Three Essays on Property Values and Water Quality

By

Jiarui Zhang

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Agricultural and Applied Economics)

at the

UNIVERSITY OF WISCONSIN-MADISON

2022

Date of final oral examination: 05/10/2022

The dissertation is approved by the following members of the Final Oral Committee:
Daniel Phaneuf, Professor, Agricultural and Applied Economics
Xiaodong Du, Associate Professor, Agricultural and Applied Economics
Sarah Johnston, Assistant Professor, Agricultural and Applied Economics
Dominic Parker, Associate Professor, Agricultural and Applied Economics
Christoph Nolte, Assistant Professor, Earth and Environment

Dedicated to my family:
Mom, Dad, Husband, and Son
For their endless love and support.

Acknowledgements

The completion of my dissertation could not have been possible without the help and support from many people. I want to take this opportunity to express my gratitude to everyone who helped me during this process.

First, I want to say special thank you for my advisor, Daniel Phaneuf, for always being supportive and encouraging for both my research and my life. You guided me into this exciting research area of property values and water quality and helped me with any challenges throughout my research. You have always been patiently answering my questions, commenting on my essays, giving me helpful suggestions. Having you as my advisor is the luckiest thing during my PhD.

I also want to express my gratitude to my committee members for thoughtful comments and valuable suggestions. I want to show my sincere thanks to Sarah Johnston, Xiaodong Du, Dominic Parker, and Christoph Nolte. I obtained insightful feedback from every conversation we had during seminars or individual talks. Your useful comments helped my research develop from early ideas to my current work.

In addition, I would like to thank my coauthor, Blake Schaeffer, who provided the comprehensive satellite dataset for my first chapter. The first talk with him at the AERE conference led me to my current research. I enjoyed learning frontier studies on water quality from him. Also, I am grateful for my coauthors of my third chapter. I would like to acknowledge their invaluable assistance and insights. It is great pleasure for me to work with all of you.

I would also like to thank my cohort in the AAE department for their insightful comments and helpful suggestions on my research. I also enjoyed every moment we got together for coffee, beer, or delicious food. Your company and encouragement make my PhD a happy journey.

My special thanks go to my parents Huimin and Hong, who give me endless support and love. You have confidents in me for every choice I made. You are always there for me every time I need help. I also want to thank my mother-in-law Zupei and father-in-law Jun who helped me take care of my son during my PhD. Their selfless help enables me to have more time and energy to pursue my PhD. In addition, I want to thank my friend Wenqi. She has always been there for me through my ups and downs. I am fortunate to have your company along the way.

Finally, I would like to express my deepest gratitude for my husband Chenxiao Guan for his unconditional love and constant support that keep me confident. I feel so lucky to start a family with him during my PhD and stand together through thick and thin. You have always been there, wherever, and whenever that was. I also want to thank my son Ethan Guan, for bringing me the motivation to work hard and overcome any challenges -- to become a mother you can be proud of.

Abstract

United States has a rich source of inland lakes that provide ecosystem services like aesthetic amenity and recreational opportunities for aquatic activities like fishing and swimming. The ability of lakes providing ecological and economic benefits largely depends on water quality. Due to the importance of water quality, past studies have estimated the value of water quality using multiple water quality measurements with different approaches. However, the benefits of water quality improvement remain unclear and underestimated. In my dissertation, I estimate the value of water quality through different perspectives.

We examine how the frequency of cyanobacterial harmful algal blooms (cyanoHABs) generates economic costs through the mechanism of residential property values. We assemble nearly two decades worth of nationwide data on property sales near US inland lakes along as well as satellite-derived measures of the annual frequency of cyanoHABs in over 2,000 large lakes during the years 2008-2011. We combine these data sources to estimate broad scale hedonic property models to recover the marginal willingness to pay to reduce the frequency of cyanobacterial blooms in seven climate regions across the United States. We find heterogeneity in the marginal cost of cyanoHABs, with a 10-percentage point increase in annual occurrence reducing average home values for near-shore properties by 3.5 percent in the Upper Midwest, 3.8 percent in the South, 3.3 percent in the Southeast, and 4.3 percent in the Northeast. We find null or inconclusive results for other regions. We use our estimates to illustrate the household-, lake-, and regional-level impacts of counterfactual changes in the frequency of cyanoHABs.

In the second chapter, I develop a residential sorting model to estimate the amenity and recreational values of water quality. I estimate two types of MWTP for the improvement of water quality associated with both values for lakefront households, as well as the MWTP associated with the recreational value for non-lakefront households. My results suggest that for a 0.1m increase in Secchi depth, the marginal amenity value ranges from $1,442 to $3,195 for lakefront households with annual income from $22,000 to $960,100. I also find that the surface area of lake and the travel distance affect the marginal recreational value of water quality.

In the third chapter, I evaluate the benefits of water quality capitalized in housing markets throughout the US and provide compelling evidence that homeowners place a premium on improved lake water quality. The value of water quality decays with distance from the waterbody. The willingness-to-pay estimate to restore the water quality to pristine level for the contiguous United States is $27 billion. This study helps policymakers to incorporate water quality valuation in environmental decision making.

Contents

Chapter 1

Property values and cyanobacterial algal blooms: evidence from satellite monitoring of inland lakes

With Daniel J. Phaneuf and Blake A. Schaeffer

1.1 Introduction

Inland lakes and reservoirs provide a range of ecosystem services that includes aesthetic amenities, recreation opportunities, and municipal drinking water. Nutrient pollution is among the major threats to these services in the United States, where 40 percent of lakes have excess phosphorus and 35 percent have excess nitrogen (US EPA, 2016). In lakes and reservoirs, excess nutrients can lead to low dissolved oxygen, changes in species composition and abundance, higher turbidity and, most visibly, an increase in the frequency and duration of algal blooms. There are many types of algal blooms, and some are harmful to human, animal, and environmental health. These harmful algal blooms (HABs) can cause dramatic disruptions to aquatic food webs, generate unattractive scum on the water surface and shorelines, affect drinking water treatment, and in some cases generate toxins that can affect human health. In freshwater systems, cyanobacteria are the main type of toxin generating HAB (Graham et al. 2017) and cyanobacterial bloom frequency, magnitude, extent, and duration may change due to a combination of warming temperatures and nutrient pollution (Paerl and Huisman, 2008; Hallegraeff et al. 2021). At the same time, several high visibility bloom events and increased appreciation of their consequences have heightened awareness of cyanobacterial blooms as a water quality problem (Brooks et al. 2016; Rashidi et al. 2021). This in turn has highlighted the importance of understanding the economic costs of HABs (Berdalet et al. 2017).

In this paper we estimate economic damages in property markets from cyanobacterial HABs (cyanoHABs) using nationwide data on home sales near inland lakes combined with satellite-derived measures of cyanoHAB frequency (Clark et al 2017; Coffer et al 2021) for over 600 large lakes in 30 continental US states. We conduct our analysis at a large spatial scale by exploiting two nationwide datasets. For property transactions we use the ZTRAX database from Zillow, which provides records on millions of home sales across the US. For water quality we assembled satellite-derived measures of annual cyanoHAB frequency at >2,000 of the largest lakes across the country, for the years 2008 to 2011. Our main research design uses a split sample approach in which we first estimate how distance to the nearest lake capitalizes into home prices. We identify the cutoff point beyond which distance to a lake does not affect home prices and use this to distinguish properties that receive amenity services from the nearest lake versus properties that are far enough away that amenity services do not capitalize. In our main regressions we then use this distinction to estimate models that identify how the price premium for proximity – the capitalized value of shoreline amenity services – is affected by water quality.

There is a large literature focused on estimating the relationship between property values and water quality in both inland and coastal contexts. For lake- and estuary-focused studies researchers have estimated capitalization effects for both waterfront and near-shore properties (Leggett and Bockstael, 2000; Walsh et al. 2011) using environmental measures such as water clarity (Weng et al. 2020), chlorophyl a (Liu et al. 2017), dissolved oxygen (Netusil et al. 2014), and other indicators such as invasive species (Horsch and Lewis, 2009). This literature is reviewed by Nichols and Crompton (2018) and Guignet et al. (2021) conduct a meta-analysis of findings related to water clarity. Most of the literature focuses on study areas consisting of a small number of counties (Walsh et al. 2017) or a single (Weng et al. 2020) or small group (Wolf and Klaiber, 2017) of waterbodies. In general researchers have found negative associations between property values and water pollution for homes close to lakes, with magnitudes inversely related to distance from the shoreline. Our paper fits in this literature in that we examine the marginal implicit price of changes in the frequency of cyanoHABs for homes near inland lakes across the country.

We conduct separate analyses for each of seven climate regions in the US using the same specification and identification assumptions. Consistent with much of the existing literature, we find that the distance from shore for which a proximity premium exists is relatively small, ranging between 100m and 400m across our seven regions. In our main regressions we find strongly robust evidence that cyanoHABs decrease home values in two of our regions. Specifically, in the Upper Midwest a 10-percentage point increase in the annual rate of cyanobacterial blooms decreases home values by 3.5 percent for properties within 100m of shore. In the South, a similar increase in the bloom frequency decreases prices by 3.8 percent for homes within 300m of shore. In the Northeast and Southeast, we find somewhat robust evidence of a negative relationship between cyanoHABs and home values. A 10-percentage point increase in bloom frequency reduces prices by 4.3 percent in the Northeast and 3.3 percent in the Southeast for homes within 400m of the closest lake. For the remaining regions (Ohio Valley, West, and Northwest) we find little evidence of an association between home values and cyanoHABs.

With this paper we make several contributions. First, with the notable exception of Wolf and Klaiber (2017), there are few property value studies focused explicitly on estimating how cyanobacterial algal blooms in inland lakes affect home prices. We augment the Ohio-based evidence in Wolf and Klaiber with a larger scale application demonstrating widespread but heterogeneous impacts across diverse regions of the country. Second, we show how large-scale analysis of property values and water quality can be conducted by leveraging the ZTRAX data and emerging technologies for monitoring water quality via satellite images. This allows us to move beyond the case study scale to simultaneously examine large and diverse regions of the country using comparable data sources, specifications, and identification assumptions. In this regard, our study complements Moore et al. (2020) and Keiser and Shapiro (2019), who also present large-scale studies of property values and water quality, though with substantially different data environments. This approach implies our results can be used in broad scale integrated assessment models, and we illustrate this potential in the paper. Finally, we explore a novel split sample identification strategy that focuses on a specific value generating mechanism – the effect of water quality on shoreline proximity premia – and for larger lakes exploits satellite-generated continuous spatial variation in water quality outcomes, unencumbered by in situ monitoring networks.

The following section provides background on cyanoHAB impacts and measurement along with a literature discussion. Section 3 describes our data in detail and section 4 presents our main specification and identification assumptions. This is followed by a discussion of main results and robustness checks in section 5. The final two sections provide discussion and conclusions.

1.2 Background

Eutrophication from excess loading of nutrients resulting from human activities is a major threat to freshwater ecosystems in the United States (Smith and Schindler, 2009). The 2012 National Lake Assessment (NLA) indicated that 35 percent of US lakes have excess nitrogen and 40 percent have excess phosphorous (USEPA, 2016). A comparison of the 2007 and 2012 NLAs showed that nutrient pollution worsened between the two assessments. Nutrient pollution can have dramatic impacts on lake ecosystems in the form of reduced water clarity, dissolved oxygen depletion, changes in aquatic species composition and abundance, fish kills, and a general deterioration of aesthetic aspects.

Excess nutrients may also change the temporal frequency, spatial extent, and magnitude of algal blooms (Huisman et al. 2018). There are many types of algae and some of them are potentially harmful to human, animal, and environmental health. Harmful algal blooms (HABs) are environmental events that occur when algal populations achieve sufficiently high density, resulting in negative environmental or health consequences (Smayda, 1997). The most common freshwater HAB is photosynthetic prokaryote cyanobacteria (cyanoHAB), also referred to as blue-green algae (Graham et al. 2017).

CyanoHABs are of concern for several reasons. In addition to aesthetic and ecosystem consequences, cyanoHABs produce several cyanotoxins that impact human and animal health. The most commonly studied of these cyanotoxins is microcystin with numerous variants (USEPA, 2014) that can cause gastrointestinal distress, dermatitis, and liver problems in humans when ingested (Massey et al. 2018). Exposure can come through contaminated drinking water (Falconer and Humpage, 2005), food consumption (Ibelings and Chorus, 2007), or accidental ingestion during recreation (Chorus et al. 2000). The extent of potential effects is large. Graham et al. (2017, Figure 1) report that 43 US states have implicated cyanoHABs in human and animal illness and death. Several large-scale acute events, especially involving cyanobacterial blooms in Lake Erie, have attracted national attention. For example, a 2014 bloom near Toledo led to do-not-use advisories for the city's water system and several subsequent hospitalizations (Wolf and Klaiber, 2017). Local-scale bloom events have led to a number of beach closures and health advisories. For example, in August 2019, 23 states issued closures, warnings, and advisories for 240 waterbodies based on algae or algae-related toxins (USEPA, 2019).

