ABSTRACT

DOMANSKI, ADAM MATTHEW. Essays in Nonmarket Valuation with Applications to Environmental Economics. (Under the direction of Roger von Haefen and Raymond Palmquist).

Non-market valuation techniques are vital instruments of policy. Providing quantitative welfare estimates validates public policy decisions. This dissertation independently addresses three aspects of nonmarket valuation: hedonic valuation of industrial accidents, limitations in recreational demand discrete choice modeling, and opportunities to expand urban planning research techniques.

A hedonic estimation of the 2006 Apex, NC chemical fire finds that this unexpected event impacted public welfare, but only mildly and for a limited time period after the accident. Important factors seemed to be the lack of long-term contamination and the zero chance of future contamination.

A mixed logit latent class estimation of recreational fishing trips in Wisconsin demonstrates the welfare implications of the random sampling of alternatives. This technique strategically reduces estimation time, but is difficult to implement in models where the independence of irrelevant alternatives (IIA) hypothesis is not present. By estimating a multi class model using the EM algorithm, IIA is reintroduced in the model. This research evaluates the impact of various sample sizes on welfare estimates.

The final research chapter applies the mixed logit model in the context of housing demand and urban planning. An exploratory estimation of a multi-class model using past research as a guide for choice set construction shows promise for this method's applicability to practitioners.

Essays in Nonmarket Valuation with Applications to Environmental Economics

by
Adam Domanski

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Economics

Raleigh, North Carolina

2010

APPROVED BY:

Dr. Laura O. Taylor	Dr. Xiaoyong Zheng
Dr. Roger H. von Haefen Co-Chair of Advisory Committee	Dr. Raymond B. Palmquist Co-Chair of Advisory Committee

DEDICATION

To my entire family.

BIOGRAPHY

Born in Fairfax, Virginia on June 29th, 1983 to Piotr and Maria Domanski, Adam grew up in Potomac, Maryland along with his older brother Mark. Having graduated from Georgetown Preparatory School in 2001, Adam went to the College of William and Mary, graduating in 2005 with a Bachelor of Arts in Economics and a minor in Chemistry. He entered North Carolina State University the following fall to pursue a PhD in Economics. During his tenure there he served as head teaching assistant for an undergraduate introductory course in economics, taught undergraduate environmental economics, and worked on various research projects ranging from recreational water quality to beach closure.

Always striving to live a balanced life, Adam fit as much travel into his graduate career as possible while also completing several triathlons, marathons, and an Ironman.

ACKNOWLEDGEMENTS

I would like to first and foremost thank my parents for the guidance, support, and love they have shown me. I would also like to thank my brother for his always candid perspective on life. I attribute much of my personality and outlook to my participation in Boy Scouts and my attendance of a Jesuit secondary school. I have also benefitted from the strong friendships I have made along all my different roads in life.

Professionally, I would like to thank Roger von Haefen for his high standards and continued support throughout my graduate education, as well as the guidance and help that I received from other faculty at NC State, specifically Ray Palmquist, Laura Taylor, Xiaoyong Zheng, David Flath, Dan Phaneuf, David Hyman, Tamah Morant, Wally Thurman, Mitch Renkow, and all the rest. I would also like to thank Chris Timmins, Marty Smith, and Rob Hicks.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. Introduction

Non-market valuation techniques are important tools for environmental economic analysis. Whether studying consumer choice for public goods, implications of public policy, or the welfare effects of environmental disasters, non-market valuation provides a quantitative estimate which can help society more easily understand and address the most difficult of environmental problems. This dissertation seeks to expand the current state of knowledge in non-market valuation by addressing several methodological limitations, providing solutions, and demonstrating the importance of environmental economic analysis. Three related topics are organized into unique chapters.

Chapter two demonstrates the value of non-market valuation to measuring the social impact of an environmental catastrophe. In October of 2006, a chemical storage facility ignited and burned uncontrollably in the town of Apex, North Carolina. The fire led to a town-wide evacuation, potential short and long term environmental contamination, new legislation issued by the North Carolina Assembly, and a number of lawsuits and compensation claims against the operators of the facility. The event raises a number of policy, academic, and practical questions which can be answered using economic techniques. Aside from the direct impact, did the fire create any long term effects in people's willingness to reside or work in Apex? Do individuals trust information provided by environmental management agencies such as the EPA? How are the costs of the fire distributed across Apex residents, the company managing the warehouse, other hazardous waste facilities, and residents across the State of North Carolina? Using hedonic regression techniques and an analysis of the events surrounding the fire, chapter two seeks to answer these questions.

By running a difference-in-differences estimation on twenty different spatial and temporal specifications, a set of parameter results are recovered which provide an indication of the impact of the fire. Since the number of impacted houses is small (due to the relative low density of the study area), statistically significant results are difficult to obtain. Combined with the short contamination period, quick response by state and federal agencies and little indication of long term offsite chemical contamination, there does not appear to be any long term impact in the housing market. Any direct effect on housing prices was limited to a two mile radius and six month time period after the fire, however this dissipated within a year.

Chapter three addresses a specific limitation in discrete choice recreational demand analysis. This popular technique is useful when modeling a consumer's choice amongst many alternatives. The quality of the results is dependent upon, among many things, the quality of the data set, the capacity of the model to address unobserved differences in taste, and the ability to accurately replicate the consumer's actual choice scenario. In attempting to achieve accurate estimates, researchers often find themselves faced with hundreds or thousands of potential alternatives. These large choice set issues are prevalent in the literature and although random sampling methods can be easily applied in a conditional logit model, doing so in a mixture model is more burdensome. Chapter three evaluates large choice set problems in the literature and presents a theoretically viable method for sampling from a large choice set in a mixture model. By using the latent class method solved by the recursive expectation-maximization (EM) algorithm and taking advantage of the independence of irrelevant alternatives assumption, random sampling can produce theoretically and empirically consistent results in a mixture model. Using a recreational fishing dataset from Wisconsin, chapter three demonstrates and tests this method. Latent class results are compared to conditional logit results for up to six classes and the effects of random sampling are tested on 50%, 25%, 12.5%, 5%, 2%, and 1% random sample sizes. For robustness, the results of 200 random samples are tested for the conditional logit model, and 10 random samples are tested for the latent class models. Due to the possibility of a non-globally concave likelihood function, 10 starting values are used with each of the random samples in the latent class model. Five different welfare scenarios are compared, and the confidence intervals across sample sizes give indication to the consistency of the estimates.

The full set of results gives practical guidance to the tradeoff between time-saved versus efficiency-lost when estimating models with progressively smaller choice sets. Depending on the needs of the researcher, random sampling in a conditional logit model can produce consistent estimates down to a 1% sample size with a 99% reduction in estimation time. Random sampling in a mixed logit model (using the methodology described) can produce reasonably consistent estimates down to a 5% sample size with up to an 84% reduction in estimation time. Although this is performed in an environmental context, its applicability is much broader.

Chapter four takes guidance from chapter three and applies the latent class model to a housing demand context similar to chapter two. This section uses a latent class housing model to recover heterogeneous preferences for a smart growth development policy. Smart growth is an urban planning model that has grown in popularity over the last twenty years and is primarily centered on focusing new development in compact urban centers, oftentimes through infill development. Such a change in the urban and suburban landscape could be favored by a segment of the population that prefers urban housing options, but disliked by others that enjoy traditional suburban living. A latent class model that flexibly accounts for unobserved preference heterogeneity could be useful for measuring the welfare effect of a smart growth policy. Discrete choice housing techniques are difficult to use and oftentimes run into large choice set problems, which may be why there has been limited implementation of mixture models in the housing literature.

By using a latent class model estimated via the EM algorithm, three classes of latent preferences are recovered. Welfare estimates suggest that there may indeed be polarized preferences for smart growth. Comparisons show that the conditional logit model understates the willingness to pay for one subset of the population while overstating it for the rest. Although no sensitivity analysis is performed and limited data is used, these results indicate the viability of both the econometric technique and future research in the field.

These next three chapters are presented as three independent research papers tied together by the nonmarket valuation techniques they use and the environmental issues they address. Each is self contained in terms of technique, data, estimation, tables, figures, and references. Following these three chapters is a conclusion which reiterates the ties and applicability of this research.

2. Property Value Effects of the 2006 Apex Chemical Fire

2.1. Introduction

On October 5, 2006 at approximately 10:30 PM, a chemical spill and several explosions occurred at the EQ Industrial Services storage facility in the Raleigh, North Carolina suburb of Apex. The resulting chemical fire burned for three days and caused the evacuation of more than 17,000 local residents – approximately half of the town's population. The chemicals stored at the facility included commercial hazardous waste materials such as oil-based paint and fuels, industrial acids, spent solvents, lead, mercury, arsenic, DDT, and vinyl chloride. When ignited, the fire burned uncontrollably and emitted a toxic cloud over the town of Apex, sending some fire fighters and elderly residents to hospitals from exposure to the fumes (Beckwith and Ovaska 2006). Once the fire was extinguished, evacuated residents were allowed to return home. Approximately one month later, the EPA certified that there was no residual chemical contamination above federal allowable standards. Following the fire, regulations were quickly passed by the North Carolina State Legislature which sought to prevent an incident such as the Apex fire from occurring again. These regulations primarily concerned commercial hazardous waste handlers. This paper analyzes the localized impact of the fire on housing prices.

This research fills a gap in the hazardous waste literature. Much previous research (including McClelland et al, 1990; Michaels and Smith, 1990; Ketkar, 1992; Smolen et al.,1992) use cross sectional analysis with site closure or cleanup as an instrument to test for a disamenity from hazardous waste. However, these instruments are not truly random shocks and a concern is that few are able to fully control for endogeneity in the closure or cleanup decision. The Apex fire is a natural experiment that allows us to examine how individuals react to unexpected contamination.

The Apex fire was a unique event for many reasons. As the evidence suggests, the fire that led to the evacuation was a completely unanticipated and isolated incident. Any local property value impact can be attributed to the fire. Additionally, the incident is a realization of the fears of living next to a hazardous chemical treatment and storage facility.

The direct cost of evacuating a community and combating a blaze of such severity are now well known. This paper contributes a discussion of how a catastrophic event enters into the hedonic property value function.

This paper uses observed home sales in Wake County, North Carolina during a two and half year period before and after the fire. Using a hedonic difference in differences approach, the implicit price of proximity to the contamination generated by the fire is recovered. For robustness, twenty different specifications varying in terms of the spatial and temporal effects of the fire are run. A test of a two mile impact radius over a six month period after the fire shows a statistically significant average marginal implicit price reduction of $6,483, or equivalently, a 2.8% price reduction during this time horizon and in this spatial area.

The paper will proceed as follows. Section one describes Apex, EQ Industrial Services, and the incident at hand. Section two provides a discussion of the costs of the Apex fire. Section three details a hedonic regression to recover the public welfare effect of the fire, and section four presents empirical results followed by a conclusion.

2.2. Background

Apex, NC is a city located within the Raleigh-Cary, NC metropolitan statistical area. The town was ranked #14 in CNN-Money Magazine's 2007 *Top 100 Places to Live*. The city comprises 10.6 square miles approximately 12 miles southwest of Raleigh, North Carolina in central Wake County. In 2006, the city had a population of 26,300, a median family income of $91,030, and a median home price of $226,942. For comparison, North Carolina and the US had a median family income of $52,336 and $57,612 respectively and median home prices of $136,149 and $232,200 respectively. With an air quality index (AQI) of 62.7 (AQI=100 is the national standard; smaller is better), two personal crime incidents per 1000 (Raleigh has six; Washington D.C. has 14), and a median commute time of 25.1 minutes, residents enjoy a high quality of life based on many metrics. Apex's median age is 32, 68.2% of residents are married, and 84% of residents completed at least some college (American Community Survey Estimates, 2006).

2.2.1. EQ Industrial Services

EQ is a Michigan-based company which specializes in providing environmental management services (EQ, 2006). As of October, 2006, the company had 11 facilities across five states providing services ranging from wastewater treatment and chemical solvent recycling to hazardous and PCB waste disposal. One of these facilities was located in Apex, NC (EQNC) which served primarily as a hazardous waste transfer facility where various hazardous chemicals were collected and stored before being shipped off to other treatment facilities.

The EPA estimates that in the U.S. over 40 million tons of hazardous waste are produced each year by dry cleaners, auto repair shops, hospitals, exterminators, photo processing centers, chemical manufacturers, electroplating companies, petroleum refineries, etc. (EPA RCRA Hazardous Waste Report, 2005). Defined as a substance that is “ignitable, corrosive, reactive, or contains certain amounts of toxic chemicals,” hazardous wastes pose substantial risks when transported, treated, or stored.

In North Carolina in 2005, 431 Large Quantity Generators (those generating 1000kg or more of wastes in a calendar month) reported generating 384,112 tons of hazardous waste. This amount does not include hazardous waste generated by 1,823 Small Quantity Generators and 4,498 Conditionally Exempt Small Quantity Generators which are not required to report quantities produced. Hazardous waste is treated at Treatment, Storage and Disposal Facilities and in 2005 there were 50 facilities managing 373,208 tons of hazardous waste in North Carolina (North Carolina Hazardous Waste Services, 2006).

Although some chemicals were treated at EQNC, its primary role was as a Waste Shipper. Waste not treated in North Carolina is shipped to facilities in other states. In 2005 there were 427 hazardous Waste Shippers in the state transporting 106,536 tons. Of these, EQNC was the third largest, shipping 6,346 tons of waste. (EPA RCRA Hazardous Waste Report, 2005)

In the United States, the Resource Conservation and Recovery Act (RCRA) regulates current management of hazardous wastes. The law emerged in 1976 as amendments to the

1965 Solid Waste Disposal Act and gave the EPA greater flexibility in dealing with hazardous wastes. The hazardous waste program, under RCRA Subtitle C, establishes a system for controlling hazardous waste from the time it is generated until its ultimate disposal. Although RCRA sets guidelines for managing waste, the EPA delegates the primary responsibility of implementing the RCRA hazardous waste program to individual states through the State Authorization Rulemaking Process.

North Carolina manages hazardous wastes through the Division of Waste Management (DWM) under the Department of Environmental and Natural Resources (DENR). The DWM has strict guidelines for safe handling of waste and has procedures for monitoring and enforcement. In the years prior to 2007 (when new statutes were introduced), hazardous waste facilities in the state were monitored and inspected on a weekly basis and fined for handling violations. In 2005, 33 North Carolina facilities were assessed civil penalties with fines totaling $648,087. (Hazardous Waste Section; Civil Penalty Assessments, 2005). From 2004 to 2006, EQNC was found in violation of North Carolina code eight times for infractions ranging from minor safety to dangerous fire hazards. (See Table 2.6). However, at the time, none of these records were officially published and according to news reports and firsthand accounts, residents surrounding the facility were unaware of the risks of an accident. (Coleman, 2006c).

2.2.2. Apex Fire[2.1]

On the evening of October 5th, chemicals stored at the facility caught fire and burned for three days. Later investigations concluded that the blaze was caused by solid chlorine-based pool chemicals which interacted with chemical oxygen generators (of the type used on aircraft in emergencies) which had not been deactivated. As the facility lacked any automatic fire suppression system, the fire quickly spread to other areas where flammables,

[2.1] This narrative of the Apex Fire is a condensed composite history collected from many mainstream news articles published in the months following the incident. The primary information sources are Associated Press (2006a), Associated Press (2006b), Bartelt, G. (2006), Beckwith, R. T. and S. Ovaska (2006), Bowens, D. (2006), Coleman, T. (2006), Coleman, T. and W. Rawlins (2006), Janelle, C. (2006), and WRAL.com (2006).

corrosives, laboratory wastes, paints, and pesticides were stored. The fire and resulting chemical cloud led the city to order the evacuation of over 17,000 residents. Three hundred firefighters from four counties responded and the incident was reported heavily in local and national news.

Those evacuated sought shelter with friends and family members, in local hotels and at local emergency shelters. Within two days of the fire, chemical analyses of the surrounding area detected no toxins at levels sufficient to cause long-term health problems. Air monitoring by the EPA found that, except in the immediate vicinity of the facility, concentrations of volatile organic compounds were below any level of concern. A phased re-entry plan was instituted and within 44 hours of the fire, all residents were allowed to return home. (Wake County, 2006). The full timeline of incidents surrounding the fire is shown in Table 2.1.

[Table 2.1 - Apex Fire Timeline]

Following an investigation of the fire, state regulators fined EQ $553,225 in March of 2007 for five violations. By October 17th, 2007, EQ agreed to pay a lower amount of $441,944 with the condition that it give up its right to store hazardous waste in the state. In addition to the fines paid to the state, EQ also compensated the city of Apex $200,000 for costs incurred with evacuation and fire suppression and allowed residents to file reimbursement claims for evacuation costs (hotel stays, etc.). In October of 2008, EQ settled a class action civil suit with local residents and business. According to the settlement, EQ agreed to pay each household that evacuated and filed a claim up to $750 and businesses could receive a maximum of $2,000 if they could prove economic loss.

On June 20th, 2007, the North Carolina General Assembly passed House Bill 36 / S.L. 2007-107 which incorporated the recommendations of the Hazardous Materials Task Force set up in response to the Apex fire. The major modifications to existing North Carolina code included requiring commercial hazardous waste facilities to demonstrate financial responsibility for dealing with accidents and off-site contamination, seek input and work more closely with local authorities, provide 24-hour security around facilities, as well as several other minor regulations.

2.3. Direct and Implicit Costs of the Fire

This next segment of the paper will analyze all the private and public costs associated with the 2006 Apex chemical fire. Within the scope of this particular accident, each of the costs and benefits were borne by different segments of society. EQNC, the residents directly affected by the fire, and the Town of Apex were the groups primarily affected.

We begin by analyzing the costs borne by EQNC. Although the fire was large, direct property damage did not extend beyond the EQNC facility. No neighboring properties caught fire nor did explosions shatter any nearby windows. All direct property damage occurred only to the company. The costs of fire suppression were originally borne by the town of Apex and the fire departments of neighboring communities that participated, but were later billed to and paid by EQNC. Thus, EQNC bore the full costs of property damage. Evacuation costs were initially borne by residents but were later paid by EQNC. Initially, the company was reimbursing all hotel and transportation claims from evacuation, but eventually compensated each resident $750. All of the costs mentioned so far were borne by EQNC, have a short term impact, and occurred during or immediately after the fire. None of these costs reflect a health or contamination hazard.

Following the incident, there was much local concern and speculation about long-term environmental contamination. Although the evacuation prevented most residents from dangerous exposure to the cloud of fumes, the homes and surrounding neighborhoods still may have been in contact with the chemicals. If contaminants had settled onto city surfaces, playground equipment, private property, or had seeped into the ground water, residents that returned to live in the area could have been exposed to long term health risks.

Concerned residents could minimize the risk of exposure by either altering their daily behavior (spending less time outside in their neighborhood) or by moving to a new home. This risk of exposure introduced a disamenity of the home. The price of any homes that are in close proximity to the EQNC facility and sold after the fire reflect this new disamenity. To assess the economic impact, a difference in differences approach can be adopted. An experiment can be constructed in which the prices of homes sold before and after the incident and inside and outside a theoretical contamination area can be compared while controlling for structural and neighborhood characteristics. Making standard assumptions about a fixed supply of houses and market equilibrium in the housing market, any resulting property value impact would represent a localized public disamenity occurring because of the fire. This can occur because of a fear of continued chemical contamination from the fire, fear of another fire in the future (causing future contamination), or some undefined stigma from the site. These costs would be borne by households within the affected area and could be quantified using traditional hedonic analysis. It is important to note that the change in the property values does not reflect any costs related to the evacuation or health costs related to initial contamination.

2.4. Hedonic Theory

Hedonic analysis is based on the theory of economic rents; that the market price for a given asset equals the present discounted value of all future rents generated from the asset. When an environmental event alters the productivity of an asset, the change in productivity will be capitalized into its price. The expectation of a future change in environmental quality will also affect current prices.

The seminal hedonic paper of Rosen (1974) formalized how heterogeneous demands for characteristics of a good interact with firms' heterogeneous supply functions. The matching of suppliers and consumers results in equilibrium in the market and any price differentials can be explained by differences in the characteristics of a good. Thus the marginal willingness to pay or implicit prices of characteristics can be recovered.

Ridker and Henning (1967) were one of the first to use residential property value data to estimate the welfare effects of changes in an environmental attribute. They assumed a residential unit consists of a 'bundle' of attributes (housing characteristics as well as a measure of air quality), each of which has an implicit market price. The market value of the unit is a combination of the values of these attributes. Using observed home prices and attribute data they were able to estimate the implicit price of air pollution. While this paper looked at the effect of sulfur compounds in the air, much subsequent research (McClelland et al. 1990, Michaels and Smith 1990, Ketkar 1992, Smolen et al. 1992, etc.) has used proximity to hazardous waste sites and changes to the public's information set as explanatory variables in the hedonic price function.

There has been interest in the hedonic literature on the impact of Locally Unwanted Land Uses (LULUs). LULUs can be loosely categorized as any land use which creates a negative local externality. This impact is most prominently measured through price effects on surrounding property. The LULU class can encompass landfills, power plants, incinerators, factories, brownfields, hazardous waste facilities, etc. The hazardous waste externality is, however, quite different than many other LULUs. Coal-fired power plants, paper mills, and landfills are often associated with a direct, quantifiable public bad (noticeable via smell, sight, or health effect), but a hazardous waste facility's externality might only occur from the public's perceived risk of a future contamination. Quantifying this externality is a difficult task, primarily since some of the effect can be caused by misperceptions, incomplete information, and risk averse responses to uncertainty.

[Table 2.2 - Classes of Research]

The past LULU studies can be categorized into two classes based upon the probability (p) of public contamination: 'emitting LULUs' (p = 1) and 'risky LULUs' ($0<p<1$). Associated papers are categorized in Table 2.2. "Emitting LULU" papers have analyzed localized externalities which pose an ongoing harm to the environment around it – e.g. toxic sites or landfills which produce an odor or some other visible disamenity. Residents in close proximity are directly contaminated by the emitting site and are fully aware of the contamination. Blomquist (1974) evaluated coal-burning power plants and found that a 10% distance increase results in a 0.9% value increase; Kohlhase (1991) analyzed a hazardous waste site which was contaminating surface and ground water and discovered that a one mile increase is worth $2,364; Nelson et al. (1992) and Reichert et al. (1992) looked at landfills and found that a one-mile increase is worth approximately $5,000 and that opening a landfill decreased surrounding property values $2,924 respectively; and Kiel and McClain (1995) found that increasing distance to an incinerator is worth from $2,283 to $8,100 per mile. Each of these papers falls into the emitting LULU category because they evaluate a public externality associated with an ongoing and readily visible contamination.

Risky LULU papers have attempted to quantify the perceived risk of a disamenity and analyze the consequences of the non-emitting site becoming an emitting one – e.g. via an explosion or leak. Residents in close proximity to these sites have some probability of contamination less than one but greater than zero. McClelland et al. (1990), Michaels and Smith (1990), Ketkar (1992), and Smolen et al. (1992) evaluated non-emitting hazardous waste sites and found that increasing distance or cleaning up a site led to welfare benefits ranging from $115 to $14,200. Non-emitting sites still generate external costs, but individual welfare is based more upon a perceived risk of future contamination.

Within the relevant literature, it is common practice to study the impact of elimination of a LULU. The change in property values resulting from a site closure is used to quantify the negative welfare effect of that LULU. These results are then used to predict the effects of closing or cleaning up other similar sites. The site closure is assumed to be an exogenous shock, but in many instances this serves as a weak assumption. Even though site closure reflects the end of any physical contamination, the event is sometimes the result of community or political pressure and may have been anticipated by local property owners. Studying the dates that the future closure of a site is announced has similar issues. The local community may have expected future closure/cleanup, may have had faith in government protection, or may have had faith in the legal system to deliver a lucrative victim lawsuit should health effects occur. These forces may already be reflected in surrounding property values. Some papers have addressed this identification issue directly. Greenstone and Gallagher (2008) look at Superfund sites that were closely ranked on the national priorities list, but due to budgetary constraints, only some were cleaned up. Primarily directed at assessing the cost-effectiveness of the EPA's 'Superfund' program, the paper used representative homes in a census block and median housing values. This type of quasi-experimental technique has not yet been used to address the impact of hazardous waste contamination on a micro-data level.

A subset of the risky LULU class of the literature analyzes the timing of accidents at risky sites, causing them to become emitting sites. Specifically, do individuals have accurate expectations and full information about potential risks and their associated potential negative impacts? A study of the Apex fire fits into this category, as does research analyzing the Three Mile Island nuclear accident (Nelson 1981, Gamble and Downing 1982), the Fairfax, VA petroleum pipeline leak (Simons 1999), and the PEPCON explosion (Carroll et al; 1996).

The latter incident is the most similar in scope to the Apex fire. In 1988, a welder's torch caused an explosion at a PEPCON chemical plant in the Las Vegas valley. This facility manufactured ammonium perchlorate (a volatile component of solid rocket fuel) and was located adjacent to other industrial and residential areas. The explosion caused extensive but repairable damage to nearby property. The PEPCON facility was not rebuilt; however there remained another operational ammonium perchlorate factory two-thirds of a mile away. Analysis showed that the explosion caused a reduction in nearby property values by 17.6%, but after the explosion, property values rebounded by 38%. They concluded that new home buyers were unaware that some hazard still existed.

Unfortunately, nuclear power plants, petroleum pipelines (and to a lesser degree ammonium perchlorate factories) are all visible sources of potential environmental contamination. An individual who purchases a home next to a nuclear power plant is well aware of the facility's proximity and most likely has some notion of the risk of contamination (due to awareness of historical incidents at Three Mile Island and Chernobyl). If an accident occurred in the future, the price shock would be smaller than if an accident was completely unanticipated. Nearby property owners may have had realistic expectations of a future disaster which would already be reflected in housing prices. Thus, when studying hazardous waste, it is similarly necessary to conduct a study which utilizes a truly random shock. Due to the unanticipated nature of the incident, the Apex Fire is a suitable natural experiment.

