fuel units and wind turbines, the company may intentionally cut off the wind generation to boost the market clear price for his production from fossil fuel units. The inelasticity assumed in this paper may not cause substantial bias in the short run analysis like this paper. It is critically important in the long run analysis such as investment decisions and then can affect the long-run performance of either the carbon tax or the renewable energy.

Bibliography

[1] AWEA, 2013. "AWEA U.S. Wind Industry Fourth Quarter 2012 Market Report", January, 2013. American Wind Energy Association.

[2] Bolinger, Mark, and Ryan Wiser. 2009. "Wind Power Price Trends in the United States: Struggling to Remain Competitive in the Face of Strong Growth." *Energy Policy* 37 (3): 1061–1071.

[3] Borenstein, Severin, James B Bushnell, and Frank A Wolak. 2002. "Measuring Market Inefficiencies in California's Restructured Wholesale Electricity Market." *American Economic Review* 92 (5) (December): 1376–1405. doi:10.1257/000282802762024557.

[4] Bushnell, James. 2010. "Building Blocks: Investment in Renewable and Non-renewable Technologies." *Harnessing Renewable Energy in Electric Power Systems: Theory, Practice, Policy*: pp159.

[5] Bushnell, James B., and Catherine Wolfram. 2005. "Ownership Change, Incentives and Plant Efficiency: The Divestiture of US Electric Generation Plants." *working paper*

[6] Bushnell, James, and Yihsu Chen. 2009. "Regulation, Allocation, and Leakage in Cap-and-Trade Markets for CO2." *Resources and Energy Economics*, forthcoming

[7] Campbell, Arthur. 2008. "Hot Air? When Government Support for Intermittent Renewable Technologies Can Increase Emissions." *working paper*

[8] Cullen, Joseph A. 2010. "Measuring the Environmental Benefits of Wind-generated Electricity." *American Economic Journal: Economic Policy*, forthcoming.

[9] Energy, G. E. 2008. *Analysis of Wind Generation Impact on ERCOT Ancillary Services Requirements*. Prepared for the Electricity Reliability Council of Texas. Schenectady, New York: GE Energy.

[10] EIA, 2012. "US wind generation increases 27% in 2011", March 2012. <http://www.eia.gov/todayinenergy/detail.cfm?id=5350>

[11] EIA, 2013. "Levelized Cost of New Generation Resources in the Annual Energy Outlook 2013", January 28^{th}, 2013. <www.eia.gov/forecasts/aeo/electricity_generation.cfm>

[12] EPA, US. 2013 "Standards of Performance for Greenhouse Gas Emission for New Stationary Sources: Electricity Utility Generation Units" March 27^{th}, 2013.

[13] ERCOT, 2007-2011, "15 minutes interval generation by fuel report", <planning.ercot.come/reports/demand-energy/>

[14] Harrison Fell and Joshua Linn, 2012. "Renewable Electricity Policies, Heterogeneity, and Cost Effectiveness",*working paper*

[15] FERC, 2012. "Form No. 714 - Annual Electric Balancing Authority Area and Planning Area Report", August, 2012.<http://www.ferc.gov/docs-filing/forms/form-714/data.asp>

[16] Fowlie, Meredith L. 2009. "Incomplete Environmental Regulation, Imperfect Competition, and Emissions Leakage." *American Economic Journal: Economic Policy* 1 (2) (August 1): 72–112. doi:10.2307/25760041.

[17] Kaffine, D. T., B. J. McBee, and J. Lieskovsky. 2011. "Empirical Estimates of Emissions Avoided from Wind Power Generation." *working paper (under review)*,

[18] Nordhaus, William D. 2010. "Economic Aspects of Global Warming in a post-Copenhagen Environment." *Proceedings of the National Academy of Sciences* 107 (26) (June 29): 11721–11726. doi:10.1073/pnas.1005985107.

[19] Novan, Kevin M. 2010. "Shifting Wind: The Economics of Moving Subsidies from Power Produced to Emissions Avoided". *Working paper*.

[20] Potomac Economics, 2011. "2010 State of the Market Report for the ERCOT Wholesale Electricity Markets"

[21] Valentino, Lauren, Viviana Valenzuela, Audun Botterud, Zhi Zhou, and Guenter Conzelmann. 2012. "System-Wide Emissions Implications of Increased Wind Power Penetration." *Environmental Science & Technology* 46 (7): 4200–4206.

[22] Wiser, Ryan. 2008. "Renewable Portfolio Standards in the United States-A Status Report with Data Through 2007." http://escholarship.org/uc/item/1r6047xb.pdf.

Table 3.1-1 Composition of Generation by Quartile (%)

	No Carbon Tax		**25$/CO2 ton**		**50$/CO2 ton**		**75$/CO2 ton**	
	Coal	Gas	Coal	Gas	Coal	Gas	Coal	Gas
1st Quartile	100	0	23	77	0	100	0	100
2nd Quartile	11	89	30	70	14	86	4	96
3rd Quartile	5	95	56	44	73	27	38	62
4th Quartile	0	100	0	100	21	79	57	43

Table 3.1-2 Carbon Dioxide Reduction in Typical Summer/Winter Day (1,000 tons)

	Present			25 $/Co2 ton			50 $/Co2 ton			75 $/Co2 ton		
	No Wind	Wind	Diff	No Wind	Wind	Diff	No Wind	Wind	Diff	No Wind	Wind	Diff
A typical day in Jun - Aug												
Coal	458	454	-4	383	357	-26	296	266	-30	156	129	-27
Gas	319	290	-29	348	325	-23	388	375	-13	471	453	-18
Total	776	743	-33	728	682	-46	685	637	-48	633	583	-50
A typical day in Dec - Feb												
Coal	432	432	0	297	247	-50	143	97	-46	47	27	-20
Gas	199	174	-25	257	247	-10	327	320	-7	385	362	-23
Total	633	608	-25	552	497	-55	474	420	-54	433	389	-44

Table 3.4-1 Hourly Demand in ERCOT (2009.6-2010.6)

Zone	Mean	Std. Dev.	Min	Max
2009.6 - 2009.12				
West	2188	333	1591	3138
North	14382	3917	7533	25871
South	10810	2872	5642	17929
Houston	10445	2666	5914	17630
ERCOT	37825	9577	21390	63400
2010.1 - 2010.6				
West	2407	363	1734	3591
North	13344	3318	7430	24018
South	9721	2316	5806	16613
Houston	9606	2108	5968	17152
ERCOT	35078	7863	21770	60789
2009.6 - 2010.6				
West	2288	364	1591	3591
North	13907	3691	7430	25871
South	10311	2687	5642	17929
Houston	10060	2462	5914	17630
ERCOT	36566	8938	21390	63400

Table 3.4-2 Fossil Fuel Generation Portfolio in ERCOT from CEMS (Unit: MW)

Zone	**Coal**		**Gas**			**Gas (%)**
	Stream	Total	Stream	Combined Cycle	Combustion	
West	1	23	2	10	11	
	702	3262	675	1628	959	82
North	15	83	35	37	11	
	10543	21035	9361	10503	1171	67
Houston	4	75	11	43	21	
	2685	15691	4005	9629	2057	85
South	12	79	15	41	23	
	6791	16224	4288	10883	1053	70
ERCOT	32	260	63	131	66	
	20721	56212	18329	32643	5240	73

Table 3.4-3 ERCOT zonal fossil fuel generations (in MWH)

Variable	Mean	Std. Dev.	Min	Max
		Coal		
ERCOT	13302	1849	6698	17737
West	400	235	0	654
North	6877	845	3427	8650
Houston	2049	435	578	2485
South	3976	811	1231	6061
		Gas		
ERCOT	15821	7295	4205	39956
West	549	533	0	2482
North	5556	3084	159	16197
Houston	5138	1602	2813	11264
South	4577	2380	801	10745
		Fossil Fuel		
ERCOT	29122	8452	12341	54153
West	949	659	0	3127
North	12433	3547	4956	23678
Houston	7187	1820	3741	13700
South	8553	2801	2618	15755

Table 3.4-4 Summary Statistics of Fossil Fuel Generators

		CEMS		**EPA-EIA**		**EIA**	
	Unit	Coal	Gas	Coal	Gas	Coal	Gas
Number of Units	#	32	260	32	255	32	255
Capacity	MWH	596(179)	216(153)	488(153)	202(147)	488(153)	202(147)
Heat Input	mmBTU/MWH	9.81(0.51)	10.99(3.68)	11.17(2.66)	12.88(15.16)	12.14(153)	10.96(5.58)
Co2 Emission Rate	tons/MWH	1.04(0.04)	0.69(0.76)	1.19(2.66)	0.74(0.81)	1.19(2.66)	0.74(0.81)
So2 Emission Rate	lbs/MWH	2.68(1.82)	0.01(0.01)	2.90(1.91)	0.01(0.02)	2.90(1.91)	0.01(0.02)
Nox Emission Rate	lbs/MWH	1.27(0.63)	1.25 (1.90)	1.41(0.66)	1.53(2.57)	1.41(0.66)	1.53(2.57)

1. The statistics are in the format of "mean(standard deviation)"
2. There are 5 gas units which report almost zero net generation and thus they are excluded from the sample used in our model.
3. The only difference between EPA-EIA and EIA is the way to construct the heat rates. We do not change other variables.
4. These averages are arithmetic averages. If using unit capacity as weights, the weighted average of Nox emission rate for gas units is still lower than that for coal units.

Table 3.4-5 Transmission Flow and Physical Limits

	Energy Flow				Transmission Limit			
Line	Mean	Std.Dev	Min	Max	Mean	Std.Dev	Min	Max
		2009.6 - 2009.12				2009.6 - 2009.12		
W_N	39	309	-797	982	1019	262	499	2506
N_W	-39	309	-982	797	790	150	46	2314
S_N	-437	326	-1313	732	711	231	130	1238
N_S	437	326	-732	1313	1284	132	771	1528
N_H	1792	652	-619	3266	3248	135	2186	3589
		2010.1 - 2010.6				2010.1 - 2010.6		
W_N	561	1001	-1532	2374	2347	275	602	6944
N_W	-561	1001	-2374	1532	1610	585	0	7170
S_N	-152	358	-1264	760	992	164	0	1352
N_S	152	358	-760	1264	1376	128	184	1610
N_H	1445	686	-730	2984	3176	265	1824	5577
		2009.6 - 2010.6				2009.6 - 2010.6		
W_N	278	761	-1532	2374	1628	714	499	6944
N_W	-278	761	-2374	1532	1166	579	0	7170
S_N	-306	369	-1313	760	840	247	0	1352
N_S	307	369	-760	1313	1326	138	184	1610
N_H	1633	690	-730	3266	3215	208	1824	5577

Table 3.5-1 Baseline Comparison with Different Heat Rate Calculation

EPA-EIA	COAL GENERATION						COAL EMISSION					
	ERCOT	WEST	3ZONE	NORTH	HOUSTON	SOUTH	ERCOT	WEST	3ZONE	NORTH	HOUSTON	SOUTH
MODEL	120.56	4.06	116.5	61.74	18.35	36.41	135.49	4.67	130.83	70.73	18.99	41.11
ERCOT	124.41	3.79	120.62	65.26	19.27	36.09	140.83	4.35	136.48	75.61	19.97	40.9
DIF(%)	-3	7	-3	-5	-5	1	-4	7	-4	-6	-5	1
	GAS GENERATION						GAS EMISSION					
	ERCOT	WEST	3ZONE	NORTH	HOUSTON	SOUTH	ERCOT	WEST	3ZONE	NORTH	HOUSTON	SOUTH
MODEL	146.5	5.94	140.55	63.67	22.18	54.7	63.63	2.83	60.8	27.07	10.17	23.56
ERCOT	142.26	5.31	136.95	52.78	42.45	41.72	72.14	2.59	72.71	24.47	23.95	21.13
DIF(%)	3	12	3	21	-48	31	-12	9	-16	11	-58	12
	FOSSIL GENERATION						FOSSIL EMISSION					
	ERCOT	WEST	3ZONE	NORTH	HOUSTON	SOUTH	ERCOT	WEST	3ZONE	NORTH	HOUSTON	SOUTH
MODEL	267.06	10	257.05	125.41	40.53	91.11	199.12	7.5	191.63	97.8	29.16	64.67
ERCOT	266.67	9.1	257.57	118.04	61.72	77.81	212.97	6.94	209.19	100.08	43.92	62.03
DIF(%)	0	10	0	6	-34	17	-7	8	-8	-2	-34	4
EIA	COAL GENERATION						COAL EMISSION					
	ERCOT	WEST	3ZONE	NORTH	HOUSTON	SOUTH	ERCOT	WEST	3ZONE	NORTH	HOUSTON	SOUTH
MODEL	120.42	4.16	116.26	61.55	18.08	36.63	136.45	4.79	131.67	71.6	18.72	41.35
ERCOT	124.41	3.79	120.62	65.26	19.27	36.09	140.83	4.35	136.48	75.61	19.97	40.9
DIF(%)	-3	10	-4	-6	-6	1	-3	10	-4	-5	-6	1
	GAS GENERATION						GAS EMISSION					
	ERCOT	WEST	3ZONE	NORTH	HOUSTON	SOUTH	ERCOT	WEST	3ZONE	NORTH	HOUSTON	SOUTH
MODEL	146.64	7.52	139.11	46.71	50.54	41.86	73.37	3.58	69.79	19.55	31.54	18.7
ERCOT	142.26	5.31	136.95	52.78	42.45	41.72	72.14	2.59	69.55	24.47	23.95	21.13
DIF(%)	3	42	2	-12	19	0	2	38	0	-20	32	-12
	FOSSIL GENERATION						FOSSIL EMISSION					
	ERCOT	WEST	3ZONE	NORTH	HOUSTON	SOUTH	ERCOT	WEST	3ZONE	NORTH	HOUSTON	SOUTH
MODEL	267.06	11.68	255.37	108.26	68.62	78.49	209.82	8.37	201.46	91.15	50.26	60.05
ERCOT	266.67	9.1	257.57	118.04	61.72	77.81	212.97	6.94	206.03	100.08	43.92	62.03
DIF(%)	0	28	-1	-8	11	1	-1	21	-2	-9	14	-3

Table 3.5-2 Percentage Co2 Emission Reduction (partial) Matrix in ERCOT

	Wind Capacity				
CO2 Price	0%	100%	chg in 1st 100%	200%	chg. in 2nd 100%
$0	5.4	0	-5.4	-4.1	-4.1
$20	-11.2	-18.4	-7.2	-23.3	-4.9
$25	-16.2	-23.4	-7.2	-28.3	-4.9
$30	-19.8	-27.2	-7.4	-32	-4.8
$35	-23.1	-30.5	-7.4	-35.3	-4.8
$40	-25.7	-33	-7.3	-37.7	-4.7
$45	-27.5	-34.6	-7.1	-39.3	-4.7
$50	-28.8	-35.8	-7	-40.4	-4.6
$55	-29.7	-36.6	-6.9	-41.1	-4.5
$60	-30.4	-37.2	-6.8	-41.7	-4.5
$65	-30.9	-37.7	-6.8	-42.1	-4.4
$70	-31.4	-38.1	-6.7	-42.5	-4.4

Table 3.5-3 Zonal Carbon Dioxide Emission Reduction (in %)

	Wind Capacity									
CO2 Price	0%	100%	chg. in 1st 100%	200%	chg. in 2nd 100%	0%	100%	chg. in 1st 100%	200%	chg. in 2nd 100%
	West					Houston				
$0	28.6	0	-28.6	-32.6	-32.6	3.2	0	-3.2	-2.6	-2.6
$20	15.1	-12.5	-27.6	-42.5	-30	-6	-11.8	-5.8	-15.2	-3.4
$25	15.5	-12.5	-28	-42.1	-29.6	-11.1	-17	-5.9	-20.4	-3.4
$30	13	-14.3	-27.3	-43.4	-29.1	-13.8	-19.9	-6.1	-23.2	-3.3
$35	10.3	-17.3	-27.6	-45.8	-28.5	-15.1	-21.4	-6.3	-25	-3.6
$40	7.3	-18.5	-25.8	-46.8	-28.3	-16	-22.9	-6.9	-26.7	-3.8
$45	7	-18.8	-25.8	-47.1	-28.3	-17.4	-24.8	-7.4	-29.1	-4.3
$50	7.4	-18.3	-25.7	-46.8	-28.5	-19	-26.8	-7.8	-31.3	-4.5
$55	7.9	-17.8	-25.7	-46.4	-28.6	-20.4	-28.6	-8.2	-33.2	-4.6
$60	8.4	-17.3	-25.7	-46.1	-28.8	-21.8	-30.3	-8.5	-35.1	-4.8
$65	9.1	-16.8	-25.9	-45.8	-29	-23.2	-32	-8.8	-37	-5
$70	9.6	-16.4	-26	-45.4	-29	-24.5	-33.6	-9.1	-38.7	-5.1
	North					South				
$0	3.6	0	-3.6	-1.7	-1.7	6.6	0	-6.6	-5.1	-5.1
$20	-22	-27.1	-5.1	-28	-0.9	-2.9	-11.6	-8.7	-20.4	-8.8
$25	-26.5	-31.7	-5.2	-32.8	-1.1	-9.3	-17.8	-8.5	-26	-8.2
$30	-30.5	-36.2	-5.7	-37.4	-1.2	-13.1	-21.4	-8.3	-29.5	-8.1
$35	-35.1	-40.9	-5.8	-42	-1.1	-16.1	-24.2	-8.1	-32.2	-8
$40	-38.5	-43.8	-5.3	-45	-1.2	-19	-27.1	-8.1	-34.7	-7.6
$45	-39.9	-44.7	-4.8	-45.9	-1.2	-22	-29.9	-7.9	-36.8	-6.9
$50	-40.6	-45.2	-4.6	-46.4	-1.2	-24	-31.5	-7.5	-38	-6.5
$55	-41.3	-45.7	-4.4	-46.7	-1	-25.1	-32.3	-7.2	-38.5	-6.2
$60	-41.7	-45.9	-4.2	-46.9	-1	-25.8	-32.8	-7	-38.7	-5.9
$65	-41.9	-45.9	-4	-46.8	-0.9	-26.4	-33	-6.6	-38.7	-5.7
$70	-41.9	-45.8	-3.9	-46.7	-0.9	-26.8	-33.2	-6.4	-38.7	-5.5