Warming temperatures and increasing nutrient pollution may lead to increased frequency, extent, and magnitude of cyanoHABs (Taranu et al. 2015; Paerl and Huisman, 2008). The 2012 NLA detected microcystin in 39 percent of sampled lakes – an increase of 9.5 percent over the 2007 assessment. For direct measures of cyanobacteria, 8.3 percent more lakes were in the most disturbed conditions in 2012, relative to 2007 (USEPA, 2016). While these summaries are too limited in power to conclude the changes are trends, it is noteworthy that among the many water quality parameters assessed in the National Lake Assessments, nutrients and cyanobacteria were the only two indicators that worsened between 2007 and 2012.

The abundance of cyanoHABs (Coffer et al. 2020) combined with their health and ecological consequences have generated interest in the widespread measurement of cyanoHAB events and their economic costs. Satellite remote sensing is of considerable importance for the former, given the ability to provide spatial and temporal detection and monitoring across wide areas. While satellite remote sensing cannot detect toxins directly (Stumpf et al. 2016), it can identify cyanobacterial blooms and quantify abundance (Kutser, 2009). Satellites measure radiation at the sensor and then the spectral signatures are used to estimate various water quality parameters, such as cyanobacteria (IOCCG, 2018; 2021). Previous studies provide a comprehensive review of past, present, and new satellite sensors available for deriving water quality measures in estuaries and inland waters (Dörnhöfer and Oppelt, 2016, Tyler et al., 2016). Recently the Interstate Research and Technology Council[1], American Water Works Association[2], and World Health Organization[3] have incorporated the use of satellites for monitoring cyanobacteria into their guidance. There are trade-offs with using satellites to detect cyanobacteria, in that monitoring is dependent on the size and shape of lakes and reservoirs. Clark et al. (2017) and then Urquhart and Schaeffer (2020) identified the lakes and reservoirs resolvable by satellites that could then be used to quantify the temporal frequency (Clark et al 2017; Coffer et al. 2021), spatial extent (Urquhart et al 2017), magnitude (Mishra et al 2019), and occurrence (Coffer et al. 2020) of cyanoHABs across more than 2,000 continental US lakes and reservoirs. The algorithm used to detect cyanobacteria has been demonstrated against *in situ* cell measures (Lunetta et al. 2015), chlorophyll-a measures (Seegers et al., 2021), state reported toxins (Mishra et al. 2021), recreational advisories (Schaeffer et al 2018; Whitman et al., in review), and phenological trends of cyanobacteria (Coffer et al 2020).

Economic costs of cyanoHABs in freshwater systems arise through human and animal health threats, diminished recreation opportunities, enhanced water treatment, reduced commercial uses, altered ecosystems, and negative aesthetic impacts. A small literature has begun to measure these costs. For example, Jones (2019) examines how infant health (proxied by birth weight and gestation time) near a Michigan lake is affected by a quasi-random reduction in microcystin, while Wolf et al. (2017), Wolf et al. (2019), and Zhang and Sohgen (2018) consider the recreation related costs of HABs in Lake Erie. Smith et al. (2019) also examine the tourism and recreation cost of algal blooms in Lake Erie. Stroming et al. (2020) measure how satellite-based monitoring of recreation risks can generate health benefits and Schinck et al. (2020) and L'Ecuyer-Sauvageau et al. (2019) use stated preference surveys to measure the

[1] https://hcb-1.itrcweb.org/
[2] https://www.awwa.org/Store/Product-Details/productId/6745
[3] https://www.who.int/publications/m/item/toxic-cyanobacteria-in-water---second-edition

drinking water benefits and ecosystem services benefits, respectively, of reducing cyanobacteria in Quebec waterbodies. Finally, in the paper that is closest to ours, Wolf and Klaiber (2017) estimate the property value impacts of cyanoHABs in four lakes across six Ohio counties.

While the literature on the costs of cyanoHABs is relatively small, there is a large literature on valuing water quality more generally. Many of these studies use property values to estimate the benefits of reducing pollution in freshwater and coastal systems. This literature is reviewed by Nichols and Crompton (2018) and Guignet et al. (2021) provide a meta-analysis of water quality/property value gradients. Studies can be classified based on the extent of impacts considered (waterfront versus waterfront and non-waterfront), types of water pollution analyzed, the scale of analysis, and identification assumptions. Earlier studies primarily focused on waterfront properties (e.g., Leggett and Bockstael, 2000) while more recent research has shown that impacts can extend to non-waterfront properties as well (e.g., Walsh et al. 2011). The most common quality measure used in the literature is water clarity, often measured by Secchi depth (e.g., Weng et al. 2020) and recently proxied using light attention (Walsh et al. 2017). Other indicators examined include, for example, dissolved oxygen (Netusil et al. 2014), invasive species (Horsch and Lewis, 2009; Zhang and Boyle, 2010), chlorophyll a (Liu et al. 2017; Weng et al. 2020), and combinations of pollutants or summary indices (Tuttle and Heintzelman, 2015). Finally, most of the literature consists of case studies examining a single watershed (e.g., Poor et al. 2007) or single or groups of counties with multiple waterbodies (Walsh et al. 2017; Wolf and Klaiber, 2017). Exceptions include Moore et al. (2020), who estimate the effect of water clarity on 113 lakes across the United States, and Keiser and Shapiro (2019), who estimate the impact of Clean Water Act investments on nationwide property values using average home values in census tracts located within 25 miles of rivers.

Identification in water quality hedonics is challenging since natural or quasi-experiments are typically not available. While water quality monitoring data exhibit substantial variation due to seasonality and idiosyncratic conditions during measurement, the underlying ecological conditions in waterbodies change slowly over time. As a result, most studies rely on variation across space and/or variation in the timing of home sales, while controlling for potentially correlated unobservables using various fixed effects specifications. For example, Wolf and Klaiber (2017) attach the average concentration of microcystin readings taken two months prior to the sale date, in order to assign water quality to property transactions near lakes in their study area. In a regression that includes census block group, year of sale, and month of sale fixed effects, they then estimate separate marginal effects for properties adjacent to lakes and those within 600 meters, relative to an effect that decays continuously with distance from the shore. With this design the authors compare price changes for near-shore homes that sold in the same block group and month cells in different years, to price changes for farther-from shore homes that sold in the same block group and month cells in different years. The key identifying assumption in Wolf and Klaiber is that within-year variation in water quality around the time of sale is the salient attribute in the market. Moore et al. (2020) use a cross sectional instrumental variables design whereby not-visible nitrogen and phosphorus concentrations serve as instruments for observable water clarity. The key identifying assumption is that nutrient concentrations are uncorrelated with other unmeasured but salient-to-home-buyer characteristics of lakes.

1.3 Data

1.3.1 Satellite cyanobacteria detection

For our analysis we assembled information from across the United States on lake water quality for the years 2008 to 2011. Satellite data were obtained from the National Aeronautical Space Administration (NASA) Ocean Biology Processing Group (OBPG) with quality assurance flags for the Envisat satellite Medium Resolution Imaging Spectroradiometer (MERIS) sensor (NASA, 2021). Full resolution (300m×300m at nadir, defined as the point directly below the satellite on the Earth surface) images were processed by NASA OBPG using their standard satellite ocean color software package SeaWIFS Data Analysis System (SeaDAS), with an updated Shuttle Radar Topography Mission waterbody data shapefile in Albers Equal Area and modifications by Urquhart and Schaeffer (2020) to correct inaccuracies such as missing lakes. Quality assurance flagging and masking included cloud cover, cloud shadow, and glint and mixed pixels were already applied by the NASA SeaDAS Level 3 processing. Water pixels were extracted using the National Hydrology Database (NHD) polygon dataset for each inland lake (McKay et al. 2012). All NHD features classified as lakes and reservoirs were selected using US Environmental Protection Agency's 2012 National Lakes Assessment (NLA) site evaluation guidelines (Schaeffer et al. 2018). Any water pixel directly adjacent to land was added to the land adjacency flag as described in Urquhart and Schaeffer (2020) to remove potential land adjacency effects such as bottom reflectance. Lakes in the NHD shapefile with a minimum of three satellite water pixels remaining after the land adjacency quality assurance flag was applied were considered resolvable waterbodies. Daily snow and ice data were obtained from the National Snow and Ice Data Center then masked to the US boundary polygon shapefile from Urquhart and Schaeffer (2020).

Cyanobacterial abundance was characterized for each valid satellite pixel using the CI-cyano algorithm initially described by Wynne et al. (2008, 2010) and later updated by Lunetta et al. (2015). Spectral bands centered at 665nm, 681nm, and 709nm are used to assess cyanoHAB and those centered at 620nm, 665nm, and 681nm are used as exclusion criterion to prevent the quantification of non-cyanobacterial blooms. The progression of the CI-cyano algorithm is detailed in Coffer et al. (2020).[4] Here, a pixel is initially classified as cyanoHAB detection if the algorithm returns a CI-cyano index value of at least 0.0001, indicating cyanobacteria in concentrations above the detection limit of the sensor.

We then compute annual cyanobacterial frequencies for each pixel as the proportion of weekly satellite composites exhibiting cyanoHAB presence, relative to the total number of weekly satellite composites that contained a valid measurement (i.e., ones not quality flagged and discarded) for the given pixel, following Clark et al. (2017) and Coffer et al. (2021). For this calculation, assignment of cyanoHAB was thresholded to detections above a CI-cyano value of 0.001, which roughly translates to 100,000 cells per milliliter – the level formerly used by the World Health Organization (WHO) to denote high risk levels (Clark et al. 2017). Pixel estimates primarily characterize the center of the lake, and

[4] See also https://oceancolor.gsfc.nasa.gov/projects/cyan/.

narrow reaches are unobservable using 300m satellite imagery, which results in the loss of smaller lakes as well as more narrow portions of resolvable lakes. These satellite data excluded measures along the land-water interface and a one-pixel buffer from the shoreline, an area where blooms may accumulate (Gons et al., 2005). Satellite remote sensing is impeded by glint, cloud cover, ice cover, and smoke so there are days or weeks when detection is not possible. These weeks are not counted in the denominator for measuring frequency.

The final dataset for merging with the property value data contains the monthly frequency of cyanoHAB occurrence (the proportion of weeks in the month with detection) above a CI-cyano value of 0.001 in every pixel in our resolvable lakes for the four years between 2008 and 2011. Table 1.1 shows summary statistics for the 2,370 lakes broken out by regions of the country that we include in our property value analysis (Upper Midwest, Northeast, Northwest, Ohio Valley, South, Southeast, and West) and an aggregate category of lakes in regions we do not use. The labeled regions largely correspond to US climate regions as defined the National Centers for Environmental Information (Karl and Koss, 1984) and so we refer below to these groupings as climate regions. CyanoHAB frequency in the table is shown as the average proportion of occurrences in all the pixels in the region over all four years of data, converted to percentages for easier interpretation. There is clear heterogeneity across regions in lake attributes. Lakes are larger on average in the Southeast and West and the frequency of cyanoHABs is lowest in the West and Southeast.

1.3.2 Property value data

We draw our transactions data from Zillow's ZTRAX[5] database, which records millions of spatially explicit property transactions in the US. The database has two components. The first component consists of assessment data that contains property attributes, including the coordinates for geolocation, whereas the second component consists of transactions data that logs sale prices and dates. A home with

[5] Zillow Transaction and Assessment Dataset. Information on the data can be found at http://www.zillow.com/ztrax.

repeated sales may appear multiple times in the transaction dataset but only once in the assessment dataset. We merge the two data files using a unique property identifier and non-arm length sales records for single-family residential homes. During the merge we also dropped multi-property transactions. This provides a pooled cross-sectional dataset that we merge with the lakes data.

To construct our analysis dataset, we pulled transactions for homes that sold between 2000 and 2017 within 10km of lakes resolved in the MERIS dataset. Ten kilometers was selected to provide both lakefront and non-lakefront properties and based on existing studies, which show that distance to shore capitalization typically goes to zero beyond a few thousand meters. Each transaction was assigned a nearest lake and the distance to the nearest shoreline was calculated using the NHD spatial dataset (McKay et al. 2012). Lakes that were not resolved by the MERIS dataset were not considered when selecting the transaction data and linking properties to their nearest lake. We then examined the number of transactions attached to each individual lake and set the cutoff for inclusion in the analysis at a minimum of 100 transactions during the four years for which we observe water quality data. We maintain transactions even when property attributes are missing, since incomplete attribute records are common in the ZTRAX data (see Table 1.8). This resulted in a study area consisting of 633 lakes spread over 30 states, divided into seven climate regions. These regions and the number of lakes in each state included in our analysis are shown in Figure 1.2.