2.5. Empirical Analysis

The empirical analysis of the Apex fire will measure the effect of the incident on local property values. Using hedonic theory, the price of a house can be decomposed into its individual components.

$$\ln P_i = \alpha X_i + \phi t_i + \varphi S_i + \delta(S_i \cdot T_i) + \varepsilon_i$$

Where

- P_i ~ Observed sale price
- X_i ~ Structural and Neighborhood characteristics
- t_i ~ Time fixed effects (3 month periods)
- S_i ~ Spatial treatment dummy
- T_i ~ Time treatment dummy
- ε_i ~ i.i.d. error term.

A difference-in-differences (DID) approach is used. Ashenfelter and Card (1985) introduced an early application of the method and its use has since become widespread. In a basic DID estimation, outcomes are observed for two groups (a 'control' and a 'treatment') for two time periods. During the first time period, neither group is exposed to a treatment while in the second time period, only one of the groups is exposed. The inclusion of affected and unaffected groups in this way removes biases that could be the result of permanent differences between the groups as well as from comparisons over time (Imbens and Wooldridge, 2009). In our case, repeat sales of the same home in both time periods is not observed, however, controlling for housing characteristics allows different homes to be compared across time.

For our analysis, δ, the coefficient on the spatial and time treatment interaction variable is the key parameter. This is a dummy for homes that are in the treatment area during the time period after the accident. The spatial variable controls for all homes that are in the treatment area during the entire time horizon. The time fixed effects control for time trends and cover three month time periods over the entire study period. Our analysis will test whether the parameter δ is statistically different than zero, thus signifying a potential property value impact from the fire.

Data for local property sales was obtained from the Wake County Assessor's Office and represents all home sales during a five-year period between May 20th, 2004 and May 20th, 2009. Observations were restricted to single-family homes zoned for residential use and owned by individuals. The study area encompasses all of Wake County and includes 84,415 observations. Summary statistics are presented in Table 2.3.

[Table 2.3 - Summary Statistics]

Structural characteristics were chosen which were uncorrelated and expected to have a direct impact on housing value, or proxy for some other characteristic which would impact housing value. Variables were either continuous/count (basement %, assessed quality, age at time of sale, heated area, and lot size) or dummy variables (city water, city sewer, natural gas, fireplaces, stories, style, and baths).

Most of the neighborhood characteristics were calculated from the Wake County Assessor's data. Housing density was calculated using the number of other homes within a one mile radius of each home across all years. Distance to the center of Research Triangle Park and distance to Raleigh's Central Business District are both calculated in linear miles. Mean commute time is collected from the 2000 United Stated Decennial Census and represents the mean commute time for the Census Block Group within which the home is located.

Twenty different specifications were tested, each with different treatment effect-time frame combinations. Ex-ante, the extent of the spatial and temporal impact of the fire was not obvious. Thus it is important to test a broad range of specifications to recover any information from the data. There are four basic specifications of the treatment effect, each used with four different time frames. The spatial categories are: 1) evacuation zone, 2) one-mile radius, 3) two-mile radius, and 4) linear distance in miles. The distance measure is truncated at five miles in the estimation. Treatment time frame specifications include 1) two and a half years, 2) one year, 3) six months, and 4) three months following the fire. Three month time fixed effects were also used beginning on January 1st of every year. The spatial treatment areas are displayed in Figure 2.1.

[Figure 2.1 - Treatment Areas Map]

Looking at the number of homes sold during each treatment time frame relative to the rest of Wake County can give one indication of the fire's impact on the housing market. This information is displayed in Figure 2.2. Each of the graphs display the percent average number of homes sold per month in the treatment area relative to all of Wake County for a set number of time frames. The time frame designations are two and half years before the fire, from the fire until three months after, three to six months after, six months to one year after, and one year to two and a half years after the fire. These large time frames are used instead of a strict monthly average for two reasons. This grouping smoothes out month to month variation and also provides better comparison to each of the treatment time periods used later in this study. The percentage of homes sold in the treatment area is compared to all of Wake County to account for any seasonal effects on the housing market.

[Figure 2.2 - Percent of Homes Sold in Treatment Area by Treatment Time Frame]

As can be seen in Figure 2.2, there is a relative decline in the percentage of affected homes sold immediately after the fire. There then appears to be a rebound effect, which later falls to a long run average. A possible explanation is that some buyers delayed their purchase after the fire until the long-term effects were better known, and then entered the market after the next time period. Note that none of these graphs give any indication to a shift in prices or change in the absolute number of homes sold. Standard errors were calculated by bootstrapping. The absolute number of home sales were resampled 50 times and the results show that none of the treatment time frames had a statistically significant number of home sales compared to the entire county.

[Table 2.4 - First Stage Hedonic Results of Property Characteristics]

The hedonic regression results are shown in Table 2.4. These parameter estimates reflect all of Wake County without any treatment interactions. Note that the three month time fixed effects parameter estimates are not shown, but do reflect the trend in the housing market during this five year study period. The coefficients on the structural and neighborhood characteristics are consistent with expectations in terms of sign and magnitude. Variables that we would expect to have a positive impact on property value have a positive sign (Basement %, City Water, Natural Gas, Assessed Quality, Fireplaces, # of Baths, Heated Area, and Lot Size) and those that we would expect to have a negative impact have a negative sign (Townhouse, Ranch, Split Level, Age at Time of Sale, and Mean Commute).

To study the effect of each of the treatments, all of the variables in Table 2.4 are used in a hedonic regression along with treatment variables for each individual specification. The parameter estimates on the spatial and temporal interactions from each run are displayed in Table 2.5. Note that the coefficients on the structural, neighborhood, and time variables are not reported, however they do not change much among specifications. Across all runs, parameters on structural characteristics, on average, vary less than 1%, with the exception of the coefficients on basement, sewer, and gas, which vary less than 10%. The parameters on neighborhood characteristics which do not overlap with the various treatment specifications (Raleigh, Garner, and Morrisville) do not change more than 5% on average. The parameters on Cary, Apex, Holly Springs, Fuquay Varina do vary across the spatial treatments, but do not vary more than 2% when compared within spatial treatments (and across different time treatments).

[Table 2.5 - Interaction of Spatial and Time-Dependant Variables]

Analysis of the spatial dummy interactions (the Evacuation Zone, 1 mile radius, and 2 mile radius specifications) shows consistency in sign. All spatial treatments in the 2.5 year post-fire time horizon have a positive sign, while the treatments all have negative signs in the 12 month, six month, and three month time horizons. In each spatial specification, the six and three month time horizons have coefficients that are within 10% of each other and more than twice as large in magnitude as the respective 12 month coefficients. This suggests that the region surrounding the site of the Apex fire experienced a drop in property values after the fire, which did not begin to rebound until six months later. The coefficients are not broadly significant, however this may be due to the small number of observations included in each treatment interaction.

[Figure 2.3 - Coefficients on Six Month Time Horizon]

A comparison of the mileage treatments shows a similar result. Homes situated farther from the site of the fire had a diminished property value (negative coefficient on the mileage variable) 1 and 2.5 years after the fire, but had a higher property value up to six months after. The quadratic specification is positive in the first degree, but negative in the second. A comparison of the various spatial specifications for the six month time horizon is shown in Figure 2.3. All the various six-month spatial specifications point to the same general result: there was a decrease in property values surrounding the fire that diminished with distance. Of these results, the only statistically significant (at the 95% level) result is the 2 mile radius. Since a semi-log hedonic specification is used, the implicit price is represented by $\partial P / \partial F_i = \delta_i \cdot P$. For the 224 homes that sold within two miles and six months of the Apex fire, the average marginal implicit price is -\$6,483, with a standard deviation of \$3,790. This amounts to an average 2.8% price reduction.

The only other treatments that produced a statistically significant result (at the 95% level) were the 2 mile radius, miles, and quadratic miles specifications over the 2.5 year time horizon. The parameters represent a positive property value impact on homes closer to the fire site. This can be explained by the continued growth in Wake County, the increase in popularity of the Apex area, and the fact that there are no long term health implications from the fire. The mileage variables also measure the impact on all homes in Wake County and thus have limiting explanatory power on a small subset of observations.

2.6. Conclusion

As best can be concluded from the results, there is no long term property value impact resulting from the Apex fire. Although the accident was very large, all testing by the EPA showed no offsite contamination. As a result, the current cost of contamination is equal to zero and we would expect no property value impact (as is seen). Additionally, since EQ is not allowed to operate at the facility any longer, the future probability of contamination is zero and there is no expectation of future contamination.

There was, however, a short term property value impact following the fire. This lasted approximately six months and diminished over the course of the year following the fire. By 2.5 years after the fire, property values had fully recovered and had actually grown. This short term negative impact may reflect a combination of uncertainty concerning the potential health implications for residents, the possibility of the site being rebuilt and activities resumed, and possibly a partial stigma towards the site. All of these effects seem to have diminished with time. It is important to note that in the long run, the financial loss accrued only to those that chose to relocate immediately following the fire while at the same time, residents who moved in received a benefit from the long term rebound in property values. This impact was a localized externality, specifically affecting only homes within two miles of the fire. The spread of the chemical contamination was well known and reported. It is clear that homes greater than two miles away were not affected by the contaminants from the fire.

Furthermore, it appears that there was a short term reduction in the number of homes sold in the impacted area immediately after the fire, which later rebounded to long term levels. This indicates that either, 1) homes might have sold slower and spent more time on the market during the period immediately following the fire, or 2) homeowners that were planning on selling may have waited to enter the market until uncertainty following the fire settled. Information on how many homes were on the market is not currently available so these conclusions cannot be verified empirically.

2.7. References

American Community Survey. (2006). United States Department of Commerce. Bureau of the Census.

Ashenfelter, O. and D. Card. (1985). “Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs.” Review of Economics and Statistics, 67(4), pp. 648-660.

Associated Press. (2006a). “N.C. Bars Hazardous Waste at Michigan Company’s Site.” WLNS. 10/26/2005. <http://www.wlns.com/Global/story.asp?S=5590661&nav=0RbQ>.

Associated Press. (2006b). “EQ to Begin Removing Hazardous Waste Out of Apex.” FOXCarolina. 11/09/2006. <http://www.fox21.com/Global/story.asp?S=5657373&nav=2KPp>.

Bartelt, G. (2006). “Phase I of Chemical Cleanup Almost Complete.” News 14 Carolina. 10/25/2006. <http://rdu.news14.com/content/headlines/?ArID=93179&SecID=2>.

Beckwith, R. T. and S. Ovaska. (2006). "Apex’s Fierce Fire Burns On." The News and Observer. 10/07/2006. <http://www.newsobserver.com/1360/story/495767.html>.

Blomquist, G. (1974). “The Effect of Electric Utility Power Plant Location on Area Property Value.” Land Economics, 50(1) pp. 97-100.

Bowens, D. (2006). “Governor’s Task Force to Take Closer Look at Hazardous Storage Facilities.” WRAL. 10/24/06. <http://www.wral.com/news/10143782/detail.html>.

Boxalla, P., W. Chanb, M. McMillanc. (2005). “The Impact of Oil and Natural Gas Facilities On Rural Residential Property Values: A Spatial Hedonic Analysis.” Resource and Energy Economics, 27(3), pp. 248-269.

Brasington, D., and D. Hite. (2005). “Demand for Environmental Quality: A Spatial Hedonic Analysis.” Regional Science and Urban Economics, 35(1), pp. 57-82.

Carroll, T., T. Clauretie, J. Jensen, M. Waddoups. (1996). “The Economics Impact of a Transient Hazard on Property Values: The 1988 Pepcon Explosion in Henderson, Nevada.” Journal of Real Estate Finance and Economics, 13, pp. 143-167.

Chattopadhyay,S., J.B. Braden and A. Patunru. (2005). “Benefits of Hazardous Waste Cleanup: New Evidence from Survey and Market-Based Property Value Approaches.” Contemporary Policy Issues, 23, pp. 357–375.

Civil Penalty Assessments. (2005). Hazardous Waste Section. North Carolina Department of Environment and Natural Resources.

Coleman, T. (2006). “Few Answers for an Angry Apex.” The News and Observer. 10/18/2006. <http://www.newsobserver.com/1360/story/499875.html>.

Coleman, T. and W. Rawlins. (2006). "Errors Extend Beyond Apex Site." The News and Observer. 12/03/2006. <http://www.newsobserver.com/1360/story/517398.html>.

EPA RCRA Hazardous Waste Report. (2005). United States Department of Commerce. Environmental Protection Agency.

EQ North Carolina. (2006). 12/08/2006. The Environmental Quality Company. <http://www.eqonline.com/locations/EQ_North_Carolina.asp>.

Flower, P. and W. Ragas, (1994). “The Effects of Refineries on Neighborhood Property Values.” Journal of Real Estate Research, 9(3), pp. 319-338.

Gamble, H.B. and Downing, R.H., (1982). “Effects of Nuclear Power Plants on Residential Property Values.” Journal of Regional Science, 22(4), pp. 457-478.

Gayer, T, J.T. Hamilton and W.K. Viscusi. (2001) “Private Values of Risk Tradeoffs at Superfund Sites: Housing Market Evidence on Learning About Risk.” Review of Economics and Statistics, 82, pp. 439-451.

Greenstone, M. and J. Gallagher. (2008). “Does Hazardous Waste Matter? Evidence from the Housing Market and the Superfund Program.” Quarterly Journal of Economics, 123(3), pp. 951-1003.

Ihlanfeldt, K. and L. Taylor. (2004). “Externality Effects of Small-Scale Hazardous Waste Sites: Evidence from Urban Commercial Property Markets.” Journal of Environmental Economics and Management, 47(1), pp. 117-139.

Imbens, G., and J. Wooldridge. (2009). “Recent Developments in the Econometrics of Program Evaluation.” Journal of Economic Literature, 47(1), pp. 5-86.

Janelle, C. (2006). “NC Plant Fire May Have Started Where Chlorination Tablets Stored.” WIStv. 10/30/2006. <http://www.wistv.com/Global/story.asp?S=5609831>.

Ketkar, K. (1992). “Hazardous Waste Sites and Property Values in the State of New Jersey”. Applied Economics, 24, pp. 647-659.

Kiel, K., K. McClain. (1995). “House Prices During Siting Decision Stages: The Case of an Incinerator from Rumor through Operation”. Journal of Environmental Economics and Management, 28, pp. 241-255.

Kiel, K. and M. Williams. (2007). “The Impact of Superfund Sites on Local Property Values: Are All Sites the Same?” Journal of Urban Economics, 60(1), pp. 170-192.

Kohlhase, J.E. (1991). “The Impact of Toxic Waste Sites on Housing Values”. Journal of Urban Economics, 30, pp. 1-26.

McClelland, G., W. Schulze, B. Hurd. (1990). “The Effect of Risk Beliefs on Property Values: A Case Study of a Hazardous Waste Site.” Risk Analysis, 10(4), pp. 485-497.

Messer, K., W. Schulze, K. Hackett, T. Cameron, and G. McClelland. (2006). “Can Stigma Explain Large Property Value Losses? The Psychology and Economics of Superfund.” Journal Environmental and Resource Economics, 33(3), pp. 299-324.

Michaels, R. G. and V. K. Smith. (1990). “Market Segmentation and Valuing Amenities with Hedonic Models: The Case of Hazardous Waste Sites.” Journal of Urban Economics, 28, pp. 223-242.

Nelson, A., J. Genereux, M. Genereux. (1992). “Price Effects of Landfills on House Values.” Land Economics, 68(4), pp. 359-365.

Nelson, J.P., (1981). “Three Mile Island and Residential Property Values: Empirical Analysis and Policy Implication.” Land Economics, 57(3), pp. 363-372.

North Carolina Hazardous Waste Services. (2006). “Hazardous Waste Management Report.” North Carolina Department of Environment and Natural Resources.

Reichert, A., M. Small, S. Mohanty. (1992). “The Impact of Landfills on Residential Property Values.” Journal of Real Estate Research, 7, pp. 297-314.

Ridker, R. G. and J. A. Henning. (1967). “The Determinants of Residential Property Values with Special Reference to Air Pollution.” Review of Economics and Statistics, 49, pp. 246-257.

Rosen, S. (1974). “Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition.” Journal of Political Economy, 82(1), pp. 43-55.

Simons, R. (1999). “The Effect of Pipeline Ruptures on Noncontaminated Residential Easement-Holding Property in Fairfax County.” Appraisal Journal, July, 255-263.

Smolen, G., G. Moore, and L. Conway. (1992) “Economic Effects of Hazardous Chemical and Proposed Radioactive Waste Landfills on Surrounding Real Estate Values.” Journal of Real Estate Research, 7(3), pp. 283-296.

US Census Bureau. (2000). “County and City Data Book: 2000.” US Census Bureau. <http://www.census.gov/prod/www/ccdb.html>. Accessed 4/10/2008.

Wake County. (2006). Wake County Emergency Management. http://www.wakegov.com/em/. Accessed 12/4/2006.

WRAL.com. (2006). “Tests Show No Contamination from EQ Fire.” WRAL. 11/17/2006. <http://www.wral.com/news/10344296/detail.html>.

2.8. Tables and Figures

Table 2.1
Apex Fire Timeline

Date		Incident
October	**5**	Fire breaks out; Town of Apex, NC evacuated.
	8	Fire extinguished; Residents return home.
	23	Governor establishes task force – will analyze the impact on human health, economy, and the environment (Bowens 2006).
	25	Phase 1 of cleanup complete: all remaining chemicals isolated and scrap metal sent to local recycling facility; chemicals to be sent to site in Michigan. Start of Phase 2 cleanup: testing for remnant hazardous material (Bartlet 2006).
	26	NC regulators do not permit EQ from using site for future storage of new hazardous waste (Associated Press 2006a).
	30	Investigators discover cause of fire (Janelle 2006).
November	**9**	EQ begins removing waste to Michigan and Ohio (Associated Press 2006b).
	17	Tests show no offsite contamination from fire (WRAL.com 2006).
December	**3**	Safety records published. (Coleman and Rawlins 2006).

Table 2.2
Classes of Research

Emitting **(p=1)**	**Risky** **(1>p>0)**	**Accidents**
Blomquist (1974)	Flower and Ragas (1994)	Nelson (1981)
Kohlhase (1991)	McClelland et al. (1990)	Gamble and Downing (1982)
Nelson et al. (1992)	Michaels and Smith (1990)	Simons (1999)
Reichert et al. (1992)	Ketkar (1992)	Carroll et al. (1996)
Kiel and McClain (1995)	Smolen et al. (1992)	
Brasington and Hite (2005)	Kiel and Williams (2007)	
Gayer et al. (2001)	Ihlandfeldt and Taylor (2004)	
Boxall et al. (2005)	Messer et al. (2006)	

Table 2.3
Summary Statistics

Variable	Mean	Std. Dev	Min	Max
Sale Price (in thousands)	248.8	161.0	25.5	5000
Basement %	0.102	0.289	0	1
Assessed Quality	0.791	0.094	0.45	1.095
Age at Time of Sale (days)	4054	5735	1	69929
Heated Area (sq. ft.)	2163	910	275	18789
Lot Size (Acres)	0.329	0.452	0.01	10.14
City Water	0.884	0.320	0	1
City Sewer	0.841	0.366	0	1
Natural Gas	0.796	0.403	0	1
One Fireplace	0.864	0.343	0	1
Multiple Fireplaces	0.021	0.143	0	1
No Fireplace	0.115	0.319	0	1
One Story	0.464	0.499	0	1
Two Story	0.530	0.499	0	1
Three or More Story	0.006	0.078	0	1
Conventional	0.760	0.427	0	1
Townhouse	0.199	0.399	0	1
Ranch	0.019	0.136	0	1
Split Level	0.022	0.148	0	1
One Full Bath	0.060	0.238	0	1
Two Full Baths	0.735	0.441	0	1
Three or More Full Baths	0.205	0.403	0	1
Raleigh	0.063	0.242	0	1
Cary	0.064	0.245	0	1
Apex	0.042	0.201	0	1
Holly Springs	0.039	0.193	0	1
Garner	0.070	0.255	0	1
Fuquay Varina	0.059	0.235	0	1
Morrisville	0.051	0.219	0	1
Housing Density (# of homes within 1 mile)	73	415	1	2115
Distance to Center of Research Triangle Park	15.50	5.89	2.70	33.61
Distance to Raleigh Central Business District	9.83	4.04	0.36	21.37
Mean Commute (minutes)	29.42	5.21	16.92	44.52

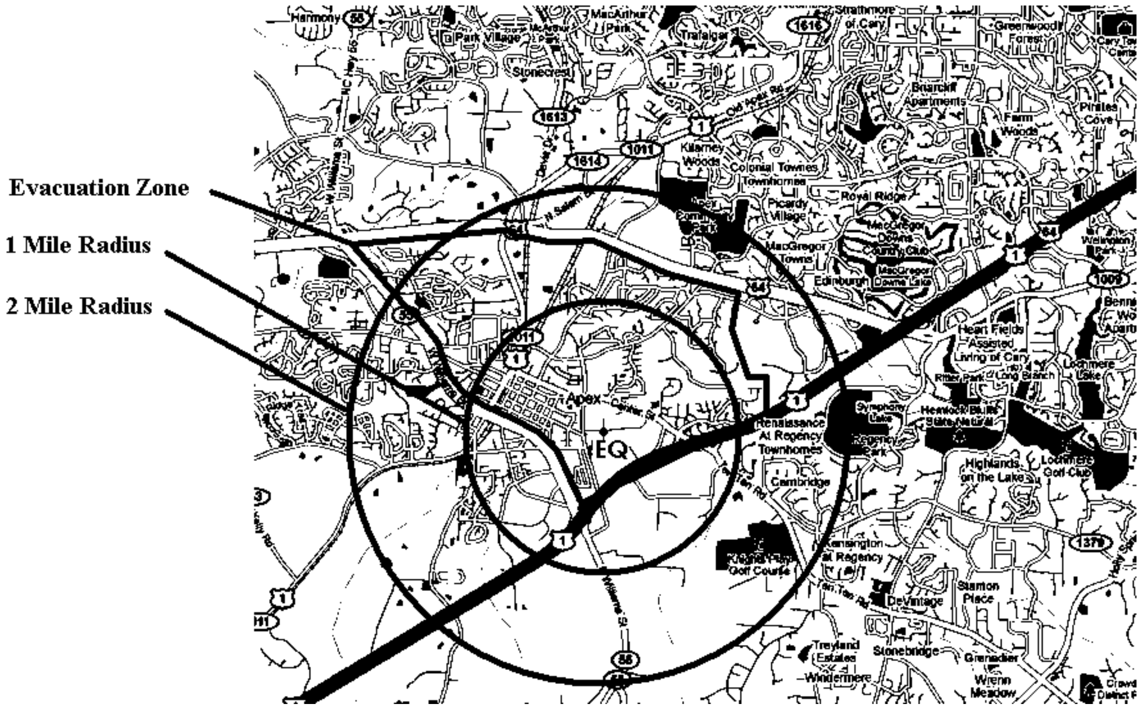
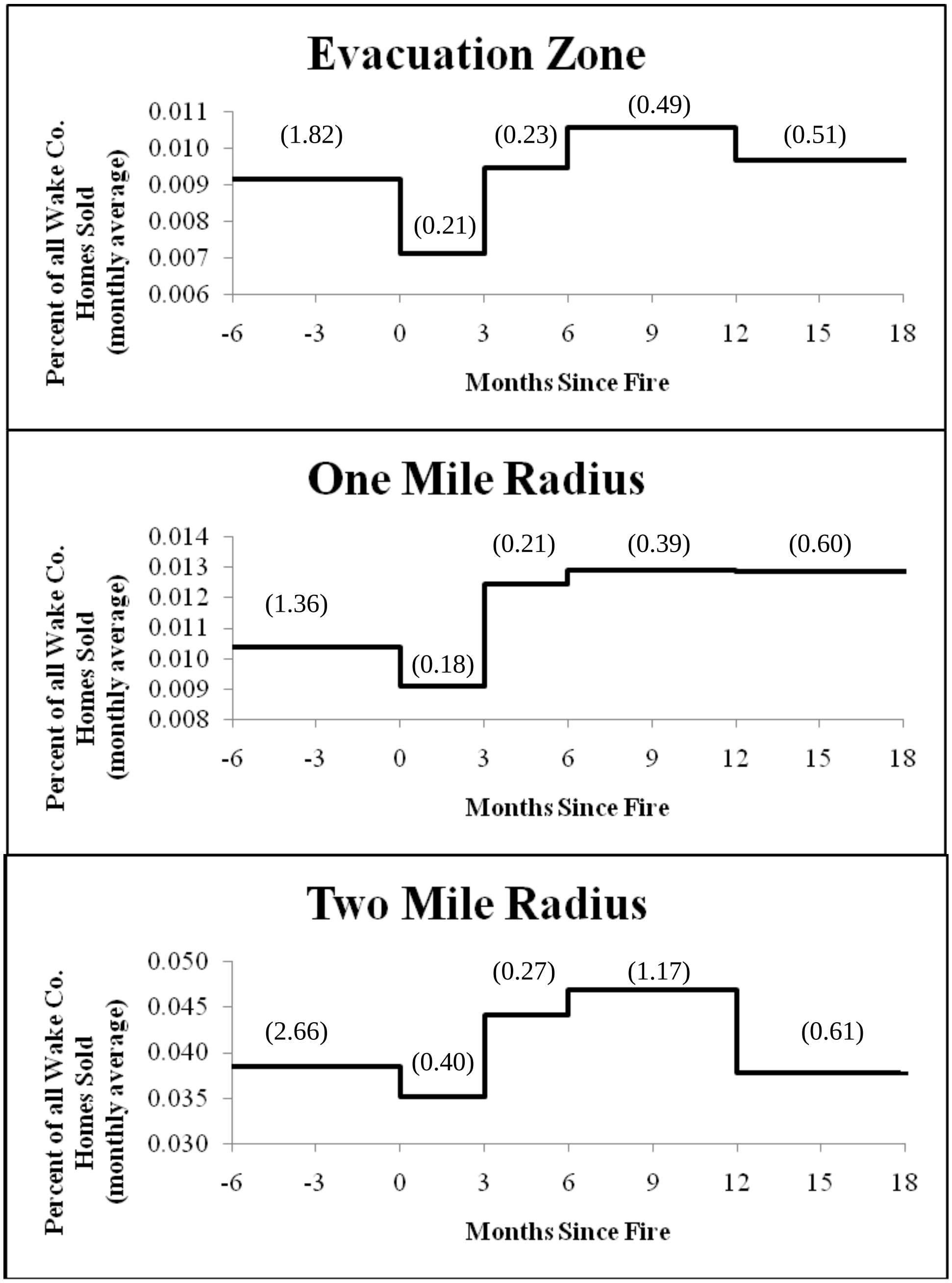


Figure 2.1
Treatment Areas Map



* Bootstrap standard errors of absolute home sales in parentheses.