Table 3.5-4 Zonal Co2 Emission Reduction (in million tons)

	Wind Capacity									
CO2 Price	0%	100%	chg in 1st 100%	200%	chg in 2nd 100%	0%	100%	chg in 1st 100%	200%	chg in 2nd 100%
	West					Houston				
$0	2.39	0	-2.39	-2.73	-2.73	1.63	0	-1.63	-1.29	-1.29
$20	1.26	-1.05	-2.31	-3.56	-2.51	-3.02	-5.92	-2.9	-7.65	-1.73
$25	1.3	-1.05	-2.35	-3.52	-2.47	-5.59	-8.56	-2.97	-10.25	-1.69
$30	1.09	-1.2	-2.29	-3.63	-2.43	-6.94	-9.98	-3.04	-11.67	-1.69
$35	0.86	-1.45	-2.31	-3.83	-2.38	-7.57	-10.77	-3.2	-12.58	-1.81
$40	0.61	-1.55	-2.16	-3.92	-2.37	-8.06	-11.52	-3.46	-13.44	-1.92
$45	0.59	-1.57	-2.16	-3.94	-2.37	-8.75	-12.48	-3.73	-14.63	-2.15
$50	0.62	-1.53	-2.15	-3.92	-2.39	-9.53	-13.46	-3.93	-15.74	-2.28
$55	0.66	-1.49	-2.15	-3.88	-2.39	-10.27	-14.37	-4.1	-16.71	-2.34
$60	0.7	-1.45	-2.15	-3.86	-2.41	-10.97	-15.21	-4.24	-17.66	-2.45
$65	0.76	-1.41	-2.17	-3.83	-2.42	-11.66	-16.08	-4.42	-18.61	-2.53
$70	0.8	-1.37	-2.17	-3.8	-2.43	-12.33	-16.89	-4.56	-19.45	-2.56
	North					South				
$0	3.28	0	-3.28	-1.58	-1.58	3.96	0	-3.96	-3.07	-3.07
$20	-20.05	-24.68	-4.63	-25.53	-0.85	-1.74	-6.98	-5.24	-12.23	-5.25
$25	-24.13	-28.91	-4.78	-29.93	-1.02	-5.58	-10.66	-5.08	-15.61	-4.95
$30	-27.79	-33.01	-5.22	-34.13	-1.12	-7.87	-12.83	-4.96	-17.71	-4.88
$35	-31.98	-37.24	-5.26	-38.24	-1	-9.69	-14.52	-4.83	-19.36	-4.84
$40	-35.08	-39.88	-4.8	-40.99	-1.11	-11.4	-16.28	-4.88	-20.84	-4.56
$45	-36.34	-40.71	-4.37	-41.8	-1.09	-13.22	-17.93	-4.71	-22.08	-4.15
$50	-37.05	-41.23	-4.18	-42.25	-1.02	-14.39	-18.91	-4.52	-22.8	-3.89
$55	-37.63	-41.62	-3.99	-42.6	-0.98	-15.05	-19.39	-4.34	-23.11	-3.72
$60	-37.99	-41.8	-3.81	-42.71	-0.91	-15.49	-19.67	-4.18	-23.23	-3.56
$65	-38.15	-41.8	-3.65	-42.67	-0.87	-15.84	-19.82	-3.98	-23.26	-3.44
$70	-38.23	-41.71	-3.48	-42.59	-0.88	-16.11	-19.93	-3.82	-23.26	-3.33

Table 3.5-5 Zonal So2 Emission Reduction (in million lbs.)

	Wind Capacity									
CO2 Price	0%	100%	chg in 1st 100%	200%	chg in 2nd 100%	0%	100%	chg in 1st 100%	200%	chg in 2nd 100%
	West					Houston				
$0	0.39	0	-0.39	-1.19	-1.19	0.19	0	-0.19	-0.19	-0.19
$20	-1.35	-1.88	-0.53	-2.64	-0.76	-9.18	-11.69	-2.51	-13.36	-1.67
$25	-1.69	-2.28	-0.59	-2.86	-0.58	-15.41	-18.31	-2.9	-20.08	-1.77
$30	-2.17	-2.7	-0.53	-3.14	-0.44	-20.19	-23.15	-2.96	-24.9	-1.75
$35	-2.72	-3.27	-0.55	-3.56	-0.29	-23.49	-26.56	-3.07	-28.4	-1.84
$40	-3.16	-3.59	-0.43	-3.79	-0.2	-26.63	-29.96	-3.33	-31.85	-1.89
$45	-3.36	-3.73	-0.37	-3.88	-0.15	-29.87	-33.17	-3.3	-35.02	-1.85
$50	-3.49	-3.81	-0.32	-3.93	-0.12	-32.59	-35.74	-3.15	-37.44	-1.7
$55	-3.59	-3.87	-0.28	-3.96	-0.09	-34.69	-37.63	-2.94	-39.22	-1.59
$60	-3.66	-3.91	-0.25	-3.99	-0.08	-36.44	-39.13	-2.69	-40.63	-1.5
$65	-3.71	-3.93	-0.22	-4	-0.07	-37.93	-40.39	-2.46	-41.72	-1.33
$70	-3.75	-3.95	-0.2	-4	-0.05	-39.2	-41.42	-2.22	-42.64	-1.22
	North					South				
$0	1.02	0.00	-1.02	-0.45	-0.45	0.33	0	-0.33	-0.79	-0.79
$20	-109.56	-118.04	-8.48	-120.00	-1.96	-34.13	-39.62	-5.49	-44.74	-5.12
$25	-136.56	-145.30	-8.74	-147.51	-2.21	-47.86	-53.46	-5.6	-58.18	-4.72
$30	-158.34	-168.86	-10.52	-171.40	-2.54	-57.47	-62.97	-5.5	-67.43	-4.46
$35	-181.69	-192.41	-10.72	-194.41	-2.00	-64.65	-69.98	-5.33	-74.87	-4.89
$40	-199.44	-207.66	-8.22	-209.94	-2.28	-70.67	-76.42	-5.75	-81.06	-4.64
$45	-208.44	-215.30	-6.86	-217.20	-1.90	-77.66	-82.92	-5.26	-86.76	-3.84
$50	-214.40	-220.57	-6.17	-222.09	-1.52	-82.57	-87.32	-4.75	-90.67	-3.35
$55	-218.86	-224.38	-5.52	-225.63	-1.25	-86.2	-90.42	-4.22	-93.25	-2.83
$60	-222.09	-227.01	-4.92	-228.03	-1.02	-88.99	-92.71	-3.72	-95.18	-2.47
$65	-224.56	-228.93	-4.37	-229.87	-0.94	-91.33	-94.5	-3.17	-96.66	-2.16
$70	-226.67	-230.37	-3.70	-231.22	-0.85	-93.29	-95.94	-2.65	-97.92	-1.98

Table 3.5-6 Zonal Nox Emission Reduction (in million tons)

	Wind Capacity									
CO2 Price	0%	100%	chg in 1st 100%	200%	chg in 2nd 100%	0%	100%	chg in 1st 100%	200%	chg in 2nd 100%
	West					Houston				
$0	3.74	0	-3.74	-5.98	-5.98	0.93	0	-0.93	-0.75	-0.75
$20	-2.23	-6.33	-4.1	-10.86	-4.53	-1.03	-2.64	-1.61	-3.64	-1
$25	-3.02	-7.36	-4.34	-11.39	-4.03	-1.97	-3.72	-1.75	-4.69	-0.97
$30	-4.37	-8.5	-4.13	-12.17	-3.67	-2.25	-4.1	-1.85	-5.13	-1.03
$35	-5.85	-10.12	-4.27	-13.38	-3.26	-2.29	-4.26	-1.97	-5.3	-1.04
$40	-6.88	-10.71	-3.83	-13.87	-3.16	-2.04	-4.28	-2.24	-5.4	-1.12
$45	-6.94	-10.71	-3.77	-13.88	-3.17	-1.91	-4.35	-2.44	-5.6	-1.25
$50	-6.88	-10.55	-3.67	-13.8	-3.25	-1.83	-4.41	-2.58	-5.78	-1.37
$55	-6.72	-10.38	-3.66	-13.69	-3.31	-1.7	-4.44	-2.74	-5.88	-1.44
$60	-6.52	-10.21	-3.69	-13.57	-3.36	-1.68	-4.55	-2.87	-6.07	-1.52
$65	-6.25	-10.04	-3.79	-13.44	-3.4	-1.67	-4.75	-3.08	-6.34	-1.59
$70	-5.99	-9.84	-3.85	-13.32	-3.48	-1.65	-4.91	-3.26	-6.55	-1.64
	North					South				
$0	2.18	0	-2.18	-1.03	-1.03	2.18	0	-2.18	-1.87	-1.87
$20	-40.8	-45.91	-5.11	-46.94	-1.03	-10.61	-14.99	-4.38	-19.28	-4.29
$25	-50.7	-55.93	-5.23	-57.15	-1.22	-16.57	-20.84	-4.27	-24.82	-3.98
$30	-58.74	-64.74	-6	-66.09	-1.35	-20.29	-24.41	-4.12	-28.33	-3.92
$35	-66.8	-72.95	-6.15	-74.18	-1.23	-23.12	-27.14	-4.02	-31.14	-4
$40	-73.18	-78.56	-5.38	-80.01	-1.45	-25.56	-29.65	-4.09	-33.37	-3.72
$45	-75.88	-80.76	-4.88	-82.2	-1.44	-27.69	-31.59	-3.9	-34.97	-3.38
$50	-77.21	-82	-4.79	-83.35	-1.35	-28.88	-32.6	-3.72	-35.78	-3.18
$55	-78.1	-82.81	-4.71	-84.15	-1.34	-29.45	-33.02	-3.57	-36.09	-3.07
$60	-78.52	-83.1	-4.58	-84.41	-1.31	-29.77	-33.25	-3.48	-36.2	-2.95
$65	-78.63	-83.07	-4.44	-84.39	-1.32	-29.94	-33.28	-3.34	-36.17	-2.89
$70	-78.66	-82.93	-4.27	-84.28	-1.35	-30.04	-33.3	-3.26	-36.14	-2.84

Table 3.5-7 Summary Statistics about Operation of Representative Generation Units

	Coal Unit				Combined Cycle Gas Unit			
Carbon Tax	Running Hrs #	% of Hrs %	Avg. Price $/MWH	Avg.Profit $/MWH	Running Hrs #	% of Hrs %	Avg. Price $/MWH	Avg.Profit $/MWH
Status Quo Wind Capacity								
0$	9284	98	35.2	13.4	3139	33	36.3	4.4
20$	7513	79	48.2	5.9	6347	67	45.2	4.5
25$	6124	65	52.8	5.4	7384	78	48.7	4.9
30$	5140	54	57.4	4.9	7982	84	52.3	5.6
35$	4467	47	62.1	4.5	8548	90	56.1	6.3
40$	3934	42	66.9	4.2	9077	96	59.8	7.1
45$	3377	36	71.8	4	9214	97	63.1	8
50$	2675	28	77.1	4.3	9269	98	66.4	9
55$	2258	24	82.4	4.5	9295	98	69.6	9.9
60$	1997	21	87.6	4.5	9326	98	72.8	10.8
65$	1747	18	92.7	4.6	9346	99	75.9	11.6
70$	1525	16	97.9	4.7	9369	99	79	12.4
60% Increase of Wind Capacity								
0$	9284	98	34.8	13.0	3006	32	36.1	4.3
20$	7484	79	48.0	5.8	6429	68	44.9	4.2
25$	6090	64	52.6	5.3	7479	79	48.4	4.6
30$	5037	53	57.3	4.8	8088	85	52	5.3
35$	4404	46	61.9	4.4	8647	91	55.8	6
40$	3845	41	66.8	4.1	9189	97	59.4	6.7
45$	3219	34	71.6	3.9	9289	98	62.7	7.6
50$	2487	26	77.1	4.2	9346	99	65.9	8.4
55$	2113	22	82.3	4.4	9381	99	69	9.3
60$	1822	19	87.5	4.4	9399	99	72.1	10.1
65$	1578	17	92.7	4.5	9407	99	75.2	10.9
70$	1375	15	97.8	4.6	9415	99	78.3	11.6
Doubled Wind Capacity								
0$	9282	98	34.6	12.8	2969	31	35.9	4.2
20$	7466	79	48	5.7	6416	68	44.8	4.1
25$	6053	64	52.6	5.2	7480	79	48.3	4.5
30$	5003	53	57.3	4.8	8102	85	51.9	5.2
35$	4362	46	61.9	4.3	8680	92	55.7	5.8
40$	3784	40	66.7	4.1	9178	97	59.3	6.5
45$	3151	33	71.6	3.9	9295	98	62.5	7.4
50$	2419	26	77	4.2	9360	99	65.7	8.3
55$	2050	22	82.2	4.3	9384	99	68.8	9.1
60$	1777	19	87.4	4.4	9409	99	71.9	9.9
65$	1515	16	92.6	4.5	9425	99	74.9	10.6
70$	1307	14	97.9	4.6	9431	100	78	11.3

1. The heat rate of the coal unit used in the calculation is 10.62mmBTU/MWH, the co2 emission rate is 1.02 tons/MWH.
2. The heat rate of the gas unit used in the calculation is 7.477mmBTU/MWH, the co2 emission rate is 0.458 tons/MWH.
3. The cost to emit other typical pollutants from fossil fuel generators, such as SO2 and Nox, are not included in the calculation.

Figure 3.1-1 The Cumulative Marginal Cost Curve of Fossil Fuel Generation

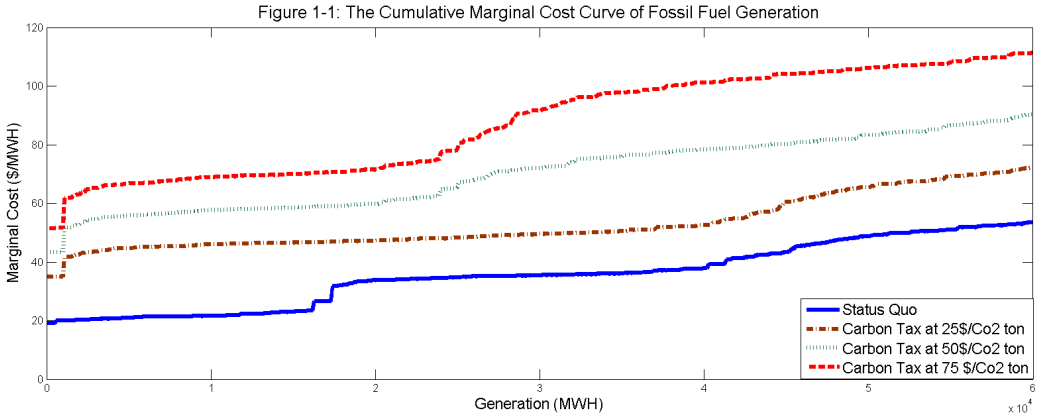
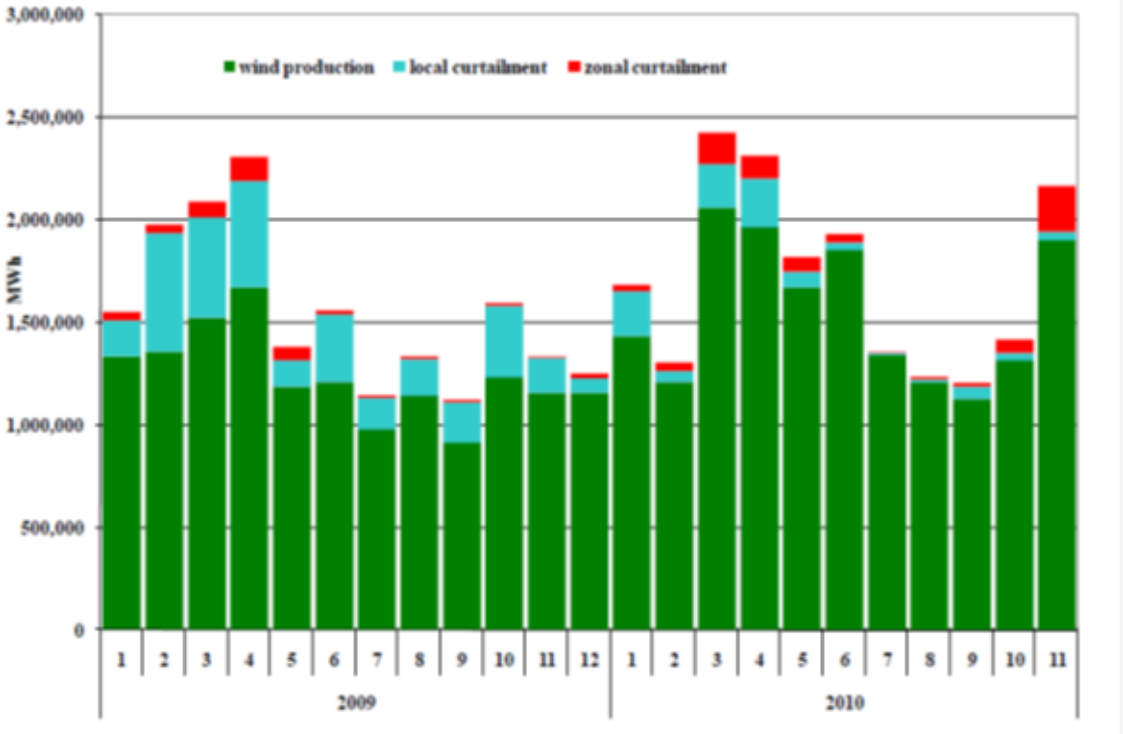