After linking transactions to their nearest lakes, we assigned each sale a water quality value using the pixel-level cyanoHAB frequency data. To take advantage of the spatial variation in water quality within lakes we implement an assignment algorithm that allows different neighborhoods on the same lake to experience different water quality levels. We define neighborhoods using census block groups. We then compute the frequency of cyanoHAB occurrence for each 300m × 300m pixel during each year of our water quality data. To implement our assignment, we compute the distance from the centroid of each block group within 10km of a lake to the centroid of each pixel in the nearest lake and compute a distance-weighted average of cyanoHAB frequency. Transactions for lakefront and non-lakefront property are then assigned water quality based on their census block group and year of sale.[6]

More formally, our assignment algorithm is as follows. Suppose there are N_l resolved pixels in lake *l*. Let d_{njl} denote the distance from the center of block group *j* to the center of pixel *n* in lake *l*. The vector of these distances for block group *j* is

$$\left(d_{1jl},\ldots,d_{N_l jl}\right). \tag{1.1}$$

Furthermore, let $cyanoHAB_{nlt}$ denote the cyanoHAB frequency in pixel *n* of lake *l* at year *t*. Then, the weighted average cyanoHAB frequency for a transaction in block group *j* during year *t* is:

$$cyanoHAB_{jt} = \sum_{n=1}^{N_l} \frac{\frac{1}{d_{njl}}}{\left(\frac{1}{d_{1jl}}+\ldots+\frac{1}{d_{N_l jl}}\right)} cyanoHAB_{nlt}. \tag{1.2}$$

In areas with closely clustered lakes, the same block group may be in the same 10km buffer of multiple lakes. For a property in the overlapping area, it was assigned the block group cyanoHAB frequency for its nearest lake.

After merging the water quality and property value data files, our primary analysis dataset consists of transactions occurring between 2008 and 2011 within 10km of the 633 lakes that have at least 100 observed sales during the four years of our water quality data. We dropped homes that sold for less than $10,000 or more than the 99.9th percentile of the price distribution in our data but retained transactions with missing attributes to arrive at our analysis sample. Table 1.2 shows summary statistics at the transaction level, broken out by climate regions. Differences in mean lake characteristics relative to Table 1.1 arise due to different units of observation (lakes with more transactions will be weighted more heavily in the averages in Table 1.2) and differences in the sets of lakes represented in the data.

As discussed in detail below, our analysis uses within-lake spatial and temporal variation in

[6] We use distance weighting of all pixels, rather than a cutoff threshold that excludes more distance pixels, because satellite navigation may not be accurate enough for assignment based on single or small groups of pixels. Indeed, best practice involves using a box of pixels as the unit of analysis for generating sub-lake statistics (Bailey and Werdell, 2006; Patt, 2002).

cyanoHAB frequency to identify an amenity/property value gradient. Figures 1.3 and 1.4 show this variation broken out by climate regions. The histograms in Figure 1.3 use a lake-year as the unit of observation. For a lake-year combination we compute the mean and standard deviation of cyanoHAB frequency assigned to block groups in the lake-year and construct the ratio of standard deviation to mean. A higher value for this ratio indicates a larger amount of within lake-year spatial variation in cyanoHAB frequency. The tall bar near zero in the histograms shows that many lake-years have little spatial variation – these likely represent smaller lakes with fewer pixel arrays – while the long right tale shows that other lake-years provide potentially useful spatial variation in cyanoHAB frequency. The histograms in Figure 1.4 display temporal variation. Here the unit of observation is a census block group. For each census block group, we compute the mean and standard deviation of cyanoHAB frequency across the four years of our water quality data. The figure shows the distribution of the standard deviation to mean ratio for census block groups in each climate region. A higher value for this ratio indicates more over-time variation in cyanoHAB frequency. The distribution of ratios is concentrated between 0 and 1 in most regions with a large proportion of the mass in the middle of the range. We conclude from this that useable year over year variation in cyanoHAB frequency is present in our water quality data.

1.4 Model specification

Past literature (e.g., Walsh et al. 2011) has shown that there is a proximity premium for properties close to lake shorelines and that this premium goes to zero beyond a threshold. Our main empirical objective is to measure how this proximity premium is affected by cyanobacterial blooms in the lake. To this end we first estimate the threshold for the proximity premium for each region in our study area. Specifically, we examine models of the following type separately for each of our seven climate regions:

$$\ln P_{it} = \sum_{d=1}^{99} \gamma_d D_{it}^{d} + \theta_j + \tau_{ts} + \xi_m + \varepsilon_{it}. \tag{1.3}$$

In equation (1.3), P_{it} is the nominal sale price of home i sold during year t and e_{it} is the error term. We control for fine scale spatial and temporal variation in home prices by including census block group (q_j), sale year-by-state (t_{ts}), and month of sale (x_m) fixed effects.[7] Our primary variables of interest are the D_{it}^{d} indicators that record the property's distance-from-lake position using 99 discrete distance bins. The bins are defined as one-hundred-meter increments out to 9900m – i.e. (0, 100m), (100, 200m), (200m, 300m), …, (9800m, 9900m) – with homes 10km from shore as the left-out category. This specification allows the proximity premium to vary non-parametrically across the distance bins and, by estimating models separately for the different climate regions, across the full study area. We expect g_d to be positive for bins nearest the shore, taper off for bins farther away, and finally be non-positive for bins beyond a threshold.

We estimate equation (1.3) using the years of property value data for which we do not have cyanoHAB data. That is, we estimate the proximity premium parameters $g_1,...,g_{99}$ using sales from the years 2000-2007 and 2012-2017 for homes within 10km of the lakes that are summarized in Table 1.2. This split sample approach allows us to define a price premium threshold for each climate region in our study area based on estimates of g_d. Specifically, for our main research design we define a proximity premium threshold for each climate region as the upper interval bound of the furthest from shore distance bin that is positive and statistically significant. Given this threshold we define properties that are within the proximity premium threshold as 'treated' homes that receive shoreline amenity services, and homes outside of the threshold as 'control' homes that do not receive shoreline amenity services.

To estimate the effect of water quality on the proximity premium we use the following specification for our main analysis:

$$\ln P_{it} = \alpha T_{it} + \beta T_{it} \times cyanoHAB_{jlt} + \theta_j + \tau_{ts} + \xi_m + \varepsilon_{it}, \qquad (1.4)$$

where the fixed effects notation follows from equation (1.3). We rely on spatial fixed effects (census

[7] In the log-linear specification the time fixed effects serve both to deflate nominal prices to a baseline year (the left-out category) and to capture real time trends in prices. Our year-by-state fixed effects therefore capture spatially differentiated inflation and real price trends. With these we avoid taking a stand on a specific geographic deflator and minimize the risk that an inaccurate assumption is correlated across space or time with changes in water quality levels.

block groups) to control for local public goods such as nearby school quality, distance to urban centers, neighborhood demographics, and other spatially varying determinants of home prices. The spatial resolution of our fixed effects is finer than other candidate choices (lake, school district, census tract) and generally provides multiple spatial units per lake.[8] Our state-by-year fixed effects allow macroeconomic effects on home prices to vary non-parametrically at the state level. Finally, the month of sale fixed effects control for seasonality in real estate markets. This may be important given that cyanoHAB frequency can only be assessed during ice-free months and there may be systematic intra-year time variation in occurrence. We do not include property attributes in our primary specification, due to the large number of missing values in the ZTRAX data. For example, Appendix Table 1.8 shows that the availability of lot size, total rooms, and total bedrooms is uneven across our regions and that applying a consistent triage rule – particularly for structure size – would eliminate a large amount of information. We explore the consequences of this omission in robustness checks described below and discuss the identification assumptions needed for consistency when attributes are missing in the Appendix.

The new variables in (4) are as follows. The indicator T_{it}=1 denotes that the home is 'treated' – i.e., property *i* sold during year *t* is within the proximity premium zone for its climate region. Its parameter *a* measures the average price difference for treated homes relative to untreated homes; it captures the price premium for near-shore homes due both to location and any systematic differences in the property attributes of near- versus off-shore homes. The variable $cyanoHAB_{jlt}$ is given in equation (1.2); recall that it is the percentage of occurrences of cyanbacterial blooms in the nearest lake *l*, assigned to property *i* based on its census block group and year of sale. Our primary interest is the coefficient *b*, which measures how the price of a treated home is affected by a one percentage point change in the occurrence of cyanoHABs.

Using equation (1.4), we estimate *b* by comparing year over year changes in average prices for treated versus control homes, within the same census block group. Specifically, including block group

[8] For our main analysis samples, the mean (median) number of census block groups attached to specific lakes varies from 24 (11) in the Upper Midwest to 59 (25) in the South.

fixed effects in the specification implies we are relying on the temporal variation in *cyanoHAB* shown in Figure 1.4 to identify *b*. Consistent estimation requires that unobserved time varying shocks to home prices within a census block group are uncorrelated with changes over time in the cyanoHAB frequency assigned to census blocks. Our inability to include attributes in our main specification specifically implies that time varying changes in (unobserved) attributes among homes selling in different time periods in the same census block group need to be uncorrelated with changes in cyanoHAB frequency over time. In the Appendix we provide additional detail on this identification assumption and arguments for its plausibility in our context.

1.5 Results

1.5.1 Main Results

Table 1.3 contains selected estimation results for equation (1.3), broken out across our study regions. The samples include sales within 10km of the lakes subsequently used in our primary analysis, for the years from 2000 to 2008 and 2012 to 2017 (the years for which we do not have water quality data). The specifications include dummy variables for each 100m increment in distance from shore out to 10km, though we only display estimates out 1000m. All parameters beyond this threshold are statistically insignificant from the left out category of 10km. The use of census block group fixed effects implies that identification is based on comparing home sale prices in different distance bins, within the same census block group.

We find heterogeneity across climate regions in the spatial extent of the proximity premium zone. The subset of positive and significant parameter estimates in Table 1.3 suggests proximity premium zones out to 100m in the Upper Midwest and Ohio Valley and out to 400m in the Northeast and Southeast. The remaining three regions fall in between, with lakeshore proximity capitalizing out to 200m in the Northwest and West and out to 300m in the South. As expected, the magnitudes of the positive and significant coefficients decrease with distance from the shore, illustrating the inverse distance gradient observed in prior studies.

We use the results from these regressions to assign ‘treatment’ and ‘control’ status to homes in our primary sample of sales occurring during 2008-2011 (the years for which we have cyanoHAB data). Homes within the proximity premium are considered treated by shoreline amenity services while homes beyond the region-specific threshold serve as controls that do not receive proximity-based shoreline amenity services. We believe this split sample approach to accommodating the distinctiveness of near-shore properties is consistent with existing empirical literature identifying the systematic difference, but preferable to jointly estimating treatment status and the cyanobacteria gradient from the same set of data.[9] In addition, the split sample approach allows us to exploit the many years of property sales data for which we do not observe water quality outcomes.

Table 1.4 summarizes findings from each climate region for our baseline model given in equation (1.4). The top two rows of the table show estimates for the main parameters of interest and the lower rows repeat the sample size and lake counts from Table 1.2, along with the treatment distance cutoff and the proportion of treated observations in each region. Our estimation samples omit sales that occur within 100m of our treatment threshold to avoid false assignment of treated and control status for these buffer homes.[10] Our primary interest is the parameter b. The log-linear specification implies that a one percentage point increase in the frequency of cyanoHAB occurrence changes treated home values by $100\times b$ percent. More generally, b measures how the price premium for having access to shoreline amenities – as defined by a distance threshold – is affected by the frequency of cyanobacterial algal blooms in the lake. If cyanoHABs are a disamenity we expect $b<0$.

Across our seven climate regions we find statistically and economically significant evidence that the frequency of cyanoHAB occurrence negatively affects home prices in four cases. Point estimates suggest that a 10-percentage point increase in cyanoHAB occurrence decreases near-shore home values

[9] We examine specifications that jointly estimate proximity premiums and cyanoHAB gradients in the robustness checks below.

[10] For example, the threshold for treatment is 200m in the Northwest and so we exclude homes sold between 200m and 300m of the nearest lake. This accounts for the slight difference in observations between Tables 1.2 and 1.4.

3.5 percent in the Upper Midwest, 4.3 percent in the Northeast, 3.8 percent in the South, and 3.3 percent in the Southeast. Point estimates for the West and Northwest are small and statistically insignificant. It is noteworthy that these two regions have comparatively low average percentage of cyanobacterial bloom occurrence (15.87 and 16.06 percent, respectively – see Table 1.2) and so low salience may explain the null results. Finally, for the Ohio Valley, we have a counterintuitive result: a higher frequency of cyanoHAB occurrence is positively associated with home values and statistically significant at the ten percent level. We investigate this unexpected result and the robustness of our findings in the following subsection.

1.5.2 Robustness

To probe the validity of our main results we consider two categories of robustness checks. The first maintains the primary split sample research design while examining how results change under a variety of alternative assumptions related to linking environmental quality to transactions, variations on how control variables enter the specification, different sample restrictions, and different fixed effects strategies. The second category of robustness examines a model that simultaneously estimates the distance from shore parameters and the cyanoHAB effects.