Figure 2.2
Percent of Homes Sold in Treatment Area by Treatment Time Frame

Table 2.4
First Stage Hedonic Results of Property Characteristics

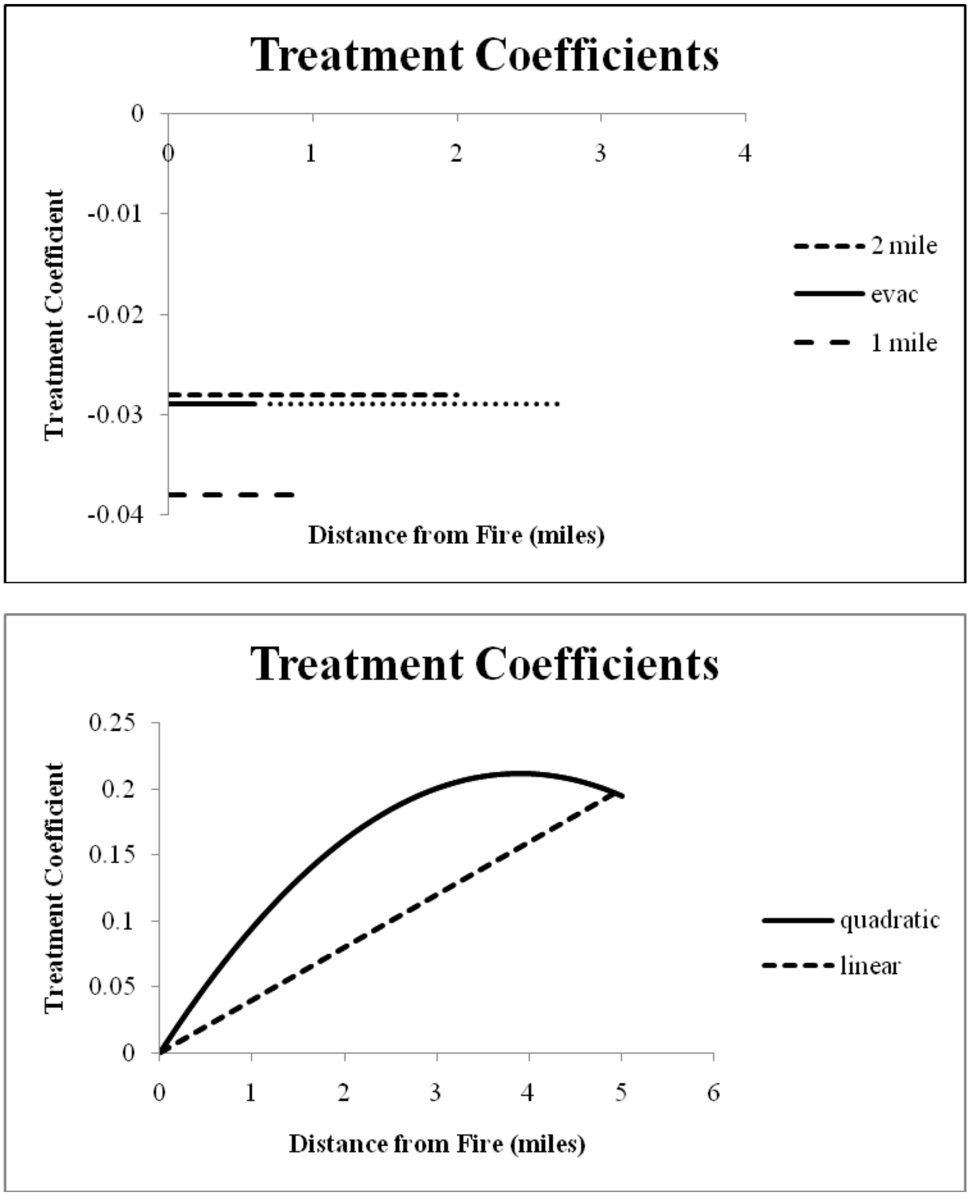
OLS on Log Sale Price		**R-Squared = 0.8915**	
Number of Observations = 84415		**Root MSE = .17108**	
Variable	**Parameter**	**Variable**	**Parameter**
Structural Characteristics		Neighborhood Characteristics	
Basement %	**0.007**	Raleigh (as opposed to unincorporated)	**0.168**
	(0.002)		**(0.003)**
Assessed Quality	**2.228**	Cary (as opposed to unincorporated)	-0.003
	(0.01)		(0.003)
Age at Time of Sale (log days)	**-0.016**	Apex (as opposed to unincorporated)	-0.005
	(0.001)		(0.003)
Heated Area (log sq. ft.)	**0.638**	Holly Springs (as opposed to unincorporated)	**0.011**
	(0.003)		**(0.003)**
Lot Size (Acres)	**0.082**	Garner (as opposed to unincorporated)	**-0.058**
	(0.002)		**(0.003)**
City Water	**0.014**	Fuquay Varina (as opposed to unincorporated)	0.000
	(0.003)		(0.003)
City Sewer	**-0.011**	Morrisville (as opposed to unincorporated)	**-0.049**
	(0.004)		**(0.003)**
Natural Gas	**0.015**	Housing Density (# of homes within 1 mile/100)	**0.001**
	(0.003)		**(0.000)**
Multiple Fireplaces (as opposed to one fireplace)	**0.101**	Distance to Center of Research Triangle Park (miles)	**-0.004**
	(0.004)		**(0.000)**
No Fireplace (as opposed to one fireplace)	**-0.056**	Distance to Raleigh Central Business District (miles)	**0.001**
	(0.002)		**(0.000)**
Two Story (as opposed to one story)	**-0.053**	Mean Commute (minutes)	**-0.005**
	(0.001)		**(0.000)**
Three or More Story (as opposed to one story)	**-0.050**		
	(0.008)	Constant	**5.897**
Townhouse (as opposed to conventional)	**-0.120**		**(0.022)**
	(0.002)		
Ranch (as opposed to conventional)	**-0.052**		
	(0.005)		
Split Level (as opposed to conventional)	**-0.095**		
	(0.004)		
Two Full Baths (as opposed to one bath)	**0.102**		
	(0.003)		
Three or More Full Baths (as opposed to one bath)	**0.141**		
	(0.004)		

NOTE: Time fixed effects not reported; Significance at the 95% level in **bold**. Some SEs are truncated.

Table 2.5
First Stage Hedonic Results
Interaction of Spatial and Time-Dependant Variables

	3 months 11/1/06-1/31/07	6 months 11/1/06-5/3/07	1 year 11/1/06-11/1/07	2.5 years 11/1/06-5/18/09
Evacuation Zone	-0.027	-0.029	-0.011	0.03
	(0.041)	(0.026)	(0.019)	(0.016)
	[18]	[48]	[107]	[245]
Distance < 1 mile	-0.041	-0.038*	-0.015	0.027
	(0.037)	(0.023)*	(0.017)	(0.015)
	[23]	[63]	[131]	[326]
Distance < 2 miles	-0.028	-0.028**	-0.006	0.028**
	(0.019)	(0.012)**	(0.009)	(0.008)**
	[89]	[224]	[475]	[958]
Distance in miles (linear, max=5)	-0.054**	0.038**	0.040**	0.054**
	(4.04E-04**)	(1.05E-03**)	(7.70E-04**)	(4.04E-04**)
Distance in miles (quadratic, max=5)	-0.099**	0.101**	0.109**	0.142**
	(3.97E-03**)	(0.014**)	(0.011**)	(5.12E-03**)
	9.21E-03**	-0.013**	-0.014**	-0.180**
	(8.05E-04**)	(2.91E-03**)	(2.14E-03**)	(1.04E-03**)

NOTE: Property characteristic and time fixed effect parameter results not reported; Results significant at the 90% and 95% level are marked with * and ** respectively. Standard errors reported in parentheses. Number of home sales included in specification in brackets. Total number of observations = 84415.



NOTE: The nearest boundary of the evacuation zone is 0.65 miles away from the fire and the farthest boundary is 2.68 miles away.

Figure 2.3
Coefficients on Six Month Time Horizon

Table 2.6
EQ Environmental Services, Apex Violation History

July 20, 2004 -- Fire hazard: At least 15 drums of a corrosive poison in a bay with caustic wastes and cyanides. Company warned.

Sept. 7, 2004 -- Safety violation: A drum of hazardous waste not closed properly, three cartons of flammable waste not marked hazardous and a container of solid hazardous waste mislabeled a liquid. Company warned.

January 3, 2005 -- Fire hazard: A large tote of caustic waste stored with a tote and a 55-gallon drum filled with acid. Company warned because it didn't do enough to prevent a reaction in case of a spill.

June 21, 2005 -- Safety violation: Four containers of acid stored with an empty tote used as a container for caustics. No immediate danger, but possible fire risk if somebody had filled the empty tote. Company warned.

July 26 and 27, 2005-- Gas leak: A tanker-truck loaded at EQ leaks ammonia gas. State says company should have done more to clean up the problem. Company fined $22,500.

Sept. 27, 2005: Safety Violation: Drum of waste not labeled hazardous. Company warned.

Dec. 21, 2005 -- Fire hazard: A 30-gallon drum containing a caustic material stored with containers of acid.

Dec. 28, 2005: Safety Violation: Workers forget to label a 55-gallon drum of waste. Company fined $9,500 for this and a previous violation.

3. Estimation and Welfare Analysis from Mixed Logit Recreation Demand Models with Large Choice Sets

3.1. Introduction

Environmental economists frequently use discrete choice models to analyze recreation site selection decisions. These models have proven valuable when modeling scenarios where agents make extensive margin decisions from large choice sets and substitution is important. However, when the individual's choice set becomes very large (on the order of hundreds or thousands of alternatives), computational limitations can make estimation with the full choice set difficult if not intractable. McFadden (1978) suggested that using a sample of alternatives in a conditional logit framework can obviate computational difficulties and produce consistent estimates. His approach has been widely used throughout the literature (e.g., Parsons and Kealy 1992; Feather 1994; Parsons and Needelman 1992). When implementing the sampling of alternatives approach, researchers typically assume that unobserved utility is independently, identically distributed type I extreme value. The assumption implies that the odds ratio for any two alternatives does not change with the addition of a third alternative. This property is known as the independence of irrelevant alternatives, or IIA. Although necessary for consistent estimation under sampling of alternatives, it is well known that IIA is often a restrictive and inaccurate characterization of choice.

In recent years, applied researchers have developed several innovative models that exploit recent computational advances to relax IIA. Perhaps the most notable and widely used is the mixed logit model. Mixed logit models generalize the conditional logit model by introducing unobserved preference heterogeneity through the parameters (Train 1998). This variation allows for richer substitution patterns and thus makes the mixed logit model a powerful and attractive tool for discrete choice modeling. However, adopting a mixed logit approach comes with a significant cost – there is no proof that the sampling of alternatives approach within the mixed logit framework will generate consistent parameter estimates. Consequently, researchers adopting the sampling of alternatives approach are forced to choose either asymptotic unbiasedness and restrictive substitution patterns with conditional logit or asymptotic bias and more flexible substitution patterns with mixed logit.[3.1]

Additionally in a mixed logit model, preference heterogeneity is often introduced through analyst-specified parametric distributions for the random parameters. The researcher's choice of error distribution thus becomes an important modeling judgment. In practice the normal distribution is often employed, although its well known restrictive skewness and kurtosis properties raise the possibility of misspecification. Alternative parametric mixing distributions have been proposed (e.g., truncated normal, log normal, triangular, uniform), but in each case misspecification remains a concern (see Hess and Rose, 2006).

Both of these problems can be overcome through the use of a finite mixture or latent class model estimated via the expectation-maximization (EM) algorithm. The latent class approach probabilistically assigns individuals to classes, where preferences are heterogeneous across – but homogeneous within – classes. This approach allows the researcher to recover separate preference parameters for each consumer type without assuming a parametric mixing distribution. Latent class models can be conveniently estimated with the recursive EM algorithm. Doing so transforms estimation of the non-IIA mixed logit model from a one-step computationally intensive estimation into a more feasible recursive estimation of IIA conditional logit models. By reintroducing the IIA property at each maximization step of the recursion, sampling of alternatives can be used to produce consistent parameter estimates (von Haefen and Jacobsen, unpublished). The implications of this estimation strategy for welfare analysis have not been investigated in the recreation demand context previously.

In this paper, we empirically evaluate the welfare implications of this novel estimation strategy for large choice set problems using a recreational dataset of Wisconsin anglers. The Wisconsin dataset is attractive for this purpose because it includes a large

[3.1] See McConnell and Tseng (200) and Nerella and Bhat (2004) for a discussion and empirical evaluation of the consistency of estimates from a sampled mixed logit.

number of recreational destination alternatives (569 lakes in total) that raises difficult but not insurmountable computational challenges for estimation with the full choice set. By comparing estimates generated with the full choice set to estimates generated with samples of alternatives of different sizes, we can gauge the benefits and costs of sampling of alternatives in terms of reduced estimation run time and efficiency loss. Our strategy for quantifying these benefits and costs involves repeatedly running latent class models on random samples of alternatives of different sizes (in particular, 50%, 25%, 12.5%, 5%, 2%, and 1% sample sizes). Our results suggest that for our preferred latent class specification, using a sample of alternatives that is 12.5% of the full choice set will generate on average a 75% time savings and 51% increase in the 95% confidence intervals for the five willingness to pay measures we construct. We also find that the efficiency losses for sample sizes as small as 5% of the full set of alternatives may be sufficiently informative for policy purposes, and that smaller sample sizes often generate point estimates with very large confidence intervals.

The chapter proceeds as follows. Section two summarizes the conditional and mixed logit models. Section three describes large choice set problems in discrete choice modeling. Section four details the latent class model estimated via the EM algorithm as well as sampling of alternatives in a mixture model. Section five presents our empirical application with the Wisconsin angler dataset. Section six concludes with a discussion of directions for future research.

3.2. The Discrete Choice Model

This section reviews the conditional logit model, the IIA assumption and the mixed logit model with continuous and discrete mixing distributions. We begin by briefly discussing the generic structure of discrete choice models. Economic applications of discrete choice models employ the random utility maximization (RUM) hypothesis and are widely used to model and predict qualitative choice outcomes (McFadden 1974). Under the RUM hypothesis, utility maximizing agents are assumed to have complete knowledge of all factors that enter preferences and determine choice. However, the econometrician’s knowledge of these factors is incomplete, and therefore preferences and choice are random from her perspective. By treating the unobserved determinants of choice as random draws from a distribution, the probabilities of choosing each alternative can be derived. These probabilities depend in part on a set of unknown parameters which can be estimated using one of many likelihood-based inference approaches (see Train (2003) for a detailed discussion).

More concretely, the central building block of discrete choice models is the conditional indirect utility function, U_{ni}, where n indexes individuals and i indexes alternatives. A common assumption in empirical work is that U_{ni} can be decomposed into two additive components, V_{ni} and ε_{ni}. V_{ni} embodies the determinants of choice such as travel cost, site characteristics, and demographic / site characteristic interactions that the econometrician observes as well as preference parameters. In most empirical applications, a linear functional form is assumed, i.e., $V_{ni} = \beta_n' x_{ni}$ where x_{ni} are observable determinants of choice and β_n are preference parameters that may vary across individuals. ε_{ni} captures those factors that are unobserved and idiosyncratic from the analyst's perspective. Under the RUM hypothesis, individual n selects recreation site i if it generates the highest utility from the available set of J alternatives (indexed by j). This structure implies the individual's decision rule can be succinctly stated as:

$$\text{alternative } i \text{ chosen } \textit{iff}$$
$$\beta_n' x_{ni} + \varepsilon_{ni} > \beta_n' x_{nj} + \varepsilon_{nj}, \forall j \neq i.$$

3.2.1. Conditional Logit

Different distributional specifications for β_n and ε_{ni} generate different empirical models. One of the most widely used models is the conditional logit which arises when $\beta_n = \beta,\ \forall n$ and each ε_{ni} is an independent and identically distributed (iid) draw from the type I extreme value distribution with scale parameter μ. The probability that individual n prefers alternative i takes the well known form (McFadden 1974):

$$P_{ni} = \frac{\exp(\beta' x_{ni} / \mu)}{\sum_j \exp(\beta' x_{nj} / \mu)} = \frac{\exp(\beta' x_{ni})}{\sum_j \exp(\beta' x_{nj})},$$

where the second equality follows from the fact that β and μ are not separately identified and thus, with no loss in generality, μ can be normalized to one.

The conditional logit model embodies the IIA property (which means that the odds ratio for any two alternatives is unaffected by the inclusion of any third alternative). To see this, consider the ratio of probabilities for sites i and k:

$$\frac{P_{ni}}{P_{nk}} = \frac{\dfrac{\exp(\beta' x_{ni})}{\cancel{\sum_j \exp(\beta' x_{nj})}}}{\dfrac{\exp(\beta' x_{nk})}{\cancel{\sum_j \exp(\beta' x_{nj})}}} = \frac{\exp(\beta' x_{ni})}{\exp(\beta' x_{nk})}.$$

Because P_{ni} and P_{nk} share the same denominator, the observable attributes for all other alternatives drops out, and thus the odds ratio will not change with the addition of any third alternative. This property of the conditional logit model is a direct result of the independent type I extreme value assumption.

IIA's restrictive implications for behavior can best be appreciated in terms of substitution patterns resulting from the elimination of a choice alternative. Consider the case of a recreational site closure due to an acute environmental incident. Assume the closed site has unusually high catch rates for trout. Intuitively, it seems plausible that the individuals who previously chose the closed site have a strong preference for catching trout and would thus resort to visiting other sites with relatively high trout catch rates when their preferred site is closed. IIA and the conditional logit model predict, however, that individuals would shift to other sites in proportion to their selection probabilities. In other words, those sites with the highest selection probabilities would see the largest increase in demand even if they do not have relatively high trout catch rates.

To generate more realistic substitution patterns, environmental economists have frequently used nested logit models where those sites with common features (e.g., high trout catch rates) are grouped into common nests that exhibit greater substitution effects (Ben-Akiva 1973; Train et al. 1987; Parsons and Kealy 1992; Jones and Lupi 1997; Parsons and Hauber 1998; Shaw and Ozog 1999; and Parsons et al. 2000). With these models, the angler's decision can be represented as a sequence of choices. For example, an angler's choice could be modeled as an initial decision of lake or river fishing, a conditional choice of target species, and a final choice of recreation site. Although the nested logit assumes that within each nest the IIA assumption holds, it relaxes IIA across different nests.

Despite its ability to relax IIA, the nested logit suffers from significant shortcomings. One shortcoming is the sensitivity of parameter and welfare estimates to different nesting structures (Kling and Thomson 1996). Another arises because all unobserved heterogeneity enters preferences through an additive error term, a characteristic shared by the conditional logit model. Although *observed* preference heterogeneity can be introduced by interacting observable demographic data with site attributes, the fact that *unobserved* heterogeneity enters preferences additively limits the analyst's ability to allow for general substitution patterns along multiple dimensions (e.g., catch rates, boat ramps, and water quality). This point seems especially relevant for situations where preferences for attributes are diverse or polarized.

3.2.2. Mixed Logit

To relax IIA and introduce non-additive unobserved preference heterogeneity, applied researchers frequently specify a mixed logit model (Train 1998; McFadden and Train 2000). Mixed logit generalizes the conditional logit by introducing unobserved taste variations for attributes through the coefficients. This is accomplished by assuming a probability density function for β_n, $f(\beta_n \mid \theta)$, where θ is a vector of parameters. Introducing preference heterogeneity in this way results in correlation in the unobservables for sites with similar attributes and thus relaxes IIA. Conditional on β_n, the probability of selecting alternative *i* in the mixed logit is:

$$P_{ni}(\beta_n) = \frac{\exp(\beta_n x_{ni})}{\sum_j \exp(\beta_n x_{nj})}.$$

The probability densities for β_n can be specified with either a continuous or discrete mixing distribution. With a continuous mixing distribution, the unconditional probability of selecting alternative *i* is:

$$P_{ni} = \int P_{ni}(\beta_n) f(\beta_n \mid \theta) d\beta_n.$$

When the dimension of β_n is moderate to large, analytical or numerical solutions for the above integral are generally not possible. However, P_{ni} can be approximated via simulation (Boersch-Supan and Hajivassiliou 1990; Geweke et al. 1994; McFadden and Ruud 1994). This involves generating several pseudo-random draws from $f(\beta_n \mid \theta)$, calculating $P_{ni}(\beta_n)$ for each draw, and then averaging across draws. By the law of large numbers, this simulated estimate of P_{ni} will converge to its true value as the number of simulations grows large.

In practice, a limitation with the continuous mixed logit model is that the mixing distribution often takes an arbitrary parametric form. Several researchers have investigated the sensitivity of parameter and welfare estimates to the choice of alternative parametric distributions (Revelt and Train 1998; Train and Sonnier 2003; Rigby et al. 2008; Hess and Rose 2006). The consensus finding is that distribution specification matters. For example, Hensher and Greene (2003) studied the welfare effect of a mixed logit model with lognormal, triangular, normal, and uniform distributions. Although the mean welfare estimates were very similar across the normal, triangular, and uniform distributions, the lognormal distribution produced results that differed by about a factor of three. In addition, although the mean welfare estimates were similar across the three tested distributions, the standard deviations varied by as much as 17 percent.

Concerns about arbitrary distributional assumptions have led many environmental economists to specify discrete or step function distributions that can readily account for different features of the data. The unconditional probability is the sum of logit kernels weighted by class membership probabilities:

$$P_n = \sum_{c}^{C} S_{nc}\left(\delta, z_n\right) P_n(\beta_c) .$$

where $S_{nc}\left(\delta, z_n\right)$ is the probability of being in class c (c = 1,…,C) and δ and z_n are parameters and observable demographics that influence class membership, respectively. If the class membership probabilities are independent of z_n, then the mixing distribution has a nonparametric or "discrete-factor" interpretation (Heckman and Singer 1984; Landry and Liu 2009). More commonly in environmental economics, however, the class membership probabilities depend on observable demographics which parsimoniously introduces additional preference heterogeneity. In these cases, the class probabilities typically assume a logit structure:

$$S_{nc}\left(\delta, z_n\right) = \frac{\exp\left(\delta_c z_n\right)}{\sum_{l=1}^{C} \exp\left(\delta_l z_n\right)}.$$

where $\delta = [\delta_1, \ldots, \delta_C]$.

3.3. Large Choice Sets

The specification of the choice set is vital to the effective implementation of any discrete choice model. Choice set definition deals with specifying the objects of choice that enter an individual's preference ordering. In practice, defining an individual's choice set is influenced by the limitations of available data, the nature of the policy questions addressed, the analyst's judgment, and economic theory (von Haefen 2008). In recreation demand applications, the combination of these factors in a given application can lead to large choice set specifications (Parsons and Kealy 1992, Parsons and Needelman 1992, Feather 2003) that raise computational issues in estimation.[3.2] This section reviews commonly used strategies for addressing the computational issues raised by large choice set specifications with a detailed treatment of the sampling of alternatives method.

[3.2] Here we are abstracting from the related issue of consideration set formation (Manski 1977; Horowitz 1991), or the process by which individuals reduce the universal set of choice alternatives down to a manageable set from which they seriously consider and choose. Consideration set models have received increased interest in recent environmental applications despite their significant computational hurdles (Haab and Hicks 1997; Parsons et al. 2000; von Haefen 2008). Nevertheless, to operationalize these models the analyst must specify the universal set from which the consideration set is generated as well as the choice set generation process. In many applications, the universal set is often very large.

3.3.1. Background

There are three generic strategies for addressing the computational issues raised by large choice sets: 1) aggregation, 2) separability, and 3) sampling. Solutions (1) and (2) require the analyst to make additional assumptions about preferences or price and quality movements within the set of alternatives.

Aggregation methods make the assumption that alternatives can be grouped into representative choice options. For a recreational demand context, similar recreation sites can be treated as one; in housing, a group of homes in a given sub-development can be aggregated. This methodology can be effective but is problematic in that the success of estimation is entirely dependent on the assumptions made in the aggregation. McFadden (1978) and Ben-Akiva and Lerman (1985) have both shown that this technique can produce biased estimates if the utility variance and composite size within aggregates is not accounted for. Although the composite size is commonly observed or easily proxied in recreation demand applications, the utility variance depends on unknown parameters and is therefore unknown and difficult to proxy by the analyst. Kaoru amd Smith (1990), Parsons and Needleman (1992) and Feather (1994) empirically investigate the bias arising from ignoring the utility variance with a recreation data set. In some cases, they find large differences between disaggregated and aggregated models, but their results suggest no clear direction of bias from aggregation. Similarly, Lupi and Feather (1998) consider a partial site aggregation method where the most popular sites and those most important for policy analysis will enter as individual site alternatives, while the remaining sites are aggregated into groups of similar sites. Their empirical results suggest =partial site aggregation can reduce but not eliminate aggregation bias.