Figure 3.4-1 The Curtailment of Wind Generation in West Zone



Source: Potomac Economics, 2010 state of market report of ERCOT wholesale electricity market, figure 28

Figure 3.4-2 the Evolution of Wind Farms in ERCOT (2000-2011)

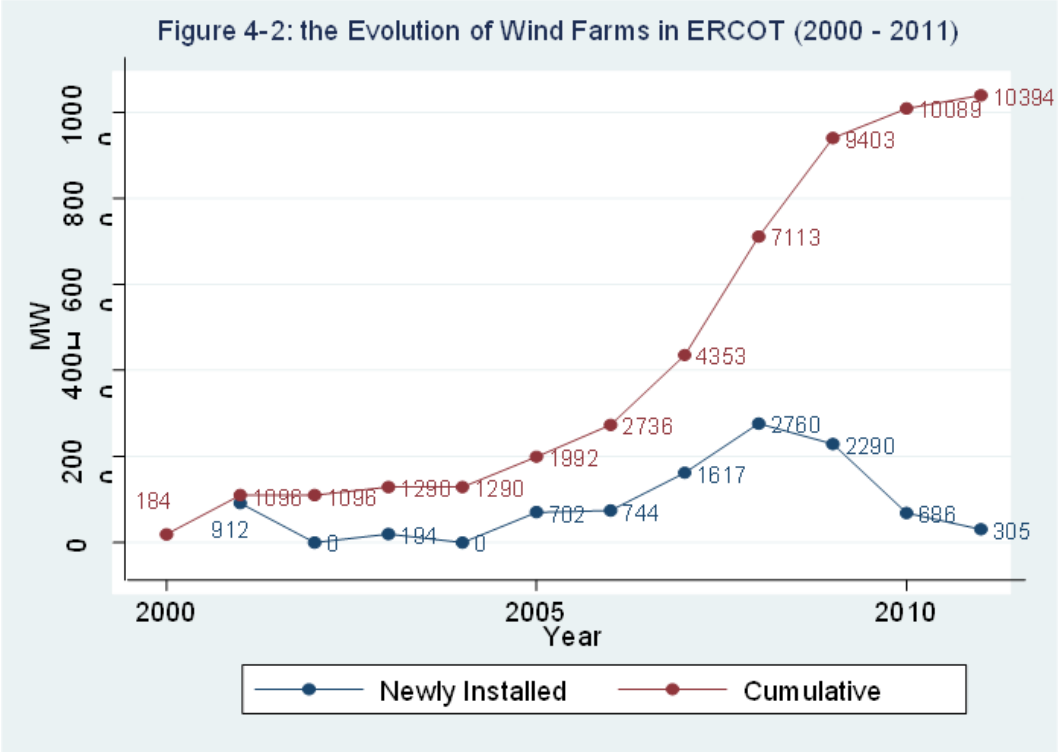


Figure 3.4-3 Average Hourly Electricity Demand and Wind Generation (MWH)

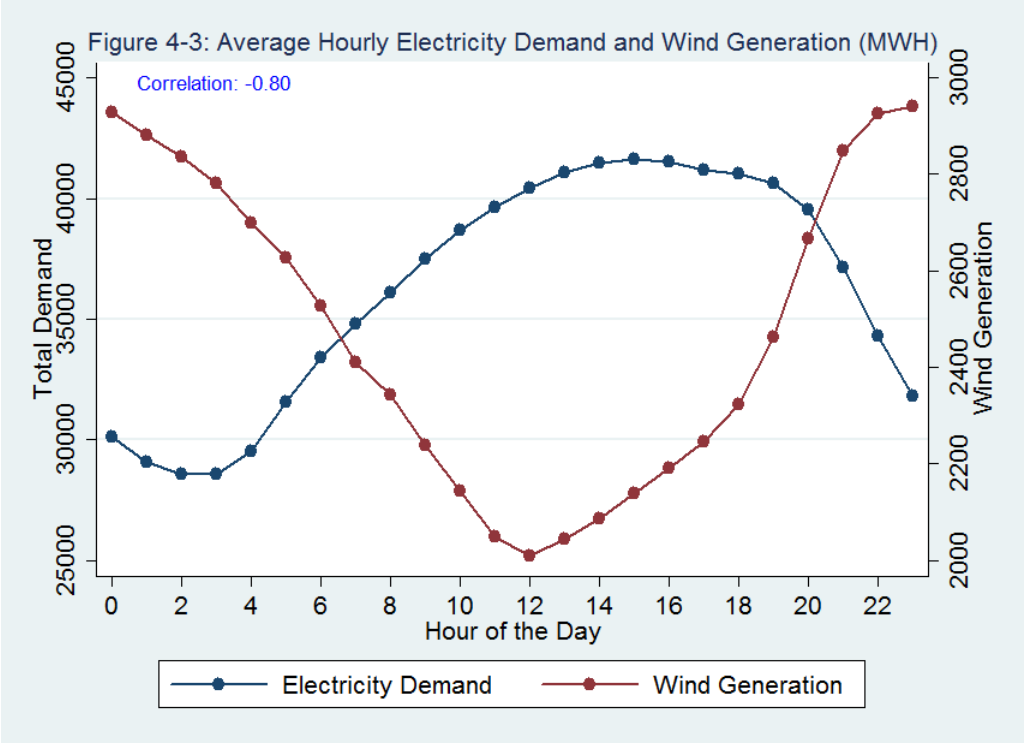


Figure 3.5-1 Share of Generation from Gas Units in ERCOT

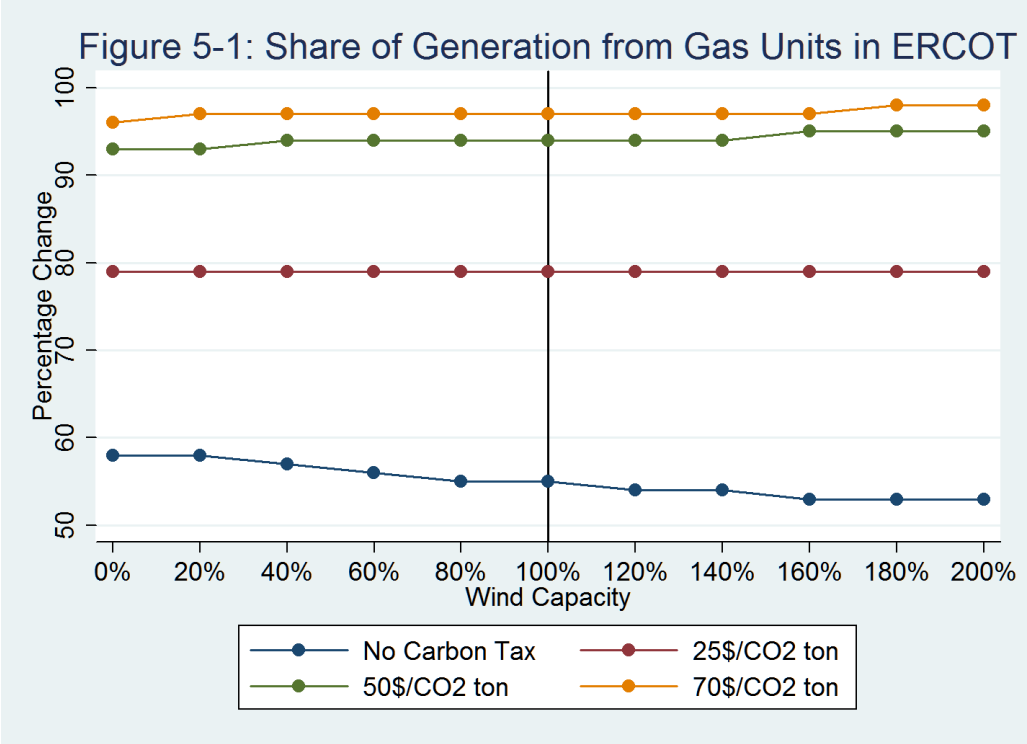


Figure 3.5-2 Zonal Share of Gas Generation in ERCOT

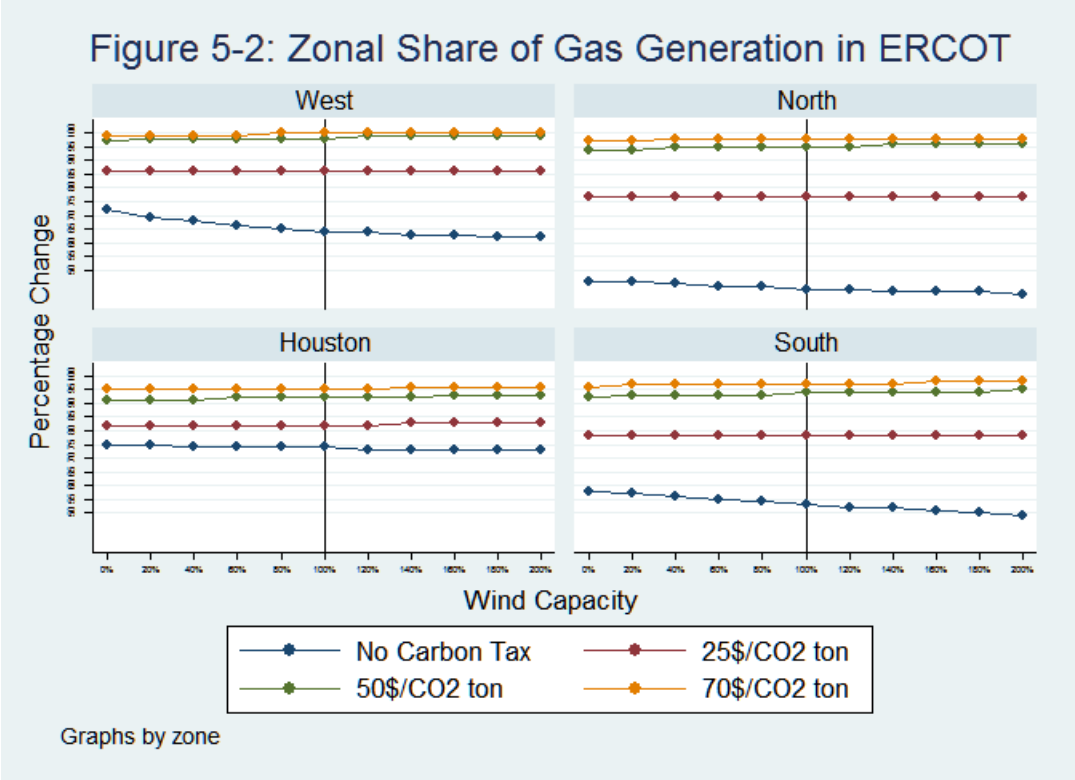


Figure 3.5-3 Reduction of CO2 Emission in ERCOT

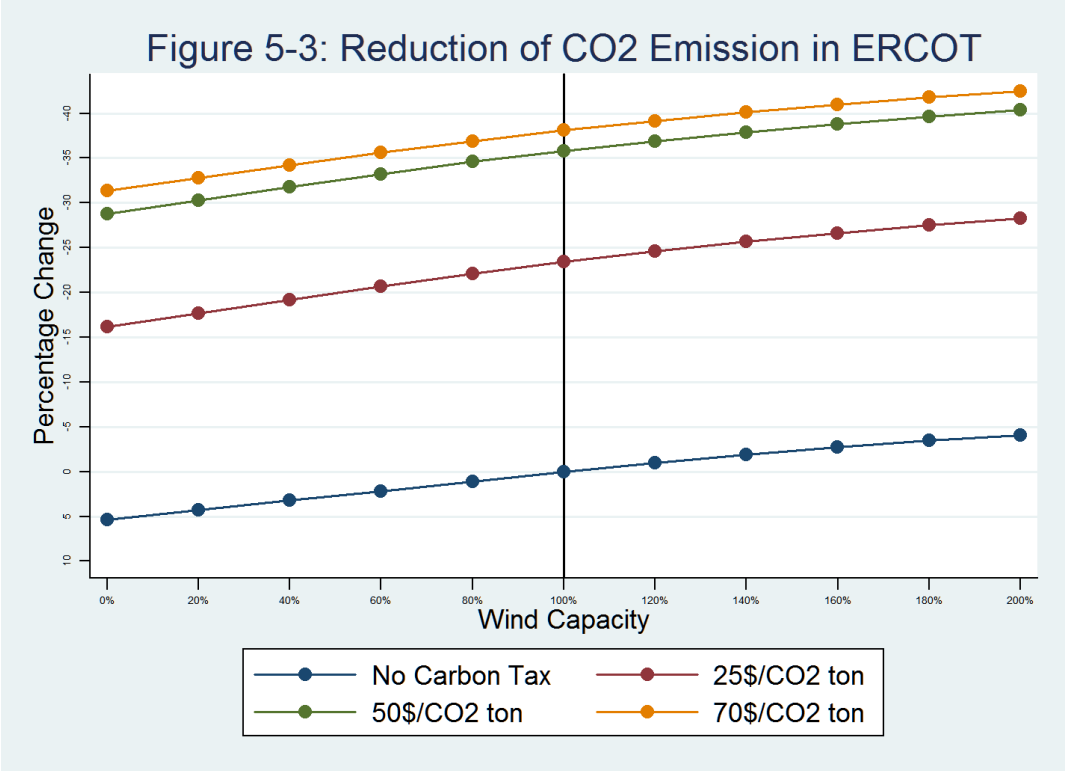


Figure 3.5-4 Contour Graph of Emission Reduction on Carbon-Wind Plane (%)