Table 1.5 collects results from several examples of the first category of robustness checks. The top row repeats the main findings from Table 1.4, the remaining rows represent different model variations, and the columns show estimates for the coefficient of interest – the interaction between ‘treated’ status and cyanoHAB frequency – for each region. Models A1-A3 examine robustness to alternative strategies for assigning water quality to transactions. In property value regressions that do not rely on quasi-experimental variation it may be that assignment decisions on environmental quality affect results. Here, the baseline specification uses variation in cyanoHAB frequency at the census block group level, for the calendar year of sale. Results in all regions are robust to alternative decisions, including using a simple lake/year average (no within-lake spatial variation; model A1); using water quality data for the twelve months prior to sale (A2)[11]; and using data from the months during the year of sale that blooms are most common (A3).

In models B1-B3 we test our identification assumptions by adding additional controls. Model B1 uses a more traditional differences in differences-type specification whereby we add a term to equation (1.4) and estimate

$$\ln P_{it} = \alpha T_{it} + \beta_0 cyanoHAB_{jlt} + \beta T_{it} \times cyanoHAB_{jlt} + \theta_j + \tau_{ts} + \xi_m + \varepsilon_{it}. \tag{1.5}$$

That is, we also include the overall average effect of cyanoHAB frequency in the specification. Table 1.5 shows that the estimates of *b* when the additional term is included are largely unchanged from the baseline model.[12] Models B2 and B3 use a generalization of equation (1.4) of the form

$$\ln P_{it} = \sum\nolimits_{d=1}^{D} \alpha_d D_{id} T_{it} + \beta T_{it} \times HAB_{jlt} + \theta_j + \tau_{ts} + \xi_m + \varepsilon_{it}, \tag{1.6}$$

where D_{id} indicates either the year of sale (B2) or one of three size categories for the nearest lake (B3). These specifications allow the proximity premium to nonparametrically vary across years and lake size, respectively. Our results for the Upper Midwest and South regions are fully robust to these changes, while findings in other regions are generally similar, with some changes in statistical significance. Notably, the Northeast region is no longer economically or statically significant for model B3, which is relevant because it partially relaxes one of the identification assumptions described in the Appendix.

In models C1-C4 we examine robustness to the exclusion of property attributes. Among the omitted attributes, lot size has the best coverage, and so we first use it to investigate robustness to attribute omission. Specifically, models C1 and C2 limit the sample to only include observations with

[11] In this specification we assign cyanoHAB based on the average of the 12 months prior to the date of sale. This means we are not able to use sales data from 2008 – lacking cyanoHAB data for 2007, we are not able to assign water quality to these transactions. This reduction in sample size and general robustness of year of sale versus 12 months prior to sale assignment strategy led us to favor the former as our main result.

[12] Estimates of b_0 are: -0.0005*** (UM), -0.0015*** (NE), 0.0006* (NW), -0.0009*** (OV), 0.0003 (S), 0.0019*** (SE), 0.0020*** (W). We caution against interpreting these as valid estimates of the average effect of cyanoHAB on home prices outside of the proximity threshold, given the strong identification assumption that would be needed for consistent estimation. For example, properties out to 10km away from the nearest lake in our sample may be in closer proximity to other lakes, which confounds the assignment of a single lake quality effect for homes comparatively far from our sample lakes.

valid lot size available. For this sample, model C1 includes lot size as a control and model C2 omits lot size as a control. Within a region the estimates are similar for the two specifications and qualitatively match our main results. Models C3 and C4 limit the sample to observations with lot size, total rooms, and total bedrooms available. Once again the within region estimates for attributes included (C3) and omitted (C4) specifications are similar, with notable departure from our main results in regions with large sample reductions. Based on these analyses we conclude that omitting attributes does not fundamentally change the estimates for a given sample but eliminating transactions so as to include attributes may result in selected samples and unreliable results.

Model C5 is our final restricted sample robustness check. Here we present results that only use two years of data (2010 and 2011) to eliminate concerns about using housing market data during the financial crisis. Our findings are robust to this sample restriction.

Finally, in Table 1.9 we present estimates from a series of specifications that use alternative fixed effect configurations with our main samples and research design. Model D1 includes year-by-lake (rather than year-by-state) fixed effects without additional spatial controls. While estimates are qualitatively similar in some regions (UM, S), differences and anomalies in other regions suggest that more spatially resolute controls are needed for consistent estimation. Models D2-D4 incrementally include more resolute temporal and spatial fixed effects, with D3 and D4 including both the block group fixed effects from equation (1.4) *and* year-by-lake or year-by-census tract time controls, respectively. Our main findings are strongly robust to these additional fixed effects.

This discussion shows that, conditional on our research design, our findings are largely robust. Specifically, we conclude from Tables 1.4, 1.5, and 1.9 that there is strong evidence of a stable link between home prices and cyanoHAB frequency in the Midwest and South regions, and less robust but still strongly indicative evidence for the Northeast and Southeast. The range of models for the Northwest and West suggest potential null effects, while the consistently counterintuitive results for the Ohio Valley suggest there are flaws in our research design and/or data limitations for this region.

For our second category of robustness checks we use our four years of data with water quality measures (2008-2011) to estimate a model that recovers both proximity premia *and* the water quality response in a single step. The specification is

$$\ln P_{it} = \sum_{d=1}^{10} \gamma_d D_{it}^d + \sum_{d=1}^{10} \beta_d D_{it}^d \times cyanoHAB_{jlt} + \theta_j + \tau_{ts} + \xi_m + \varepsilon_{it}. \tag{1.7}$$

Estimation jointly recovers a nonparametric gradient capturing the proximity effects (the g_d parameters) and a set of distance-varying effects of cyanoHAB frequency on the proximity premium (the b_d parameters). We include ten distance bins defined by 100m increments out to 1000m, so that the left-out category in equation (1.7) is homes more than 1km from shore. This specification allows us to examine the sensitivity of our results to pre-determining the spatial extent of the 'treatment' area, as well as distance-based heterogeneity in the treatment effect.

Estimates for the b_d terms are shown in Table 1.6 and the remaining parameters are shown in Appendix Table 1.10. The results confirm the robustness of our findings for the Upper Midwest and South regions. For both, the cyanoHAB effect is negative and significant out to 200m from the nearest lake and largely insignificant for the other distance bins. The magnitudes of the effects mirror the baseline models and are intuitively decreasing with distance from shore. The null results for the NW are likewise consistent with our findings from the primary research design. The results for the NE and SE display sign-intuitive point estimates that are largely insignificant statistically. We interpret these results as underpowered rather than contrary to our main findings – and illustrative of the efficiency advantages of our two-step approach. For the West region we again find no evidence for a cyanoHAB effect, though a counterintuitive estimate arises for the less than 100m distance bin.

The results for the Ohio Valley region are the one substantive departure from our findings using the primary research design. Tables 1.6 and 1.10 suggest that the proximity premium and cyanoHAB effects may exist out to 600m from shore, though the latter is insignificant for less than 100m. To understand this discrepancy, we note that the 100m treatment threshold in our main specification implies there are less than 950 'treated' sales for the OV sample; other regions have 2-17 times more treated observations in their main specifications (see the sample sizes and proportions in Table 1.4). We speculate that the OV dataset as configured for the main specification is therefore more prone to anomalous results and/or failure of identification assumptions.

More generally, the results in Table 1.6 suggest that the cutoff distance for the treatment area may impact estimates of the treatment effect in some cases, and so in the bottom half of Appendix Table 1.9 we present additional robustness checks that vary the size of the treatment zone. Model E1 does not differentiate between treated and control properties in measuring the cyanoHAB gradient, so that the parameter of interest measures the average effect of cyanoHAB frequency on *all* properties within 10km of the nearest lake. The small and in some cases counterintuitive results show that we will misrepresent the negative impact of cyanoHAB frequency on home values if we do not consider distance-based heterogeneity. Models E2 and E3 increase by 100m and double the threshold, respectively (these are equivalent for UM and OV). The estimates are qualitatively similar to the main findings in Table 1.4. Models E4 and E5 decrease by 100m and halve the threshold, respectively (these are equivalent for NW and W). The findings from these changes are as expected given earlier results, though with a loss of economic and statistical significance for the SE region.

Finally, robustness checks F focus only on the OV region and present results for different treatment thresholds. These estimates confirm the patterns in Table 1.6 and show that the counterintuitive impact of observed sales in the <100m distance bin are reduced when additional observations are used to identify an average effect.

1.6 Discussion and Application

1.6.1 Main findings

By exploiting nationwide data on property values and cyanoHABs we draw several conclusions. First, we find strong evidence for the Upper Midwest and South regions that the frequency of cyanoHABs decreases the capitalized value of amenity and recreation services for near-shore homes. This conclusion is robust to alternative implementation choices within our main split sample research design and the same findings emerge using an alternative, single sample research design. Importantly, these findings emerge for the two most data-rich regions: the South and (especially) the Upper Midwest regions have the largest number of resolvable lakes (112 and 219, respectively) and available transactions, algal blooms that are common enough to be salient, and a large amount of useable variation in cyanoHAB frequency.

Second, conditional on our preferred split sample design, we find little evidence that cyanoHABs impact near-shore home values in the West and Northwest regions. These null results emerge for regions that have comparatively small average cyanoHAB frequency (15.9, 16.1 percent of days, respectively), and so this type of freshwater amenity problem may not be as salient in these areas. At the same time, satellite measurements of cyanoHAB frequency in areas with thick property markets are available at a comparatively small number of lakes in the West and Northwest (49 and 41 respectively), and so our analysis may be underpowered in these regions.

Third, we find some evidence that cyanoHAB frequency impacts the size of the proximity premium for lakes in the Northeast and Southeast regions. The findings for these regions are largely robust to specification decisions within our preferred split sample design but not robust to the alternative single-sample design. For the Northeast region both our preferred and alternative research designs provide evidence that cyanoHABs affect prices, but they differ somewhat in magnitudes, statistical significance, and the distance from shore out to which impacts exist. For the Southeast region, the alternative design generates statistically insignificant estimates for all distance bins, though the point estimates are negative and comparable in magnitude to the primary design for distances out to 200m.

Finally, our findings for the Ohio Valley region are counterintuitive for our main research design but plausible for our alternative research design. In our main design the ‘treatment’ zone for the Ohio Valley extends only out to 100m – meaning that only 0.85 percent of our observations are treated. We believe this comparatively small collection of near shore sales occurred in a way that invalidates our identification assumptions. In the alternative design we recover intuitive estimates for distance bins beyond 100m and out to 600m; likewise, using alternative cutoffs in the primary design that include these sales among the treated units generates plausible results. Nonetheless, we emphasize that estimates from the OV are too unstable to be considered anything beyond illustrative.

More generally, our results demonstrate how nationwide data on water quality and property values can be used to estimate regional models that generate heterogeneous findings, which can be directly compared due to their common data sources, specifications, and identification assumptions. This allows us to assess how ex-ante defined identification assumptions perform in different regional contexts. On this, we find that our research design based on observational variation and fixed effects heavy specifications performs exceptionally well in data-rich regions such as the Upper Midwest and South, delivers plausibly null or underpowered results for the West, and Northwest, provides suggestive evidence for the Northeast and Southeast, and generates inconsistent findings for the Ohio Valley. From a weight of evidence perspective, we conclude that our approach can generate valid inference for diverse regions across wide spatial scales but acknowledge that identification challenges can arise due to idiosyncrasies in local conditions and data environments.

1.6.2 Application

Valid inference across broad spatial scales is important for understanding the economic benefits of reducing nutrient pollution. Our results are especially useful for building large-scale integrated assessment models (IAMs) that connect policy actions at points in space to the fate and transport of nutrient emissions, changes in ecosystem services at downstream waterbodies, and economic values from those changes. To illustrate this potential, we use estimates from the Upper Midwest, Northeast, South, and Southeast regions to predict the household level, lake-level, and regionally aggregated economic benefits from representative changes in cyanoHAB frequency.

We examine two types of scenarios for each region. First, we predict the economic value of reducing cyanoHAB detection by one week per year in each lake. This change is plausibly marginal and so we rely on standard hedonic property value theory (see Phaneuf and Requate, 2017, chapter 18) to link it to marginal willingness to pay. Second, we predict the annualized capitalization impact of a 25 percent reduction in the frequency of blooms in each lake. We refer to this as a capitalization effect since price responses to non-marginal changes require additional assumptions to be interpreted as welfare measures.

For each scenario we complete the following steps to predict values:

i. Beginning with the full set of satellite resolved lakes in Table 1.1, use the ZTRAX assessment data to count the number of residential homes H_l within the region-specific treatment threshold on each satellite-resolved lake.

ii. Use the ZTRAX transactions data to identify satellite-resolved lakes with home sales within the treatment threshold during 2013-2016. Compute the average sale price $\overline{P}_l$ for homes sold in proximity to each lake.

iii. Compute the lake-specific baseline cyanoHAB percentage using the four years of available satellite data. Compute the change in cyanoHAB percentage (DHAB_l) for each lake for the scenario.[13]

iv. Compute the annual household-level value of the improvement using

$$WTP_h = \overline{P}_l \times \Delta HAB_l \times \beta_l \times 0.05,$$

where b_l is the estimate from Table 1.4 corresponding to the region where lake l is and we have used a 5 percent annualization rate to convert purchase price into a one-year value.

v. Compute the annual lake-level value of the improvement using

$$WTP_l = H_l \times WTP_h.$$

vi. Compute the regional value of the improvement using

$$WTP^r = \sum_{l=1}^{L_r} WTP_l,$$

where L_r is the number of resolved lakes with observed property transactions in region r.