Separability assumptions allow the researcher to selectively remove alternatives based on whether recreation sites support the type of recreation considered. For example, it is common in recreation demand analysis to focus on just boating, fishing, or swimming behavior. In these cases, sites that do not support a particular recreation activity are often eliminated. Likewise, recreation studies frequently focus on day trips, which implies a geographic boundary to sites that can be accessed with a day trip. Empirical evidence by

Parsons and Hauber (1998) on the spatial boundaries for choice set definition suggests that after some threshold distance, adding more alternatives has a negligible effect on estimation results. Nevertheless, even if the separability assumptions that motivate shrinking the choice set are valid, the remaining choice set can still be intractably large, particularly when sites are defined in a disaggregate manor..

3.3.2. Sampling of Alternatives

The third common solution is to use a sample of alternatives the decision maker faces in estimation. By selecting a random subset of the relevant alternatives, the researcher is left with a much more computationally tractable choice set. McFadden (1978) proved the approach will produce consistent estimates as long as the resulting choice probability ratios do not change due to the elimination of choice alternatives. This is feasible within the standard logit model due to the IIA assumption. Sampling has been successfully utilized and demonstrated in the literature (Parsons and Kealy 1992; Sermons and Koppelman 2001; Waddell 1996; Bhat et al. 1998; Guo and Bhat 2001; Ben-Akiva and Bowman 1998, von Haefen and Jacobsen, unpublished).

When faced with a very large choice set, randomly sampling from alternatives can simplify the computational process while still producing consistent estimates as long as the uniform conditioning property holds (McFadden 1978). This necessary condition requires that each alternative has an equal probability of being included in the sampled choice set. More formally, uniform conditioning states that if there are two alternatives, i and j which are both members of the full set of alternatives C and both have the possibility of being an observed choice, the probability of choosing a sample of alternatives D (which contains the alternatives i and j) is equal, regardless of whether i or j is the chosen alternative.

Random parameter models, as shown earlier, can account for preference heterogeneity and in some cases provide for an improvement in fit over the conditional logit model. However, when faced with a large choice set, the continuous distribution method cannot provide consistent estimates when sampling from alternatives. Recall that the mixed logit probability is represented by:

$$P_{ni} = \int L_n(\beta) f(\beta \mid \theta) d\beta$$

The relative probability of choosing alternative *i* over *i** is:

$$\frac{P_{ni}}{P_{ni*}} = \frac{\int \frac{\exp(\beta_n x_{ni})}{\sum_j \exp(\beta_n x_{nj})} f(\beta \mid \theta) d\beta}{\int \frac{\exp(\beta_n x_{ni*})}{\sum_j \exp(\beta_n x_{nj})} f(\beta \mid \theta) d\beta}$$

The denominators are inside the integral and therefore do not cancel. The resulting relative choice probabilities depend on the other alternatives and IIA does not hold.

McConnell and Tseng (2000) perform an empirical analysis on beach use and recreational fishing to evaluate sampling of alternatives. Since there is no theoretical foundation for sampling of alternatives in a continuous distribution mixture model, they seek to broaden the understanding of the mixed logit model through empirical evidence. Their results with a recreation data set consisting of a relatively small choice set suggest that sampling in a continuous distribution mixture model does not alter the results significantly or systematically, although their results should be interpreted cautiously given the same number of sampling replications they consider. Nerella and Bhat (2004) perform a similar analysis with simulated data. They analyze the effect of sample size on the empirical accuracy and efficiency of multinomial and mixed multinomial models. Their results suggest that analysts can, in practice, use a 12.5% sample in the conditional logit model and down to a 25% sample in the mixed logit model.

3.4. Sampling in a Mixture Model

Utilizing the latent class model estimated via the recursive expectation-maximization (EM) algorithm, sampling of alternatives can generate theoretically consistent estimates (von Haefen and Jacobsen, unpublished). This section describes use of the EM algorithm to estimate a latent class model and addresses the issues of model selection, and computation of standard errors.

3.4.1. EM Algorithm

The EM algorithm (Dempster et al. 1977) is an alternative estimation method for recovering parameter estimates from likelihood-based models where maximum likelihood estimation is computationally difficult. The EM algorithm has become a popular tool with estimation problems involving incomplete data (McLachlan and Krishnan 1997) as well as mixture estimation (Bhat 1997; Train 2008). The method also facilitates the consistent sampling of alternatives as shown by von Haefen and Jacobsen (unpublished).

Assuming that an unknown parameter (in this case the latent class membership) is represented as a value in some parameterized probability distribution, the EM algorithm is a recursive procedure which begins with the expectation or "E" step: specifying the expected value of unknown parameters (class membership) given some known parameters (our current guess of the model parameters). The maximization or "M" step follows: the known parameters are then updated given the expected values of the unknown parameters. The steps are then repeated until convergence, defined as a pre-determined small change in the parameter estimates between iterations (Train 2008). This methodology represents an improvement over gradient-based methods by its ability to transform the computationally difficult maximization of a log of sums into a simpler recursive maximization of the sum of logs.

Given our log-likelihood function:

$$LL_n = \ln\left(\sum_c S_{nc}(\delta_c) L_n(\beta_c^t)\right)$$

and some fixed set of starting values for the parameters $\bar{\phi}^t = (\beta_c^t, \delta^t)$, the EM algorithm lets us iteratively calculate a new value for the parameters:

$$\phi^{t+1} = \arg\max_{\phi} \sum_{n=1}^{N} \sum_{c=1}^{C} h_{nc}(\bar{\phi}^t) \ln\left(S_{nc}(\delta) L_n(\beta_c)\right)$$

where t represents the iteration number and N is the number of observations. Since the right hand portion of the equation can be rewritten as

$$\ln\left(S_{nc}(\delta) L_n(\beta_c)\right) = \ln\left(S_{nc}(\delta)\right) + \ln\left(L_n(\beta_c)\right),$$

the maximization can be performed independently for each set of parameters. Beginning with the "E" step at iteration t, using various starting values, the probability (weight) of individual n belonging to class c conditional on the parameters β_c^t and δ^t is:

$$h_{nc}(\phi^t) = \frac{S_{nc}\left(\delta^t\right)L_n(\beta_c^t)}{\sum_{l=1}^{C} S_{nl}\left(\delta^t\right)L_n(\beta_l^t)}$$

A maximization is then performed to update the individual class probability dependent on individual specific variables treating the weights from the previous step as given:

$$\delta^{t+1} = \arg\max_{\delta} \sum_{n=1}^{N}\sum_{c=1}^{C} h_{nc}(\phi^t)\ln S_{nc}\left(\delta\right)$$

Another maximization is performed to update the conditional probability parameters, again treating the weights as fixed; independently for each class:

$$\beta_c^{t+1} = \arg\max_{\beta} \sum_{n=1}^{N} h_{nc}(\phi^t)\ln L_n(\beta_c)$$

The weights are then recalculated using the new parameter values, and the entire process is repeated until convergence[3.4]. Each successive maximization uses the prior parameters ϕ^t and individual-specific class probabilities to form the weights used in the maximization the new parameter values. The previously computationally burdensome estimation has now been transformed into a recursive conditional logit estimation for each class and choice probability. By breaking the mixed logit non-IIA model into a series of conditional logit IIA models, sampling of alternatives can be reintroduced at each recursive step. The maximization procedure calculates a conditional logit (IIA) likelihood function for each class independently, keeping the individual weights fixed from the previous step.

It should be noted that when using this method, as with any mixture model, convergence may be at a local instead of a global maximum because the unconditional

[3.4] Note that when sampling, the full choice set is still used when updating the weights while the fixed sampled choice sets are used when maximizing the conditional likelihood function.

likelihood is not globally concave. To address this, it is often necessary to use multiple starting values.

3.4.2. Model Selection

A practical issue with latent class models is the selection of the number of classes. Traditional specification tests (likelihood ratio, Lagrange multiplier, and Wald tests) are inappropriate in this context because increasing the number of classes also increases the number of variables to be estimated. These tests ignore the potential of overfitting the model. Throughout the latent class literature a variety of information criteria statistics have been used. In general form (Hurvich and Tsai 1989), the information criteria statistic is specified as $IC = -2*LL + A*\gamma$ where LL is the log likelihood of the model at convergence, A is the number of estimated parameters in the model, and γ is a penalty constant. There are a number of different information criteria statistics that differ in terms of the penalty associated with adding parameters, represented by the penalty constant γ.

[Table 3.1 – Information Criteria Statistics]

In each case, the optimal model is that which gives the minimum value of the respective information criteria. Roeder et al. (1999) and Greene and Hensher (2003) suggest using the Bayesian Information Criteria (BIC). One advantage of the BIC over traditional hypothesis testing is that it performs better under weaker regularity conditions than the likelihood ratio test (Roeder et al. 1999). Alternatively, many past papers (Meijer and Rouwendal 2006; Desarbo et al. 1992; Morey et al. 2006) have used the Akaike Information Criteria (AIC) (Akaike 1974). Other papers have compared the various information criteria (Thacher et al. 2005; Scarpa and Thiene 2005), but there is no general consensus in the literature for using one test over the others.

As observed previously in Hynes et al. (2008), it is possible in practice for the analyst to select overfitted models when using only the AIC or BIC. For example, specified models with many parameters can generate parameter estimates and standard errors that are implausibly large. In our subsequent empirical exercise, parameter estimates for specific classes diverged from estimates for other classes by three orders of magnitude in some cases, while at the same time being coupled with a very small latent class probability. Results like these suggest overfitting and the need for a more parsimonious specification. Since the CAIC or crAIC penalize the addition of parameters more severely, they may be more useful to applied researchers if evidence of overfitting arises.

3.4.3. Standard Errors

Calculation of the standard errors of parameter estimates from the EM algorithm can be cumbersome since there is no direct method for evaluating the information matrix (Train 2008). There is a large statistical literature addressing various methods of calculating standard errors based upon the observed information matrix, the expected information matrix, or on resampling methods (Baker 1992, Jamshidian and Jennrich 2002, Meng and Rubin 1991). Train (2008) provides additional discussion on the different strategies available for calculating standard errors.

An aspect of the EM algorithm that can be exploited is the fact that the score of the log-likelihood is solved at each maximization step. Ruud (1991) uses this observed log-likelihood at the final step of the iteration to compute estimates of the standard errors. Derived by Louis (1982), the information matrix can be approximated with the outer-product-of-the-gradient formula:

$$\hat{I} = N^{-1}\sum_{n=1}^{N} g(\hat{\varphi})g(\hat{\varphi})'$$

where g is the score vector generated from the final step of the EM algorithm. This estimation of the information matrix is a common method for recovering standard errors and is the simplest method for doing so with the EM algorithm.[3.5]

[3.5] A limitation with this approach to estimating standard errors is that because it does not estimate the empirical Hessian directly and thus cannot be used to construct robust standard errors.

3.5. Empirical Example

This section describes the empirical example including information about the data set along with a comparison of results from the conditional logit and latent class models.

3.5.1. Data

An empirical illustration is performed with data from the Wisconsin Fishing and Outdoor Recreation Survey. Conducted in 1998 by Triangle Economic Research, this dataset has been investigated previously by Murdock (2006) and Timmins and Murdock (2007). A random digit dial of Wisconsin households produced a sample of 1,275 individuals who participated in a telephone and diary survey of their recreation habits over the summer months of 1998. 513 individuals reported taking a single day trip to one or more of 569 sites in Wisconsin (identified by freshwater lake or, for large lakes, quadrant of the lake). Of the 513 individuals, the average number of trips was 6.99 with a maximum of 50. Each of the 569 lake sites had an average of 6.29 visits, with a maximum of 108. In many ways this is an ideal dataset to evaluate the consistency of sampling of alternatives; it is large enough that a researcher might prefer to work with a smaller choice set to avoid computational difficulties, but small enough that estimation of the full choice set is still feasible for comparison. Table 3.2 presents summary statistics.

[Table 3.2 – Summary Statistics]

The full choice set is estimated with both a conditional logit model and several latent class specifications. The parameter results are evaluated and information criteria are used to compare improvements in fit across specifications. The same estimation is then also performed on randomly sampled choice sets equal to 50%, 25%, 12.5%, 5%, 2%, and 1%[3.6] of the non-selected alternatives. The sampling properties of the conditional logit model will be used to benchmark the latent class results. Since we use the outer product of the gradient to recover standard errors in the latent class model, we will use the same method with the conditional logit model.

[3.6] 285, 142, 71, 28, 11, and 6 alternatives respectively.

3.5.2. Conditional Logit Results

Estimation code was written and executed in Matlab and Gauss. In contrast to the latent class model, the likelihood function for the conditional logit model is globally concave so starting values will effect run times but not convergence. A complicating factor with our dataset is that individuals make multiple trips to multiple destinations. For consistency of the parameter estimates, it is necessary to generate a random sample of alternatives for each *individual-site visited pair*. For a sample size of M, M-1 alternatives were randomly selected and included with the chosen alternative. Two hundred random samples were generated for each sample size.

[Figure 3.1 – Estimation Time: Conditional Logit Model]

The primary benefit from sampling of alternatives is to reduce the computational burden of estimation. An analysis of sampling's effect on estimation run times suggests an almost linear relationship with diminishing returns at very small samples. Figure 3.1 shows the average estimation time of the sampled model relative to the estimation time of the full model. Cutting the sample by an additional 50% in any model roughly equates to a 56% reduction in estimation time, however the marginal time saved decreases with sample size.

[Table 3.3 –Parameter Estimates: Conditional Logit Model]

The parameter estimates and standard errors for each of the sample sizes are shown in Table 3.3. The means of the estimates and means of the standard errors from the 200 random samples are reported. Two log-likelihood values are reported in this table: the "sampled log-likelihood" (SLL) and the "normalized log-likelihood" (NLL). In any sampled model, a smaller set of alternatives will generally result in a larger log-likelihood. This number, however, is not useful in comparing goodness-of-fit across different sample sizes. The NLL is reported for this reason. After convergence is reached in a sampled model, the parameter estimates are used with the full choice set to compute the log-likelihood. A comparison of the NLL across samples shows that, when sampling, the reduction in information available in each successive sample reduces goodness of fit, as expected. A decrease in the sample size also increases the standard errors of the NLL reflecting the smaller amount of information used in estimation.

The parameters themselves are sensible (in terms of sign and magnitude) in the full model and relatively robust across sample sizes. Travel cost and small lake are negative and significant, while all fish catch rates and the presence of boat ramps are positive and significant, as expected. The standard errors for the parameters generally increase as the sample size drops, reflecting an efficiency loss when less data is used. In the smallest samples, this decrease in fit is enough to make parameters that are significant with the full choice set insignificant.

Table 3.3 suggests that parameter estimates are somewhat sensitive to sample size, but the welfare implications of these differences is unclear. To investigate this issue, welfare estimates for five different policy scenarios are constructed from the parameter estimates summarized in Table 3.3. The following policy scenarios are considered (see Table 3.4): 1) infrastructure construction[3.7], 2) an increase in entry fees[3.8], 3) an urban watershed management program, 4) an agricultural runoff management program[3.9], and 5) a fish stocking program[3.10]. Note that general equilibrium congestion effects are not considered here (Timmins and Murdock 2007), but these scenarios can be augmented or modified to fit any number of policy proposals.

[Table 3.4 – Welfare Scenarios]

The methodology used to calculate WTP is the log-sum formula derived by Hanemann (1978) and Small and Rosen (1981). Given our constant marginal utility of income $\beta_p^* f\left(y - p_j\right) = \beta_p^*\left(y - p_j\right)$ and a price and attribute change from (p^0, q^0) to (p^1, q^1), the compensating surplus is

$$CS = \frac{1}{\beta_p}\left(\max{}_j\left(-\beta_p p_j^1 + \beta_q q_j^1 + \varepsilon_j\right) - \max{}_j\left(-\beta_p p_j^0 + \beta_q q_j^0 + \varepsilon_j\right)\right)$$

[3.7] Supposing that a boat ramp was constructed at each Wisconsin lake that did not have one (27% of sites)
[3.8] $5 increase in entry fees at all state-managed sites (defined by being in a state forest or wildlife refuge); approximately 23% of sites.
[3.9] Supposing that a storm water or non-point source pollution management policy could improve the quality of water and increase the catch rate by a uniform 5% across all fish species at affected sites.
[3.10] Fish stocking program where the catch rate of trout is increased by 25% across all sites that currently contain trout.

and for our iid type 1 extreme value case, the expected consumer surplus has a closed form

$$E(CS) = \frac{1}{\beta_p}\left(\ln\left(\sum_j^J \exp(-\beta_p p_j^1 + \beta_q q_j^1) \right) - \ln\left(\sum_j^J \exp(-\beta_p p_j^0 + \beta_q q_j^0) \right) \right).$$

The full choice set is used for computation of WTP estimates.

[Figure 3.2 – Welfare Results: Conditional Logit Model]

Figure 3.2 summarizes the performance of the welfare estimates across different sample sizes using box-and-whisker plots. To construct these plots, mean WTP, 95%, and 75% confidence intervals (CIs)s for each unique sample were first calculated. Note that all CIs were constructed using the parametric bootstrapping approach suggested by Krinsky and Robb (1986). The plots contain the mean estimates of these summary statistics across the 200 random samples that were run. As the plots suggest, there is a loss of precision and efficiency with smaller sample sizes. Depending on the welfare scenario, there are modest upward or downward deviations relative to the full sample specification, however there is no consistent trend across scenarios.

For concreteness, consider the welfare effects of building a boat ramp at every site that does not have one (scenario one). The results from the full choice set model indicate that the average recreational fisherman in the dataset would be willing to pay an additional $0.70 per trip to fund the construction of a boat ramp at the 156 sites without one, with the 95% CI between $0.63 and $0.76 per trip. A researcher could have similarly run one eighth of the sample size and would expect to find a mean WTP of $0.65 per trip with the 95% CI between $0.56 and $0.73 per trip.

[Table 3.5 – Increase in Range of 95% CI of WTP Estimates Compared to Full Choice Set: Conditional Logit Model]

The loss in precision from sampling identified in Figure 3.2 comes with a significant benefit – a reduction in run time. To quantify the tradeoff between precision and run time, Table 3.5 reports the change in the range of the 95% CI across sample sizes in comparison to that of the model utilizing the full choice set. The 75% CI range is not reported, but did exhibit similar behavior. The variation in CI ranges across the five policy scenarios is relatively small, so Table 3.5 only reports mean CI ranges. The results strongly suggest that for samples as small as 12.5%, the time savings are substantial while the precision losses are modest. For example, the 50% sample size estimates were generated with a 56% time savings and resulted in 6% larger CIs. Similarly, the 12.5% sample generated results with a 90% time savings while CIs were 33% larger. By contrast for sample sizes below 12.5%, the marginal reductions in run times are small while the loss in precision is substantial. For example, moving from a 12.5% to a 5% sample of alternatives reduces run times by less than ten percent but more than doubles the loss in precision. More strikingly, moving from a 5% to a 1% sample of alternatives generates a one percent run time savings but increases CI ranges more than threefold.

3.5.3. Latent Class Results

A similar evaluation of sampling was conducted with the latent class model . For these models, convergence was achieved at the iteration in the EM algorithm where the parameter values did not change. Since the likelihood function is not globally concave and there is the possibility of convergence on a local minimum, a total of ten starting values were used for each fixed sample, the largest SLL of which was determined to be the global maximum[3.11]. The travel cost parameter was fixed across all classes but the remaining site characteristic parameters were random.

[Figure 3.3 – Estimation Time: Latent Class Model]

To provide a useful comparison to the conditional logit results presented earlier, an equivalent sample selection process was used. Ten independent samples were taken for each successive sample size, using the same randomization procedure as in the conditional logit model. The average computation time is shown in Figure 3.3 and the estimation time increases substantially with each additional class. Sampling provides a decrease in relative

---

[3.11] It may be advantageous in some situations to use more starting values to ensure convergence on a global minimum, but due to the computational burden in estimation and the large number of runs conducted, we limited ourselves to just ten starting values. This may be defensible in our situation because we do have good starting values from our full choice set model where we considered 25 sets of starting values.

runtime on a similar scale as in the conditional logit model, with diminishing time savings at much smaller sample size.

Additionally, sampling did not seem to decrease runtimes uniformly as in the conditional logit models. Convergence problems lead to increased estimation times when sampling in an overspecified model (specifically, the five and six class models). In a correctly specified model (three classes), each successive sample reduces overall runtime. Model selection issues are described in the next section.

For relatively small sample sizes with large numbers of classes, convergence was sometimes elusive. This may be the result of a specific random sample chosen having an insufficient variation in the site characteristics data to facilitate convergence. Additional runs were able to eventually produce random samples that were able to converge, however there are sample selection concerns with these results. The properties of the random samples that do not converge were not examined and the existence, source, and magnitude of any bias remain an avenue of further study.

[Table 3.6 – Information Criteria]

Model selection is performed using the information criteria described earlier. Using the NLL, the various decision rules produce different results. The CAIC indicates that five or more classes were optimal with the full choice set.[3.12] However, a careful inspection of the parameter estimates suggests possible over-fitting of the model. With five classes, the EM algorithm is attempting to estimate 87 parameters from 513 sets of individual choices. With a large number of classes, the parameters for one class for certain variables diverge dramatically. For example, one class in the five class model has parameters of -35, -86, and 29 for the forest, trout, and musky variables respectively. By contrast, the mean parameter estimates across the other four classes are 0.71, 0.50, and 4.70, respectively. This empirical finding may be the result of the 5-class model attempting to use a single class to account for a handful of anomalous outlying observations. Thus, in our view, the more appropriate

[3.12] The AIC and BIC indicated the same number of optimal classes as the CAIC.

decision criterion is the crAIC which incorporates the greatest penalty for an increased number of parameters.[3.13]

Recall that all the LL values are calculated with the parameter estimates from each sampled model using the full choice set. This methodology simplifies comparison between models that use different numbers of classes as well as sample sizes. We previously found that the three-class model is preferred to the conditional logit model when using the full choice set. We also know that reducing the sample size while keeping the number of classes fixed reduces computation time dramatically.

[Figure 3.4 – Distribution of Parameter Estimates]

The latent class model delivers three sets of parameter estimates for each of the indirect utility parameters. The distributions of select parameter estimates (boat ramp, urban, and trout catch rate) are shown in Figure 3.4. As can be seen, the latent class model can recover preference distributions that do not necessarily resemble commonly used continuous parametric distributions (e.g., normal, uniform).

[Figure 3.5 – Welfare Results: Latent Class Model]

The stability of the WTP estimates across sampling in the mixed model is analyzed in Figure 3.4 using the results from the optimal model as determined by the crAIC. Using the same policy scenarios as in the conditional logit model, WTP estimates (mean, 95%, and 75% CIs) are constructed for each individual in each class. They are then weighted by the individual latent class probabilities and summed together to produce a single value for each run. In the sampled models, the welfare estimates reported are the weighted average of ten random samples.

Interpreted, the results indicate that the average Wisconsin fisherman would be willing to pay $0.85 per trip for the proposed agricultural runoff management program (and postulated 5% increase in catch rates), with the 95% CI between $0.73 and $0.99.[3.14] The researcher could conversely have run a 5% random sample of alternatives and recovered a

[3.13] Scarpa and Thiene (2005) found similar results and drew similar conclusions.
[3.14] The CI represents the interval that is likely to include the parameter estimate.

mean WTP of $0.86 per trip, with the 95% CI being between $0.70 and $1.04. In comparison with the conditional logit model, the latent class WTP estimates are larger in magnitude for scenarios two, four, and five, while smaller for scenarios one and three.

[Table 3.7 – Increase in Range of 95% CI of WTP Estimates Compared to Full Choice Set: Latent Class Model]

Table 3.7 shows the change in the range of the mean 95% CIs across sample sizes in comparison to that of the model utilizing the full choice set. The results show that sampling can produce reasonably reliable WTP estimates down to the 5% level.[3.15] Relative to using the full choice set, the 50% choice set model's 95% CI is, on average for all considered policy scenarios, 10% wider, and this value becomes 28%, 51%, and 76% for the 25%, 12.5%, and 5% samples respectively[3.16]. At the 2% and 1% levels, the CIs for some of the samples are extremely large. In this model, WTP estimates at sample sizes below 5% could be considered unreliable. Once again, dependent on the needs of the researcher, an improvement in computation time is traded off with a lack in precision. Ultimately of course, a researcher's total run time is conditional on starting values, convergence criteria, and the number of random samples estimated.

3.6. Conclusion

This paper has investigated the welfare implications of sampling of alternatives in a mixed logit framework. By employing the EM algorithm, estimation of latent class mixed logit models can be broken down into a recursive conditional logit estimation for each class. Within each class, IIA holds and thus allows for sampling of alternatives.

We have empirically investigated the performance of the conditional logit and latent class models under various sample sizes in a recreational demand application. Our results suggest that there is modest efficiency loss and significant time savings for the conditional logit models estimated with samples of alternatives as low as 12.5% of the full choice set.

[3.15] This value is 1% for the conditional logit model.
[3.16] These values are 6%, 16%, 33%, and 76% respectively for the conditional logit model.

Smaller sample of alternative sizes generate small reductions in run times at the cost of substantial increases in the range of confidence intervals. Depending on the needs of the researcher however, results may be useful down to the 1% sample size.

For the latent class models, the results reported in this paper suggest that sampling of alternatives significantly reduces run times and performs very well for samples of alternatives as low as 25%. Although sample sizes may be useful down to the 5% level, they come with relatively small time savings and substantial losses in precision. Estimates from sample sizes below 5% were found to be unreliable in our application.

Although a broad selection of samples sizes were tested, it is unclear whether the percent sample or absolute sample size play the predominant role in the efficiency loss and time savings of sampling. Other research (von Haefen and Jacobson, unpublished) indicates through a Monte Carlo exercise that absolute sample size may be more important. Additional tests performed with an alternative dataset may be necessary to recover this result.

Certain lessons for the practitioner should be noted. As with any mixture model, the latent class model may be sensitive to starting values. We considered 10 sets of starting values for each specification due to computational limitations, but a greater number of starting values would be preferred to ensure estimation of the global minima. More starting values will increase total computation time so the researcher will have to exercise judgment in this regard. Additionally, although the consistency and efficiency of estimates at small samples will depend on the data set, our results suggest that extremely small samples (below 5%) should be avoided. See other work by von Haefen and Jacobsen (unpublished) for a further discussion and Monte Carlo analysis.