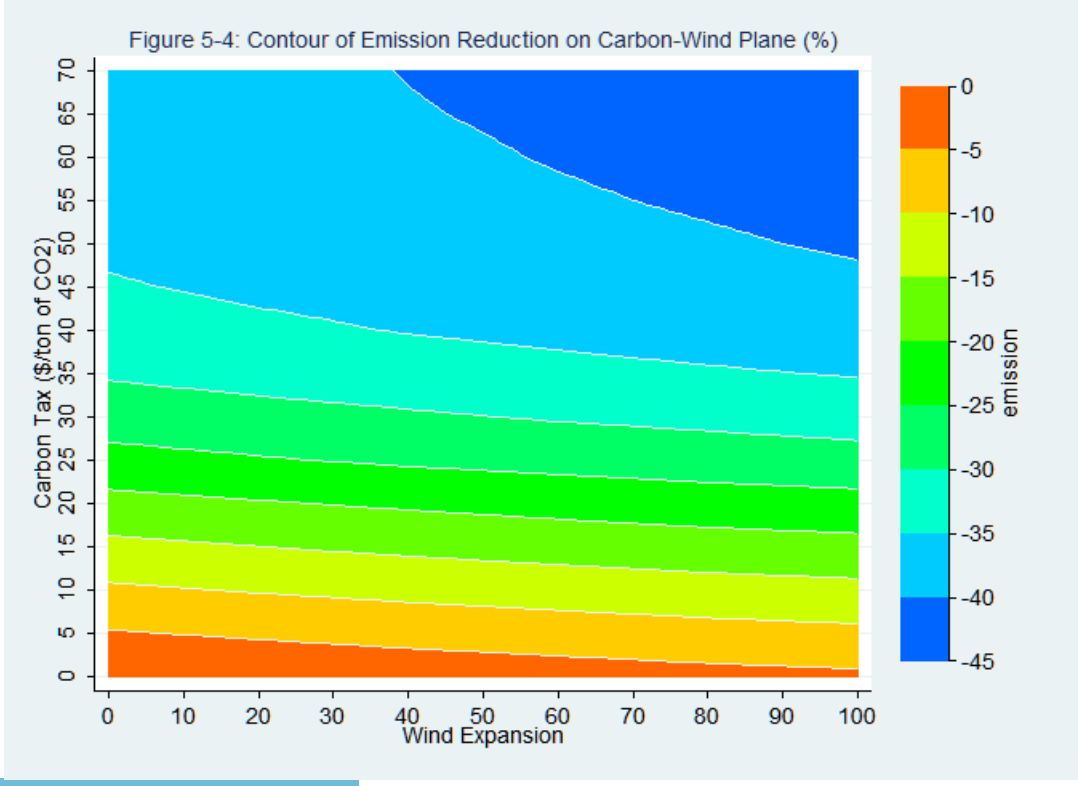


Figure 3.5-5 Distribution of Market Price with Status Quo Wind Capacity

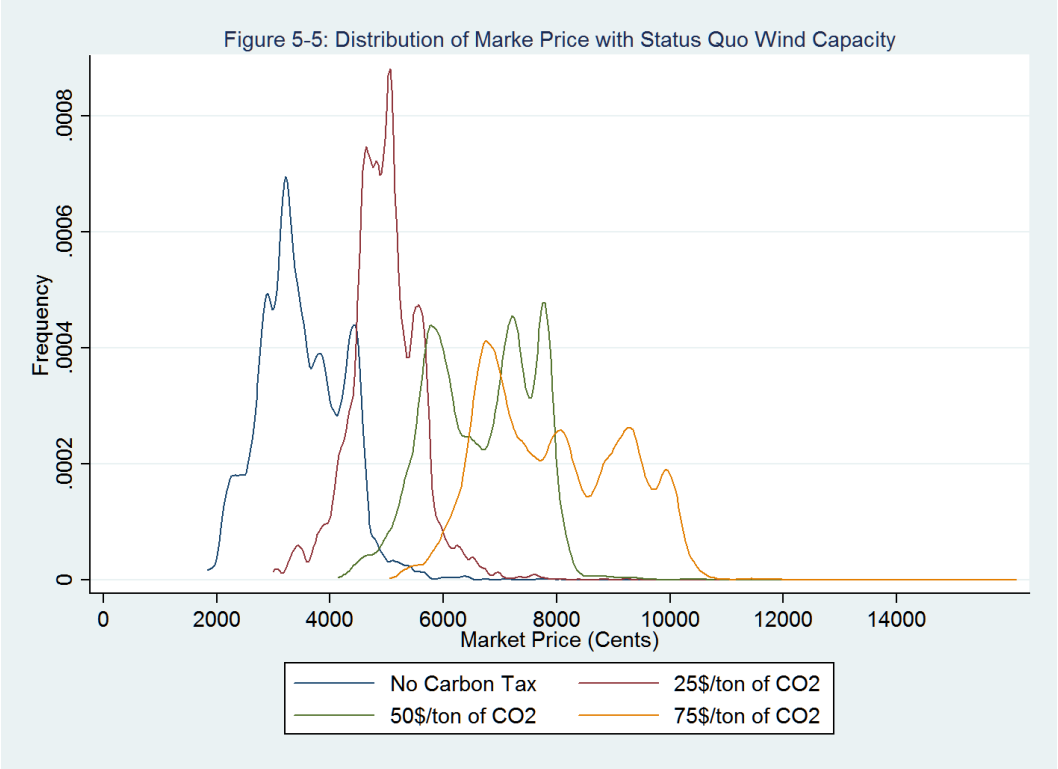
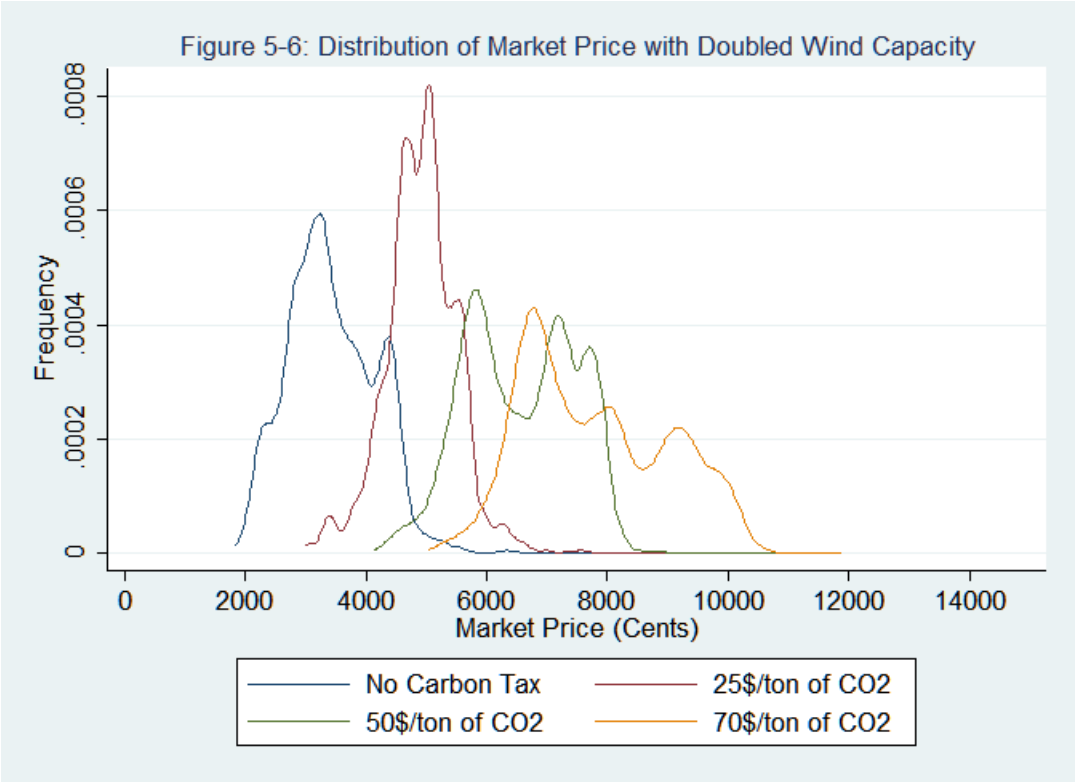


Figure 3.5-6 Distribution of Market Price with Doubled Wind Capacity



APPENDIX A. Additional Material for Chapter 2

Appendix A

The purpose of this appendix is to discuss in greater detail the alternative aggregated logit models described in the body the paper and the issues associated with parameter identification. We start, however, by listing some basic properties of the Gumbel Distribution that are then used in subsequent sections.

A.1 Properties about Gumbel Distribution

The logit model is based on the Extreme Type I (or Standard Gumbel) Distribution, used to characterize the unobserved portion of the utility that an individual derives from choosing an alternative. In addition to yielding a close form for the associated choice probabilities, the Gumbel distribution has other useful properties, including the following three properties:

- Property 1: If $X \sim Gumbel(0, \mu)$, then $X/\mu \sim Gumbel(0, 1)$; i.e., the standard Gumbel distribution.
- Property 2: If $X \sim Gumbel(0, \mu)$, then $X + A \sim Gumbel(A, \mu)$.
- Property 3: If $X_j \sim Gumbel(A_j, \mu)$ for $j = 1, \ldots, J$ and X_j is independent of $X_k \ \forall k \neq j$, then $Y \equiv \max(X_1, X_2, \ldots, X_J) \sim Gumbel(A_Y, \mu)$, where $A_Y \equiv \mu \cdot ln\left[\sum_{j=1}^{J} exp(A_j/\mu)\right]$.

  Proof: The cdf for X_j is given by:

$$F_{X_j}(x_j) = exp\left(-exp\left[-\frac{(x_j - Aj)}{\mu}\right]\right) \quad j = 1, \ldots, J. \tag{A.1}$$

Therefore, the cdf for Y is given by

$$\begin{aligned}
F_Y(y) &= \prod_{j=1}^{J} F_{X_j}(y/\mu) \\
&= \prod_{j=1}^{J} exp\left(-exp\left[-\frac{(y-Aj)}{\mu}\right]\right) \\
&= exp\left(-\sum_{j=1}^{J} exp\left[-\frac{(y-Aj)}{\mu}\right]\right) \\
&= exp\left(-exp(-y/\mu)exp\left[ln\left\{\sum_{j=1}^{J} exp\left[\frac{Aj}{\mu}\right]\right\}\right]\right) \\
&= exp\left(-exp(-y/\mu)exp\left[\frac{A_Y}{\mu}\right]\right) \\
&= exp\left(-exp\left[-\frac{(y-A_Y)}{\mu}\right]\right)
\end{aligned}$$

which is the $Gumbel(A_Y, \mu)$ distribution. A special case of Property 3 arises when $A_j = 0\ \forall j$, in which case $A_Y = \mu \cdot ln(J)$.

A.2 The Logit Model

Consider the simplest repeated logit model, where the utility that individual i receives from choosing access point j along segment s on choice occasion t is given by.

$$U_{isjt} = \begin{cases} \epsilon_{isjt} & \text{if } s = j = 0 \\ V_{isj} + \epsilon_{isjt} & \text{otherwise} \end{cases} \tag{A.2}$$

where $V_{isj} = \alpha_{sj} + \beta C_{isj}$, V_{i00} has been normalized to zero for the stay-at-home option, and the ϵ_{isjt} are iid Type I Extreme Value random variables capturing unobserved attributes impacting these conditional utilities.

The purpose of this section is twofold. First, we argue that, even when access-point trip data has been aggregated to segment-level information, the parameters of the above model are apparently identified. While a formal proof of identification is not provided, we are able to show that the model satisfies necessary conditions for identification in the form of a rank condition for the score function. Moreover, a series of Monte Carlo exercises substantiate the claim of identification. Second, while the access-point level ASC's (i.e., the α_{sj}) are identified,

they are only *poorly* identified, depending upon the assumed logistic structure of the model. This makes the use of access-point level ASC's tenuous and, to the extent that segments are relatively uniform in their underly attributes, it may be preferable to allow for segment level ASC's only.

To examine identification in the repeated logit model, we derive the associated gradients and Hessian. Using the model in (A.2), the contribution of individual i to the likelihood function given by:

$$\begin{aligned}\mathcal{L}_i(\boldsymbol{n}_i) &= \sum_{s=0}^{S} n_{is\bullet} ln\left(\sum_{j\in A_s} P_{isj}\right) \qquad \text{(A.3)}\\ &= \sum_{s=0}^{S} n_{is\bullet} ln(P_{is\bullet})\\ &= \left\{\sum_{s=1}^{S} n_{is\bullet} V_{is\bullet}\right\} - T\cdot ln\left[1+\sum_{r=1}^{S}\sum_{k\in A_r} exp(V_{irk})\right], \qquad \text{(A.4)}\end{aligned}$$

where $\boldsymbol{n}_i = (n_{i0\bullet},\ldots,n_{iS\bullet})$ and $n_{is\bullet} = \sum_{t=1}^{T} y_{is\bullet t}$ denotes the total number of times aggregate alternative s is chosen across the T choice occasions, with

$$V_{is\bullet} \equiv ln\left[\sum_{j\in A_s} exp(V_{isj})\right]. \qquad \text{(A.5)}$$

The gradients for the ASC's are given by:

$$\frac{\partial \mathcal{L}_i}{\partial \alpha_{sj}} = \sum_{r=1}^{S}\sum_{k\in A_r} \frac{\partial \mathcal{L}_i}{\partial V_{irk}}\frac{\partial V_{irk}}{\partial \alpha_{sj}}. \qquad \text{(A.6)}$$

But

$$\begin{aligned}\frac{\partial \mathcal{L}_i}{\partial V_{irk}} &= n_{ir\bullet}\frac{exp(V_{irk})}{\left[\sum_{j\in A_r} exp(V_{irk})\right]} - T\cdot\frac{exp(V_{irk})}{\left[1+\sum_{s=1}^{S}\sum_{j\in A_s} exp(V_{isj})\right]} \qquad \text{(A.7)}\\ &= \left(n_{ir\bullet} - T\cdot P_{ir\bullet}\right) P_{irk|r}\end{aligned}$$

where

$$P_{isj|s} \equiv \frac{exp(V_{isj})}{\left[\sum_{j\in A_s} exp(V_{isj})\right]} \qquad \text{(A.8)}$$

denotes the probability that access point $j \in A_s$ is chosen, given that segment s has been chosen and

$$P_{is\bullet} \equiv \frac{\left[\sum_{j\in A_s} exp(V_{isj})\right]}{\left[1+\sum_{r=1}^{S}\sum_{k\in A_r} exp(V_{irk})\right]} \qquad \text{(A.9)}$$

denotes the probability that segment s is chosen. Substituting (A.7) into (A.6) yields:

$$\frac{\partial \mathcal{L}_i}{\partial \alpha_{sj}} = (n_{is\bullet} - T \cdot P_{is\bullet})\, P_{isj|s}. \tag{A.10}$$

The first order conditions for the ASC's becomes:

$$\begin{aligned} 0 = \frac{\partial \mathcal{L}}{\partial \alpha_{sj}} &= \sum_{i=1}^{N} \frac{\partial \mathcal{L}_i}{\partial \alpha_{sj}} \\ &= \sum_{i=1}^{N} (n_{is\bullet} - T \cdot P_{is\bullet})\, P_{isj|s}. \end{aligned} \tag{A.11}$$

Note that, in general, $\frac{\partial \mathcal{L}}{\partial \alpha_{sj}} \neq \frac{\partial \mathcal{L}}{\partial \alpha_{sk}}$ unless $P_{isj|s} = P_{isk|s}\ \forall j, k \in A_s$, which will not be the case since C_{isj} will vary across access points. Also, if we let $\hat{n}_{isj} \equiv n_{is\bullet} P_{isj|s}$, then these first order conditions can be rewritten as:

$$0 = \sum_{i=1}^{N} (\hat{n}_{isj} - T \cdot P_{isj}),$$

which is identical to the standard first order conditions for the logit model without aggregation, *except* that now the actual trips to access point j (i.e., n_{isj}) is replaced by a fitted value derived from the implicit logit model of the choice of access points along a segment. The corresponding gradient for β is given by:

$$\begin{aligned} \frac{\partial \mathcal{L}_i}{\partial \beta} &= \sum_{r=1}^{S} \sum_{k \in A_r} \frac{\partial \mathcal{L}_i}{\partial V_{irk}} \frac{\partial V_{irk}}{\partial \beta} \\ &= \sum_{r=1}^{S} \sum_{k \in A_r} (n_{ir\bullet} - T \cdot P_{ir\bullet})\, P_{irk|r} C_{irk} \\ &= \sum_{r=1}^{S} (n_{ir\bullet} - T \cdot P_{ir\bullet}) \sum_{k \in A_r} P_{irk|r} C_{irk} \\ &\approx \sum_{r=1}^{S} (n_{ir\bullet} - T \cdot P_{ir\bullet})\, C_{ir\bullet} \end{aligned} \tag{A.12}$$

where the last approximation stems from equation (24) in the body of the paper. Equation (A.12) highlights the fact that the influence of β on the log-likelihood function is driven in large part by the variation in travel costs across segments. We can also write the gradient as:

$$\frac{\partial \mathcal{L}_i}{\partial \beta} = \sum_{r=1}^{S} \sum_{k \in A_r} (\hat{n}_{irk} - T \cdot P_{irk})\, C_{irk},$$

which is the form the gradient would take without aggregation, *except* that, again, n_{irk} is replaced with the fitted value $\hat{n}_{irk}$. Using these results, the first order condition for β becomes:

$$\begin{aligned} 0 &= \frac{\partial \mathcal{L}}{\partial \beta} \\ &= \sum_{i=1}^{N} \frac{\partial \mathcal{L}_i}{\partial \beta} \\ &= \sum_{i=1}^{N} \sum_{r=1}^{S} \left(n_{ir\bullet} - T \cdot P_{ir\bullet}\right) \sum_{k \in A_r} P_{irk|r} C_{irk} \\ &\approx \sum_{i=1}^{N} \sum_{r=1}^{S} \left(n_{ir\bullet} - T \cdot P_{ir\bullet}\right) C_{ir\bullet}. \end{aligned} \tag{A.13}$$

It is clear from the above results that the score function for the aggregated logit model is identical in form to its counterpart without aggregation *except* that the access point counts (i.e., the n_{irk}'s) are replaced with fitted value (i.e., the $\hat{n}_{irk}$'s). Moreover, the scores satisfy a necessary condition for identification of the parameter vector $\phi \equiv (\alpha_{\bullet\bullet}, \beta)$, where $\alpha_{\bullet\bullet}$ is the vector of access point level ASC's; i.e., the score function is of full column rank.

Turning to the Hessian calculation, we have that:

$$\begin{aligned} \frac{\partial^2 \mathcal{L}_i}{\partial \alpha_{rk} \partial \alpha_{sj}} &= \frac{\partial}{\partial \alpha_{rk}} \left(n_{is\bullet} - T \cdot P_{is\bullet}\right) P_{isj|s} \\ &= n_{is\bullet} \frac{\partial P_{isj|s}}{\partial \alpha_{rk}} - T \frac{\partial P_{isj}}{\partial \alpha_{rk}} \end{aligned} \tag{A.14}$$

The partial derivative in the first term becomes:

$$\begin{aligned} \frac{\partial P_{isj|s}}{\partial \alpha_{rk}} &= \delta_{sr} \frac{\delta_{jk} \left[\sum_{l \in A_s} exp(V_{isl})\right] exp(V_{isj}) - exp(V_{isj}) exp(V_{isk})}{\left[\sum_{l \in A_s} exp(V_{isl})\right]^2} \\ &= \delta_{sr} \left(\delta_{jk} P_{isj|s} - P_{isj|s} P_{isk|s}\right) \qquad \text{(A.15)} \\ &= \delta_{sr} \left(\delta_{jk} - P_{isk|s}\right) P_{isj|s} \qquad \text{(A.16)} \end{aligned}$$

The partial derivative in the second term becomes:

$$\begin{aligned} \frac{\partial P_{isj}}{\partial \alpha_{rk}} &= \frac{\delta_{jk} \left[\sum_{m=1}^{S} \sum_{l \in A_m} exp(V_{iml})\right] exp(V_{isj}) - exp(V_{isj}) exp(V_{irk})}{\left[\sum_{m=1}^{S} \sum_{l \in A_m} exp(V_{iml})\right]^2} \\ &= \delta_{jk} P_{isj} - P_{isj} P_{irk} \qquad \text{(A.17)} \\ &= \left(\delta_{jk} - P_{irk}\right) P_{isj}. \qquad \text{(A.18)} \end{aligned}$$

Combining terms we get:

$$\begin{aligned}\frac{\partial^2 \mathcal{L}_i}{\partial \alpha_{rk} \partial \alpha_{sj}} &= n_{is\bullet} \delta_{sr} \left(\delta_{jk} - P_{isk|s}\right) P_{isj|s} - T \left(\delta_{jk} - P_{irk}\right) P_{isj} \\ &= \delta_{sr} \left(\delta_{jk} - P_{irk|r}\right) \hat{n}_{isj} - T \left(\delta_{jk} - P_{irk}\right) P_{isj}. \end{aligned} \tag{A.19}$$

The change from the logit model without aggregation is the addition of the first term.