[13] For the 1-week reduction scenario we use an assumption for the resolvable weeks per year for that region (43 weeks for the Upper Midwest and Northeast; 52 weeks for the South and Southeast) and compute the implied percentage point reduction from one fewer week with detection (e.g. 2.5 percentage points for the Upper Midwest). For the 25 percent reduction scenario, we reduce the percentage of occurrence at each lake by 0.25 times the baseline percentage at that lake.

These steps generate household and lake-specific estimates that are conditional on the lake being resolvable and the existence of observed sales. In general, this means our household and lake-specific estimates represent values for relatively large lakes in populated areas. The region-level aggregate figures are summations of these lakes and so only reflect estimates for the specific set of lakes.

Table 1.7 contains estimates for our two scenarios, broken out across the four regions for household, lake, and regional level benefits. Within a region, the household level estimates have variation based on differences in home values near the different lakes, while the lake level estimates have variation based on differences in average home prices and the number of residential parcels in proximity. Across regions, variation is based on these factors along with differences in the parameter estimates shown in Table 1.4 and differences in the proximity premium distances. In general, the household and lake-level estimates are approximately representative of large lakes in populated areas (satellite resolvable lakes in areas with residential property markets), while the regional totals are illustrative of magnitudes, but specific to the lakes in our data.

Our estimates suggest that cyanoHAB events are costly to nearshore homeowners and that, equivalently, their reduction would generate sizeable welfare and capitalization gains. For example, the average annual household-level figure for reducing cyanoHAB frequency by 25 percent is $337 for Upper Midwestern homeowners within 100m of shore; this translates to over $80,782 for an average lake/year in aggregate value and a likely lower-bound total of over $25.5 million across the Upper Midwest.

1.7 Concluding remarks

In this paper we have examined the economic costs of cyanobacterial harmful algal blooms using property value regressions. Pairing nationwide data on homes sales with satellite derived estimates of cyanoHAB frequency at lakes across the US allows us to present evidence for seven climate regions. Our analysis suggests that more frequent cyanoHABs decrease property values in four of the regions we examine and has null or inconclusive effects in three other regions. For the regions with identified effects, we show using counterfactual policy scenarios that reducing the frequency of cyanobacterial algal blooms could generate large economic benefits at the household, lake, and regional levels.

While our analysis uses plausible identification assumptions, there are limitations that need to be acknowledged. Most importantly, it is difficult to locate 'natural' experiments in lake applications, given the rarity of exogenous shocks that change water quality over short times periods. Our reliance on observational data and natural variation in the frequency of HABs means some anomalous results – such as our Ohio Valley findings – will occur, and the validity of even our robust findings is subject to untestable assumptions.

Further research should continue to pursue large scale estimation of lake water quality/property value gradients using the ZTRAX or similarly broad scale data. This should include examining additional environmental variables at broad scale to understand the extent to which markets capitalize different measurements. The cyanoHAB dataset we use here is currently the only operationally satellite-derived and spatially compatible quality parameter available at the continental scale for US lakes and reservoirs (see Seegers et al. 2021 for the publicly available data). This dataset could be used for future lake and reservoir algorithm development to process additional water quality indicators, such as water clarity, that can be integrated with home sales data. Likewise, efforts such as the LAGOS project[14], GLEON[15], LIMNADES[16], and Water Quality Portal[17] are providing large scale and spatially compatible *in situ* monitoring data that can be linked to property value databases. These additional data sources will allow researchers to compare a variety of identification assumptions and to assemble evidence on research designs that are most likely to provide valid estimates in the absence of quasi-experiments. At the same time, it will be useful to expand the years of water quality data linked to property markets across the country. Our analysis has been limited to the years 2008-2011 due to the current availability of the satellite data in that window. As satellite assessment of lake water quality (and aggregations of in situ

[14] https://lagoslakes.org/lagos-us-overview/
[15] https://gleon.org/data
[16] https://limnades.stir.ac.uk/Limnades_login/index.php
[17] https://www.waterqualitydata.us/

monitoring) expand in space and time availability it will be useful to revisit our analysis with more and newer data. This should include the use of finer scale imagery to resolve the smaller lakes and reservoirs left out of this study.

Other limitations relate to interpretation. First, the spatial resolution of the satellite used in this study, along with QA flags that mask the land-water interface, may stymie measures along the shoreline. One confounding factor inherent to satellite remote sensing of inland waters is straylight contamination along the land-water interface (Schaeffer et al. 2012), especially since cyanobacteria blooms may aggregate along the shoreline. In this study, the nearest pixels from land were quality flagged so our estimates may be conservative as wind advection may form scum conditions at the surface along the shoreline (Chorus and Bartram, 1999). Second, our identification strategy has focused on a specific mechanism – how the proximity premium is affected by lake water quality. This leaves other mechanisms unaddressed, including the wider recreation benefits of lake water quality improvements. Future research should exploit the availability of satellite-derived water quality data for large scale recreation demand analysis. At the same time, alternative property market research designs may allow estimation of a larger range of benefits.

Acknowledgements

Data on property values was provided by Zillow through the Zillow Transaction and Assessment Dataset (ZTRAX). More information on accessing the data can be found at http://www.zillow.com/ztrax. The results and opinions are those of the author(s) and do not reflect the position of Zillow Group. The research was partially supported by a grant from the "Valuing Environmental Benefits from Natural Resource Conservatory" program by the USDA's National Institute of Food and Agriculture (NIFA – grant # 2019-67023-29637). It was also supported by the NASA Ocean Biology and Biogeochemistry Program/Applied Sciences Program (proposals 14-SMDUNSOL14-0001 and SMDSS20-000) and by US EPA. This article has been reviewed by the CEMM and approved for publication. Mention of trade names or commercial products does not constitute endorsement or recommendation for use by the US Government. The views expressed in this article are those of the authors and do not necessarily reflect the views or policies of the US EPA.

1.8 Bibliography

Bailey, Sean W., and P. Jeremy Werdell. "A multi-sensor approach for the on-orbit validation of ocean color satellite data products." *Remote Sensing of Environment* 102, no. 1-2 (2006): 12-23.

Berdalet, E., N. Banas, E. Bresnan, M. Burford, K. Davidson, C. Gobler, B. Karlson, R. Kudela, P.T. Lim, M. Montresor, V. Trainer, G. Usup, K. Yin, H. Enevoldsen and E. Urban, eds., (2017). *Global Harmful Algal Blooms, Science and Implementation Plan*, http://www.globalhab.info/files/Science-and-implementation-plan-final5.pdf.

Brooks, Bryan W., James M. Lazorchak, Meredith DA Howard, Mari-Vaughn V. Johnson, Steve L. Morton, Dawn AK Perkins, Euan D. Reavie, Geoffrey I. Scott, Stephanie A. Smith, and Jeffery A. Steevens. "Are harmful algal blooms becoming the greatest inland water quality threat to public health and aquatic ecosystems?." *Environmental toxicology and chemistry* 35, no. 1 (2016): 6-13.

Chorus, Ingrid, Ian R. Falconer, Henry J. Salas, and Jamie Bartram. "Health risks caused by freshwater cyanobacteria in recreational waters." *Journal of Toxicology and Environmental Health Part B: Critical Reviews* 3, no. 4 (2000): 323-347.

Clark, John M., Blake A. Schaeffer, John A. Darling, Erin A. Urquhart, John M. Johnston, Amber R. Ignatius, Mark H. Myer, Keith A. Loftin, P. Jeremy Werdell, and Richard P. Stumpf. "Satellite monitoring of cyanobacterial harmful algal bloom frequency in recreational waters and drinking water sources." *Ecological indicators* 80 (2017): 84-95.

Coffer, Megan M., Blake A. Schaeffer, John A. Darling, Erin A. Urquhart, and Wilson B. Salls. "Quantifying national and regional cyanobacterial occurrence in US lakes using satellite remote sensing." *Ecological indicators* 111 (2020): 105976.

Coffer, Megan M., Blake A. Schaeffer, Wilson B. Salls, Erin Urquhart, Keith A. Loftin, Richard P. Stumpf, P. Jeremy Werdell, and John A. Darling. "Satellite remote sensing to assess cyanobacterial bloom frequency across the United States at multiple spatial scales." *Ecological Indicators* 128 (2021): 107822.

Dörnhöfer, Katja, and Natascha Oppelt. "Remote sensing for lake research and monitoring–Recent advances." *Ecological Indicators* 64 (2016): 105-122.

Falconer, Ian R., and Andrew R. Humpage. "Health risk assessment of cyanobacterial (blue-green algal) toxins in drinking water." *International journal of environmental research and public health* 2, no. 1 (2005): 43-50.

Graham, Jennifer L., Neil M. Dubrovsky, and Sandra M. Eberts. *Cyanobacterial harmful algal blooms and US Geological Survey science capabilities*. US Department of the Interior, US Geological Survey, 2016.

Gons, Herman J., Hans Hakvoort, Steef WM Peters, and Stefan GH Simis. "Optical detection of cyanobacterial blooms." In *Harmful cyanobacteria*, pp. 177-199. Springer, Dordrecht, 2005.

Guignet, Dennis, Matthew T. Heberling, Michael Papenfus, and Olivia Griot. "Property values, water quality, and benefit transfer: A nationwide meta-analysis." *Land Economics* (2021): 050120-0062R1.

Hallegraeff, Gustaaf M., Donald M. Anderson, Catherine Belin, Marie-Yasmine Dechraoui Bottein, Eileen Bresnan, Mireille Chinain, Henrik Enevoldsen et al. "Perceived global increase in algal blooms is attributable to intensified monitoring and emerging bloom impacts." *Communications Earth & Environment* 2, no. 1 (2021): 1-10.

Horsch, Eric J., and David J. Lewis. "The effects of aquatic invasive species on property values: evidence from a quasi-experiment." *Land Economics* 85, no. 3 (2009): 391-409.

Huisman, Jef, Geoffrey A. Codd, Hans W. Paerl, Bas W. Ibelings, Jolanda MH Verspagen, and Petra M. Visser. "Cyanobacterial blooms." *Nature Reviews Microbiology* 16, no. 8 (2018): 471-483.

Ibelings, Bas W., and Ingrid Chorus. "Accumulation of cyanobacterial toxins in freshwater “seafood” and its consequences for public health: a review." *Environmental pollution* 150, no. 1 (2007): 177-192.

Greb, Steven, Arnold Dekker, and Caren Binding. "Earth Observations in Support of Global Water Quality." (2018).

Bernard, Stewart, Raphael M. Kudela, Lisl Robertson Lain, and Grant Pitcher. "Observation of Harmful Algal Blooms with Ocean Colour Radiometry." (2021).

Jones, Benjamin A. "Infant health impacts of freshwater algal blooms: Evidence from an invasive species natural experiment." *Journal of Environmental Economics and Management* 96 (2019): 36-59.

Karl, Thomas, and Walter James Koss. "Regional and national monthly, seasonal, and annual temperature weighted by area, 1895-1983." (1984).

Keiser, David A., and Joseph S. Shapiro. "Consequences of the Clean Water Act and the demand for water quality." *The Quarterly Journal of Economics* 134, no. 1 (2019): 349-396.

Kutser, Tiit. "Passive optical remote sensing of cyanobacteria and other intense phytoplankton blooms in coastal and inland waters." *International Journal of Remote Sensing* 30, no. 17 (2009): 4401-4425.

Leggett, Christopher G., and Nancy E. Bockstael. "Evidence of the effects of water quality on residential land prices." *Journal of Environmental Economics and Management* 39, no. 2 (2000): 121-144.

L'Ecuyer-Sauvageau, Chloe, Charlene Kermagoret, Jerome Dupras, Jie He, Justin Leroux, Marie-Pier Schinck, and Thomas G. Poder. "Understanding the preferences of water users in a context of cyanobacterial blooms in Quebec." *Journal of environmental management* 248 (2019): 109271.

Liu, Tingting, James J. Opaluch, and Emi Uchida. "The impact of water quality in Narragansett Bay on housing prices." *Water Resources Research* 53, no. 8 (2017): 6454-6471.

Lunetta, Ross S., Blake A. Schaeffer, Richard P. Stumpf, Darryl Keith, Scott A. Jacobs, and Mark S. Murphy. "Evaluation of cyanobacteria cell count detection derived from MERIS imagery across the eastern USA." *Remote Sensing of Environment* 157 (2015): 24-34.