At the current state of research, this paper has demonstrated the practicality of sampling of alternatives in a discrete choice mixture model. By running several specifications on a recreation dataset, the applicability of the method has been illustrated as well. Future research could include a comparison against the nested logit model and continuous random parameter mixed logit model. Additional research analyzing bias in sampling techniques would also be valuable.

3.7. References

Akaike, H. (1974). "A New Look at the Statistical Model Identification." IEEE Transactions on Automatic Control, 19, pp. 716-723.

Baker, S. (1992). "A Simple Method for Computing the Observed Information Matrix When Using the EM Algorithm with Categorical Data." Journal of Computational and Graphical Statistics, 1(1), pp. 63-76.

Ben-Akiva, M. (1973). "Structure of Passenger Travel Demand Models." Ph.D. Dissertation, Department of Civil Engineering, MIT.

Ben-Akiva, M. and J. Bowman. (1998). "Integration of an Activity-based Model System and a Residential Location Model." Urban Studies, 35(7), pp. 1131-1153.

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

Bhat, C. (1997). "Recent Methodological Advances Relevant to Activity and Travel Behavior Analysis." In H.S. Mahmassani, ed., Recent Developments in Travel Behavior Research. Pergamom, Oxford.

Bhat, C., A. Govindarajan, and V. Pulugurta. (1998). "Disaggregate Attraction-End Choice Modeling: Formulation and Empirical Analysis." Transportation Research Record, 1645, pp. 60-68.

Boersch-Supan, A. and V. Hajvassiliou. (1990). "Smooth Unbiased Multivariate Probability Simulators for Maximum Likelihood Estimation of Limited Dependent Variable Models." Journal of Econometrics, 58, pp. 347-368.

Dempster, A., N. Laird, and D. Rubin. (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm." Journal of the Royal Statistical Society. Series B, 39(1), pp. 1-38.

Desarbo, W., M. Wedel, M. Vriens, and V. Ramaswamy. (1992). "Latent Class Metric Conjoint Analysis." Marketing Letters, 3(3), pp. 273-288.

Feather, P. (1994). "Sampling and Aggregation Issues in Random Utility Model Estimation." American Journal of Agricultural Economics, 76(4), pp. 772-780.

Feather, P. (2003). "Valuing Food Store Access: Policy Implications for the Food Stamp Program." American Journal of Agricultural Economics, 85(1), pp. 162-172.

Forinash, C. and F. Koppelman. (1993). "Application and Interpretation of Nested Logit Models of Intercity Mode Choice." Transportation Research Record, 1413, pp. 98-106.

Geweke, J., M. Keane, and D. Runkle (1994). “Alternative Computational Approaches to Inference in the Multinomial Probit Model.” Review of Economics and Statistics, 76, pp. 609–632.

Greene, W. and D. Hensher (2003). “A Latent Class Model for Discrete Choice Analysis: Contrasts with Mixed Logit.” Transportation Research B: Methodological 37(8), pp 681-698.

Guo, J.Y., and C.R. Bhat (2001). “Residential Location Choice Modeling: A Multinomial Logit Approach.” Technical Paper, Department of Civil Engineering, The University of Texas at Austin.

Haab, T. and R. Hicks (1997). “Accounting for Choice Set Endogeneity in Random Utility Models of Recreation Demand.” Journal of Environmental Economics and Management, 34, pp. 127-147.

Hanemann, W. (1978). “A Methodological and Empirical Study of the Recreation Benefits from Water Quality Improvement.” Ph.D. Dissertation, Department of Economics, Harvard University.

Heckman, J. and B. Singer (1984). “A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data.” Econometrica, 52, pp. 271-330.

Hensher, D. and W. Greene (2003). “The Mixed Logit Model: The State of Practice.” Transportation, 30(2), pp 133-176.

Hess, S. and J.M. Rose (2006), “Effects of Distributional Assumptions on Conditional Estimates from Mixed Logit Models.” ETH Working Paper, Arbeitsberichte Verkehrs- und Raumplanung, 423, IVT, ETH Zurich, Zurich.

Horowitz, J. and J. Louviere (1995). “What is the Role of Consideration Sets in Choice Modeling?” International Journal of Research in Marketing, 12(1), pp. 39-54.

Hurvich, C. and C. Tsai (1989). “Regression and Time Series Model Selection in Small Samples.” Biometrika, 76(2), pp. 297-307.

Hynes, S., N. Hanley, and R. Scarpa (2008). "Effects on Welfare Measures of Alternative Means of Accounting for Preference Heterogeneity in Recreational Demand Models." American Journal of Agricultural Economics, 90(4), pp. 1011-1027.

Jamshidian, M. and R. Jennrich (2002). "Standard Errors for EM Estimation." Journal of the Royal Statistical Society: Series B, 62(2), pp. 257-270.

Kling, C. and C. Thomson (1996). "The Implications for Model Specification for Welfare Estimation in Nested Logit Models." American Journal of Agricultural Economics, 78, pp. 103-114.

Landry, C. and H. Liu (2009). "A Semi-Parametric Estimator for Revealed and Stated Preference data - An Application to Recreational Beach Visitation." Journal of Environmental Economics and Management, 57(2), pp. 205-218.

Lee, B. (1999). "Calling Patterns and Usuage of Residential Toll Service under Self Selecting Tariffs." Journal of Regulatory Economics, 16(1), pp. 45-82.

Louis, T. (1982). "Finding the Observed Information Matrix when Using the EM Algorithm." Journal of the Royal Statistical Society. Series B, 44(2), pp. 226-233.

Lupi, F. and P. Feather (1998). "Using Partial Site Aggregation to Reduce Bias in Random Utility Travel Cost Models." Water Resources Research, 34(12), pp. 3595-3603.

McConnell, K. and W. Tseng (2000). "Some Preliminary Evidence on Sampling of Alternatives with the Random Parameters Logit." Marine Resource Economics, 14, pp. 317-332.

McFadden, D. and K. Train (2000). "Mixed MNL Models for Discrete Response." Journal of Applied Econometrics, 15(5), pp. 447-470.

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

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.

McFadden, D. and P. Ruud (1994). "Estimation by Simulation." Review of Economics and Statistics, 76(4), pp. 591-608.

McLachlan, GJ and T. Krishnan (1997). "The EM Algorithm and Extensions." Hoboken, New Jersey: John Wiley & Sons Inc.

Meijer, E. and J. Rouwendal (2006). “Measuring Welfare Effects in Models with Random Coefficients.” Journal of Applied Econometrics, 21(2), pp. 227-244.

Meng, X. and D. Rubin (1991). “Using EM to Obtain Asymptotic Variance-Covariance Matrices: The SEM Algorithm.” Journal of the American Statistical Association, 86(416), pp. 899-909.

Morey, E., J. Thacher, and W. Breffle (2006). “Using Angler Characteristics and Attitudinal Data to Identify Environmental Preference Classes: A Latent-Class Model.” Environmental and Resource Economics, 34(1), pp. 91-115.

Murdock, J. (2006). “Handling Unoberserved Site Characteristics in Random Utility Models of Recreation Demand.” Journal of Environmental Economics and Management, 51(1), pp. 1-25.

Nerella, S. and C. Bhat (2004). “Numerical Analysis of Effect of Sampling of Alternatives in Discrete Choice Models.” Transportation Research Record, 1894, pp. 11-19.

Parsons, G. and A. B. Hauber (1998). “Spatial Boundaries and Choice Set Definition in a Random Utility Model of Recreation Demand.” Land Economics, 74(1), pp. 32-48.

Parsons, G. and M. Kealy (1992). “Randomly Drawn Opportunity Sets in a Random Utility Model of Lake Recreation.” Land Economics, 68(1), pp. 93-106.

Parsons, G. and M. Needelman (1992). “Site Aggregation in a Random Utility Model of Recreation.” Land Economics, 68(4), pp. 418-433.

Parsons, G., A. Plantinga, and K. Boyle (2000). “Narrow Choice Sets in a Random Utility Model of Recreation Demand.” Land Economics, 76(1), pp. 86-99.

Revelt, D. and K. Train (1998). “Mixed Logit with Repeated Choices: Households’ Choices of Appliance Efficiency Level.” Review of Economics and Statistics, 80, pp. 1–11.

Rigby, D., K. Balcombe, and M. Burton (2009). “Mixed Logit Model Performance and Distributional Assumptions: Preferences and GM Foods.” Environmental and Resource Economics 42(3), pp. 279-295.

Roeder, K., K. Lynch, and D. Nagin (1999). “Modeling Uncertainty in Latent Class Membership: A Case Study in Criminology.” Journal of the American Statistical Association, 94, pp. 766-776.

Ruud, P. (1991). “Extensions of Estimation Methods Using the EM Algorithm.” Journal of Econometrics, 49(3), pp. 305-341.

Scarpa, R. and M. Thiene (2005). “Destination Choice Models for Rock Climbing in the Northeastern Alps: A Latent-Class Approach Based on Intensity of Preferences.” Land Economics, 81(3), pp. 426-444.

Sermons, M. and F. Koppelman (2001). “Representing the Differences Between Female and Male Commute Behavior in Residential Location Choice Models.” Journal of Transport Geography, 9(2), pp. 101-110.

Shaw, W. and M. Ozog (1999). “Modeling Overnight Recreation Trip Choice: Application of a Repeated Nested Multinomial Model.” Environmental and Resource Economics, 13(4), pp. 397-414.

Small, K. and H. Rosen (1981). “Applied Welfare Economics with Discrete Choice Models.” Econometrica, 49(1), pp. 105-130.

Thacher, J., E. Morey and E. Craighead (2005). “Using Patient Characteristics and Attitudinal Data to Identify Depression Treatment Preference Groups: A Latent-Class Model.” Depression and Anxiety, 21(2), pp. 47-54.

Timmins, C. and J. Murdock (2007). “A Revealed Preference Approach to the Measurement of Congestion in Travel Cost Models.” Journal of Environmental Economics and Management, 53(2), pp. 230-249.

Train, K. (2008). “EM Algorithms for Nonparametric Estimation of Mixing Distributions.” Journal of Choice Modeling 1(1), pp. 40-69.

Train, K. and G. Sonnier (2003). “Mixed Logit with Bounded Distribution of Partworths.” Working paper. University of California, Berkeley and Los Angeles.

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 Choices.” Rand Journal of Economics, 18(1), pp. 109-123.

Train K (2003). “Discrete Choice Methods with Simulation.” Cambridge: Cambridge University Press.

Train, K. (1998). “Recreation Demand Models with Taste Differences over People.” Land Economics, 74(2), pp. 230-239.

von Haefen, R. and M. Jacobsen (unpublished). “Sampling of Alternatives in a Mixture Model.” Working Paper, Department of Agricultural and Resource Economics, North Carolina State University.

von Haefen, R. (2008). "Latent Consideration Sets and Continuous Demand System Models," Environmental and Resource Economics, 41(3), pp. 363-379.

Waddell, P. "Accessibility and Residential Location: The Interaction of Workplace, Residential Mobility, Tenure, and Location Choices." Presented at the Lincoln Land Institute TRED Conference, 1996. (http://www.odot.state.or.us/tddtpau/modeling.html).

3.8. Tables and Figures

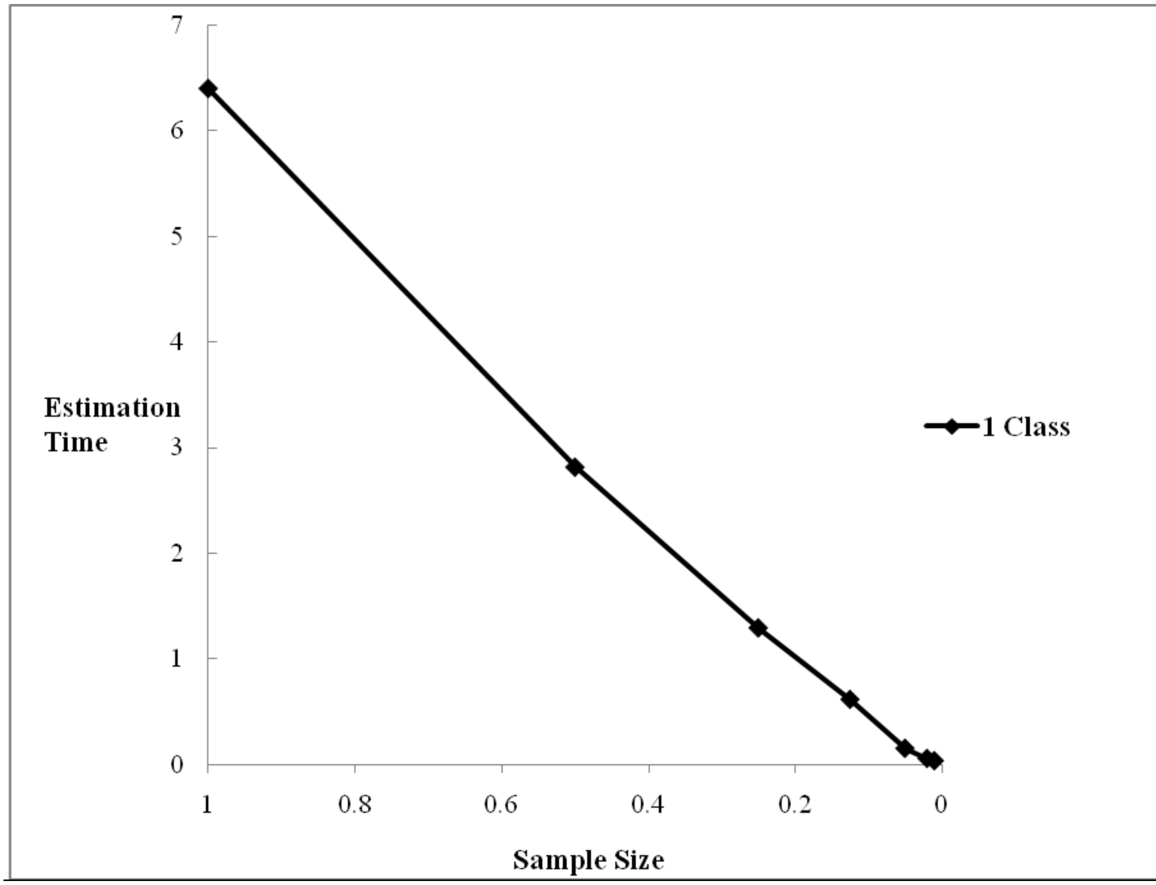
Table 3.1
Information Criteria Statistics

Information Criteria	Penalty Constant γ
Akaike Information Criteria	2
Bayesian Information Criteria	$\ln(N)$
Consistent Akaike Information Criteria	$1+\ln(N)$
Corrected Akaike Information Criteria	$2 + 2(A + 1)(A + 2)/(N - A - 2)$

* General formula: $IC = -2*LL + A*\gamma$, where LL: log likelihood, A: # of parameters, γ: penalty constant

Table 3.2
Summary Statistics

Variable	Description	Mean	St. Dev.
	Individual Summary Statistics		
trips	day trips during 1998 season	6.994	(7.182)
boat	dummy = 1 if household owns boat	0.514	-
kids	dummy = 1 if children under 14 in household	0.414	-
income	personal income	$28,991	(12,466)
	Site Summary Statistics		
tcost	round trip travel time x opp. cost of time +$0.15 x round trip miles	$100.70	(58.28)
ramp	dummy = 1 if site has at least one paved boat launch ramp	0.726	-
refuge	dummy = 1 if site is inside a wildlife area or refuge	0.056	-
forest	dummy = 1 if site is in a national, state, or county forest	0.178	-
urban	dummy = 1 if urban area on shoreline	0.179	-
restroom	dummy = 1 if restroom available	0.580	-
river	dummy = 1 if river fishing location	0.313	-
small lake	dummy = 1 if inland lake surface area <50 acres	0.172	-
trout	catch rate for brook, brown, and rainbow trout	0.094	(0.170)
smallmouth	catch rate for smallmouth bass	0.200	(0.205)
walleye	catch rate for walleye	0.125	(0.145)
northern	catch rate for northern pike	0.085	(0.057)
musky	catch rate for muskellunge	0.010	(0.022)
salmon	catch rate for coho and chinook salmon	0.009	(0.048)
panfish	catch rate for yellow perch, bluegill, crappie, and sunfish	1.579	(0.887)



* Estimation time represents the average time to convergence (in minutes) for 200 Monte Carlo runs on a 3 GHz 64 bit dual processor with 8 GB of RAM.

Figure 3.1
Estimation Time
Conditional Logit Model

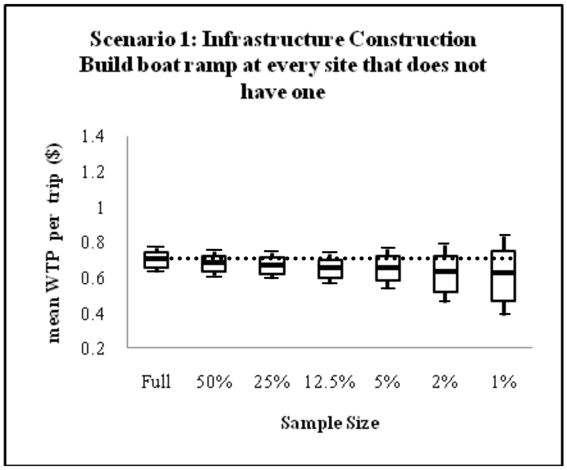
Table 3.3
Parameter Estimates: Conditional Logit Model

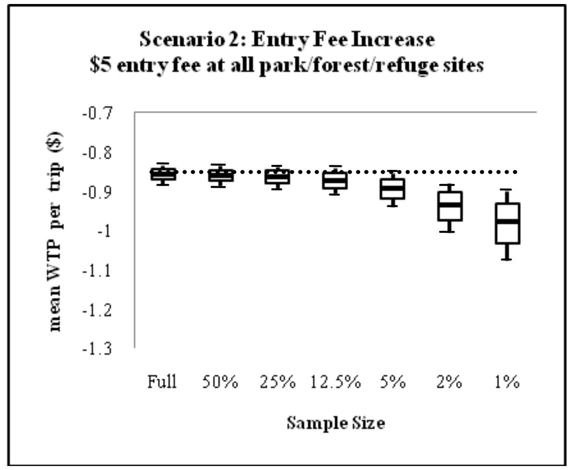
Sample Size	50% [285]	25% [142]	12.5% [71]	5% [28]	2% [11]	1% [6]	50% [285]
SLL	-13257	-10901	-8640	-6568	-4172	-2314	-1414
se		(42)	(56)	(65)	(75)	(73)	(63)
NLL	-13257	-13264	-13274	-13294	-13344	-13432	-13542
se		(4)	(6)	(11)	(27)	(53)	(105)
Variable							
tcost	**-10.07**	**-10.026**	**-9.907**	**-9.7**	**-9.303**	**-8.84**	**-8.6**
	(0.412)	**(0.407)**	**(0.4)**	**(0.393)**	**(0.39)**	**(0.401)**	**(0.434)**
ramp	**0.421**	**0.407**	**0.397**	**0.38**	**0.371**	0.349	0.343
	(0.176)	**(0.177)**	**(0.178)**	**(0.179)**	**(0.184)**	(0.195)	(0.217)
refuge	0.165	0.169	0.177	0.19	0.202	0.229	0.24
	(0.194)	(0.195)	(0.196)	(0.199)	(0.211)	(0.233)	(0.27)
forest	0.152	0.142	0.131	0.124	0.128	0.174	0.229
	(0.171)	(0.174)	(0.177)	(0.182)	(0.185)	(0.202)	(0.227)
urban	-0.068	-0.084	-0.092	-0.094	-0.08	-0.024	0.032
	(0.117)	(0.118)	(0.119)	(0.122)	(0.134)	(0.157)	(0.186)
restroom	0.149	0.149	0.144	0.143	0.147	0.174	0.211
	(0.13)	(0.131)	(0.131)	(0.13)	(0.134)	(0.148)	(0.168)
river	-0.013	-0.015	-0.02	-0.032	-0.07	-0.18	-0.266
	(0.297)	(0.299)	(0.305)	(0.319)	(0.348)	(0.399)	(0.458)
small lake	**-0.789**	**-0.789**	**-0.776**	**-0.765**	**-0.724**	**-0.707**	**-0.702**
	(0.161)	**(0.164)**	**(0.168)**	**(0.175)**	**(0.185)**	**(0.209)**	**(0.236)**
trout	**1.651**	**1.674**	**1.728**	**1.781**	**1.975**	**2.49**	**2.972**
	(0.566)	**(0.571)**	**(0.584)**	**(0.605)**	**(0.668)**	**(0.797)**	**(0.933)**
smallmouth	**0.943**	**0.948**	**0.972**	**0.996**	**1.05**	**1.149**	**1.174**
	(0.359)	**(0.362)**	**(0.371)**	**(0.376)**	**(0.369)**	**(0.385)**	**(0.427)**
walleye	**2.69**	**2.652**	**2.605**	**2.54**	**2.457**	**2.433**	**2.364**
	(0.379)	**(0.376)**	**(0.374)**	**(0.38)**	**(0.405)**	**(0.479)**	**(0.561)**
northern	**2.659**	**2.536**	**2.328**	2.013	1.505	0.908	0.658
	(0.935)	**(0.955)**	**(0.998)**	(1.07)	(1.215)	(1.472)	(1.746)
musky	**5.361**	**6.136**	**6.809**	**7.417**	**8.375**	**9.158**	**9.769**
	(1.346)	**(1.752)**	**(2.069)**	**(2.324)**	**(2.585)**	**(2.803)**	**(3.376)**
salmon	**7.733**	**7.852**	**7.968**	**8.043**	**8.139**	**7.848**	**7.545**
	(1.384)	**(1.405)**	**(1.441)**	**(1.503)**	**(1.635)**	**(1.88)**	**(2.132)**
panfish	**0.763**	**0.769**	**0.78**	**0.789**	**0.804**	**0.814**	**0.814**
	(0.189)	**(0.189)**	**(0.191)**	**(0.194)**	**(0.201)**	**(0.223)**	**(0.253)**

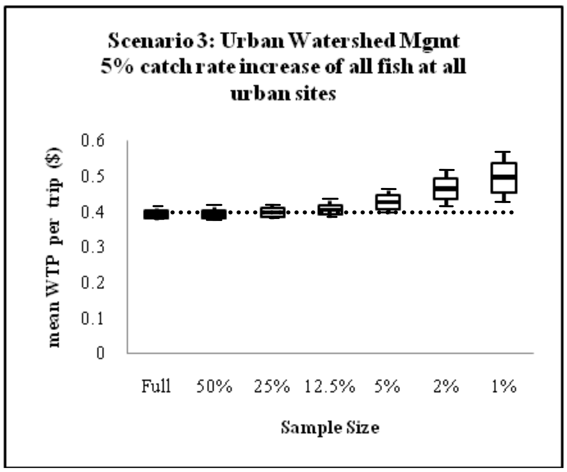
* Results for the sampled models represent the mean of 200 random samples; clustered non-robust standard errors in parentheses; **bold** indicates significance at the 5% level; "NLL" is the log-likelihood calculated at the parameter values for the entire choice set for comparison purposes.

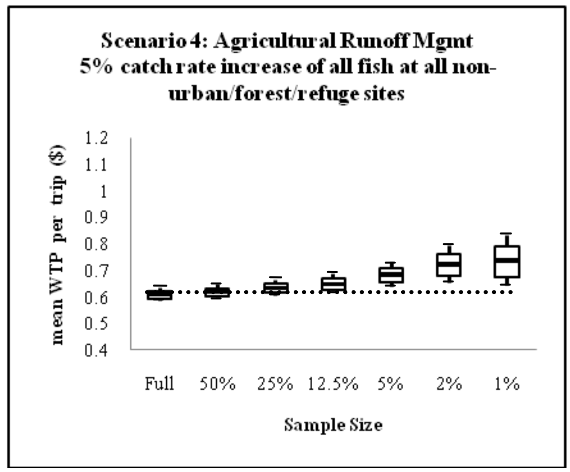
Table 3.4
Welfare Scenarios

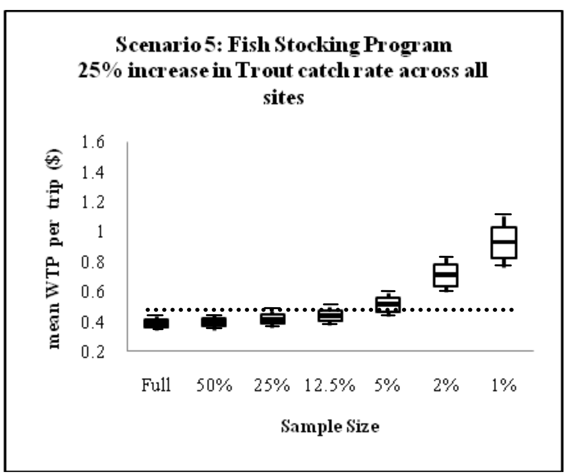
Scenario	Impacted Characteristics	Affected Sites
Infrastructure Construction	Build boat ramp at every site that does not have one	27.4%
Entry Fee Increase	$5 entry fee at all park/forest/refuge sites	23.4%
Urban Watershed Management	5% catch rate increase for all fish at all urban sites	17.9%
Agricultural Runoff Management	5% catch rate increase for all fish at all non-urban/forest/refuge sites	30.1%
Fish Stocking Program	25% increase in Trout catch rate across all sites	98.4%











* Mean WTP of 200 unique samples, the mean of the mean, 95th, and 75th CIs of which are reported.

* Method: Small and Rosen (1981); Hanemann (1978) using the parameter estimates from the sample size specified, constructed with the full choice set.

* Outer-product-of-the-gradient method used for calculating SEs.