Finally, we have:

$$\begin{aligned}\frac{\partial^2 \mathcal{L}_i}{\partial \beta \partial \alpha_{rk}} &= \frac{\partial^2 \mathcal{L}_i}{\partial \alpha_{rk} \partial \beta} \\ &= \frac{\partial}{\partial \alpha_{rk}} \sum_{s=1}^{S} \sum_{j \in A_s} \left(n_{is\bullet} - T \cdot P_{is\bullet}\right) P_{isj|s} C_{isj} \\ &= \sum_{s=1}^{S} \sum_{j \in A_s} \left(n_{is\bullet} \frac{\partial P_{isj|s}}{\partial \alpha_{rk}} - T \cdot \frac{\partial P_{isj}}{\partial \alpha_{rk}} \right) C_{isj} \\ &= \sum_{s=1}^{S} \sum_{j \in A_s} \left(\frac{\partial^2 \mathcal{L}_i}{\partial \alpha_{rk} \partial \alpha_{sj}} \right) C_{isj} \\ &= \sum_{s=1}^{S} \sum_{j \in A_s} \left[\delta_{sr} \left(\delta_{jk} - P_{irk|r}\right) \hat{n}_{isj} - T \left(\delta_{jk} - P_{irk}\right) P_{isj} \right] C_{isj}. \end{aligned} \tag{A.20}$$

Verifying the identification of the parameter vector $\phi \equiv (\alpha_{\bullet\bullet}, \beta)$ on the basis of the implied Hessian matrix is difficult. However, there are several arguments that suggest identification indeed holds. First, the column of the Hessian matrix associated with β is almost surely not collinear with the other columns, as it is a travel cost weighted average of those other columns, with the travel costs varying by individual and site. Moreover, the source of any identification problem is likely to lie with columns associated with the ASC's from the same segment. This is because $\frac{\partial^2 \mathcal{L}_i}{\partial \alpha_{rk} \partial \alpha_{sj}}$ is unchanged from the disaggregated model when $r \neq s$. That is, the concern would lie with the diagonal sub-matrices from the full Hessian, defined by:

$$\begin{aligned}\mathcal{H}_s &\equiv \frac{\partial^2 \mathcal{L}_i}{\partial \alpha_{s\bullet} \partial \alpha'_{s\bullet}} \\ &= \mathcal{A}_s - \mathcal{B}_s \end{aligned} \tag{A.21}$$

where the $(j,k)^{th}$ elements of $\mathcal{A}_s$ and $\mathcal{B}_s$ are given, respectively, by:

$$\mathcal{A}_{s,(j,k)} = n_{is\bullet} \left(\delta_{jk} - P_{isk|s}\right) P_{isj|s} \tag{A.22}$$

and

$$\mathcal{B}_{s,(j,k)} = T\left(\delta_{jk} - P_{isk}\right) P_{isj}. \tag{A.23}$$

Note that $\mathcal{A}_s$ and $\mathcal{B}_s$ have the same structure, with the only difference being that $\mathcal{A}_s$ focuses on trips with the segment, while $\mathcal{B}_s$ focuses on all trips. We know that $\mathcal{B}_s$ is nonsingular, since it is a diagonal block of the Hessian matrix in the case of disaggregated data, suggesting that $\mathcal{A}_s$ is also nonsingular. This, of course, does not rule out the possibility that the difference the two matrices might be singular. Second, numerous Monte Carlo simulation exercises (available from the authors upon request) demonstrate that the α_{sj} can be recovered, despite aggregation of trips to the segment level, when the true underlying data generating process is indeed a logit model.

At the same time, it is also clear that the α_{sj}'s are only *poorly* identified, depending on the structure of the model for identification. This can be seen if we consider a linear probability model instead of the logit specification. In this case, the aggregated choice probabilities become

$$\begin{aligned} P_{is\bullet} &\equiv \sum_{j \in A_s} V_{isj} \\ &= \sum_{j \in A_s} \alpha_{sj} + \beta \sum_{j \in A_s} C_{isj} \\ &= \alpha_{s\bullet} + \beta C_{is} \end{aligned} \tag{A.24}$$

where $\alpha_{s\bullet} \equiv \sum_{j \in A_s} \alpha_{sj}$ and $C_{is} \equiv \sum_{j \in A_s} C_{isj}$. The aggregated choice probabilities can be written as functions of the aggregated access-point ASC's (i.e., $\alpha_{s\bullet}$) and β, precluding identification of the individual access point ASC's. Similar problems emerge if access point characteristics are included in the model, as this corresponds to a restriction of the form $\alpha_{sj} = \alpha_s + \gamma X_{sj}$, where X_{sj} denotes a segment level attribute. In this case, *gamma* will only be identified via the assumed structure of the logit model.

Given the fact that access-point ASC's are only structurally identified, it would seem prudent to incorporate ASC's only at the level in which the data are available (i.e., the segment level), particularly if the segments themselves are relatively uniform in quality. This would

correspond to the restriction:

$$\alpha_{sj} = \alpha_s \ \forall j \in A_s. \tag{A.25}$$

The first order conditions in this case become:

$$\begin{aligned} 0 = \frac{\partial \mathcal{L}}{\partial \alpha_s} &= \sum_{j \in A_s} \frac{\partial \mathcal{L}}{\partial \alpha_{sj}} \\ &= \sum_{i=1}^{N} (n_{is\bullet} - T \cdot P_{is\bullet}) \end{aligned}$$

and

$$\begin{aligned} 0 &= \frac{\partial \mathcal{L}}{\partial \beta} \\ &\approx \sum_{i=1}^{N} \sum_{r=1}^{S} (n_{ir\bullet} - T \cdot P_{ir\bullet})\, C_{ir\bullet}, \end{aligned}$$

where

$$P_{is\bullet} = \frac{exp(V_{is\bullet})}{1 + \sum_{n=1}^{S} exp(V_{in\bullet})} \tag{A.26}$$

and

$$V_{is\bullet} = \alpha_s + \beta C_{is\bullet}, \tag{A.27}$$

with

$$C_{is\bullet} \equiv \frac{1}{\beta} ln[\sum_{j \in A_s} exp(\beta C_{isj})] \approx \sum_{k \in A_s} P_{isk|s} C_{isk}. \tag{A.28}$$

As noted in the body of the text, the aggregated choice model has the same basic logit structure, except that the travel cost is a probability weighted average of the access-point travel costs. The first order conditions are, likewise, analogous. In particular, the first order conditions for the ASC's are such that, as is case in a standard logit setting, the model will be mean fitting (i.e., the predicted number of trips to a given segment will precisely equal the actual number of trips).

Nested Logit Models

We consider in the paper two nested logit specifications, where the conditional utility received by individual i in visiting access point j along segment s on choice occasion t is given by

(A.2), where now $V_{isj} = \alpha_s + \beta C_{isj}$; i.e., there are segment level alternative specific constants. The first specification allows for a nest for all of the trip alternatives, whereas the second model treats each segment as a separate nest.

Specification 1: Trip Nest

In the first specification, all of the segments (and their associated access points) are grouped together in a single nest. In this case, the choice probability for access point j becomes:

$$P_{isj} = exp(\tilde{V}_{isj}) \left[\sum_{r=1}^{S} \sum_{k \in A_r} exp(\tilde{V}_{irk}) \right]^{\theta-1} \left\{ 1 + \left[\sum_{r=1}^{S} \sum_{k \in A_r} exp(\tilde{V}_{irk}) \right]^{\theta} \right\}. \tag{A.29}$$

where

$$\tilde{V}_{isj} = \frac{V_{isj}}{\theta} = \frac{\alpha_s}{\theta} + \frac{\beta C_{isj}}{\theta} = \tilde{\alpha_s} + \tilde{\beta} C_{isj}. \tag{A.30}$$

The choice probability for aggregate site s becomes

$$\begin{aligned} P_{is\bullet} &= \sum_{j \in A_s} exp(\tilde{V}_{isj}) \left[\sum_{r=1}^{S} \sum_{k \in A_r} exp(\tilde{V}_{irk}) \right]^{\theta-1} \left\{ 1 + \left[\sum_{r=1}^{S} \sum_{k \in A_r} exp(\tilde{V}_{irk}) \right]^{\theta} \right\} \\ &= exp(\tilde{V}_{is\bullet}) \left[\sum_{r=1}^{S} \sum_{k \in A_r} exp(\tilde{V}_{ir\bullet}) \right]^{\theta-1} \left\{ 1 + \left[\sum_{r=1}^{S} \sum_{k \in A_r} exp(\tilde{V}_{ir\bullet}) \right]^{\theta} \right\} \end{aligned} \tag{A.31}$$

where

$$\begin{aligned} \tilde{V}_{is\bullet} &= ln \left[\sum_{j \in A_s} exp(\tilde{V}_{isj}) \right] \\ &= \tilde{\alpha_s} + ln \left[\sum_{j \in A_s} exp(\tilde{\beta} C_{isj}) \right] \\ &= \tilde{\alpha_s} + \tilde{\beta} \tilde{C}_{is} \end{aligned} \tag{A.32}$$

with

$$\begin{aligned} \tilde{C}_{is} &= \frac{1}{\tilde{\beta}} ln \left[\sum_{j \in A_s} exp(\tilde{\beta} C_{isj}) \right] \\ &\approx \sum_{j \in A_s} P_{isj|s} C_{isj}. \end{aligned} \tag{A.33}$$

Note that the structure of the choice probability in (A.31) is identical to that for the access point probabilities, except that the site cost is now a nonlinear function (index) of the access

point level travel cost in (A.29). Note that, even if the travel cost are identical along aggregate segment (i.e., $C_{isj} = C_{isk} = C_{is}\ \forall j,k \in A_s$), the travel cost parameter β will be identified by variation in travel cost across sites. Indeed, in this case, C_{is} is no longer a function of $\tilde{\beta}$ and the model reduces to a standard nested logit model.

Specification 2: Segment Nests

In the second specification, the access points within each segment form distinct nests. In this case, (A.29) is replaced with

$$P_{isj} = exp(\breve{V}_{isj}) \left[\sum_{k\in A_s} exp(\breve{V}_{isk})\right]^{\theta_s - 1} \left\{1 + \sum_{r=1}^{S} \left[\sum_{k\in A_r} exp(\breve{V}_{irk})\right]^{\theta_r}\right\}, \tag{A.34}$$

where

$$\breve{V}_{isj} = \frac{V_{isj}}{\theta_s} = \frac{\alpha_s}{\theta_s} + \frac{\beta C_{isj}}{\theta_s} = \breve{\alpha}_s + \breve{\beta}_s C_{isj}. \tag{A.35}$$

Now, (A.31) becomes

$$\begin{aligned} P_{is\bullet} &= \sum_{j\in A_s} exp(\breve{V}_{isj}) \left[\sum_{k\in A_s} exp(\breve{V}_{isk})\right]^{\theta_s - 1} \left\{1 + \sum_{r=1}^{S} \left[\sum_{k\in A_r} exp(\breve{V}_{irk})\right]^{\theta_r}\right\} \\ &= \left[\sum_{k\in A_s} exp(\breve{V}_{isk})\right]^{\theta_s} \left\{1 + \sum_{r=1}^{S} \left[\sum_{k\in A_r} exp(\breve{V}_{irk})\right]^{\theta_r}\right\} \\ &= \exp(\breve{V}_{is\bullet}) \left\{1 + \sum_{r=1}^{S} exp(\breve{V}_{ir\bullet})\right\} \end{aligned} \tag{A.36}$$

where

$$\begin{aligned} \breve{V}_{is\bullet} &= ln \left\{\left[\sum_{k\in A_s} exp(\breve{V}_{isk})\right]^{\theta_s}\right\} \\ &= \theta_s ln \left\{\left[\sum_{k\in A_s} exp(\breve{V}_{isk})\right]\right\} \\ &= \theta_s \breve{\alpha}_s + \theta_s ln \left[\sum_{j\in A_s} exp(\breve{\beta}_s C_{isj})\right] \\ &= \theta_s \breve{\alpha}_s + \theta_s \breve{\beta}_s \breve{C}_{is} \\ &= \alpha_s + \beta \breve{C}_{is} \end{aligned} \tag{A.37}$$

with

$$\breve{C}_{is} = \frac{1}{\breve{\beta}_s} ln \left[\sum_{j \in A_s} exp(\breve{\beta}_s C_{isj}) \right]. \tag{A.38}$$

Note that the segment choice probabilities look much like a standard logit model in this case, *except* for the fact that the segment cost is a nonlinear function of the access point costs. Identification of θ_s is equivalent to identification of β_s and hence is tied to the nonlinear relationship in (A.38).

Normal Error Component Mixed Logit Models The paper considers two aggregated versions of the normal error component logit mixture models (NECLM). We draw on results in Walker *et al.,* (2) in examining issues of identification for these models. In both models, we assume that $V_{isj} = \alpha_s + \beta C_{isj}$; i.e., there are segment level alternative specific constants.

Specification 1: Trip Nest

In the first model, the error component $\tau_{it} \sim \mathcal{N}(0, \sigma_\tau^2)$ is shared by all trip alternatives (i.e., all access points and segments), so that:

$$U_{isjt} = \begin{cases} \epsilon_{isjt} & s = j = 0 \\ V_{isj} + \tau_{it} + \epsilon_{isjt} & \text{otherwise,} \end{cases} \tag{A.39}$$

where the ϵ_{isjt}'s are distributed *i.i.d.* Gumbel$(0, 1)$. The utility associated with segment s is then given by

$$\begin{aligned} U_{is\bullet t} &= \max_{j \in A_s} (V_{isj} + \epsilon_{isjt} + \tau_{it}) \\ &= \max_{j \in A_s} (V_{isj} + \epsilon_{isjt}) + \tau_{it} \\ &= ln \left(\sum_{j \in A_s} exp(V_{isj}) \right) + \epsilon_{is\bullet t} + \tau_{it} \\ &= \alpha_s + \beta C_{is\bullet} + \epsilon_{is\bullet t} + \tau_{it} \end{aligned} \tag{A.40}$$

where $\epsilon_{is\bullet t}$'s are distributed *i.i.d.* Gumbel$(0, 1)$, with the third equality following from Property 3 of Gumbel distributions given the normalization of $\mu = 1$.

The aggregated conditional utility functions define a new choice model with the similar nest structure as the disaggregated model except that the choice now is correspondent to

the segment level choice instead of access point level choice, and the new travel cost variable becomes a weighted average of original travel cost variables. The disaggregated model is no doubt identified, thus the new aggregated model is also identified.

Specification 2: Segment Nest

In specification 2, the error component $\tau_{ist} \sim \mathcal{N}(0, \sigma_\tau^2)$ is shared by all trip alternatives within the same segment, so that:

$$U_{isjt} = \begin{cases} \epsilon_{isjt} & s = j = 0 \\ V_{isj} + \tau_{ist} + \epsilon_{isjt} & \text{otherwise,} \end{cases} \tag{A.41}$$

where the ϵ_{isjt}'s are distributed *i.i.d.* Gumbel$(0, 1)$. The utility associated with segment s is then given by

$$\begin{aligned} U_{is\bullet t} &= \max_{j \in A_s} (V_{isj} + \epsilon_{isjt} + \tau_{ist}) \\ &= \max_{j \in A_s} (V_{isj} + \epsilon_{isjt}) + \tau_{ist} \\ &= ln \left(\sum_{j \in A_s} exp(V_{isj}) \right) + \epsilon_{is\bullet t} + \tau_{ist} \\ &= \alpha_s + \beta C_{is\bullet} + \epsilon_{is\bullet t} + \tau_{ist} \end{aligned} \tag{A.42}$$

where $\epsilon_{is\bullet t}$'s are distributed *i.i.d.* Gumbel$(0, 1)$

This is one of *the alternative specific variance model* discussed in page 1103 - 1107 of Walker *et al.,* (2). We follow their procedures to check the order and rank condition.

- Order Condition

  If there are S (river) segments, the order condition tells that there are $\frac{(S-1)S}{2}$ parameters in the error terms can be identified. For a model with S nests, one for each river segment, the number of unknown parameters about error terms, i.e., normal error terms representing the nest structures (τs) and extreme Type I errors (ϵs), is $S + 1$. Thus, when S is bigger enough, say $s \geq 4$, the maximum number of estimable parameters ($\frac{(4-1)\times 4}{2} = 6$) is larger than the total number of unknown parameters of error terms ($4 + 1 = 5$).

- Rank Condition

Here, we assume S and for each segment s there are J_s access points. Based on the above derivation, the new choice model looks like

$$U_{i1\bullet t} = \alpha_1 + \beta C_{i1\bullet} + \sigma_1 \tau_{i1} + \epsilon_{i1\bullet t}$$

$$U_{i2\bullet t} = \alpha_2 + \beta C_{i2\bullet} + \sigma_2 \tau_{i2} + \epsilon_{i2\bullet t}$$

$$\dots$$

$$U_{iS\bullet t} = \alpha_S + \beta C_{iS\bullet} + \sigma_S \tau_{iS} + \epsilon_{iS\bullet t}$$

$$U_{i0t} = \epsilon_{i0t}$$

where the $\alpha_s + \beta C_{is\bullet}, \forall s = 1, 2, \dots, S$ in the conditional utility function are the abbreviations for the terms $ln\left(\sum_{j \in A_s} exp(V_{isjt})\right)$, $s = 1, 2, \dots, S$.

The utility difference by subtracting U_{i0t} will be

$$\begin{bmatrix} \alpha_1 + \beta C_{i1\bullet} + \sigma_1 \tau_{i1} + \epsilon_{i1\bullet t} - \epsilon_{i0t} \\ \alpha_2 + \beta C_{i2\bullet} + \sigma_2 \tau_{i2} + \epsilon_{i2\bullet t} - \epsilon_{i0t} \\ \dots \\ \alpha_S + \beta C_{iS\bullet} + \sigma_S \tau_{iS} + \epsilon_{iS\bullet t} - \epsilon_{i0t} \end{bmatrix}$$

The covariance matrix thereafter is

$$\begin{bmatrix} \sigma_1^2 + 2\mu g & \mu g & \dots & \mu g \\ \mu g & \sigma_2^2 + 2\mu g & \dots & \mu g \\ \dots & \dots & \dots & \dots \\ \mu g & \mu g & \dots & \sigma_S^2 + 2\mu g \end{bmatrix}_{S \times S}$$

where g is the variance of a standard Gumbel distribution, $g = \pi^2/6$. The matrix has the same off-diagonal element of μg and diagonal elements, (s, s), of $\sigma_s^2 + 2\mu g, s = 1, 2, \dots, S$. There are $S+1$ unique elements in this matrix, $(\sigma_1^2 + 2\mu g, \dots, \sigma_S^2 + 2\mu g, \mu g)$. The Jacobian matrix of this vector with respect to four unknown parameters, $(\sigma_1^2, \dots, \sigma_S^2, \mu)$ is

$$\begin{bmatrix} 1 & 0 & \dots & 2g \\ 0 & 1 & \dots & 2g \\ \dots & \dots & \dots & \dots \\ 0 & \dots & 1 & 2g \\ 0 & \dots & 0 & g \end{bmatrix}_{([S+1] \times [S+1])}$$

The rank of this matrix is clearly $S + 1$, which means, according to Walker *et al.,* (2), the model can only identify up to S unknown parameters of error terms and we need to normalize $\mu = 1$ to identify other S σ_s.