Massey, Isaac Yaw, Fei Yang, Zhen Ding, Shu Yang, Jian Guo, Muwaffak Al-Osman, Robert Boukem Kamegni, and Weiming Zeng. "Exposure routes and health effects of microcystins on animals and humans: A mini-review." *Toxicon* 151 (2018): 156-162.

McKay, Lucinda, Timothy Bondelid, Tommy Dewald, Craig Johnston, Richard Moore, and Alan Rea. “NHDPlus Version 2: User Guide,” (2012) prepared for US EPA Office of Water.

Mishra, Sachidananda, Richard P. Stumpf, Blake A. Schaeffer, P. Jeremy Werdell, Keith A. Loftin, and Andrew Meredith. "Measurement of cyanobacterial bloom magnitude using satellite remote sensing." *Scientific reports* 9, no. 1 (2019): 1-17.

Mishra, Sachidananda, Richard P. Stumpf, Blake Schaeffer, P. Jeremy Werdell, Keith A. Loftin, and Andrew Meredith. "Evaluation of a satellite-based cyanobacteria bloom detection algorithm using field-measured microcystin data." *Science of The Total Environment* 774 (2021): 145462.

Moore, Michael R., Jonathan P. Doubek, Hui Xu, and Bradley J. Cardinale. "Hedonic price estimates of lake water quality: Valued attribute, instrumental variables, and ecological-economic benefits." *Ecological Economics* 176 (2020): 106692.

National Aeronautics and Space Agency. Medium Resolution Imaging Spectroradiometer (MERIS) Data, Goddard Space Flight Center, Ocean Ecology Laboratory, Ocean Biology Processing Group, NASA OB.DAAC (2021), https://oceancolor.gsfc.nasa.gov/projects/cyan/.

Netusil, Noelwah R., Michael Kincaid, and Heejun Chang. "Valuing water quality in urban watersheds: A comparative analysis of Johnson Creek, Oregon, and Burnt Bridge Creek, Washington." *Water Resources Research* 50, no. 5 (2014): 4254-4268.

Nicholls, Sarah, and John Crompton. "A comprehensive review of the evidence of the impact of surface water quality on property values." *Sustainability* 10, no. 2 (2018): 500.

Patt, Frederick S. *Navigation algorithms for the SeaWiFS mission*. NASA Center for AeroSpace Information, 2002.

PaeRL, H. W., and J. HuIsMan. "Blooms like it hot. science, v. 320." (2008): 57-8.

Phaneuf, Daniel J., and Till Requate. *A course in environmental economics: theory, policy, and practice*. Cambridge University Press, 2016.

Poor, P. Joan, Keri L. Pessagno, and Robert W. Paul. "Exploring the hedonic value of ambient water quality: a local watershed-based study." *Ecological Economics* 60, no. 4 (2007): 797-806.

Rashidi, Hamidreza, Helen Baulch, Arshdeep Gill, Lalita Bharadwaj, and Lori Bradford. "Monitoring, managing, and communicating risk of harmful algal blooms (HABs) in recreational resources across Canada." *Environmental Health Insights* 15 (2021): 11786302211014401.

Schaeffer, Blake A., Sean W. Bailey, Robyn N. Conmy, Michael Galvin, Amber R. Ignatius, John M. Johnston, Darryl J. Keith et al. "Mobile device application for monitoring cyanobacteria harmful algal blooms using Sentinel-3 satellite Ocean and Land Colour Instruments." *Environmental modelling & software* 109 (2018): 93-103.

Schinck, Marie-Pier, Chloé L'Ecuyer-Sauvageau, Justin Leroux, Charlène Kermagoret, and Jérôme Dupras. "Risk, drinking water and harmful algal blooms: a contingent valuation of water bans." *Water Resources Management* 34, no. 12 (2020): 3933-3947.

Seegers, Bridget N., P. Jeremy Werdell, Ryan A. Vandermeulen, Wilson Salls, Richard P. Stumpf, Blake A. Schaeffer, Tommy J. Owens, Sean W. Bailey, Joel P. Scott, and Keith A. Loftin. "Satellites for long-term monitoring of inland US lakes: The MERIS time series and application for chlorophyll-a." *Remote Sensing of Environment* 266 (2021): 112685.

Smayda, Theodore J. "Harmful algal blooms: their ecophysiology and general relevance to phytoplankton blooms in the sea." *Limnology and oceanography* 42, no. 5part2 (1997): 1137-1153.

Smith, Robert B., Brad Bass, David Sawyer, David Depew, and Susan B. Watson. "Estimating the economic costs of algal blooms in the Canadian Lake Erie Basin." *Harmful Algae* 87 (2019): 101624.

Smith, Val H., and David W. Schindler. "Eutrophication science: where do we go from here?." *Trends in ecology & evolution* 24, no. 4 (2009): 201-207.

Stroming, Signe, Molly Robertson, Bethany Mabee, Yusuke Kuwayama, and Blake Schaeffer. "Quantifying the human health benefits of using satellite information to detect cyanobacterial harmful algal blooms and manage recreational advisories in US Lakes." *GeoHealth* 4, no. 9 (2020): e2020GH000254.

Stumpf, Richard P., Timothy W. Davis, Timothy T. Wynne, Jennifer L. Graham, Keith A. Loftin, Thomas H. Johengen, Duane Gossiaux, Danna Palladino, and Ashley Burtner. "Challenges for mapping cyanotoxin patterns from remote sensing of cyanobacteria." *Harmful algae* 54 (2016): 160-173.

Taranu, Zofia E., Irene Gregory-Eaves, Peter R. Leavitt, Lynda Bunting, Teresa Buchaca, Jordi Catalan, Isabelle Domaizon et al. "Acceleration of cyanobacterial dominance in north temperate-subarctic lakes during the Anthropocene." *Ecology letters* 18, no. 4 (2015): 375-384.

Tyler, Andrew N., Peter D. Hunter, Evangelos Spyrakos, Steve Groom, Adriana Maria Constantinescu, and Jonathan Kitchen. "Developments in Earth observation for the assessment and monitoring of inland, transitional, coastal and shelf-sea waters." *Science of the Total Environment* 572 (2016): 1307-1321.

Tuttle, Carrie M., and Martin D. Heintzelman. "A loon on every lake: A hedonic analysis of lake water quality in the Adirondacks." *Resource and Energy Economics* 39 (2015): 1-15.

Urquhart, Erin A., Blake A. Schaeffer, Richard P. Stumpf, Keith A. Loftin, and P. Jeremy Werdell. "A method for examining temporal changes in cyanobacterial harmful algal bloom spatial extent using satellite remote sensing." *Harmful algae* 67 (2017): 144-152.

Urquhart, Erin A., and Blake A. Schaeffer. "Envisat MERIS and Sentinel-3 OLCI satellite lake biophysical water quality flag dataset for the contiguous United States." *Data in brief* 28 (2020): 104826.

US Environmental Protection Agency, (2019). "Freshwater HAB newsletter," https://www.epa.gov/sites/production/files/2019-09/documents/habs-newsletter-aug-2019.pdf, retrieved June 4, 2021.

US Environmental Protection Agency, (2016). "National Lakes Assessment 2012: A Collaborative Survey of Lakes in the United States," EPA 841-R-16-113, Washington, DC.

US Environmental Protection Agency, (2014). "Cyanobacteria and cyanotoxins: information for drinking water systems," Office of Water, EPA-810F11001.

Walsh, Patrick J., J. Walter Milon, and David O. Scrogin. "The spatial extent of water quality benefits in urban housing markets." *Land Economics* 87, no. 4 (2011): 628-644.

Walsh, Patrick, Charles Griffiths, Dennis Guignet, and Heather Klemick. "Modeling the property price impact of water quality in 14 Chesapeake Bay Counties." *Ecological economics* 135 (2017): 103-113.

Weng, Weizhe, Kevin J. Boyle, Kaitlin J. Farrell, Cayelan C. Carey, Kelly M. Cobourn, Hilary A. Dugan, Paul C. Hanson, Nicole K. Ward, and Kathleen C. Weathers. "Coupling Natural and Human Models in the Context of a Lake Ecosystem: Lake Mendota, Wisconsin, USA." *Ecological Economics* 169 (2020): 106556.

Whitman, Peter, Blake Schaeffer, Wilson Salls, Megan Coffer, Sachidananda Mishra, Bridget Seegers, Keith Loftin, Richard Stumpf, and P. Jeremy Werdell. "A validation of satellite derived cyanobacteria detections with state reported events and recreation advisories across US lakes." (2022).

Wolf, David, and H. Allen Klaiber. "Bloom and bust: Toxic algae's impact on nearby property values." *Ecological economics* 135 (2017): 209-221.

Wolf, David, Will Georgic, and H. Allen Klaiber. "Reeling in the damages: Harmful algal blooms' impact on Lake Erie's recreational fishing industry." *Journal of environmental management* 199 (2017): 148-157.

Wolf, David, Wei Chen, Sathya Gopalakrishnan, Timothy Haab, and H. Allen Klaiber. "The impacts of harmful algal blooms and E. coli on recreational behavior in lake erie." *Land Economics* 95, no. 4 (2019): 455-472.

Wynne, Timothy T., Richard P. Stumpf, Michelle C. Tomlinson, and Julianne Dyble. "Characterizing a cyanobacterial bloom in western Lake Erie using satellite imagery and meteorological data." *Limnology and Oceanography* 55, no. 5 (2010): 2025-2036.

Wynne, Timothy T., Richard P. Stumpf, Michelle C. Tomlinson, and Julianne Dyble. "Characterizing a cyanobacterial bloom in western Lake Erie using satellite imagery and meteorological data." *Limnology and Oceanography* 55, no. 5 (2010): 2025-2036.

Zhang, Congwen, and Kevin J. Boyle. "The effect of an aquatic invasive species (Eurasian watermilfoil) on lakefront property values." *Ecological Economics* 70, no. 2 (2010): 394-404.

Zhang, Wendong, and Brent Sohngen. "Do US anglers care about harmful algal blooms? A discrete choice experiment of Lake Erie recreational anglers." *American Journal of Agricultural Economics* 100, no. 3 (2018): 868-888.

1.9 Figures and tables

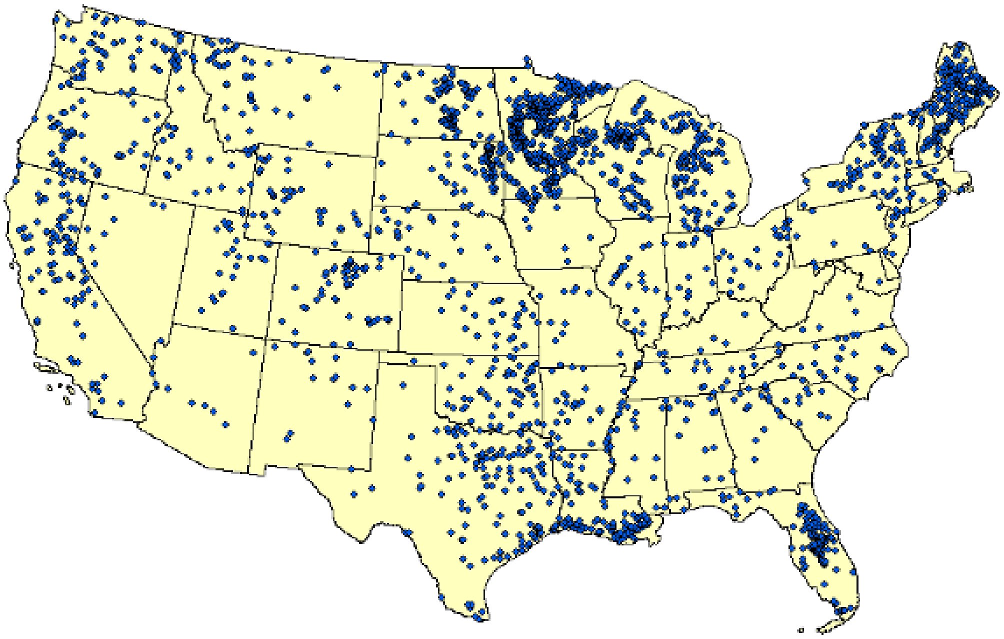

Figure 1.1: Geographical distribution of all satellite resolved lakes. 2,370 resolved lakes for which satellite measures of cyanoHAB frequency are available.