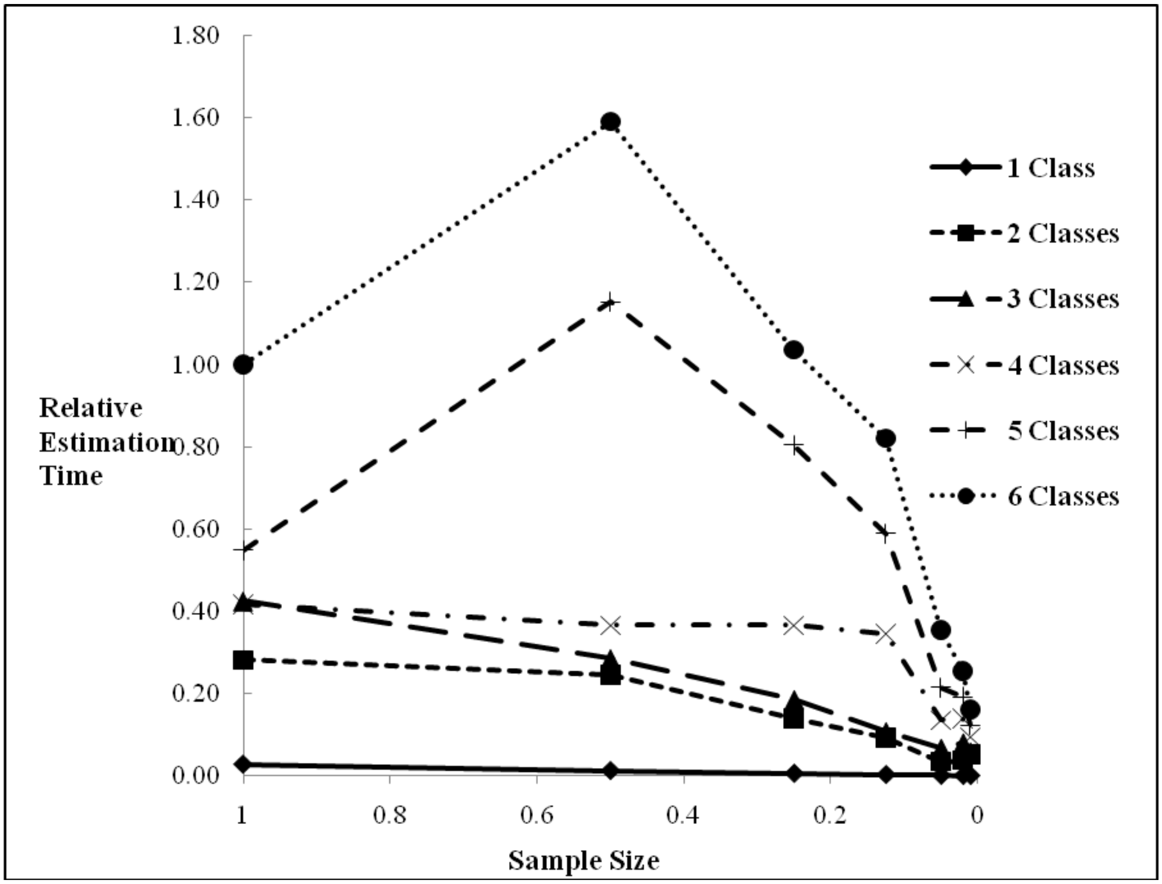
* The dashed line represents the mean WTP estimate for the full sample model.

Figure 3.2
Welfare Results
Conditional Logit Model

Table 3.5
Increase in Range of 95% CI of WTP Estimates
Conditional Logit Model

	Sample Size					
	50% [285]	**25% [142]**	**12.5% [71]**	**5% [28]**	**2% [11]**	**1% [6]**
Efficiency Loss	6%	16%	33%	76%	165%	272%
Percent Error	1%	3%	6%	13%	28%	42%
Time Savings	56%	80%	90%	98%	99%	99%

* Compared to the full choice set. Absolute sample size in brackets. Efficiency Loss is the percent increase in the range of the 95% CI. Percent Error is the percentage deviation of the mean WTP.



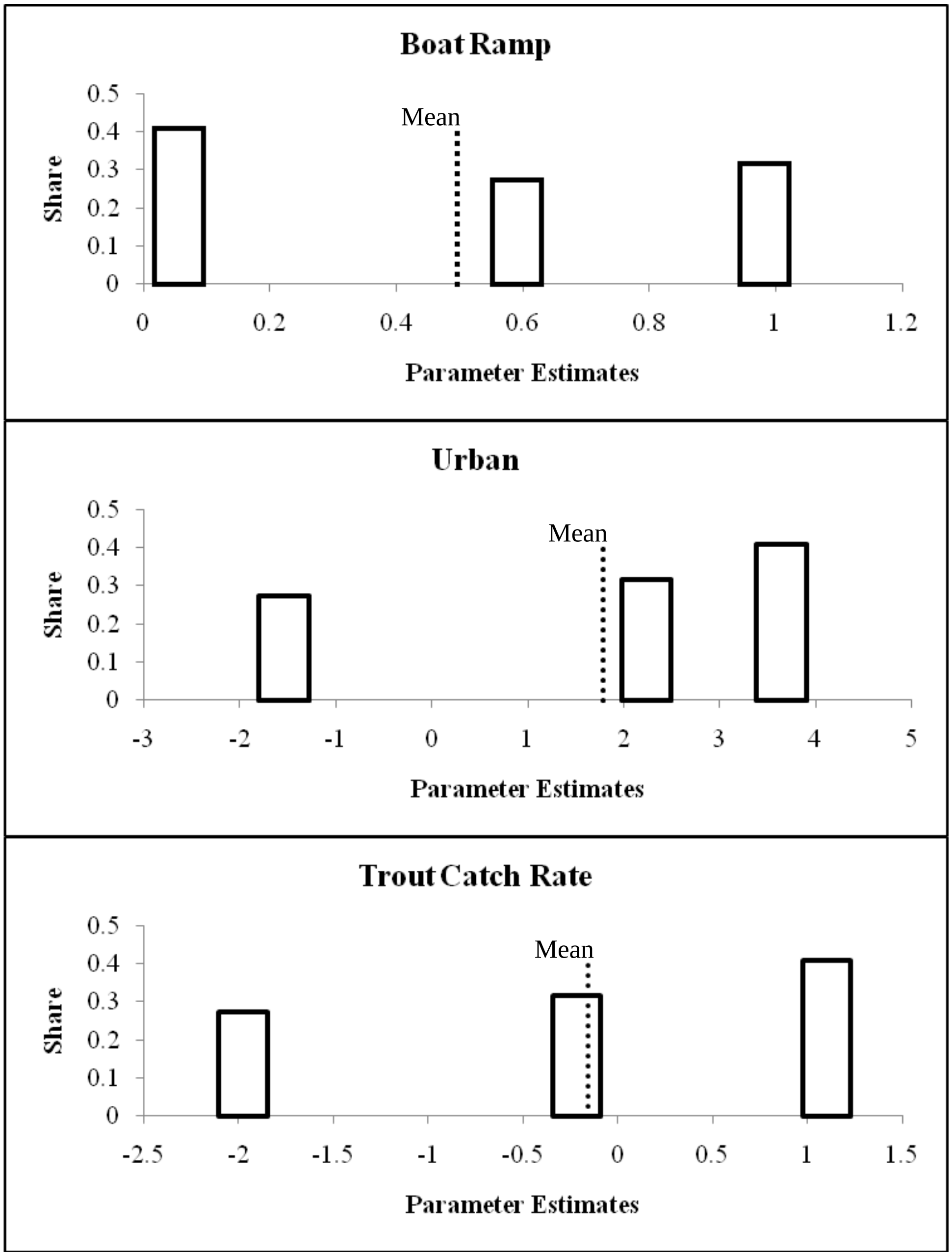
* Estimation Time represents the time to convergence (in minutes) for one random sample with one set of starting values on a 3 GHz 64 bit dual processor with 8 GB of RAM.

Figure 3.3
Estimation Time
Latent Class Model

Table 3.6
Information Criteria

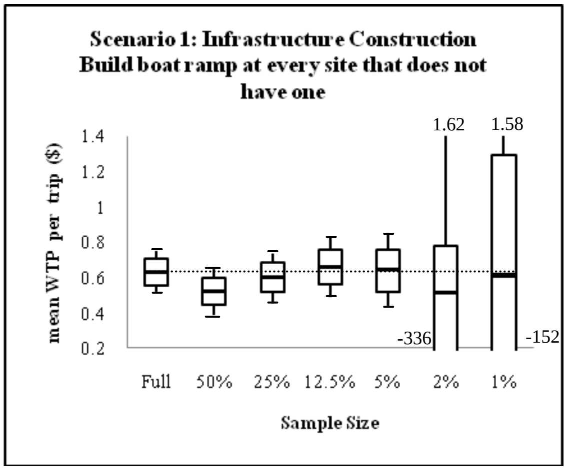
	# of classes	Sample Size						
		Full	**50%**	**25%**	**12.5%**	**5%**	**2%**	**1%**
NLL	**1**	-13257	-13264	-13274	-13294	-13344	-13432	-13542
	2	-12416	-12425	-12447	-12471	-12586	-12905	-13264
	3	-12025	-12042	-12066	-12084	-12220	-12526	-12872
	4	-11724	-11734	-11757	-11805	-12010	-12311	*DNC*
	5	-11432	-11486	-11532	-11592	-11748	-12232	*DNC*
CAIC	**1**	26623	26637	26657	26697	26797	26973	27193
	2	25071	25089	25132	25180	25411	26049	26767
	3	24420	24453	24500	24538	24809	25422	**26113**
	4	23947	23967	24013	24110	24518	25122	*DNC*
	5	**23494**	**23601**	**23693**	**23813**	**24127**	**25093**	*DNC*
crAIC	**1**	26560	26574	26594	26634	26734	26910	27130
	2	25063	25081	25124	25172	25403	26040	26759
	3	**24765**	**24798**	**24846**	**24883**	**25154**	**25767**	**26458**
	4	25141	25161	25207	25304	25712	26316	*DNC*
	5	26260	26367	26459	26579	26892	27859	*DNC*

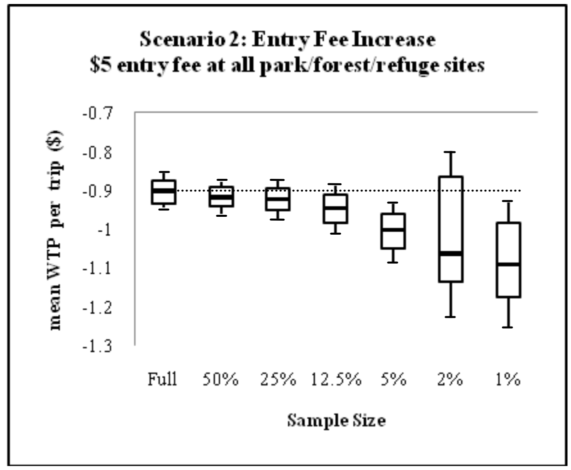
* CAIC, and crAIC calculated using the LL calculated with the full choice set; Mean of 200 and ten random samples reported for the one and multiple class models respectively; Optimal # of classes outlined and in **bold** defined by the minimum of the information criteria; *DNC* = model did not converge.

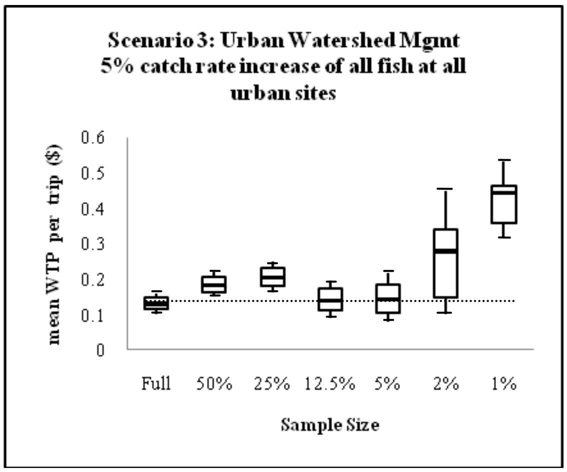


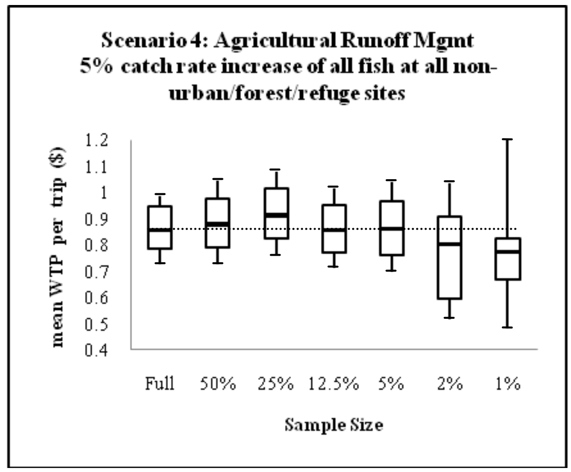
* Select parameters. Results from the three-class model with the full choice set; best of ten starting values. Class share is the mean of individual class shares calculated using the individual specific parameters. The dashed line represents the parameter mean.

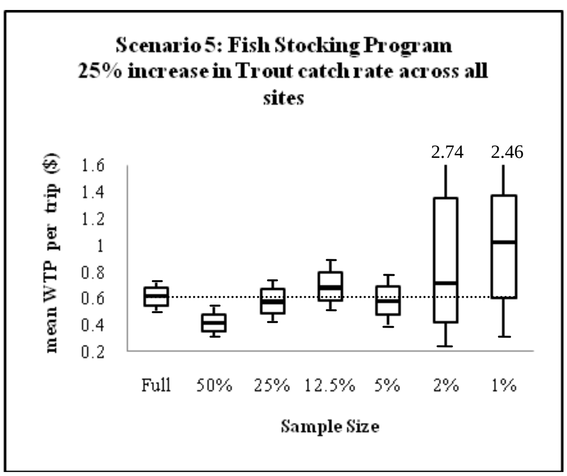
Figure 3.4
Parameter Estimate Distributions











* Mean WTP of ten unique samples, the mean of the mean, 95th, and 75th CIs of which are reported.

* Method: Small and Rosen (1981); Hanemann (1978) using the parameter estimates from the sample size specified, constructed with the full choice set.

* The dashed line represents the mean WTP estimate for the full sample model.

Figure 3.5
Welfare Results
Latent Class Model (crAIC)

Table 3.7
Increase in Range of 95% CI of WTP Estimates
Latent Class Model

	Sample Size					
	50% [285]	**25% [142]**	**12.5% [71]**	**5% [28]**	**2% [11]**	**1% [6]**
Efficiency Loss	10%	28%	51%	76%	84632%	18360%
Percent Error	19%	15%	6%	6%	34%	60%
Time Savings	33%	56%	75%	84%	81%	86%

* Compared to the full choice set. Absolute sample size in brackets. Efficiency Loss is the percent increase in the range of the 95% CI. Percent Error is the percentage deviation of the mean WTP.

4. Preferences For Property Spacing Due To Smart Growth Development Practices

4.1. Introduction

Smart growth is an urban planning model that concentrates growth in a city center and promotes compact communities with public and alternative (walking/biking) transportation options, mixed-use developments, and neighborhood schools. Aside from creating communities with a unique sense of place, smart growth policies prevent cities from expanding into outlying undeveloped land. The model has been promoted as an alternative to the suburban and urban sprawl that characterizes many US cities (Smart Growth Network 2003).

Smart growth development can be achieved either by the creation of newly-planned developments or the conversion of existing ones. Current neighborhoods might be altered by changing street layouts and encouraging mixed-use infill development. A converted community will feature new employment and commercial centers, enhanced transportation options, and an overall denser population with smaller spacing between residences.

Although a number of smart growth development programs have been implemented over the last twenty years (and they remain popular), there is a limited understanding of how these changes in public policy have been received by the general population. Past research has characterized the dynamics of urban clusters and how smart growth communities might naturally develop (Brueckner 2001, Fujita and Thisse 2002, Burchfield et al. 2006), how individuals respond to urban social interactions (Helsley and Strange 2007) and how smart growth impacts neighboring land uses (Handy 2005). Using a latent class model of housing choice, this paper will focus on testing for heterogeneous preferences of homeowners for increased housing density caused by implementation of a smart growth policy on existing neighborhoods. Such a model can guide policymakers to target certain communities or city areas that would be more responsive to implementation of a smart growth development policy.

This chapter uses a limited set of variables to characterize smart growth. The variables that more directly affect welfare in an urban setting may be ancillary to the ones used in this example. Crime, noise, and pollution (all traditionally associated with urban neighborhoods) may play an important role in determining residential household choice. This study does not address those impacts directly. Rather, the primary goal of this paper is to use a proxy for a smart growth development policy to test the viability of a latent class housing choice model that incorporates unobserved preference heterogeneity.

Recent urban economics research has found growing evidence of increases in developed and undeveloped land fragmentation along with substantial spatial heterogeneity (Irwin and Bockstael 2007). There has also been a recent increase in the amount of urban infill development (Burchfeld et al. 2006; and Sheppard 2006). The growth in urban land use could be attributed to increasing population and income, changes in transportation costs, and the changing value of agricultural land that lies on the urban-rural threshold. Incorporation of preference heterogeneity in property value models can contribute to a greater understanding of how individuals choose amongst a wide variety of housing options available on the market.

When analyzing housing choice, there is a large potential for the existence of preference heterogeneity across consumers. Individuals' choices for different housing types may be correlated with characteristics such as age, family size, or pure personal preference. Discrete choice mixture models have the ability to account for observed and unobserved preference heterogeneity more flexibly, and can be used effectively to measure the welfare impact of a smart growth development policy. Implementation of a discrete choice mixture model is unfortunately difficult when dealing with large choice sets. Such a computational problem commonly arises with city-wide housing models. Researchers are often faced with a very large set of possible alternatives and are forced to make restrictive assumptions in order to ease the computational burdens of estimating a very large choice set. However, recent innovations in discrete choice modeling provide opportunities to overcome these limitations in a housing context.

One such method is the use of the expectations-maximization (EM) algorithm to estimate a latent class discrete choice model. The model provides an opportunity to avoid the restrictive aggregation, separability, or IIA[4.1] assumptions commonly used to overcome large choice set problems while more flexibly accounting for unobserved preference heterogeneity.

This paper uses 2005-07 property value data from Wake County, North Carolina to demonstrate the applicability of a latent class model on observed housing transactions data while testing for heterogeneous preferences of housing density, property spacing, and distances from urban and employment centers. There are a wide variety of housing options available in the data set. Using a latent class discrete choice property value model, three classes of preference are recovered. These preferences can be broadly categorized as urban, rural-suburban fringe, and suburban sub-development dwellers. Welfare estimates for a hypothetical smart growth policy suggest that individuals in the first two classes would benefit, but not the third. Policymakers could use such a model to target the location of smart growth corridors in a community or target expenditure more efficiently. Population-wide welfare estimates are smaller, albeit still positive, for a latent class discrete choice model relative to a conditional logit model.

The rest of this paper proceeds as follows. Section two introduces the theory behind smart growth policies. Section three describes the current state of the literature as it relates to property value models including hedonics and discrete choice modeling (logit and random parameters). It also describes recent innovations in the use of latent class modeling in a housing context. Section four presents the dataset and the methodology used to construct individual choice sets. Section five presents estimation and welfare results. Section six finishes with a conclusion and discussion of further research.

4.2. Smart Growth

Smart growth is an effort, through the use of public and private subsidies, zoning restrictions, and investment in public infrastructure, to create a supportive environment for refocusing a share of regional growth within central cities and inner suburbs. At the same time, a share of growth is taken away from the rural and undeveloped portions of the

---

[4.1] Independence of irrelevant alternatives metropolitan area. The ten tenets of smart growth are listed in Table 4.1 (Smart Growth Network 2003).

[Table 4.1]

The principles of the policy cover a wide swath of urban development issues. From transportation, open space, recreation, and employment options, smart growth is a broad topic that addresses many policy angles. This paper will focus on those aspects concerning consumer housing; specifically where heterogeneous preferences may arise. The third tenet of smart growth proposes creating a range of housing opportunities and choices. Understanding preferences for various housing types can help guide policymakers in creating livable communities that cater to the entire population. Incorporating preference heterogeneity into models of housing choice can prove important when designing growth and development policies in a community.

4.3. Property Value Models

The environmental economics literature has empirically used differences in observed property values to estimate the impact of changes in local environmental quality. Air quality, hazardous waste, water quality, crime, and numerous other applications have been studied. The preeminent vehicle for doing so is the hedonic method. Using a least-squares approach, this method can produce marginal willingness to pay estimates for certain environmental attributes (Palmquist, 2005).

By using observed sale prices and information on the characteristics of the house and neighborhood, researchers can decompose the price of a home and recover the contribution of each of the characteristics. The hedonic price schedule in a market can be determined and the resulting parameters can be used to measure the effect of marginal changes in environmental quality. The price of the home can be characterized by the function:

$$p_i = \beta x_i + \varepsilon_i$$

Where

- p_i ~ Observed sale price
- x_i ~ Housing and neighborhood characteristics
- ε_i ~ error term.

In practice, the error term might encompass unobserved or omitted variables which play into the consumer's utility function, or by search/transactions costs. If the former is the case, the coefficient estimates might be subject to omitted variable bias.

An alternative method is the discrete choice random utility maximization (RUM) model. In this method, a homeowner's behavioral decision is modeled by analyzing the purchased home against a set of alternatives (McFadden 1978). This approach assumes that the house chosen is the one which maximizes the individual's utility. By comparing the characteristics of the chosen home against others not chosen, a model can be developed that predicts how varying combinations of characteristics can change the choice probabilities. The choice characteristics are often similar to those used in the hedonic model but now include net income as a determinant of the choice. The sale prices of homes included in the choice set are assumed to be representative of market prices available to all consumers (Palmquist, 2005).[4.2] The utility consumer *n* receives from choosing alternative *i* is

$$u_{ni} = U(x_i, y_n - p_i)$$

where y_n is the individual's household income. The consumer then chooses alternative *i* out of all other options *j* because $u_{ni} > u_{nj}$. If search costs are assumed to be constant across the choice set (perhaps via an assumption of a near-perfectly competitive market for the services of a real estate agent), the discrete choice housing model would be an appropriate technique.

A potential source of difficulty with this method arises from the definition of the choice set of the consumer. A large number of houses may be on the market at any one time and computation of the entire available choice set can be burdensome and time consuming.

[4.2] According to Palmquist (2005), this is appropriate if the error in the hedonic equation is due to omitted characteristics, but less so if search costs come into play.

Fortunately, when using the conditional logit model, the alternatives can be randomly sampled to produce a smaller representative choice set (as shown by McFadden 1978). Discrete choice methods have been employed in various contexts in past housing literature.

A number of papers represent the housing decision as a discrete choice across communities. Nechyba and Strauss (1998) estimate a nested model where individuals first choose the level of housing services (i.e. size of the house) and then choose the community based upon local public characteristics (e.g. crime rate, public school expenditure, etc.) and the community entry cost (the price of an equally sized house minus the tax rate). Using the nested logit model, the authors are able to sample from alternatives to limit the size of the choice set and find a significant impact from local school spending, community entry prices (housing prices and local tax rates), the degree of commercial activity, the distance from a metropolitan area, and crime. This model of discrete community choice is used elsewhere as well (Ioannides and Zabel 2004). The conditional logit model serves as an alternative estimation technique to the hedonic model, however employing the mixed logit model in a housing context allows for greater flexibility, which will be discussed later in this paper.

Chattopadhyay et al. (2005) compare the welfare estimates of a hedonic and discrete choice specification for the cleanup of a Superfund site. They use traditional transaction data for a hedonic model and stated preference survey data for a RUM model and find similar willingness to pay estimates.

There is also a long history of using transaction data to construct choice scenarios (Ellickson 1981; Quigley 1985; Palmquist and Israngkura 1999; Yates and Mackay 2006; Banzhaf and Smith 2007). Many of these authors have specifically addressed the issue of selecting a choice set that is small enough to be computationally feasible. Earlier papers performed a nested logit discrete choice analysis (similar to the models of neighborhood choice discussed earlier). Other papers have used various separability assumptions. Banzhaf and Smith (2007) perform a meta analysis of various choice sets defined by limitations across time, geographic space, and price. They compare 24 different specifications of the true choice set, draw five samples from each combination of boundaries, and then compute welfare measures using the true choice set. Their results provide strong guidance on how to construct choice sets in discrete choice housing models. In its simplest form, they suggest that variation across space is more important than time. Their work, however, is restricted to the conditional logit model which has a limited ability in accounting for unobserved preference heterogeneity.

An additionally innovative property value modeling technique is the sorting model of Bayer and Timmins (2007). Fundamentally, this class of research uses a discrete choice model while allowing unobserved variation in each choice alternative. Additionally, the total probability that any individual *n* will choose alternative *i*, is equal to one for each alternative. Using alternative specific constants and a contraction mapping technique (Berry, Levinsohn, and Pakes, 1995), a vector of housing prices is estimated.

Other innovations in the discrete choice literature have introduced random preference heterogeneity into the logit model. The mixed logit model assumes that individuals may have different tastes which are unobservable to the researcher (McFadden and Train, 2000). By defining preferences not as a single parameter but as a distribution of parameters, a better fitting model can be constructed and more precise and informative welfare analyses can be conducted. This model also relaxes this aforementioned IIA assumption. Doing so allows the researcher to loosen related behavioral restrictions, but is limiting because she can no longer sample from alternatives to reduce the choice set size. The use of any discrete choice random utility model in a housing application necessarily implies a large choice set. Unfortunately, this may eliminate the possibility of using a random parameter mixed logit model.

A recent innovation has used a discrete mixture model to allow consistent sampling of alternatives. Preferences are defined by a discrete distribution and the model is estimated via the EM algorithm. By assuming that preferences vary across different types of individual, but not within those types, IIA can be reintroduced to allow random sampling. An independent conditional logit estimation is performed for each class of individual. By assuming that IIA holds within class, and using share parameters to define the distribution of the classes across the population, the sampling of alternatives technique can produce consistent estimates. The method is detailed in the next section.