- Equity Condition

  With the simply formal of the covariance matrix of utility differences, the equity condition is satisfied easily with the normalization of $\mu = 1$.

Appendix B

Description of Pseudo Data Set

For a typical simulation, we generate the pseudo data set using the following steps.

1. Generating access points along each river segments

   Locate all the river access points and household locations in a 400 by 400 surface to mimic the territory of Iowa. For the S river segments, we first randomly generate S coordinate pairs $(x_s, y_s), s = 1, 2, \ldots, S$ for the central point, also the middle access point of each river segment.[1] Those points are assumed to be uniformly distributed in the interval of $[50, 350]$.[2]

   Once we have the coordinates of the midpoint of each segment, $(x_s, y_s), s = 1, 2, \ldots, S$, the direction vector for each segment will be generated with the following rules.

$$\begin{aligned} \delta_{xs} &\sim U[0, 25] \\ \delta_{ys} &= \sqrt{25^2 - \delta_{xs}^2} \end{aligned} \tag{A.43}$$

   With these direction vectors and if there are $J = 2k+1$(k is a positive integer and equals 1 in this simulation.) access points along the straight river segment s, the access points for this segment will be

$$\begin{aligned} x_{sj} &= x_s + \frac{(j-k)\delta_{xs}}{k} j = 1, 2, \ldots, J \\ y_{sj} &= y_s + \frac{(j-k)\delta_{ys}}{k} j = 1, 2, \ldots, J \end{aligned} \tag{A.44}$$

   We also consider the more realistic situations (K and C) when river segments are kinked, this nonlinear river segments are achieved by generating $2k$ small direction vectors $(\delta_{xs}^m, \delta_{ys}^m), m = 1, 2, \ldots, 2k$.[3] Then the access points for the segment s except for the middle point are,[4]

$$\begin{aligned} x_{sj} &= x_s + sign(j-k)\delta_{xs}^j, j = 1, 2, \ldots, 2k \\ y_{sj} &= y_s + sign(j-k)\delta_{ys}^j, j = 1, 2, \ldots, 2k \end{aligned} \tag{A.45}$$

[1] For simplicity, we assume there are same odd number of access points along each river segment. Specifically, $J_s = 3$ in all the simulations

[2] This arrangement will easily restrict the river segment in the interior of the 400 by 400 surface.

[3] There will be some restrictions on the scale of these vectors to make sure those access points are still in the 400 by 400 surface

[4] $sign(0) = -1$

In doing so, the points, $(x_{sj}, y_{sj}), j = 1, \ldots, J_s$, will not form a line by the randomness of direction vectors,$(\delta_{xs}^m, \delta_{ys}^m), m = 1, 2, \ldots, 2k$. In this application, we assume there are three access points along each river segment (k=1).

2. Generating the location of households

We distinguish two scenarios for the distribution of households. In the first scenario called uniform scenario (U), let there are I households, the location of the i^{th} household is given by $x_i^h, y_i^h, i = 1, 2, \ldots, I$, where

$$\begin{aligned} x_i^h &\sim U[0, 400] \\ y_i^h &\sim U[0, 400] \end{aligned} \tag{A.46}$$

In the scenario called population center scenario(C), 60 percent of the households are distributed around the first two river segments, while other households are uniformly located in the other area of 400 by 400 surface.[5] The objective of having population centers is to check the possible effects of different population distributions on the estimation results.

3. Travel distances and travel costs.

The one-way travel distance for household i to the segment-entry point combination $sj, s = 1, 2, \ldots, S$ and $j = 1, 2, \ldots, J_s$, is given by

$$d_{isj} = \sqrt{(x_{sj} - x_i^h)^2 + (y_{sj} - y_i^h)^2}. \tag{A.47}$$

The travel cost for household i to the segment-entry point combination sj then is calculated by the formula:

$$C_{isj} = 2d_{isj}(f + \frac{w}{3m}). \tag{A.48}$$

where f denote the cost per mile, w denotes the wage rate and m denotes the average speed (i.e., miles per hour), which is used to convert distance into time.[6] The similar

[5]Using S=5 as an example, we first randomly locate half of households in the whole 400 by 400 surface and randomly locate the other half in two squares with the first two river segments' middle points as the center and a edge length of 50. By doing so, we approximately achieve the goal of having 60 percent of population around the first two river segments. For the other number of river segments, similar procedure is taken.

[6]We set $f = \$0.25, w = \$30$ and $m = 50MPH$ in the data generation process.

formula is extensively used in recreation literature to calculate the travel cost (see, e.g., (2) and (14)).

4. Generating choice variable y_{isjt} and *observed* choice variable n_{is}.

   The conditional utility for individual i to visit the river segment s through segment-entry point combination sj is assumed to take the following form

$$U_{isjt} = V_{isj} + \epsilon_{isjt} = \alpha_s + \beta C_{isj} + [\beta_w w_{sj}] + \epsilon_{isjt}. \tag{A.49}$$

   where α_s is the segment specific constant, which is generated from a uniform distribution of $[a-\tau, a+\tau]$, a is the mean and τ measures the spread. β measuring the marginal utility of income. w_{sj} is the water quality at the segment-entry point combination sj, β_w is the corresponding coefficient.[7]. ϵ_{isjt} are obtained from the type I extreme value distribution. The baseline utility of staying-at-home is normalized to 0 (i.e., $V_{i00t} = 0$). By varying the value of a and τ, the segment specific constants are chosen to maintain the percentage of households choosing staying-at-home option in the certain range.[8]

   Once having the values of U_{isjt}, we set y_{isjt} to 1 if U_{isjt} happens to the highest realized value and set other $y_{i\hat{s}\hat{j}t}, \hat{s} = 1, \ldots, S$ and $\hat{j} = 1, \ldots, J_{\hat{s}}$ and $\hat{s} \neq s, \hat{j} \neq j$ to 0. We repeat this process for all I households and T choice occasions.

   In the real data we used, we only have the segment level information rather than the subsegment level information. To mimic this situation,we aggregate households' y_{isjt} to get the segment-household variable n_{is} by the formula,

$$n_{is} = \sum_{j=1}^{J_s} \sum_{t=1}^{T} y_{isjt} \tag{A.50}$$

[7] In our Monte Carlo experiment, the β_w can not be fully recovered in some models because the perfectly collinearity between the water quality term and the segment specific constant. The two stage method suggested in Murdock (1) is not applicable here since we only have at most 20 observations about water quality, which make the estimation in second stage unstable.

[8] In Egan et al.(2009) about Iowa lake recreation, there were around 60% percent of people who did not have any lake related recreation. In our simulation, we treat this figure as a simulation target.

Appendix C

Table A.1 ASCs from the First Stage Estimation

Variable	Model 1 Agg. Prob.	Model 2 Shortest Distance	Model 3 Midpoint Proxy
ASC1	5.0168	-2.5803	0.6417
ASC2	6.4838	-2.2305	0.0695
ASC3	4.2019	-3.4634	-0.4817
ASC4	4.493	-3.7579	-1.4633
ASC5	4.545	-2.9166	0.1653
ASC6	4.2614	-3.4752	-0.4998
ASC7	3.1682	-4.5548	-1.5384
ASC8	2.9342	-5.0963	-2.3111
ASC9	3.6907	-4.4287	-1.6011
ASC10	4.0319	-4.2685	-1.9934
ASC11	4.0704	-3.9964	-1.1845
ASC12	3.1332	-4.8918	-2.2246
ASC13	4.0769	-3.7241	-0.6416
ASC14	3.6216	-4.3763	-1.6162
ASC15	4.9505	-3.3022	-1.1118
ASC16	4.3029	-3.591	-0.871
ASC17	2.9853	-4.957	-2.2786
ASC18	4.573	-3.6875	-1.486
ASC19	3.1382	-4.5025	-1.7223
ASC20	4.2786	-3.3988	-0.5855
ASC21	4.6964	-2.8534	0.348
ASC22	4.556	-3.2944	-0.1479
ASC23	5.1354	-2.7182	0.5686
ASC24	5.8183	-2.3012	0.5805
ASC25	6.272	-2.0915	0.1649
ASC26	5.4714	-2.4832	0.8248
ASC27	3.9506	-4.1008	-1.4409
ASC28	3.8002	-4.1785	-1.6687
ASC29	3.3551	-4.2817	-1.2639
ASC30	2.7924	-4.8721	-2.0278
ASC31	3.5914	-4.4605	-1.8268
ASC32	3.7831	-4.2534	-1.733
ASC33	4.1951	-3.696	-0.8774
ASC34	4.9502	-3.3018	-1.0835
ASC35	4.1013	-3.524	-0.2001

Table A.1 ASCs from the First Stage Estimation (con't)

Variable	Model 1 Agg. Prob.	Model 2 Shortest Distance	Model 3 Midpoint Proxy
ASC36	4.4308	-3.6547	-0.9388
ASC37	3.0409	-5.2913	-2.9776
ASC38	4.0977	-3.8005	-1.1647
ASC39	4.5527	-4.2716	-2.0123
ASC40	4.7399	-3.3317	-0.6854
ASC41	3.5073	-4.3099	-0.5902
ASC42	3.3274	-4.4224	-1.3207
ASC43	4.8567	-2.994	0.2139
ASC44	2.8565	-5.2535	-2.3306
ASC45	5.0799	-2.7595	0.2832
ASC46	4.8258	-2.832	0.5209
ASC47	5.5218	-2.552	0.4069
ASC48	3.9406	-4.3586	-2.0055
ASC49	5.5566	-2.3159	0.3292
ASC50	4.9957	-2.8903	0.3013
ASC51	4.1497	-3.9638	-1.1712
ASC52	5.991	-1.8865	1.1902
ASC53	5.373	-2.5519	0.1782
ASC54	4.8605	-3.0844	0.3977
ASC55	4.6052	-3.0453	0.2042
ASC56	4.7738	-2.7955	0.4751
ASC57	3.5038	-4.4903	-1.8154
ASC58	5.1499	-2.4113	0.7936
ASC59	3.3877	-4.2272	-1.4112
ASC60	5.1493	-2.462	0.2613
ASC61	5.5962	-2.7213	-0.5656
ASC62	5.0793	-2.7603	-0.2031
ASC63	5.3224	-2.6086	-0.1805
ASC64	6.2124	-1.6717	0.8552
ASC65	6.3197	-1.9378	0.9622
ASC66	5.8436	-2.0881	1.3979
ASC67	5.338	-2.8586	0.2896
ASC68	8.0218	-0.3075	1.8435
ASC69	7.6212	-0.5893	2.1892
ASC70	7.4072	-0.8687	2.2833
ASC71	6.8148	-1.3366	1.5669
ASC72	6.1517	-2.2009	1.1546
ASC73	7.4835	-0.9342	1.4901

Appendix D

Iowa River Survey 2009 Sample

Bibliography

[1] Murdock, Jennifer. (2006) "Handling unobserved site characteristics in random utility models of recreation demand", *Journal of Environmental Economics and Management*,Vol. 51, No.1 pp. 1-25.

[2] Walker, J.L., M. Ben-Akiva, and D. Bolduc (2007) "Identification of Parameters in Normal Error Component Logit-Mixture (NECLM) Models" *Journal of Applied Econometrics* Vol. 22 pp 1095-1125.

APPENDIX B. Additional Material for Chapter 3

Some Properties about Gumbel Distribution

The backbone of logit model is based on the Extreme Type I distributional assumption (*Standard Gumbel Distribution*) about the unobservable part of utility function, from researchers' perspective. In addition to having close form about choice probabilities, the standard Gumbel distribution has other properties. In order to check the identification conditions later, we list several relative properties here with brief derivation.

- Property 1: If $X \in$ *Gumbel(0,*μ*)*, then $X/\mu \in$ *Gumbel(0,1)*.
- Property 2: If $X \in$ *Gumbel(0,*μ*)*, then $X + A \in$ *Gumbel(A,*μ*)*.
- Property 3: If $x_1, x_2, \ldots, x_N \in$ *IID Gumbel(0,*μ*)*, then
  $y = \max(x_1, x_2, \ldots, x_N) \in$ *Gumbel(*$\mu ln(N), \mu$*)*.

  Proof:

$$
\begin{aligned}
\text{CDF:} F_{x_i}(X) &= e^{-e^{-X/\mu}} \in \ \textit{Gumbel(0,}\mu\textit{)},\ i = 1,2,\ldots,N \\
\text{CDF:} F_y(Y) &= \prod_{i=1}^{N} F_{x_i}(Y/\mu) \\
&= \prod_{i=1}^{N} e^{-e^{-Y/\mu}} \\
&= e^{-Ne^{-Y/\mu}} \\
&= e^{-e^{-(Y-\mu ln N)/\mu}} \in \ \textit{Gumbel(}\mu ln(N)\textit{,}\mu\textit{)}
\end{aligned}
$$

- Property 4: If $x_1, x_2, \ldots, x_N \in$ *Independent Gumbel(*A_i, μ*)*, then $y = \max(x_1, x_2, \ldots, x_N)$ will belong to *Gumbel(*$\mu ln(\sum_{i=1}^{N} exp(A_i/\mu))$*,1)*.

Following the derivation in *Property 3*, the cumulative distribution function of y will be $F_y(Y) = e^{-e^{-(Y-\mu ln(\sum_{i=1}^{N} exp(A_i/\mu))/\mu}}$), which is $Gumbel(\mu ln(\sum_{i=1}^{N} exp(A_i/\mu)), \mu)$.

Identification conditions of Normal Error Component Logit Model (NECLM)

Walker *et al.* (2) proposes three conditions to check the identifiability of parameters in the error terms. The three conditions are:

- Order condition

  The number of estimable parameters in the error terms, S, subject to

  $$S \leq \frac{J(J-1)}{2} - 1$$

  where J is the number of choice options.

- Rank condition

  The number of estimable parameters in the error terms, S, adheres to

  $$S = rank(Jocobian(vecu(\Omega_\Delta))) - 1$$

  where Ω_Δ is the covariance matrix of utility difference, $\Delta(U_j)$, which is obtained by subtracting utility of U_j from each utility, $U_k \quad k = 1, \ldots, J$. $\Delta(U_j)$ will be a matrix with rank $J-1$. *vecu* vecterizes the unique elements of Ω_Δ into a column vector. The Jacobian matrix is obtained by differentiating the vector with respect to the unknown parameters in the error term.

- Equality condition

  This condition requires there are at least one possible set of values for the unknown parameters in the error term which gives the same covariance matrix as the normalized covariance matrix.

  $$\Omega_\Delta = \Omega_\Delta^{Normalized}$$

As pointed out by the authors, the rank condition usually puts a tighter bound on the maximum number of estimable parameters and the equality condition could make the normalization process very difficult because of the complicated structure of the covariance matrix. In the

following part, we will put emphasis on the rank condition and bypass the discussion of the equality condition due to the fact the aggregation model makes the covariance matrix extremely complex.

Identification of the Aggregate Probability Model

In this part, we consider the situation in which there are $J = 5$ options in the choice set and we assume the conditional utility function of choosing each option has the following forms:

$$\begin{aligned} U_{i1} &= V_{i1} + [\sigma_t\tau_t + \sigma_l\tau_l] + \eta_{i1} \\ U_{i2} &= V_{i2} + [\sigma_t\tau_t + \sigma_l\tau_l] + \eta_{i2} \\ U_{i3} &= V_{i3} + [\sigma_t\tau_t + \sigma_r\tau_r] + \eta_{i3} \\ U_{i4} &= V_{i4} + [\sigma_t\tau_t + \sigma_r\tau_r] + \eta_{i4} \\ U_{i0} &= V_{i0} + \eta_{i0} \end{aligned} \tag{B.1}$$

where $V_{i\bullet}$ stands for the observed part of conditional utility, τ_t, τ_l, τ_r are independent standard normal errors with standard deviations of $\sigma_t, \sigma_l, \sigma_r$, respectively. $\eta_{i\bullet}$ are the *iid Gumbel(0,μ)* errors. The presence of the terms in the brackets represents different nest structures: i) standard multinomial logit model if none of normal errors show up, ii) a nest structure like case 2 in Figure 1, iii) a nest structure like case 3 in Figure 1, iv) a two-layer nest structure like case 4 in Figure 1.

The partial observation of group choice on option $j = 0, 3, 4$ is equivalent to a new choice model in which there are three options and the conditional utility functions are defined as

$$\begin{aligned} U_{i1} &= V_{i1} + [\sigma_t\tau_t + \sigma_l\tau_l] + \eta_{i1} \\ U_{i2} &= V_{i2} + [\sigma_t\tau_t + \sigma_l\tau_l] + \eta_{i2} \\ U_{i\widetilde{0}} &= \max\{U_{i0}, U_{i3}, U_{i4}\} \end{aligned}$$

where $U_{i\bullet}$ is defined as above. According to the properties about standard Gumbel distribution, the $U_{i\widetilde{0}}$ could be written as

$$U_{i\widetilde{0}} = \mu ln(\{\sum_{j=3}^{4} exp((V_{ij} + [\sigma_t\tau_t + \sigma_r\tau_r])/\mu)\} + 1) + \eta_{i\widetilde{0}}$$

where $\eta_{i\widetilde{0}}$ is a $Gumbel(0, \mu)$ random variable. Right now, the normal error terms enter into utility functions nonlinearly. The new model will not be a NECLM although it is based on a NECLM. The three conditions can no longer be used here to check the identifiability, while we will discuss the issue by following the spirit of the Walker *et al.* (2), that is, to discuss whether the covariance matrix implied in the new model have enough variation to identify the parameters in the error terms.