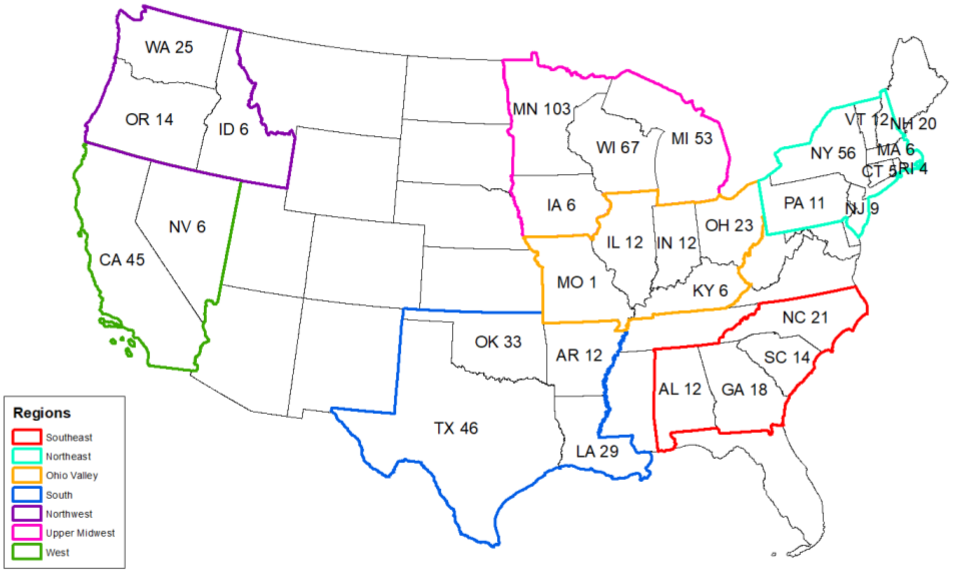

Figure 1.2: Study area broken into climate regions and showing the number of lakes in each state included in the analysis. There are 638 lakes include in our analysis.

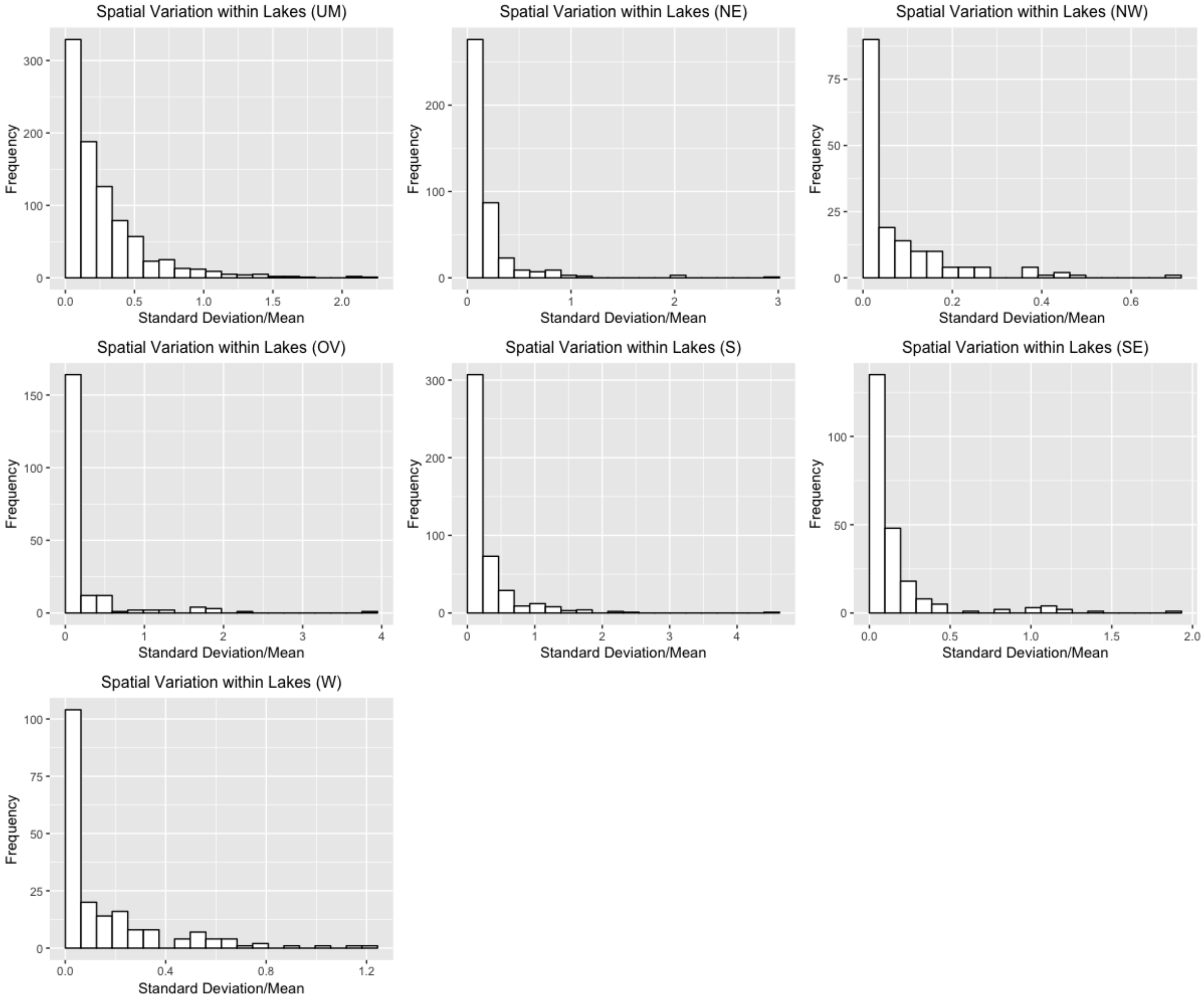

Figure 1.3: Spatial variation within lake-year cells. The unit of observation is cyanoHAB frequency for a census block group in a year for the 633 lakes include in our analysis. Figures show the distribution of the ratio of standard deviation to mean cyanoHAB frequency for the collection of census block group measures in a lake-year cell in a specific climate region. A higher value for the ratio implies more within lake-year spatial variation in water quality.

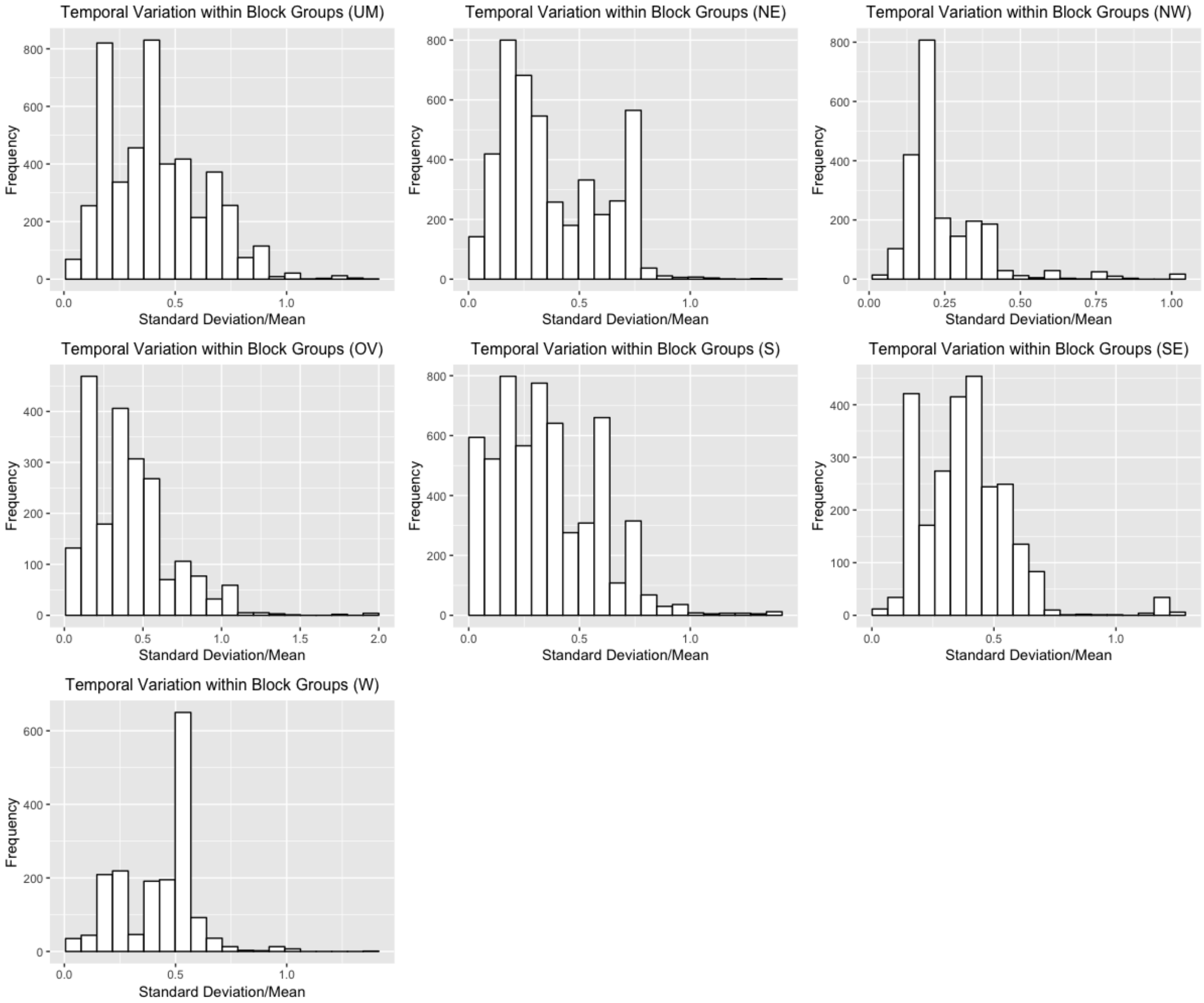

Figure 1.4: Temporal variation within census block groups. The unit of observation is cyanoHAB frequency for a census block group in a year for the 633 lakes include in our analysis. Figures show the distribution of the ratio of standard deviation to mean cyanoHAB frequency for the four years of water quality measures for a specific census block group in a specific climate region. A higher value for the ratio implies more temporal variation in water quality at the census block group level.

Table 1.1: Summary statistics for satellite-resolved lakes

Region	UM	NE	NW	OV	S	SE	W	Other	All Lakes
Number of lakes	709	144	139	76	323	77	102	800	2,370
Lake size (km^2)	15.85	25.19	26.98	33.41	39.92	52.83	48.45	30.53	28.48
	(72.32)	(68.60)	(47.61)	(61.91)	(110.75)	(68.79)	(131.80)	(182.32)	(126.40)
Lake circumference (km)	888	859	1,091	794	1,124	1,106	1,261	961	(979)
	(819)	(596)	(799)	(347)	(1,251)	(901)	(1,672	(949)	(969)
cyanoHAB frequency (%)	27.67	17.63	24.15	27.52	30.27	14.40	21.36	31.55	27.80
	(20.35)	(11.13)	(17.76)	(21.43)	(24.29)	(17.67)	(16.79)	(24.19)	(21.99)

Notes: Standard deviations in parentheses. Column headings are climate regions define as Upper Midwest (UM), Northeast (NE), Northwest (NW), Ohio Valley (OV), South (S), Southeast, and West (W). Other corresponds to lakes that are in regions we do not consider in our analysis. *cyanoHAB frequency* is the percentage of cyanobacterial cell readings taken across all four years of water quality data. The lakes are shown in Figure 1.1 and correspond to the set of waterbodies for which satellite-derived measures of cyanoHAB frequency are available.

Table 1.2: Summary statistics for primary analysis sample

Region	UM	NE	NW	OV	S	SE	W
Numbers of lakes	219	105	41	51	112	56	49
Lake size (km^2)	45 (132)	40 (102)	56 (53)	27 (47)	230 (521)	75 (64)	43 (135)
Lake circumference (km)	1,275 (1756)	987 (837)	1,796 (979)	800 (338)	3,143 (5,542)	959 (529)	1,393 (1,576)
Sale Price ($)	181,664 (165,225)	278,409 (323,386)	344,078 (275,193)	140,526 (136,198)	137,641 (116,921)	191,542 (161,690)	286,131 (190,322)
Numbers of transactions	286,039	181,732	123,283	106,897	186,632	206,256	213,414
cyanoHAB frequency (%)	22.20 (18.63)	17.39 (14.24)	16.06 (13.19)	25.13 (20.94)	25.92 (24.90)	14.61 (19.79)	15.87 (17.77)

Notes: Standard deviations in parentheses. Column headings are climate regions define as Upper Midwest (UM), Northeast (NE), Northwest (NW), Ohio Valley (OV), South (S), Southeast, and West (W). Sale price is in nominal dollars. *cyanoHAB frequency* is the percentage of cyanobacterial cell readings taken across all four years of water quality data. The lakes correspond the counts shown in Figure 1.2.