Mixture models are not entirely new to the housing literature however. Bhat and Guo (2004) construct a continuous distribution mixed logit model and use survey data to analyze taste preferences in community choice. Mohammadian et al (2005) use a similar model to evaluate developers' decisions over what type of housing (e.g. detached, condo, townhouse) to build. Walker and Li (2007) use a latent class model to recover household location decisions. Using a hypothetical housing option survey, they find three housing types: suburban dwellers, urban dwellers, and transit riders. This is the first mainstream application using latent class choice modeling for residential location. This stated preference technique provides valuable insight to the preference heterogeneity that exists across home owners. However, latent class estimation has yet to be applied to observed transactions data.

4.4. Latent Class Model

The latent class method incorporates unobserved preference heterogeneity by assuming that there is a discrete number of types (or classes) of consumers, and that each individual's membership is a latent variable. The unconditional likelihood of individual *n* choosing alternative *i* is the weighted sum of logit kernels with a discrete mixing distribution:

$$P_{ni} = \sum_c S_{nc}(\delta_c) L_n(\beta_c) \text{ , } c = 1,\ldots,C.$$

where S_{nc} is the latent class probability. The subscript c represents each discrete class of parameters. Conditional on the agent's choices, the probability that an agent is a member of class c (given the coefficients β_c and δ_c) is $h_{nc} = S_{nc}(\delta_c) L_n(\beta_c) / P_n$ (Train 2008).

Computationally, the log-likelihood function for the sample to be maximized is

$$LL = \sum_n \ln(P_{nc}) = \sum_n \ln\left(\sum_c S_{nc}(\delta_c) L_{ni}(\beta_c)\right)$$

where

$$S_{nc}(\delta_c) = \frac{\exp(\delta_c z_n)}{\sum_{c=1}^{C} \exp(\delta_c z_n)}$$

$$L_n(\beta_c) = \frac{\exp(\beta_c x_i)}{\sum_{j=1}^{J} \exp(\beta_c x_j)}$$

and x_n represents site-specific parameters and z_n represents individual specific parameters. When demographic variables are unavailable (as is often the case in observed transaction housing data), the latent class probability S_{nc} can be specified non-parametrically (as described in Train 2008). This method is similar to discrete factor analysis (Heckman and Singer 1984; Landry and Liu 2009). Given that the number of classes is fixed, any individual's latent class probability can be assumed to be equal to the share of the population that belongs to that class and S_{nc} becomes S_c. Gradient-based maximization of the log-likelihood is possible, albeit computationally burdensome, thus the EM algorithm is used to simplify the process.

The EM algorithm (Dempster et al. 1977) has become a popular tool in statistical estimation problems involving incomplete data (McLachlan and Krishnan 1997) and mixture estimation (Bhat 1997; Train 2008). Assuming that an unknown parameter (in this case the latent class membership) is represented as a value in some parameterized probability distribution, the EM algorithm is a recursive procedure which begins with the expectation or "E" step: specifying the expected value of unknown parameters given some known parameters. Following this is the maximization or "M" step: the parameters of the known values are then re-estimated given the expected values of the unknown parameters. The steps are then repeated until convergence, defined by a pre-determined small change in the parameter estimates between iterations (Train 2008).

In a housing context, given some set of starting values $\varphi^t = \beta_c^t$, the EM algorithm allows us to calculate a new set of parameters:

$$\varphi^{t+1} = \arg\max_{\varphi} \sum_n \sum_c h_{nc}(\varphi^t) \ln\left(S_c L_n(\beta_c)\right)$$

where t represents the iteration number. Since the right hand portion of the equation can be rewritten as

$$\ln\left(S_c L_n(\beta_c)\right) = \ln\left(S_c\right) + \ln\left(L_n(\beta_c)\right),$$

the maximization can be performed independently for each class. Beginning with the "E" step, the probability (weight) of individual n belonging to class c given the parameters β_c is:

$$h_{nc} = \frac{S_c L_n(\beta_c^t)}{\sum_{c^*} S_{c^*} L_n(\beta_{c^*}^t)}$$

A maximization is then performed to update the individual class probability dependent on individual specific variables using the weights from the previous step as given:

$$LL_n = \sum_c^C h_{nc} \ln\left(S_c\right)$$

Without demographic variables, the prior maximization is attained at:

$$S_c = \frac{\sum_n h_{nc}}{\sum_{c'} \sum_n h_{nc'}}$$

Another maximization is performed to update the conditional indirect utility parameters, again using the weights as fixed; independently for each class:

$$LL_n = \sum_c^C h_{nc} \ln\left(L_n\left(\beta_c\right)\right)$$

The weights are then re-estimated using the new parameter values, and the entire process is repeated until convergence. Each successive maximization takes the prior parameters and individual-specific class probabilities as fixed for the maximization of the new values.[4.3]

4.5. Empirical Analysis

An application of the latent class housing method is conducted using data obtained from the Wake County, North Carolina Tax Assessment Office. The dataset includes all

[4.3] For a more complete discussion of the methodology, see chapter three of this dissertation.

single family homes sold in Wake County between August 1, 2005 and July 31, 2007. Observations are limited to single family residential homes owned by individuals. There are 36,941 observations that fit these specifications. Summary statistics are presented in Table 4.2.

[Table 4.2]

Structural characteristics were chosen which would impact housing value, or proxy for some other characteristic. Variables were either introduced directly (basement percentage, assessed quality, age at time of sale, heated area, and lot size) or as dummy variables (city water, city sewer, natural gas, fireplaces, stories, style, and baths).

Most of the neighborhood characteristics were calculated from the Wake County data. Distance to the center of Research Triangle Park (RTP) and distance to Raleigh's Central Business District (CBD) are both calculated in linear miles. Mean commute time was collected from the 2000 United Stated Decennial Census and represents the mean commute time for the Census Block Group within which the home is located. Household income is also derived from the Decennial Census. The data includes 165 different block groups, with an average of 223 observations per block group.

To measure preferences for housing density and relative crowding, two key variables are used: density and property spacing. Housing density is calculated using the number of other single family homes within a one-mile radius of each home across all years. This represents the number of neighbors within one mile, however it might not entirely represent how crowded a neighborhood feels. Over the past thirty years, the trend in new home construction has been to build larger homes on similarly sized or smaller parcels than in the past. A neighborhood of 1,000 square foot homes on quarter acre lots will have a different feel compared to a neighborhood of 5,000 square foot homes on the same quarter acre lots; the latter will feel more crowded. Traditional regressions incorporating lot size, house size, and housing density would fail to capture this perceived crowding impact. Similarly, an increase in urban infill development will increase the housing density; however the relative change in crowding will depend on the size of the house built. To capture this housing attribute, a 'property spacing' variable is constructed which is equal to the size of the house in square feet, divided by the number of floors (to produce a footprint size), converted to acres, and then subtracted from the lot size in acres. This method does not account for all variables that contribute to the spacing between homes. Oblong lots, wide homes, and a large variation in homes throughout a neighborhood can all contribute to perceived neighborhood crowding. However, the property spacing metric described above provides a measure of the spacing between a home and its property boundary. Future research could use parcel GIS data to calculate the actual distance between the walls of neighboring homes.

Discrete choice housing models often use a measure of net income to create a basic equation of indirect utility. Various assumptions are made about the cost of home ownership and the resulting annual income is expressed as a choice variable. For our purposes, net income is calculated as the mean annual income for the census block minus the annualized home sale price using a fixed 5% annual interest rate over a thirty year time horizon. This variable is held fixed across classes and a constant marginal utility of income is assumed. The net income variable then represents annual expenditure on the composite numeraire good.

The choice sets are constructed following methodology described by Banzhaf and Smith (2007). In order to represent the consumer's choice set as closely as possible, alternatives are selected from other homes that were on the market during the same time period and that were within certain sale price and distance windows.

With respect to time windows, Banzhaf and Smith analyze the impact of 1, 3, 5, and 7 month time frames. They find that expanding the time horizon adds potential alternatives that are uncorrelated with the economic variables and hence with utility. In terms of the spatial boundary, they show that it is important to allow for sufficient variation in the variable being investigated.

Furthermore, the authors analyze the effects of sample size on their conclusions. Following McFadden (1978), we know that sampling produces consistent estimates in a conditional logit model. Banzhaf and Smith evaluate choice sets of 15, 20, 30, and 40 alternatives and conclude that using a 15 choice set is sufficient.

Following these recommendations, choice sets were constructed to include homes which sold within a three-month time window (1.5 months before and after the chosen sale date), within fifteen linear miles, and within 75% and 125% of the sale price. Wake County is approximately 40 miles wide at its widest, 38 miles from north to south, and the average commute is 29 minutes. The county is roughly partitioned into several residential (North Raleigh, Cary, Apex, Morrisville, Garner, and downtown Raleigh) and outlying areas (Holly Springs, Fuquay-Varina, Angier, Wendell, and Wake Forest) with downtown Raleigh placed in the relative center of the county. Discussions with local realtors have verified that a fifteen mile radius would accurately represent the spatial boundary of a potential homebuyer. Each choice set boundary will encompass a neighboring residential area and downtown Raleigh, where much urban and infill development is occurring. Following these boundaries, the average choice set size included 435 homes which were then randomly sampled to a 15 home choice set.[4.4] A five-mile radius dataset was also constructed, however upon inspection, many choice sets did not extend over downtown Raleigh.

Note that although information from prior research and information about the study area is used to construct the choice sets, only one specification is actually tested. No alternate specifications are considered, nor are any robustness checks performed. Secondarily, no demographic variables are used to recover individual latent class probabilities. The estimation performed can be considered exploratory and may give an indication of heterogeneity present in the dataset. The incorporation of demographic variables, tests of larger samples sizes, and estimation of alternate specifications are left as opportunities for future research.

[Table 4.3]

Estimation was performed in Matlab for up to four classes with analytically-coded gradients and numerical Hessians. A total of ten starting values on one random sample were

[4.4] Although the model only included homes sold between August 1, 2005 and July 31, 2007, a wider time frame was used in the construction of the choice sets so that homes sold on the boundary of the time frame did not have a restricted set of alternatives.

used for each of the multiple class models[4.5]. Results from the one-class model show reasonable parameters in terms of sign and significance and are listed in Table 4.3. Variables that increase the probability of a house being chosen (e.g. have a positive sign) include basement, city water, natural gas, assessed quality, lot size, number of fireplaces, number of bathrooms, size in square feet, and net income. Variables with a negative sign include townhouse, ranch, split level, age, housing density, and property spacing.

[Table 4.4]

Several latent class models were run, with various information criteria used to determine the optimal number of classes. Information tests such as the AIC, BIC, CAIC, and crAIC, take into account the increased number of variables in determining goodness of fit (which the traditional likelihood ratio test fails to do). Each of the criteria indicate an improvement in fit by using a latent class model. The class comparisons are listed in Table 4.4. As can been seen, the CAIC and crAIC decision rules generate different conclusions as to the optimal number of classes. To avoid overfitting, the most stringent of the two is always used.[4.6] In this case, the optimal number of classes is three.

Calculation of the standard errors of parameter estimates was via the Ruud (1991) method, using the observed log-likelihood at the final step of the iteration. Derived by Louis (1982), the information matrix can be approximated with the outer product of the gradient formula:

$$\hat{I} = N^{-1}\sum_{n=1}^{N} g(\hat{\varphi})g(\hat{\varphi})'$$

where g is the score vector. This estimation of the information matrix is a common method for recovering standard errors and is the simplest method for doing so with the EM algorithm[4.7].

[Table 4.5]

[4.5] Again, a viable robustness check would test the results from multiple random samples.
[4.6] For a further discussion of the various information criteria, see Thacher et al. 2005; Scarpa and Thiene 2005; Hynes et al. 2008; or chapter three of this dissertation.
[4.7] It is not possible to recover the robust covariance matrix without an approximation of the Hessian.

Select parameter estimates from the optimal model are shown in Table 4.5. Using the share parameter calculated during the 'E' step of the estimation process, the population can be segregated into three types of preferences. By comparing the sign and magnitude of parameters across the three classes[4.8], relative preferences for various community types can be inferred. However, care must be taken when interpreting parameter values. Some of these might be strict preference parameters (home size), but some may relate to a certain type of community make up (older homes in Wake County tend to be located in denser neighborhoods closer to the CBD). Without additional information, it is difficult to recover which is this case.

Of the three classes, the first is the largest, with a 40.5% share. In comparison to classes two and three, this class prefers to be closer to RTP and the Raleigh CBD. The parameter on home size is positive, but not as large relative to the other two classes, and this group prefers less spacing between their home and the next. The second class prefers homes that are closer to RTP and farther from the Raleigh CBD, prefers larger homes than class one (but not as large as class three), and likes to be close to neighbors. The third class likes to be farther from RTP and closer to the Raleigh CBD, prefers the largest homes (in comparison to classes one and two), and prefers relatively greater space between their neighbors.

Recall that since latent class probability is calculated non-parametrically, it is represented as a share value common across all observations. Individual homeowners cannot be probabilistically assigned to one of the latent preference categories. Additionally, since the supply of housing is considered fixed, the latent classes also represent the shares of housing type available to consumers. It is not possible to recover spatial boundaries for each housing type using the results of this estimation. However, by including demographic variables in the estimation of latent class probabilities (either assigned by block groups, or preferably, by matching homes to observations in the full census dataset), a county-wide map of preference type can be recovered and potentially provide the analyst or policy maker with valuable spatial segregation information. In the current estimation regardless, the parameters

---

[4.8] Recall that net income is held fixed across classes.

can still be used to recover more precise welfare estimates than from a homogenous-preference conditional logit model.

Analysis of welfare effects for a given policy scenario can easily be performed to predict its impact on different portions on the population. A hypothetical policy can be proposed, and its assumed change in neighborhood variables will be used to construct welfare estimates. As mentioned earlier, there is a growing trend for development of smart growth policies in urban centers, although these policies may be welcomed by one subset of the population but discouraged by another. Individuals who prefer to live in a surburban development might not benefit from a smart growth policy, but others might. Much of smart growth's strategy centers not only on properly designing new communities, but also transforming existing ones into more walkable and compact communities that are closer to employment centers. Expanded results could allow policymakers to target those communities.

A hypothetical policy proposal is presented that would seek to focus future growth in the current urban landscape areas instead of extending sprawl. This policy would additionally implement structural changes that would transform every community in Wake County into a smart growth community. This can be done by encouraging employment centers to locate in physical community centers, promoting more infill development, and improving community layout, thus changing the environmental characteristics of the neighborhood. In relation to our parameters, such a policy would shorten commute time, create denser neighborhoods, and decrease the spacing between homes.

Not every home in Wake County would be located in such an idyllic community in practice, however for the sake of this paper, a uniform 10% decrease in commute time and a 10% decrease in property spacing is assumed for testing purposes. Willingness to pay (WTP) estimates for each class are calculated using the log-sum method of Small and Rosen (1981) and Hanemann (1978) with the sampled choice set. Assuming a constant marginal utility of income $\beta_p^* f\left(y - p_j\right) = \beta_p^*\left(y - p_j\right)$ and an attribute change from (q^0) to (q^1), the compensating surplus is

$$CS = \frac{1}{\beta_p}\left(\max{}_j\left(-\beta_p p_j^0 + \beta_q q_j^1 + \varepsilon_j\right) - \max{}_j\left(-\beta_p p_j^0 + \beta_q q_j^0 + \varepsilon_j\right)\right)$$

and for our iid type 1 extreme value case, the expected consumer surplus has a closed form

$$E(CS) = \frac{1}{\beta_p}\left(\ln\left(\sum_j^J \exp(-\beta_p p_j^0 + \beta_q q_j^1)\right) - \ln\left(\sum_j^J \exp(-\beta_p p_j^0 + \beta_q q_j^0)\right)\right).$$

[Table 4.6]

Welfare results are shown in Table 4.6. These values reflect annual costs and benefits from the policy in dollars. The first class has a positive WTP of $2,934 per year, the second has a WTP of $44, while the third class has a WTP of -$3,027. As can be seen, the first class of individuals (the urban dwellers) would benefit from this policy while the second and third classes (the rural-suburban fringe and suburban subdevelopment dwellers) would not. The conditional logit WTP result is a positive WTP of $1,445. If we sum the three mean latent WTPs together using the class shares, we get a single positive value that is smaller than the value obtained from conditional logit estimation.

The conditional logit model suggests a uniformly implemented smart growth development policy would benefit the entire population. However, the latent class results suggest that the simple model actually understates the policy's impact for approximately 40% of the population, and overstates it for the remaining 60%. Using only conditional logit estimation to define policy direction would ignore the importance of preference heterogeneity in the population.

These results illustrate the importance of the third tenet of smart growth: creating a range of housing opportunities and choices. Different individuals (or the same individual at different stages of their life) may have preferences for different types of housing. Maintaining this diversity of choices is an important consideration. Policymakers can use this information to direct limited financial resources to develop smart growth corridors that would affect the right individuals. In Wake County, the results suggest that a non-uniform development policy should be employed.

The next stage of research could involve revisiting the data, incorporating household demographics, and estimating parameterized latent class probabilities. Aside from serving as a robustness check, these additional estimation results could probabilistically assign individual households to each of the latent classes. As mentioned earlier, this would be useful in defining spatial boundaries for each of the three classes and providing further guidance for policy.

4.6. Conclusion

The above research analyzes the impact of new development on existing communities and provides an opportunity to expand the role of the discrete choice model in the property value literature. By providing an avenue for introducing unobserved taste variation into the analysis, more informative welfare results can be obtained. Similar advances have already been implemented elsewhere and there are opportunities to apply those same innovations in the housing literature.

The empirical exercise performed above has shown that with proper construction of individual choice sets, a discrete choice property value model can yield reasonable results. Furthermore, by running a latent class model, an improvement in model fit can be attained while more flexibly accounting for unobserved preference heterogeneity. Analysis shows that the population can be partitioned into three types of residential consumers, each with different preferences for a smart growth development policy.

Understanding this preference heterogeneity can be very valuable to policymakers as they choose where in a metropolitan landscape to implement new smart growth development practices. As opposed to implementing a uniform development policy, a latent class model can dictate how to implement a non-uniform policy that takes homeowner type into consideration and may provide a more targeted and economically optimal solution.

When using this method in practice, care must be taken to ensure proper construction of representative choice sets. A sufficient number of starting values also needs to be run to ensure calculation of the global maximum of the unconditional log likelihood. Additional runs utilizing demographic data to parametrically estimate latent class probabilities are a necessary next step. This is left open for further research. It is unclear however, if block group variables[4.9] would provide sufficient spatial precision, or whether access to the full census data set would be required.

[4.9] Obtainable via publically available census data.

4.7. References

Banzhaf, H. S. and V. K. Smith. (2007). "Meta-Analysis in Model Implementation: Choice Sets and the Valuation of Air Quality Improvements." Journal of Applied Econometrics, 22(6), pp. 1013-1031.

Bayer, P. and C. Timmins. (2007). "Estimating Equilibrium Models of Sorting Across Locations." Economic Journal, Royal Economic Society, 117(518), pp. 353-374.

Berry, S., J. Levinsohn, and A. Pakes. (1995). "Automobile Prices in Market Equilibrium." Econometrica, 63, pp. 841-890.

Bhat, C. (1997). "Recent Methodological Advances Relevant to Activity and Travel Behavior Analysis." In H.S. Mahmassani, ed., Recent Developments in Travel Behavior Research. Pergamom, Oxford.

Bhat, C. and J. Guo. (2004). "A Mixed Spatially Correlated Logit Model: Formulation and Application to Residential Choice Modeling." Transportation Research Part B, 38, pp. 147-168.

Burchfield, M., H. Overman, D. Puga and M. Turner. (2006). "Causes of Sprawl: A Portrait from Space." Quarterly Journal of Economics, 121, pp. 587-633.

Chattopadhyay, S., J. Braden, and A. Patunru. (2005). "Benefits of Hazardous Waste Cleanup: New Evidence from Survey and Market Based Property Value Approaches." Contemporary Economic Policy, 23(3), pp. 357-375.

Dempster, A., N. Laird, and D. Rubin. (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm." Journal of the Royal Statistical Society. Series B, 39(1) pp. 1-38.

Ellickson, B. (1981). "An Alternative Test of the Hedonic Theory of Housing Markets." Journal of Urban Economics, 9(1), pp. 56-79.

Epple, D. and H Seig. (1999). "Estimating Equilibrium Models of Local Jurisdictions." Journal of Political Economy, 107, pp. 645-681.

Fujita, M. and J. Thisse. (2002). "Economics of Agglomeration. Cities, Industrial Location, and Regional Growth." Cambridge University Press, Cambridge.

Handy, S. (2005). “Smart Growth and the Transportation-Land Use Connection: What Does The Research Tell Us?” International Regional Science Review, 28(2), pp. 146-167.

Hanemann, W. (1978). "A Methodological and Empirical Study of the Recreation Benefits from Water Quality Improvement." Ph.D. dissertation, Department of Economics, Harvard University.

Heckman, J. and B. Singer. (1984). "A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data." Econometrica, 52, pp. 271-330.

Helsley, R. and W. Strange. (2007). “Urban Interactions and Spatial Structure.” Journal of Economic Geography, 7(2), pp. 119-138.

Hynes, S., N. Hanley, and R. Scarpa. (2008). "Effects on Welfare Measures of Alternative Means of Accounting for Preference Heterogeneity in Recreational Demand Models." American Journal of Agricultural Economics, 90(4), pp. 1011-1027.

Ioannides, Y. and J. Zabel. (2008). “Interactions, Neighborhood Selection and Housing Demand.” Journal of Urban Economics, 63(1), pp. 229-252.

Irwin, E. and N. Bockstael (2007). “The Evolution of Urban Sprawl: Evidence of Spatial Heterogeneity and Increasing Land Fragmentation.” Proceedings of the National Academy of Sciences.

Landry, C. and H. Liu. (2009). "A Semi-Parametric Estimator for Revealed and Stated Preference data - An Application to Recreational Beach Visitation." Journal of Environmental Economics and Management, 57(2), pp. 205-218.

Louis, T. (1982). “Finding the Observed Information Matrix when Using the EM Algorithm.” Journal of the Royal Statistical Society. Series B, 44(2), pp. 226-233.

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.

McFadden, D. and K. Train. (2000). "Mixed MNL Models for Discrete Response." Journal of Applied Econometrics, 15(5) pp. 447-470.

McLachlan, GJ and T. Krishnan. (1997). “The EM Algorithm and Extensions.” Hoboken, New Jersey: John Wiley & Sons Inc.

Mohammadian, A. K., Haider, M. and Kanaroglou, P. S. (2004). “Incorporating Spatial Dependencies in Random Parameter Discrete Choice Models.” Presented at 84th Annual Transportation Research Board Meeting.

Nechyba, T. and R. Strauss. (2008) "Community Choice and Local Public Services: A Discrete Choice Approach." Regional Science and Urban Economics, 28(1), pp. 51-74.

Palmquist, R. (2005). “Property Value Models.” Handbook of Environmental Economics, 2(16), pp. 763-819. Editors: K. Maler and J. Vincent.

Palmquist, R. and A. Israngkura. (1999). “Valuing Air Quality with Hedonic and Discrete Choice Models.” American Journal of Agricultural Economics, 81(5), pp. 1128-1133.

Quigley, J. (1985). "Consumer Choice of Dwelling, Neighborhood, and Public Services." Regional Science and Urban Economics, 15(1), pp. 1-23.

Ruud, P. (1991). "Extensions of Estimation Methods Using the EM Algorithm." Journal of Econometrics, 49(3), pp. 305-341.

Scarpa, R. and M. Thiene. (2005). "Destination Choice Models for Rock Climbing in the Northeastern Alps: A Latent-Class Approach Based on Intensity of Preferences." Land Economics, 81(3), pp. 426-444.

Sheppard, S. (2006), “Infill Versus Outspill: The Microstructure of Urban Expansion.” Williams College Working Papers in Economics.

Small, K. and H. Rosen. (1981). “Applied Welfare Economics with Discrete Choice Models.” Econometrica, 49(1), pp. 105-130.

Smart Growth Network. (2003). “Getting to Smart Growth II: 100 More Policies for Implementation.” International City/County Management Association.

Thacher, J., E. Morey, and E. Craighead (2005). “Using Patient Characteristics and Attitudinal Data to Identify Depression Treatment Preference Groups: A Latent-Class Model.” Depression and Anxiety, 21(2), pp.47-54.

Train, K. (2008). "EM Algorithms for Nonparametric Estimation of Mixing Distributions." Journal of Choice Modeling, 1(1), pp. 40-69.

Walker, J. and J. Li. (2007). "Latent Lifestyle Preferences and Household Location Decisions." Journal of Geographical Systems, 9, pp. 77-101.

Yates, J. and D. Mackay. (2006). “Discrete Choice Modeling of Urban Housing Markets: A Critical Review and an Application.” Urban Studies, 43(3), pp. 559-581.