Case 1: Standard Multinomial Logit Model

In this case, there are error terms in the brackets and the new choice model becomes

$$\begin{aligned} U_{i1} &= V_{i1} + \eta_{i1} \\ U_{i2} &= V_{i2} + \eta_{i2} \\ U_{i\widetilde{0}} &= ln(exp(V_{i3}) + exp(V_{i4}) + 1) + \eta_{i\widetilde{0}} \end{aligned}$$

Normalizing $\mu = 1$, this is a standard multinomial logit model with a nonlinear utility part in the newly defined group option $j = \widetilde{0}$. Another observation is that if there are ASCs in the original utility parts, $V_{i\bullet}$, the model can not identify all the ASCs in the original model. This observation also holds in other cases.

Case 2: Model with the nest structure in case 2 of Figure 1

The model defined by (B.1) with full information is identified with the normalization of $\mu = 1$ via checking the order and rank conditions.

The new choice becomes

$$\begin{aligned} U_{i1} &= V_{i1} + [\sigma_t \tau_t] + \eta_{i1} \\ U_{i2} &= V_{i2} + [\sigma_t \tau_t] + \eta_{i2} \\ U_{i\widetilde{0}} &= \underbrace{\mu ln(exp(\sigma_t/\mu\tau_t)[exp(V_{i3}/\mu) + exp(V_{i4}/\mu)] + 1)}_{\nu_{i\widetilde{0}}} + \eta_{i\widetilde{0}} \\ &= \nu_{i\widetilde{0}} + \eta_{i\widetilde{0}} \end{aligned}$$

where ν is a random variable closely related to the exponential distribution family and independent from $\eta's$ and correlated with τ_t. After some derivation, we could know the covariance

matrix of the utility difference, $\Delta(U_{i1})$, looks like

$$\begin{pmatrix} 2g\mu^2 & g\mu^2 + cov(\tau_t, \nu) \\ g\mu^2 + cov(\tau_t, \nu) & var(\nu) + \sigma_t^2 + 2g/\mu^2 + cov(\tau_t, \nu) \end{pmatrix}$$

where $g = \pi^2/6$. The number of unique element of this matrix is 3 and the rank of Jacobian matrix of the vector of unique elements with respect to the two unknown parameters, σ_t and μ, generally will be 2. Thus the rank condition suggests we need to normalize μ to 1 to make the model identified in terms of the two parameters in the error terms.

Case 3: Model with the nest structure in case 3 of Figure 1

The full information model (B.1) is also identified by checking the order and rank conditions.

The new model implied by the partial information has the form

$$\begin{aligned} U_{i1} &= V_{i1} + [\sigma_l \tau_l] + \eta_{i1} \\ U_{i2} &= V_{i2} + [\sigma_l \tau_l] + \eta_{i2} \\ U_{i\tilde{0}} &= \underbrace{\mu ln(exp(\sigma_r/\mu\tau_r)[exp(V_{i3}/\mu) + exp(V_{i4}/\mu)] + 1)}_{\nu_{i\tilde{0}}} + \eta_{i\tilde{0}} \\ &= \nu_{i\tilde{0}} + \eta_{i\tilde{0}} \end{aligned}$$

where ν is a random variable defined by a nonlinear transformation of τ_r along with unknown and fixed parameters, σ_r, V_{i3} and V_{i4}. By assumption, ν is independent of η's and τ_l. Similarly, we could get the covariance matrix of the utility difference, $\Delta(U_{i1})$ as

$$\begin{pmatrix} 2g\mu^2 & g\mu^2 \\ g\mu^2 & var(\nu) + \sigma_l^2 + 2g/\mu^2 + cov(\tau_l, \nu) \end{pmatrix}$$

where $g = \pi^2/6$. The vector of unique elements of this matrix is a 3×1 vector, $(2g\mu^2, g\mu^2, var(\nu) + \sigma_l^2 + 2g/\mu^2 + cov(\tau_l, \nu))$, but $2g\mu^2$ is perfectly linear to $g\mu^2$.

The rank of the Jacobian matrix of this vector with respect to σ_l, σ_r and μ^2 is only 2 given other variables, V_{i3} and V_{i4}. Thus we need to normalize two of three unknown parameters, $(\sigma_l, \sigma_r, \mu)$ to identify the model according the rank condition in Walker *et al.* (2). However, the aggregation model makes the error term directly depend on other variables which have a large domain, thus the covariance matrix has much more variations. If the richness of covariance matrix decides the identifiability of the model, we argue that the aggregation model is still identified with the normalization of $\mu = 1$.

Case 4: Model with the nest structure in case 4 of Figure 1

The full information model (B.1) is also identified by checking the order and rank conditions.

The new model implied by the partial information has the form

$$
\begin{aligned}
U_{i1} &= V_{i1} + [\sigma_t\tau_t + \sigma_l\tau_l] + \eta_{i1} \\
U_{i2} &= V_{i2} + [\sigma_t\tau_t + \sigma_l\tau_l] + \eta_{i2} \\
U_{i\widetilde{0}} &= \underbrace{\mu ln(exp(\sigma_t/\mu\tau_t + \sigma_r/\mu\tau_r)[exp(V_{i3}/\mu) + exp(V_{i4}/\mu)] + 1)}_{\nu_{i\widetilde{0}}} + \eta_{i\widetilde{0}} \\
&= \nu_{i\widetilde{0}} + \eta_{i\widetilde{0}}
\end{aligned}
$$

where ν is a random variable defined by a nonlinear transformation of τ_t andτ_r along with unknown and fixed parameters, $\sigma_t, \sigma_r, V_{i3}$ and V_{i4}. By assumption, ν is independent of η's and τ_l. Similarly, we could get the covariance matrix of the utility difference, $\Delta(U_{i1})$ as

$$
\begin{pmatrix}
2g\mu^2 & g\mu^2 + cov(\tau_t, \nu) \\
g\mu^2 + cov(\tau_t, \nu) & var(\nu) + \sigma_l^2 + \sigma_t^2 + 2g/\mu^2 + cov(\tau_l, \nu)
\end{pmatrix}
$$

where $g = \pi^2/6$. The vector of unique elements of this matrix is a 3×1 vector, $(2g\mu^2, g\mu^2 + cov(\tau_t, \nu), var(\nu) + \sigma_t^2 + \sigma_l^2 + 2g/\mu^2 + cov(\tau_t, \nu))$. Conditional on other variables, the rank of Jacobian matrix of this vector with respect to $(\sigma_t, \sigma_l, \sigma_r, \mu^2)$ will be three. Thus rank condition suggests we must need two normalization. Again we argue that we may identify this model with only one normalization of $\mu = 1$ due to the complicated structure of ν.

Walker *et al.* (2) focuses on the identification of unknown parameters in the error terms, the identification of other unknown parameters, such as alternative specific constants (ASCs), depends on other conditions. Murdock (1) argues that ASCs could be consistently estimated with maximum likelihood estimation technique, Babatunde *et al.* (1) also gives a Bayesian framework to estimate the ASCs in RUM models. The identification of ASCs in an aggregation RUM model have not been discussed in the literature yet and we have no intention to formally prove the identifiability of ASCs here. In the simulation work, we estimate several sets of models with/without ASCs. The results show that ASCs could be well estimated under certain circumstances and the bias of partial models with ASCs has been greatly mitigated.

APPENDIX C. Additional Material for Chapter 4

Matching Fossil Fuel Units

Since the locational information from the EPA CEMS only report the unit's NERC region information, to allocate the fossil fuel units into four ERCOT zones, we need cross check other public available information from other sources. ERCOT websites (www.ercot.com and planning.ercot.com) give the annual capacity, demand and reserve reports (CDR) which have the zonal information at the plant level, which allow us to match the majority of the units into four zones. Also Cullen (2010) reports a list of units with zonal information for his research based on the generation lists provided by the ERCOT.[1] For the remaining undecided units, I assign them to a specific zone based on the county they are located since we can find the county-zone pair appeared in the CDR reports. The following table (Table A1) shows all the operating units in the research period along with their assigned zones and fuel used.[2] The locational information is also showed in the map of figure C.1.

Net Load Adjust

EPA CEMS measures the units' gloss load and does not measure the power consumed by the power plants for running the affiliated equipment, such as pollution control devices. At the same time, the demand data from the ERCOT measures the electricity consumed by consumers which only includes the electricity in the power grid. Depending on the type of boilers and type of fuels, the energy consumed by generation facilities varies significantly. The following tables (Table **??**) show the net load (EIA) and gross load information for a subset of coal units in ERCOT. Another challenge is that CEMS does not necessarily cover the output from the second stage of the combined cycle gas units. This could be seen by cross checking the net load and gross load for combined cycle units in the sample.

For the coal units, the adjustment is straightforward if we impose the assumption that the ratio defined by annual net load over annual gloss load is constant and does not vary as output level changes. For the gas units, we still use this method to convert gloss load to net load and adjust the emission

[1] There are some newly added units after the research period of Cullen's paper, which cover the generation information until 2007.

[2] An operating unit is defined as having positive output, gross load, in CEMS reports during Jun 2009 to Jun 2010.

ratio defined as tons of pollutants per MWH output. The major pollutant considered in this paper is the carbon dioxide and other pollutants, like sulfur dioxide and nitrogen oxides, are not directly modeled. Compared with the coal unit, combined cycle gas units have more complex output structures. The combined cycle configuration could include one gas turbine and one steam turbine or multiple gas turbines and one steam turbine. Since obviously some gas units in CEMS report only cover the gross output from the gas turbine, a single ratio for those units implicitly imposes the assumption that the output from the second stage, the steam turbine, is always in the same proportion to the output from the gas turbine(s). If we accept these assumptions, we could adjust the capacity, output and emission rates accordingly.

Figure C.2 shows the total generation from coal units with/without adjustment in CEMS reports compared with the total generation by coal units from ERCOT reports. The comparison of gas units is showed in Figure C.3. It could be seen that after the adjustment the sequences of fossil fuel generation match the sequences from ERCOT report quite well, especially for gas generation. Another observation is that the difference between gross coal generation and net coal generation is larger. We also calculate the correlation between sequences of hourly generation. The correlation between the sequences of adjusted coal generation and that of ERCOT reports is 0.9809. The correlation for sequences of gas generation is 0.9974.[3]

Zonal Wind Generation

In the sample period, the ERCOT 15-minutes generation reports did not report the zonal level wind generation and instead only report the ERCOT wide wind generation. To construct the zonal level wind generation, we must make some assumptions. The first assumption we make is that the zonal wind generation can be properly approximate as the ERCOT wind generation times the installed capacity of wind farms in that zone[4]. The second assumption we made is that the unmatched part between demand and generation is exogenous and will not change in other scenarios. The unmatched part is defined as follows:

$$\epsilon_{it} = d_{it} - gen_{it}^{fossilfuel} - gen_{it}^{hydro} - gen_{it}^{nuclear} - \hat{wind}_{it} - import_{it}$$

where

- ϵ_{it} is the unmatched error in zone i at time t.

[3] The correlations between gloss output in EPA CEMS and that in ERCOT reports are 0.9889 for coal generation and 0.9977 for gas generation.

[4] In GE (2008) report, they use the meteorology models to predict the wind speed in ERCOT and transform the //wind speed into wind generation.

- d_{it} is the demand in zone i at time t.
- $gen_{it}^{fossilfuel}$ is the generation from fossil fuel generators in zone i at time t.
- gen_{it}^{hydro} is the hydro generation in zone i at time t.
- $gen_{it}^{nuclear}$ is the nuclear generation in zone i at time t.
- $\hat{wind}_{it}$ is the wind generation calculated according to the installed capacity in zone i at time t.
- $import_{it}$ is the net imported electricity from other zones in zone i at time t.

Table C.3 summarized the zonal wind generation along with the zonal demand and unmatched errors. The unmatched errors sometimes are very significant. The existence of big error is understandable given the assumptions we make before. At the same time the facts that CEMS do not cover the entire fossil fleet and the uncontrolled amount of generation from other sources are all contributed to the errors.

Results from "EPA-EIA" simulations

Table 3.5-1 shows that if "EPA-EIA" heat rates are used in the simulation, our dispatch model will divert a significant amount of gas generation from the Houston zone to surrounding zones. On the other hand, if "EIA" heat rates are used instead, the model dispatch more than observed amount of gas generation in the Houston zone. The culprit behind these differences is a group of combined heat and power gas units in the Houston zone.

Table C.4 shows their heat rates and generation information. Measured by the gross generation, these co-generation units contribute almost 60% of the total fossil fuel generation in the Houston zone and more than three quarters of the total gas generation. Another observation is that their heat rates, "EPA-EIA" heat rates, calculated with combined information from EIA and EPA, CEMS are substantially higher than the heat rates, "EIA" heat rates, calculated with information from EIA alone. With these differences, it is highly possible that "EPA-EIA" heat rates will lead to lower gas generation from these units and "EIA" heat rates tend to overestimate their generation. Checking the output profile of these units in the baseline simulations confirms our speculation. The total generation from these co-generation units in the baseline simulation with "EPA-EIA" heat rates is only 11 million MWH, well below the counterpart from the baseline simulation with "EIA" heat rates, which is 44.5 million MWH.

With the data from public sources, we are afraid that some compromise must be made to precede our research questions. In this paper, we focus more on the simulation with "EIA" heat rates. It still is better to check whether the other choice will lead to systematically different emission pattern. Table

C.5 shows the counterpart emission reduction matrix as in Table 3.5-2. The pattern shown in these two tables are essentially similar, except some minor difference.

Table C.1: EPA CEMS Units and Zonal Affiliation

EPA Plant ID	Unitid	Fuel	Zone	EPA Plant ID	Unitid	Fuel	Zone
127	1	Coal	West	3506	2	Natural Gas	North
3492	CT1	Natural Gas	West	3507	9	Natural Gas	North
3492	CT2	Natural Gas	West	3508	1	Natural Gas	North
3492	CT3	Natural Gas	West	3508	2	Natural Gas	North
3492	CT4	Natural Gas	West	3508	3	Natural Gas	North
3492	CT5	Natural Gas	West	3576	BW2	Natural Gas	North
3492	CT6	Natural Gas	West	3576	BW3	Natural Gas	North
3494	5	Natural Gas	West	3576	CE1	Natural Gas	North
3494	6	Natural Gas	West	3576	GE4	Natural Gas	North
3494	CT1	Natural Gas	West	3628	**4	Natural Gas	North
3494	CT2	Natural Gas	West	3628	**5	Natural Gas	North
3494	CT3	Natural Gas	West	3628	1	Natural Gas	North
3494	CT4	Natural Gas	West	3628	2	Natural Gas	North
3494	CT5	Natural Gas	West	3628	3	Natural Gas	North
52176	1	Natural Gas	West	4195	2	Natural Gas	North
52176	2	Natural Gas	West	4195	3	Natural Gas	North
55215	GT1	Natural Gas	West	4266	4	Natural Gas	North
55215	GT2	Natural Gas	West	4266	5	Natural Gas	North
55215	GT3	Natural Gas	West	6136	1	Coal	North
55215	GT4	Natural Gas	West	6146	1	Coal	North
56349	CT1A	Natural Gas	West	6146	2	Coal	North
56349	CT1B	Natural Gas	West	6146	3	Coal	North
56349	CT2A	Natural Gas	West	6147	1	Coal	North
56349	CT2B	Natural Gas	West	6147	2	Coal	North
298	LIM1	Coal	North	6147	3	Coal	North
298	LIM2	Coal	North	6180	1	Coal	North
3452	1	Natural Gas	North	6180	2	Coal	North
3452	2	Natural Gas	North	6243	1	Natural Gas	North
3453	6	Natural Gas	North	6243	2	Natural Gas	North
3453	7	Natural Gas	North	6243	3	Natural Gas	North

3453	8	Natural Gas	North	7030	U1	Coal	North
3476	2	Natural Gas	North	7030	U2	Coal	North
3476	3	Natural Gas	North	8063	1	Natural Gas	North
3476	4	Natural Gas	North	8063	CT1	Natural Gas	North
3476	5	Natural Gas	North	8063	CT2	Natural Gas	North
3490	1	Natural Gas	North	8063	CT3	Natural Gas	North
3490	2	Natural Gas	North	8063	CT4	Natural Gas	North
3491	3	Natural Gas	North	50109	HRSG1	Natural Gas	North
3491	4	Natural Gas	North	50109	HRSG2	Natural Gas	North
3491	5	Natural Gas	North	54817	EAST	Natural Gas	North
3497	1	Coal	North	55062	1	Natural Gas	North
3497	2	Coal	North	55062	2	Natural Gas	North
3502	1	Natural Gas	North	55062	3	Natural Gas	North
3502	2	Natural Gas	North	3469	THW31	Natural Gas	Houston
3504	1	Natural Gas	North	3469	THW32	Natural Gas	Houston
3504	2	Natural Gas	North	3469	THW33	Natural Gas	Houston
55091	STK1	Natural Gas	North	3469	THW34	Natural Gas	Houston
55091	STK2	Natural Gas	North	3469	THW41	Natural Gas	Houston
55091	STK3	Natural Gas	North	3469	THW42	Natural Gas	Houston
55091	STK4	Natural Gas	North	3469	THW43	Natural Gas	Houston
55091	STK5	Natural Gas	North	3469	THW44	Natural Gas	Houston
55091	STK6	Natural Gas	North	3470	WAP1	Natural Gas	Houston
55097	1	Natural Gas	North	3470	WAP2	Natural Gas	Houston
55097	2	Natural Gas	North	3470	WAP3	Natural Gas	Houston
55097	3	Natural Gas	North	3470	WAP4	Natural Gas	Houston
55097	4	Natural Gas	North	3470	WAP5	Coal	Houston
55132	OGTDB1	Natural Gas	North	3470	WAP6	Coal	Houston
55132	OGTDB2	Natural Gas	North	3470	WAP7	Coal	Houston
55132	OGTDB3	Natural Gas	North	3470	WAP8	Coal	Houston
55139	CTG1	Natural Gas	North	7325	SJS1	Natural Gas	Houston
55139	CTG2	Natural Gas	North	7325	SJS2	Natural Gas	Houston
55172	GT-1	Natural Gas	North	10298	CG801	Natural Gas	Houston

55172	GT-2	Natural Gas	North	10298	CG802	Natural Gas	Houston
55172	GT-3	Natural Gas	North	10298	CG803	Natural Gas	Houston
55223	GT-1	Natural Gas	North	10298	CG804	Natural Gas	Houston
55226	GT1	Natural Gas	North	10741	G102	Natural Gas	Houston
55226	GT2	Natural Gas	North	10741	G103	Natural Gas	Houston
55226	GT3	Natural Gas	North	10741	G104	Natural Gas	Houston
55226	GT4	Natural Gas	North	50137	1	Natural Gas	Houston
55230	CT-1	Natural Gas	North	50815	ENG101	Natural Gas	Houston
55230	CT-2	Natural Gas	North	50815	ENG201	Natural Gas	Houston
55320	GT-1	Natural Gas	North	50815	ENG301	Natural Gas	Houston
55320	GT-2	Natural Gas	North	50815	ENG401	Natural Gas	Houston
55480	U1	Natural Gas	North	50815	ENG501	Natural Gas	Houston
55480	U2	Natural Gas	North	50815	ENG601	Natural Gas	Houston
55480	U3	Natural Gas	North	52088	GT-A	Natural Gas	Houston
55480	U4	Natural Gas	North	52088	GT-B	Natural Gas	Houston
55480	U5	Natural Gas	North	52088	GT-C	Natural Gas	Houston
55480	U6	Natural Gas	North	55015	1	Natural Gas	Houston
3460	CBY1	Natural Gas	Houston	55015	2	Natural Gas	Houston
3460	CBY2	Natural Gas	Houston	55015	3	Natural Gas	Houston
3464	GBY5	Natural Gas	Houston	55015	4	Natural Gas	Houston
3464	GBY73	Natural Gas	Houston	55047	CG-1	Natural Gas	Houston
3464	GBY74	Natural Gas	Houston	55047	CG-2	Natural Gas	Houston
3464	GBY81	Natural Gas	Houston	55047	CG-3	Natural Gas	Houston
3464	GBY82	Natural Gas	Houston	55187	CHV1	Natural Gas	Houston
3464	GBY83	Natural Gas	Houston	55187	CHV2	Natural Gas	Houston
3464	GBY84	Natural Gas	Houston	55187	CHV3	Natural Gas	Houston
3468	SRB1	Natural Gas	Houston	55187	CHV4	Natural Gas	Houston
3468	SRB2	Natural Gas	Houston	3611	1	Natural Gas	South
3468	SRB3	Natural Gas	Houston	3611	2	Natural Gas	South
3468	SRB4	Natural Gas	Houston	3612	1	Natural Gas	South
55299	CTG1	Natural Gas	Houston	3612	2	Natural Gas	South
55299	CTG2	Natural Gas	Houston	3612	3	Natural Gas	South

55327	CTG-1	Natural Gas	Houston	3612	CT01	Natural Gas	South
55327	CTG-2	Natural Gas	Houston	3612	CT02	Natural Gas	South
55327	CTG-3	Natural Gas	Houston	3613	3	Natural Gas	South
55357	CTG1	Natural Gas	Houston	3631	CT7	Natural Gas	South
55357	CTG2	Natural Gas	Houston	3631	CT8	Natural Gas	South
55365	GT-1	Natural Gas	Houston	3631	CT9	Natural Gas	South
55365	GT-2	Natural Gas	Houston	4937	1	Natural Gas	South
55365	GT-3	Natural Gas	Houston	4939	1	Natural Gas	South
55365	GT-4	Natural Gas	Houston	4939	3	Natural Gas	South
55464	CTG1	Natural Gas	Houston	4939	4	Natural Gas	South
55464	CTG2	Natural Gas	Houston	6178	1	Coal	South
55464	CTG3	Natural Gas	Houston	6179	1	Coal	South
55464	CTG4	Natural Gas	Houston	6179	2	Coal	South
55470	EPN801	Natural Gas	Houston	6179	3	Coal	South
55470	EPN802	Natural Gas	Houston	6181	1	Coal	South
55470	EPN803	Natural Gas	Houston	6181	2	Coal	South
56806	CBY41	Natural Gas	Houston	6183	SM-1	Coal	South
56806	CBY42	Natural Gas	Houston	6648	4	Coal	South
3439	4	Natural Gas	South	7097	**1	Coal	South
3439	5	Natural Gas	South	7097	**2	Coal	South
3441	8	Natural Gas	South	7762	HRSG1	Natural Gas	South
3441	9	Natural Gas	South	7762	HRSG2	Natural Gas	South
3443	9	Natural Gas	South	7900	SH1	Natural Gas	South
3548	1	Natural Gas	South	7900	SH2	Natural Gas	South
3548	2	Natural Gas	South	7900	SH3	Natural Gas	South
3548	GT-1A	Natural Gas	South	7900	SH4	Natural Gas	South
3548	GT-1B	Natural Gas	South	7900	SH5	Natural Gas	South
3548	GT-2A	Natural Gas	South	52071	5A	Coal	South
3548	GT-2B	Natural Gas	South	52071	5B	Coal	South
3548	GT-3A	Natural Gas	South	55086	101	Natural Gas	South
3548	GT-3B	Natural Gas	South	55086	102	Natural Gas	South
3548	GT-4A	Natural Gas	South	55098	1	Natural Gas	South

3548	GT-4B	Natural Gas	South	55098	2	Natural Gas	South
3559	10	Natural Gas	South	55123	CTG-1	Natural Gas	South
3559	9	Natural Gas	South	55123	CTG-2	Natural Gas	South
3601	1	Natural Gas	South	55137	CTG-1	Natural Gas	South
3601	2	Natural Gas	South	55137	CTG-2	Natural Gas	South
3601	3	Natural Gas	South	55137	CTG-3	Natural Gas	South
3609	3	Natural Gas	South	55154	1	Natural Gas	South
3609	4	Natural Gas	South	55154	2	Natural Gas	South
3609	CGT1	Natural Gas	South	55168	CTG-1A	Natural Gas	South
3609	CGT2	Natural Gas	South	55168	CTG-1B	Natural Gas	South
3609	CGT3	Natural Gas	South	55206	CU1	Natural Gas	South
3609	CGT4	Natural Gas	South	55206	CU2	Natural Gas	South
55144	STK1	Natural Gas	South	56350	CT1A	Natural Gas	South
55144	STK2	Natural Gas	South	56350	CT1B	Natural Gas	South
55144	STK3	Natural Gas	South	56350	CT2A	Natural Gas	South
55144	STK4	Natural Gas	South	56350	CT2B	Natural Gas	South
55153	CTG-1	Natural Gas	South	56674	1	Natural Gas	South
55153	CTG-2	Natural Gas	South	56674	2	Natural Gas	South
55153	CTG-3	Natural Gas	South	56674	3	Natural Gas	South
55153	CTG-4	Natural Gas	South	56674	4	Natural Gas	South

Table C.2 Net Load and Gross Load of Coal Plants in ERCOT

Plant ID	Fuel	Gross Load	Source	Net Load	Source	Ratio(N/G)
127	Coal	3.9	EPA CEMS	3.6	EIA	0.92
298	Coal	14.0	EPA CEMS	13.0	EIA	0.93
3470	Coal	18.0	EPA CEMS	17.0	EIA	0.94
3497	Coal	9.9	EPA CEMS	9.3	EIA	0.94
6136	Coal	3.8	EPA CEMS	3.6	EIA	0.95
6139	Coal	11.0	EPA CEMS	11.0	EIA	1
6146	Coal	19.0	EPA CEMS	18.0	EIA	0.95
6147	Coal	14.0	EPA CEMS	13.0	EIA	0.93
6178	Coal	4.7	EPA CEMS	4.5	EIA	0.96
6179	Coal	12.0	EPA CEMS	11.0	EIA	0.92
6180	Coal	8.2	EPA CEMS	3.9	EIA	0.48
6181	Coal	5.8	EPA CEMS	5.6	EIA	0.97
6183	Coal	3.2	EPA CEMS	2.9	EIA	0.91
6193	Coal	7.0	EPA CEMS	6.5	EIA	0.93
6194	Coal	8.1	EPA CEMS	7.7	EIA	0.95
6648	Coal	3.3	EPA CEMS	3.0	EIA	0.91
7030	Coal	2.7	EPA CEMS	2.4	EIA	0.89
7097	Coal	7.0	EPA CEMS	6.6	EIA	0.94
7902	Coal	5.2	EPA CEMS	4.8	EIA	0.92
52071	Coal	3.7	EPA CEMS	3.3	EIA	0.89
					Average	0.91

1. The gross load and net load information is based on the year 2010 data.
2. The unit for load is million MWHs.
3. The Net Generation is from EIA ("http://www.eia.gov/electricity/data/browser//topic/0?freq=A")

Table C.3 Calculated Zonal Wind, Demand and Errors

Variable	Mean	Std. Dev.	Min	Max
West				
Wind	1969	1245	13	5070
Demand	2288	364	1591	3591
Error	-364	575	-2851	2000
North				
Wind	60	37	0	180
Demand	13907	3691	7430	25871
Error	832	704	-3103	4214
Houston				
Wind	0	0	0	0
Demand	10060	2462	5914	17630
Error	-1055	754	-3633	1989
South				
Wind	443	331	2	1383
Demand	10311	2687	5642	17929
Error	942	700	-1653	3379

Table C.4 Some Attributes about Co-Gen and Non-Cogen Combined Cycle Gas Unit

	Unit	Non-cogen Units	Co-gen Units
Sample	#	12	31
"EIA" heat rate	mmBTU/MWH	9.16(1.36)	6.58(2.22)
"EPA-EIA" Heat Rate	mmBTU/MWH	8.80(1.02)	8.90(2.28)
Gross Generation	million MWH	9.3	43
Percentage of Fossil Fuel Generation	%	12.7	58.9

Table C.5 Percentage CO2 Emission Reduction (partial) Matrix in ERCOT w/ EPA-EIA Heat Rates

	Wind Capacity				
CO2 Price	0%	100%	chg. in 1st 100%	200%	chg. in 2nd 100%
$0	5.7	0	-5.7	-4.3	-4.3
$20	-7.1	-14.6	-7.5	-19.7	-5.1
$25	-11.5	-19.3	-7.8	-24.5	-5.2
$30	-15	-23	-8	-28.2	-5.2
$35	-18.4	-26.4	-8	-31.6	-5.2
$40	-21.3	-29.3	-8	-34.4	-5.1
$45	-23.4	-31.2	-7.8	-36.2	-5
$50	-24.8	-32.5	-7.7	-37.4	-4.9
$55	-25.9	-33.5	-7.6	-38.3	-4.8
$60	-26.8	-34.2	-7.4	-38.9	-4.7
$65	-27.5	-34.7	-7.2	-39.4	-4.7
$70	-28.1	-35.2	-7.1	-39.8	-4.6

Figure C.1 Spatial Allocation and Annual Generation of CEMS Units in ERCOT

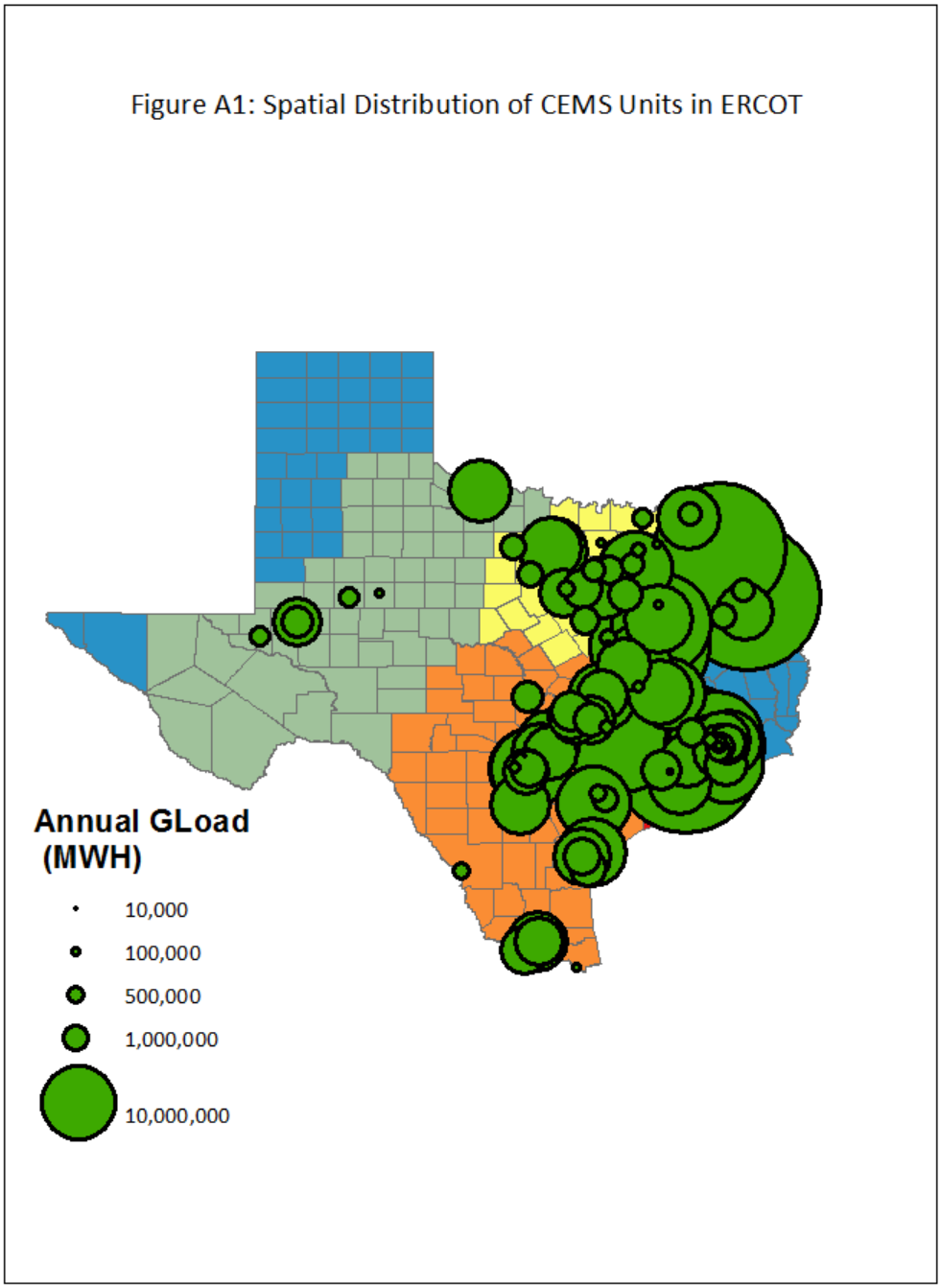


Figure C.2 Monthly Generation from Coal Units in ERCOT

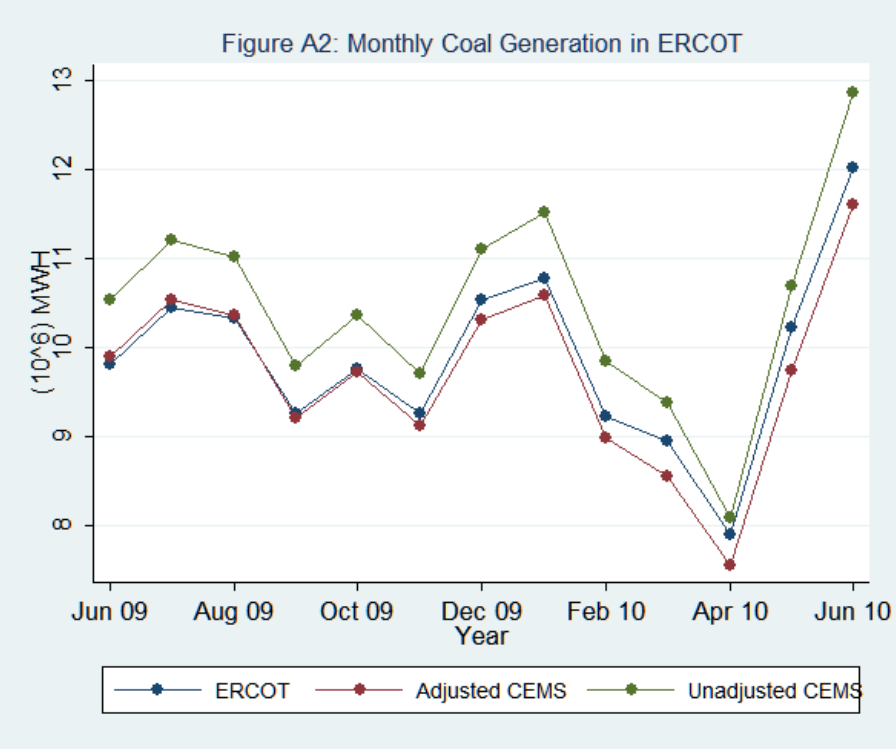


Figure C.3 Monthly Generation from Gas Units in ERCOT