4.8. Tables and Figures

Table 4.1
Tenets of Smart Growth

1	Mix land uses
2	Take advantage of compact building design
3	Create a range of housing opportunities and choices
4	Create walkable communities
5	Foster distinctive, attractive communities with a strong sense of place
6	Preserve open space, farmland, natural beauty, and critical environmental areas
7	Strengthen and direct development towards existing communities
8	Provide a variety of transportation choices
9	Make development decisions predictable, fair, and cost effective
10	Encourage community and stakeholder collaboration in development decisions

* Smart Growth Network (2003)

Table 4.2
Summary Statistics

Variable	Mean	Std. Dev	Min	Max
Sale Price (in 1,000s)	248.8	161.0	25.5	5000.0
Basement %	0.102	0.289	0	1
City Water	0.884	0.320	0	1
City Sewer	0.841	0.366	0	1
Natural Gas	0.796	0.403	0	1
Assessed Quality	0.791	0.094	0.45	1.095
Lot Size (Acres)	0.329	0.452	0.01	10.14
One Fireplace	0.864	0.343	0	1
Multiple Fireplaces	0.021	0.143	0	1
No Fireplace	0.115	0.319	0	1
One Story	0.464	0.499	0	1
Two Story	0.530	0.499	0	1
Three or More Story	0.006	0.078	0	1
Conventional	0.760	0.427	0	1
Townhouse	0.199	0.399	0	1
Ranch	0.019	0.136	0	1
Split Level	0.022	0.148	0	1
One Full Bath	0.060	0.238	0	1
Two Full Baths	0.735	0.441	0	1
Three or More Full Baths	0.205	0.403	0	1
Mean Commute (minutes)	29.42	5.21	16.92	44.52
Distance to Center of Research Triangle Park	15.50	5.89	2.70	33.61
Distance to Raleigh Central Business District	9.83	4.04	0.36	21.37
Age at Time of Sale (days)	4054	5735	1	69929
Heated Area (sq. ft.)	2163	910	275	18789
Housing Density (# of homes within 1 mile)	73	415	1	2115
Raleigh	0.063	0.242	0	1
Cary	0.064	0.245	0	1
Apex	0.042	0.201	0	1
Holly Springs	0.039	0.193	0	1
Garner	0.070	0.255	0	1
Fuquay Varina	0.059	0.235	0	1
Morrisville	0.051	0.219	0	1
Property Spacing	0.268	0.393	0.016	9.920
Net Income (in 1,000s)	47.36	19.35	-25.76	142.82

NOTE: Net Income is derived from housing price and the mean income of the census block group. A negative value reflects a high priced home in a low income neighborhood but does not necessarily reflect a negative income of that individual.

Table 4.3
Parameter Estimates: Conditional Logit Model

Log Likelihood	-98919		
CAIC	198189		
crAIC	197905		
Basement %	0.024 (0.031)	**Mean Commute (minutes)**	-0.004* (0.002*)
City Water	0.015 (0.032)	**Distance to Center of Research Triangle Park (miles)**	0.013** (0.002**)
City Sewer	-0.008 (0.036)	**Distance to Raleigh Central Business District (miles)**	0.044** (0.002**)
Natural Gas	0.015 (0.026)	**Age at Time of Sale (days/1000)**	-0.011 (0.016)
Assessed Quality	1.932** (0.132**)	**Heated Area (sq. ft./1000)**	0.226** (0.024**)
Lot Size (Acres)	2.419* (1.246*)	**Housing Density (# of homes within 1 mile/100)**	-0.003* (0.002*)
Multiple Fireplaces (as opposed to 1 fireplace)	0.037 (0.052)	**Raleigh (as opposed to unincorporated)**	0.24** (0.033**)
No Fireplace (as opposed to 1 fireplace)	-0.004 (0.02)	**Cary (as opposed to unincorporated)**	-0.037 (0.029)
Two Story (as opposed to one story)	0.017 (0.031)	**Apex (as opposed to unincorporated)**	-0.054** (0.025**)
Three or More Story (as opposed to one story)	-0.014 (0.076)	**Holly Springs (as opposed to unincorporated)**	-0.066** (0.031**)
Townhouse (as opposed to conventional)	-0.071** (0.018**)	**Garner (as opposed to unincorporated)**	-0.017 (0.024)
Ranch (as opposed to conventional)	-0.094** (0.045**)	**Fuquay Varina (as opposed to unincorporated)**	0.013 (0.028)
Split Level (as opposed to conventional)	-0.132** (0.041**)	**Morrisville (as opposed to unincorporated)**	0.133** (0.029**)
Two Full Baths (as opposed to one bath)	0.009 (0.033)	**Property Spacing (Acres)**	-2.381* (1.247*)
Three or More Full Baths (as opposed to one bath)	0.063* (0.038*)	**Net Income ($/10,000)**	0.032** (0.003**)

NOTE: ** denotes significant at the 95% level; *denotes significant at the 90% level

Table 4.4
Information Criteria Statistics

	Number of Classes			
	1	**2**	**3**	**4**
LL	-98922	-98398	-98052	-97968
CAIC	198189	197475	**197118**	197284
CrAIC	197905	196925	196318	**196259**

NOTE: Optimal # of Classes outlined and in **bold**

Table 4.5
Parameter Estimates: Three Class Model

Log Likelihood	-98919		
CAIC	198189		
crAIC	197905		
Class Share	0.405	0.280	0.315
Basement %	0.016	0.318**	-0.242**
	(0.039)	(0.035**)	(0.044**)
City Water	-0.015	-0.180**	0.219**
	(0.039)	(0.086**)	(0.003**)
City Sewer	-0.051	0.461	-0.195**
	(0.061)	(0.037)	(0.002**)
Natural Gas	0.069	-0.019	-0.145**
	(0.046)	(0.068)	(0.006**)
Assessed Quality	1.509**	3.106**	2.089**
	(0.555**)	(0.157**)	(0.025**)
Lot Size (Acres)	4.669	5.622	-0.625**
	(33.166)	(3.906)	(0.048**)
Multiple Fireplaces (as opposed to 1 fireplace)	-0.030	0.032	0.101**
	(0.102)	(0.070)	(0.002**)
No Fireplace (as opposed to 1 fireplace)	-0.014	-0.023	0.050
	(0.035)	(0.036)	(0.069)
Two Story (as opposed to one story)	0.046	0.049	-0.041
	(0.069)	(0.057)	(0.048)
Three or More Story (as opposed to one story)	-0.109	1.031*	-0.719**
	(0.224)	(0.594*)	(0.028**)
Townhouse (as opposed to conventional)	-0.106**	-0.066	-0.095*
	(0.038**)	(0.067)	(0.057*)
Ranch (as opposed to conventional)	-0.066	0.071	-0.137**
	(0.135)	(0.200)	(0.056**)
Split Level (as opposed to conventional)	-0.057	-0.400	0.034
	(0.071)	(0.085)	(0.046)
Two Full Baths (as opposed to one bath)	0.122	-0.228**	-0.084
	(1.121)	(0.043**)	(0.151)
Three or More Full Baths (as opposed to one bath)	0.180**	-0.173**	-0.043
	(0.065**)	(0.078**)	(2.446)
Mean Commute (minutes)	-0.025	0.023*	0.034**
	(0.019)	(0.012*)	(0.017**)
Distance to Center of Research Triangle Park (miles)	-0.025**	-0.035**	0.076**
	(0.003**)	(0.005**)	(0.012**)
Distance to Raleigh Central Business District (miles)	-0.040**	0.237**	-0.007**
	(0.002**)	(0.010**)	(0.000**)
Age at Time of Sale	0.149	0.247	0.285
	(0.214)	(0.363)	(0.523)

Table 4.5 Continued

Heated Area (sq. ft./1000)	0.168**	0.205**	0.372**
	(0.025**)	(0.018**)	(0.043**)
Housing Density (# of homes within 1 mile/100)	0.007	-0.022	0.008
	(0.005)	(0.019)	(0.006)
Raleigh (as opposed to unincorporated)	0.000	1.863*	-4.613*
	(0.000)	(0.258*)	(0.660*)
Cary (as opposed to unincorporated)	-3.173	-2.694	1.921
	(3.092)	(3.313)	(2.069)
Apex (as opposed to unincorporated)	-11.227**	-1.585**	2.079**
	(3.182**)	(0.521**)	(0.636**)
Holly Springs (as opposed to unincorporated)	1.221**	-13.165**	-1.592**
	(0.406**)	(6.552**)	(0.804**)
Garner (as opposed to unincorporated)	-4.211	3.053	-3.565
	(8.056)	(5.125)	(6.364)
Fuquay Varina (as opposed to unincorporated)	1.165	0.003	-3.604
	(2.492)	(0.005)	(8.178)
Morrisville (as opposed to unincorporated)	-8.721**	1.052**	1.238**
	(1.333**)	(0.200**)	(0.191**)
Property Spacing	-4.548*	-5.687**	0.696*
	(2.429*)	(2.400**)	(0.393*)
Net Income (fixed across classes)	0.043**	-	-
	(0.005**)	-	-

NOTE: ** denotes significance at the 95% level; * denotes significance at the 90% level; some parameters truncated for display purposes.

Table 4.6
Welfare Estimates

	Conditional Logit	**Latent Class**			
		1	**2**	**3**	**Composite**
Share		40.5%	28.0%	31.5%	
WTP	1445.1	2933.5	44.3	-3026.5	247.1
	(401.9)	(496.1)	(584.8)	(1066.1)	(549.1)

NOTE: Welfare Scenario: 10% decrease in commute time and a 10% increase in neighborhood density. WTP reflects annual change in welfare in $. Standard deviation of WTP estimates in parentheses.

5. Conclusion

This dissertation has addressed several opportunities for research using nonmarket valuation techniques. Quantitative results are useful, not just in academic exercises, but also in policy and practice. The three research projects presented here can at the minima, serve to demonstrate the practical applicability of quantitative methods, but also seek to expand the current state of knowledge in the field.

The hedonic method is used in chapter two to examine the welfare impact of a localized and unexpected environmental catastrophe. An explosion at a hazardous waste storage facility caused the evacuation of an entire town. Chemical analyses were used by state and federal agencies to measure the impact of the fire. Although useful in understanding the health impacts from certain levels of contamination, this type of research neglects the consumer response to the fire. By studying the implicit cost of contamination through the housing market, a welfare result can be obtained which approximates the economic impact on the residents of the affected community.

A strict hedonic estimation shows viability of the method and of the data set. Estimation of a wide set of structural and neighborhood characteristic variables produces coefficients that can be considered reasonable when comparable to those used in other hedonic studies. This set of variables is then re-estimated with a variable that defines the contamination zone.

Although the physical contamination of the fire is known, the spatial and temporal degree of the economic impact is not known *a priori*. For this reason, a wide number of specifications (space and time interactions) are tested. Every possible combination of five different contamination zone definitions along with four different time periods are run in a hedonic model. This set of 20 parameters provides additional information to the researcher when determining the impact of the fire. Although the small number of observations in the contamination zone leads to few statistically significant results, the trend in the parameters over time indicates that there was a mild short-term impact in the housing market, which subsequently diminished. The spatial range of this impact is a two-mile radius from the site, and up to six months after the fire, and is equivalent to a temporary 2.8% (mean: $6,483; standard deviation: $3,790) price reduction.

The second chapter addresses a limitation in discrete choice modeling. This technique is useful for representing a consumer's choice out of a set of alternatives. Oftentimes, this choice set can be very large and can cause difficulties in empirical estimation. The primary vehicle for discrete choice modeling is the conditional logit kernel, which assumes a type 1 iid distribution of the errors. This gives rise to the IIA assumption, which states that the relative probability of choosing any two alternatives is independent of any third alternative. Although oftentimes inappropriate when modeling behavioral decisions, the IIA assumption allows random sampling of alternatives to produce consistent estimates. Past research has proven the viability of random sampling in a conditional logit model, but there is no proof for doing so when IIA doesn't hold.

Methods to relax the IIA assumption and more flexibly model consumer choice have been developed. One of these is the mixed logit model. By assuming that consumer preferences fit a distribution, the mixed logit model allows unobserved preference heterogeneity to enter into the estimation. Unfortunately, by relaxing IIA, random sampling of alternatives may not produce consistent estimates.

Chapter three tests an alternative estimation method which reintroduces IIA into a mixture model in a limited fashion, thus allowing for random sampling. By specifying parameters as a discrete distribution and recursively estimating the resulting latent class model via the EM algorithm, we empirically show that random sampling can produce consistent estimates.

Using a dataset of recreational fisherman in Wisconsin, sampling is tested first in a conditional logit model, and then compared with the results of a latent class model. 200 random samples are tested at sample sizes of 50%, 25%, 12.5%, 5%, 2%, and 1% for the conditional logit model. The impact of sampling on five different welfare scenarios show that random sampling can produce consistent point estimates down to a 1% level with a 99% reduction in estimation time, and that the 5% level produces point estimates with CIs 76% wider than the full model at a 98% reduction in estimation time. The 12.5% level produces

WTP estimate CIs that are only 33% wider at a 90% reduction in estimation time. The choice of optimal sample size (and tradeoff between consistency and computation time) is left to the practitioner to determine.

Similar analysis was performed on the latent class model. Due to the fact that the mixed logit likelihood function may be globally non-concave (unlike the conditional likelihood), 10 different starting values are used in estimation. 10 different random samples are tested for the same set of sample sizes and welfare scenarios. The results provide two important sets of information for the practitioner. First, when choosing the optimal model, it is important to use the minimal number of classes that either the crAIC or CAIC model suggest. Over specification and convergence issues arise when attempting to estimate too many classes on limited data. Secondly, the latent class model produces consistent (a 51% increase in CI range) and useful estimates down to a 12.5% sample with a 75% reduction in estimation time. Useful estimates are also obtainable with a 5% sample, however results at the 2% and 1% levels are unreliable in our scenario. As in the conditional logit model, the needs of the researcher best determine the appropriate sample size to use.

The fourth chapter uses the latent class estimation method and applies it to property transaction data. Although primarily an exploratory piece, this paper suggests that there may be promise in using mixture models in a housing context. The data construction method uses time, space, and price windows to create choice sets for each individual property transaction. Randomly sampling from these choice sets and using a nonparametric latent class model estimated via the EM algorithm produces results that are an improvement in fit over conditional logit estimation on the same data.

The methodology is used to determine the welfare implications of smart growth. Smart growth is a development policy through which incentives and public funds are used to transform existing urban landscapes. The wide variety of housing options available in the dataset used gives indication to the existence of preference heterogeneity. Accounting for this heterogeneity could be important for policy makers trying to determine where to best allocate limited funds.

The results indicate three different classes of residential consumer, each with an independent set of preference parameters. Welfare analysis shows that, for the hypothetical smart growth policy proposal, one class of consumer would benefit, while the other two would not. The conditional logit model suggests a uniform positive benefit for the scenario. The results indicate that additional research may be viable. Future research would include demographic variables in the latent class probability and would also perform a sensitivity analysis on the estimation results.

All three research chapters demonstrate applicability of nonmarket valuation methods, but also provide opportunities for future research. There are many unstudied accidents (on the scope of the Apex fire) that have occurred over the last ten years. Similar estimation and a meta-analysis of the results matrix could provide richer information as to the effects of unexpected chemical contamination. Future research on the latent class model could involve testing similar sampling methodology on different kernels or expansion into a nested latent model. New survey data (along with Monte Carlo data) would provide additional information to practitioners seeking to implement the model. And additional exploration into latent class housing estimation is an obvious opportunity stemming from the smart growth chapter.

APPENDIX

Matlab Estimation Code

EM.m

```
clear; clc;

%
% Associated m.files: Emfunc1.m; Emfunc2.m; EMfunc3.m
% load data
% data must be in form
%   Y ~ choice outcome (NxJ)
%   X ~ choice specific explanatory variables (NxJK)
%     ~ grouped by explanatory variables
%       eg. For choice options A,B,C and explanatory variables 1,2
%           X = [X1A,X1B,X1C,X2A,X2B,X2C]
%   Z ~ individual specific explanatory variables (NxL)
%   N ~ number of observations
%   J ~ number of choice alternatives
%   K ~ number of choice specific explanatory variables per choice
%   L ~ number of individual specific explanatory variables
%
%
% This version constructed specifically for use with the wake county
% dataset: 36941 observations, 15 alternatives (including chosen
% alternative), and 30 variables

%%%USER EDIT%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%% DEFINE NUMBER OF CLASSES
for nocs=4:4  %FULL RUN
    C=nocs;
%%% Set Convergence tolerance
tol0 = 0.1;
%%% Set Maximum Number of Iterations
maxiter=200;

%%%END USER EDIT%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%


time=clock;
LLLrecord=[];
LLrecord=[];
take=0;
LLold=0;

%specific to wakeco
load wakeco.mat ;
%end specific to wakeco


[N J]=size(Y);
[N R]=size(X);
```

```
%%% Replicate choice set for repeat observations by the same
individual

Yid=zeros(N,J);
for n=1:N
    jj=0;
    for j=1:J
        if Y(n,j)>0
            jj=jj+1;
            Yid(n,j)=jj;
        end
    end
end

Ytemp=sum((Y==0==0),2);

index=(1:N)';
indextemp3=[];

csn=cumsum(Ytemp);
tmp=zeros(1,csn(end));
tmp([1; csn(1:end-1)+1])=1;
idx=cumsum(tmp)';
Ytemp2=Y(idx,:);
Yidtemp2=Yid(idx,:);
Xtemp2=X(idx,:);


for i=1:N
    indextemp2=[];
    indextemp2=[repmat(index(i,:),[Ytemp(i,:),1]),(1:Ytemp(i,:))'];
    indextemp3=[indextemp3;indextemp2];
end

X=Xtemp2;
Xtemp2=[];
Y=Ytemp2;
Ytemp2=[];
Yid=Yidtemp2;
Yidtemp2=[];
index=indextemp3;
indextemp3=[];
indextemp2=[];

[N J]=size(Y);
[N R]=size(X);



for n=1:N
```

```
    for j=1:J
        if Yid(n,j)~=index(n,2)
            Y(n,j)=0;
        end
    end
end




%RND=rand(N,J)+(Y==0==0);

%delete later
RND=load('RND');
RND=RND-(Y==0==0);
%end delete later

RNK=[];
for i=1:N
    [temp,RNKtmp]=sort(-RND(i,:),2);
    ridx(RNKtmp)=1:J;
    RNK=[RNK;ridx];
end

RNK=RNK.*(Y==0);

Ytemp=[];
Xtemp=[];


nz=RNK>(round(ssize*J)-1);

Xtemp2=[];
Ytemp2=[];
for i=1:N
    Ytemp=Y(i,:);
    Xtemp=X(i,:);
    nztemp=nz(i,:);
    nztempx=repmat(nztemp,[1,15]);
    Ytemp(nztemp)=[];
    Xtemp(nztempx)=[];
    Ytemp2=[Ytemp2;Ytemp];
    Xtemp2=[Xtemp2;Xtemp];
end

X=Xtemp2;
Y=Ytemp2;
Xtemp2=[];
Ytemp2=[];

mean(mean(X(:,1:(round(ssize*J)))))
```

```
[N J]=size(Y);
[N R]=size(X);
[N Q]=size(Z);
[Nall Jall]=size(Yall);
[Nall Rall]=size(Xall);



K=R/J;

PK=1;



%%%%Starting share values (equal shares)
S=ones(N,C)/C;



%%%%Starting beta estimates (sample partitioned into C subsamples
%   and logit estimated on each subsample)



Nsub=N/C;

%constrain output from fminunc
options                                                        =
optimset('Display','iter','GradObj','on','derivativecheck','off',
'LargeScale','on');

%estimate starting values via logit
Bout=[];
temp=0;
temp1=0;
for k=1:C
    temp=1+temp1;
    temp1=floor(Nsub+temp1);
    Xsub = X(temp:temp1,:);
    Ysub = Y(temp:temp1,:);
    [Btemp       output       hessian]       =       fminunc(@(B)
EMfunc1(B,Xsub,Ysub,J,R),B,options);
    Bout=[Bout,Btemp];
end
B0=Bout;
B=B0;
```

```
%construct recursive method



tolrecord=[];
check = 0;
iter=0;
while check == 0;
iter=iter+1;

%calculate weights

P=[];
for k=1:C

Xnum=[];
for i=1:J
    Xtemp=exp(X(:,(i):(J):R)*B(:,k));
    Xnum=[Xnum,Xtemp];
end

Xtemp=sum(Xnum,2);
Xden=repmat(Xtemp, [1,J]) ;
P(:,k)=exp(sum(Y.*log(Xnum./Xden),2));
end



hncnum=S.*P;
hncdum=sum(hncnum,2);
hncdum=repmat(hncdum,[1,C]);
hnc=hncnum./hncdum;





% Update shares

scden=sum(sum(hnc));

for i=1:C
    S(:,i)=sum(hnc(:,i))/scden;
end


%update coefficents

B2=[];
```

```
for i=1:C
    B2temp=B(1:29,i);
    B2=[B2;B2temp];
end

B2=[B2;B(30,1)];



[B2temp fval exitflag output grad hessian] = fminunc(@(B2)
EMfunc3(B2,X,Y,J,R,hnc,C),B2,options);


hessCP=hessian;

Bout=[];
k=1;
for i=1:C
    Bout(:,i)=B2temp(k:k+28,:);
    k=k+29;
end

B2tempf=repmat(B2temp(C*29+1,:),[1,C]);
Bout=[Bout;B2tempf];




%calculate LL

B=Bout;

Xnum=[];
for i=1:J
    Xtemp=exp(X(:,(i):(J):R)*B);
    Xnum=[Xnum,Xtemp];
end

[T1 T2]=size(Xnum);

Xden=[];
for i=1:C
Xtemp=sum(Xnum(:,i:(C):T2),2);
Xden=[Xden,Xtemp];
end

Xden=repmat(Xden, [1,J]) ;

L=[];
P=((Xnum./Xden));
for i=1:C
Ltemp=sum(Y.*log(P(:,i:C:T2)),2);
```

```
L=[L,Ltemp];
end

LL=sum(log(sum(S.*exp(L),2)))



tol = abs((LLold-LL));
tolrecord=[tolrecord;max(tol)];
LLold=LL;

if iter>2
if                        all(all(abs(tolrecord(iter,:)-tolrecord(iter-
1,:)))<(tol0)*10)==1;
    PK=-maxiter;
end
end


max_beta_change=max(tol)




if take ==1;
    check=1;
end

if all(all(tol<tol0))==1;
take = 1;
else
    check=0;
end;
if iter==maxiter;
    take=1;
end;
end;



% CALCULATE STANDARD ERRORS
[junk1 junk2 score ] = EMfunc3(B2,X,Y,J,R,hnc,C);

[N J]=size(Y);
[N R]=size(X);

sall=sum(Y==0==0,2);

scoresum=zeros(N,29*C+1);
k=1;
for i=1:N
    j=sall(i,1);
```

```
        scoresum(i,:)=sum(score(k:k+j-1,:),1);
        k=k+j;
    end

    score=scoresum;

    W=score'*score;

    stdPS      = sqrt(diag(inv(W)));



    stdRbstout=[];
    k=1;

    for i=1:C
        stdRbstout(:,i)=stdPS(k:k+28,:);
        k=k+29;
    end

    stdRbsttempf=repmat(stdPS(C*29+1,:),[1,C]);
    stdRbstout=[stdRbstout;stdRbsttempf];
    stdCS=stdRbstout;

    %END SE CALCULATION

    time1=clock-time;

    diary on;

    disp('NEXT')
    disp('number of classes =')
    disp(C)
    disp('number of iterations')
    disp(iter)
    disp('run time = ')
    disp(time1)
    disp('Shares =')
    disp(sum(S)/N)
    disp('Choice Specific Estimates =')
    disp(B)
    disp('Standard Errors (CSE)= ')
    disp(stdCS)
    disp('Common Log Likelihood =')
    disp(LL)

    diary off;

    end
```

EMfunc1.m

```
function [F g] = EMfunc1(B,X,Y,J,R)


Xnum=[];
for i=1:J
    Xtemp=exp(X(:,(i):(J):R)*B);
    Xnum=[Xnum,Xtemp];
end

Xtemp=sum(Xnum,2);
Xden=repmat(Xtemp, [1,J]) ;

P=((Xnum./Xden));
L=Y.*log(P);
F = -sum(sum(L));




temp2=[];
for i=1:J
    temp=(Y(:,i)-(sum(Y,2).*P(:,i)))'*X(:,(i):(J):R);
    temp2=[temp2;temp];
end

g=-(sum(temp2));



end
```

EMfunc2.m

```
function [F g score] = EMfunc2(D,Z,hnc,C)

[T1 T2]= size(Z);
Znum=ones(T1,C);
Znum(:,2:C)=exp(Z*D);
Zden=sum(Znum,2);
Zden=repmat(Zden,[1,C]);
P=Znum./Zden;
F=-sum(sum(hnc.*log(P)));


temp=[];
for i=2:C
temp2=repmat(hnc(:,i)-P(:,i),[1,4]).*Z;
temp=[temp,temp2];
end
score=temp;
```

```
gtemp=-sum(score);

g=zeros(C-1,4);
j=1;
for i=2:C
    k=i-1;
    g(k,:)=gtemp(:,j:j+3);
    j=j+4;
end


g=g';




end
```

EMfunc3.m

```
function [F g score] = EMfunc3(B2,X,Y,J,R,hnc,C)

B2temp=[];
Xnum=[];
k=1;
for j=1:C
    B2temp(:,j)=[B2(k:k+28,1);B2(29*C+1,1)];
    k=k+29;
end

for i=1:J
       Xtemp=exp(X(:,(i):(J):(R))*B2temp);
       Xnum=[Xnum,Xtemp];
end


[T1 T2]=size(Xnum);

Xden=[];
for i=1:C
Xtemp=sum(Xnum(:,i:(C):T2),2);
Xden=[Xden,Xtemp];
end



Xden=repmat(Xden, [1,J]) ;

L=[];
P=(Xnum./Xden);
for i=1:C
```

```
Ltemp=sum(Y.*log(P(:,i:C:T2)),2);
L=[L,Ltemp];
end

F = -sum(sum(L.*hnc,2));

temp3=[];
for k=1:C
temp2=zeros(T1,29);
for i=1:J
    j=(i*C-C+k);
    temp=repmat((Y(:,i)-
(sum(Y,2).*P(:,j))).*hnc(:,k),[1,29]).*X(:,(i):(J):(R-J));
    temp2=temp2+temp;
end
temp3=[temp3,temp2];
end
score=temp3;

temp4=zeros(T1,1);

for k=1:C
temp2=zeros(T1,1);
for i=1:J
    j=(i*C-C+k);
temp=((Y(:,i)-(sum(Y,2).*P(:,j))).*hnc(:,k)).*X(:,(R-J+i));
    temp2=temp2+temp;
end
temp4=temp4+temp2;
end

score =[score,temp4];

g=-sum(score);


end
```