Table 1.3: Selected estimation results for determining proximity premium threshold

Parameter	Distance	UM	NE	NW	OV	S	SE	W
g_1	100m	0.6038***	0.7206***	0.4407***	0.5236***	0.5480***	0.7751***	0.5077***
g_2	100–200m	0.0468	0.2648***	0.2124***	0.0678	0.2576***	0.3031***	0.1861**
g_3	200–300m	-0.0242	0.0846***	0.0765	-0.0511	0.1526**	0.1146**	0.0104
g_4	300–400m	-0.0355	0.0529**	-0.0128	-0.052	0.0827	0.0865*	0.0118
g_5	400–500m	-0.04	0.0341	-0.0646	-0.0171	0.0641	0.0654	0.0088
g_6	500–600m	-0.0368	0.0139	-0.0824	-0.0263	0.0583	0.0798	-0.0209
g_7	600–700m	-0.0279	0.0107	-0.1039	-0.0407	0.0475	0.0684	-0.01
g_8	700–800m	-0.0235	-0.005	-0.1154	-0.0226	0.0299	0.0516	0.0035
g_9	800–900m	-0.0183	-0.0126	-0.1234	-0.0341	0.0228	0.0731	0.025
g_{10}	900–1000m	-0.025	0.0151	-0.1326	-0.0124	0.0576	0.057	0.0427
Lakes		219	105	41	51	112	56	49
N		938,289	789,395	583,353	411,433	496,982	834,382	655,728

Notes: * p<0.1, ** p<0.05, $^{***}p$<0.01. Regressions are nominal sale price regressed on distance bin dummy variables and block group, state by year of sale, and month of sale fixed effects. Specifications include indicators for 99 distance bins in 100m increments out to 10km. Parameters for distances beyond 1000m are not show. Left out category is 10km. Samples include home sales within 10km of study lakes during the years 2000 to 2007 and 2012 to 2016 (years without cyanoHAB data). Standard errors clustered at the block group level.

Table 1.4: Baseline results

Parameter	UM	NE	NW	OV	S	SE	W
a (T)	0.7555***	0.4623***	0.4032***	0.5783***	0.3145 ***	0.4370***	0.2297***
b (T×cyanoHAB)	-0.0035***	-0.0043***	-0.0017	0.0038*	-0.0038***	-0.0033*	0.0021
Lakes	219	105	41	51	112	56	49
Treatment Cutoff	100m	400m	200m	100m	300m	400m	200m
Proportion Treated	0.0208	0.0502	0.0185	0.0085	0.0278	0.0876	0.0089
N	281,514	180,219	121,803	105,906	185,030	203,880	212,163

Notes: * p<0.1, ** p<0.05, ***p<0.01. Regressions are nominal sale price regressed on a treatment indicator, cyanoHAB percentage occurrence interacted with treatment status, and block group, state by year of sale, and month of sale fixed effects. Samples include home sales within 10km of study lakes during the years 2008 to 2011 (years with water quality data). Homes sold within the 100m distance band beyond the treatment cutoff are excluded from the sample. Standard errors clustered at the block group level.

Table 1.5: Robustness estimates for main empirical design

Model	Change	UM	NE	NW	OV	S	SE	W
Main	NA	-0.0035***	-0.0043***	-0.0017	0.0038*	-0.0038***	-0.0033*	0.0021
A1	Use lake/year average for HAB frequency	-0.0035***	-0.0044***	-0.0023	0.0038*	-0.0040***	0.0002	0.0020
A2	Use HAB frequency 12 months prior to sale date	-0.0034***	-0.0040**	-0.0028	0.0021	-0.0027*	-0.0039*	0.0023
A3	Use HAB frequency April to October	-0.0034***	-0.0048***	-0.0029	0.0036*	-0.0032***	-0.0030	0.0010
B1	Include b_0*cyanoHAB* in specification.	-0.0033***	-0.0038**	-0.0018	0.0040*	-0.0039***	-0.0035*	0.0012
B2	Allow treatment effect (*a*) to vary with sale year	-0.0036***	-0.0040**	-0.0016	0.0033	-0.0041***	-0.0033	0.0017
B3	Allow treatment effect (*a*) to vary with lake size	-0.0037***	-0.0010	-0.0020	0.0032	-0.0051***	-0.0042*	0.0021
C1	Sample of non-missing lot size sales: lot size included.	-0.0024***	-0.0042**	-0.0003	0.0019	-0.0047***	-0.0029	0.0001
C2	Sample of non-missing lot size sales: lot size omitted.	-0.0028***	-0.0042**	-0.0026	0.0023	-0.0035***	-0.0025	0.0007
C3	Sample with all three attributes: attributes included	-0.0042**	0.0049	0.0135**	0.0008	-0.0046**	-0.0024	0.0018
C4	Sample with all three attributes: attributes omitted	-0.0033*	0.0027	0.016**	0.0009	-0.0022	-0.0075***	0.0025
C5	Only use observations from 2010-2011	-0.0035**	-0.0035	-0.0043	0.0022	-0.0033*	-0.0049**	-0.0007
	Sample A2	208,780	180,219	92,215	105,906	133,595	144,647	154,802
	Sample C1, C2	224,834	173,640	114,744	80,617	140,294	165,295	186,940
	Sample C3, C4	107,419	49,005	24,052	56,138	49,638	48,821	39,783
	Sample C5	134,333	86,351	61,710	49,192	85,450	96,352	98,043

Notes: * *p*<0.1, ** *p*<0.05, ****p*<0.01. Regressions are nominal sales price regressed on cyanoHAB percentage occurrence interacted with treatment status, the fixed effects from equation (1.4), and varying controls for treatment status and distance to shore. Reported estimates are for the interaction of treatment status and cyanoHAB frequency (*b* in equation (1.4)) under different data and specification configurations. Samples for models A1, A3, and B1-B3 match samples for the name specification. Samples for A2 use only three years of sales data (2009, 2010, 2011). Samples for models C1, C2, C3, and C4 use observations with non-missing lot size attribute (C1 and C2) and non-missing lot size, total rooms, and total bedrooms (C3 and C4). The C5 sample only uses data only from 2010 and 2011. Standard errors clustered at the block group level.

Table 1.6: Selected parameter estimates from alternative empirical design

Parameter	UM	NE	NW	OV	S	SE	W
(<100m)×cyanoHAB	-0.0032***	-0.0021	0.0011	0.0014	-0.0073***	-0.0028	0.0069***
(100–200m)×cyanoHAB	-0.0015*	-0.0055**	-0.0023	-0.0037**	-0.0051***	-0.0029	0.0003
(200–300m)×cyanoHAB	0.0005	-0.0027	-0.0007	-0.0040**	-0.0011	0.0008	-0.0014
(300–400m)×cyanoHAB	0.0006	-0.0025	0.0002	-0.0034**	-0.0013	-0.0046	-0.0012
(400–500m)×cyanoHAB	-0.0005	-0.0029	0.0022	-0.0028	-0.0012	-0.0009	-0.0017
(500–600m)×cyanoHAB	-0.0013	-0.0025	0.0016	-0.0036**	-0.0008	-0.0005	-0.0008
(600-700m)×cyanoHAB	-0.0021**	-0.0008	0.0026	-0.0000	0.0011	0.0011	-0.0014
(700-800km)×cyanoHAB	0.0006	-0.0017	0.0022	-0.0005	-0.0008	-0.0016	-0.0020**
(800-900m)×cyanoHAB	-0.0004	-0.0022*	0.0026	0.0005	0.0013	0.0004	-0.0015*
(900-1000m)×cyanoHAB	0.0000	-0.0015	0.0015	0.0003	0.0011	0.0025	0.0001
Distance-from-shore dummy variables	Y	Y	Y	Y	Y	Y	Y
Lakes	219	105	41	51	112	56	49
N	286,039	181,732	123,283	106,897	186,632	206,256	213,414

Notes: * $p<0.1$, ** $p<0.05$, $^{***}p<0.01$. Regressions are nominal sale price regressed on block group, state by year of sale, and month of sale fixed effects along with distance to shore dummy variables for 100m increments out to 1000m and interactions between these dummy variables and cyanoHAB occurrence percentage. Estimates for the distance to shore dummy variables are shown in the Table 1.8. Properties sold farther than 1km from the nearest lake is the left out category. Samples include home sales within 10km of study lakes during the years 2008 to 2011 (years with cyanoHAB data). Standard errors clustered at the block group level.

Table 1.7: Application

Scenario[a]	Upper Midwest[b]	South[c]	Northeast[d]	Southeast[e]
Per house annual benefits from 1 week/year reduction in cyanobacterial detection at nearest lake	$124 (75)	$58 (31)	$148 (124)	$80 (31)
Lake-level annual benefits from 1 week/year reduction in cyanobacterial detection	$30,371 (53,767)	$69,496 (130,730)	$115,923 (257,543)	$218,246 (448,930)
Regional annual benefits from 1 week/year reduction in cyanobacterial detection at resolvable lakes with surrounding property markets	$9,566,904	$8,478,562	$14,606,248	$13,094,731
Per house annual capitalization from 25 percent reduction in cyanobacterial bloom days at nearest lake	$337 (293)	$171 (188)	$265 (269)	$102 (177)
Lake-level annual capitalization from 25 percent reduction in cyanobacterial bloom days at nearest lake	$80,782 (160,826)	$184,503 (433,957)	$166,734 (353,205)	$141,203 (253,655)
Regional annual capitalization from 25 percent reduction in cyanobacterial bloom days at resolvable lakes with surrounding property markets	$25,446,286	$22,509,413	$21,008,443	$8,472,179

Notes:

[a] annual values calculated using region-specific estimates from Table 1.4, where a 5 percent annualization rate was used to translate purchase price to annual value.

[b] calculations based on census of parcels in ‘treated’ zone (100m or closer to shore) at 315 resolvable lakes in the Upper Midwest region with 6,893 property transactions in 2013-2016 within 100m of the nearest lake. Average sale price of homes in treated zone is $307,582.

[c] calculations based on census of parcels in ‘treated’ zone (300m or closer to shore) at 122 resolvable lakes in South region with 4,413 property transactions in 2013-2016 within 300m of the nearest lake. Average sale price of homes in treated zone is $159,212

[d] calculations based on census of parcels in ‘treated’ zone (400m or closer to shore) at 126 resolvable lakes in the Northeast region with 12,727 property transactions in 2013-2016 within 400m of the nearest lake. Average sale price of homes in treated zone is $298,336.

[e] calculations based on census of parcels in ‘treated’ zone (400m or closer to shore) at 60 resolvable lakes in the Southeast region with 25,591 property transactions in 2013-2016 within 400m of the nearest lake. Average sale price of homes in treated zone is $251,928.

Table 1.8: Percent of observations with selected attributes available

Parameter	UM	NE	NW	OV	S	SE	W
Sample size	286,039	181,732	123,283	106,897	186,632	206,256	213,414
% with lot size	80	96	94	76	76	81	88
% with total rooms	45	28	21	60	35	35	22
% with total bedrooms	68	68	98	77	54	58	97

Table 1.9: Additional robustness estimates

Model	Change	UM	NE	NW	OV	S	SE	W
D1	Lake by year FEs	-0.0043***	-0.0016	-0.0113**	0.0004	-0.0058***	0.0046***	-0.0039
D2	Census tract by year FEs	-0.0040***	-0.0070***	-0.0037	0.0032**	-0.0053**	-0.0047***	0.0008
D3	Lake by year FEs and block group FEs	-0.0033***	-0.0048***	-0.0019	0.0039*	-0.0040***	-0.0036*	0.0023
D4	Census tract by year FEs and block group FEs	-0.0035***	-0.0049**	-0.0015	0.0043**	-0.0043***	-0.0041*	0.0027
E1	Assume homogeneous HAB effect across the study area	-0.0006***	-0.0015***	0.0005*	-0.0010***	0.0002	0.0019***	0.0020***
E2	Increase the treatment area by 100m	-0.0017**	-0.0043***	-0.0014	0.0006	-0.0039***	-0.0036**	0.0003
E3	Double treatment area threshold		-0.0040**	-0.0006		-0.0030***	-0.0034**	-0.0000
E4	Decrease the treatment area by 100m	NA	-0.0043**	0.0017	NA	-0.0061***	-0.0011	0.0081***
E5	Halve the treatment area threshold	-0.0051***	-0.0049**		0.0053*	-0.0056***	-0.0005	
F	Baseline OV model with different treatment thresholds		100m (baseline)	200m	400m	600m	800m	
			0.0038*	0.0006	-0.0024*	-0.0027**	-0.0023**	

Notes: * p<0.1, ** p<0.05, ***p<0.01. Regressions are nominal sales price regressed on cyanoHAB percentage occurrence interacted with treatment status, the fixed effects from equation (1.4), and varying controls for treatment status and distance to shore. Reported estimates are for the interaction of treatment status and cyanoHAB frequency (b in equation (1.4)) under different specification configurations. Standard errors clustered at the block group level.

Table 1.10: Additional parameter estimates corresponding to table 1.6

Parameter	UM	NE	NW	OV	S	SE	W
(<100m)	0.7335***	0.8489***	0.5873***	0.6493***	0.6584***	0.8852***	0.3194***
(100–200m)	0.1188***	0.3845***	0.3598***	0.2999***	0.3370***	0.4075***	0.0508
(200–300m)	-0.0542	0.1315***	0.2150***	0.1645**	0.1216***	0.1987***	-0.0126
(300–400m)	-0.0466	0.1472***	0.1400**	0.0278	0.1209***	0.1523***	-0.0640
(400–500m)	-0.0455	0.096**	0.0576	0.1048*	0.0644	0.1331***	-0.0360
(500–600m)	0.0048	0.0275	0.0558	0.1316**	0.0544	0.1092**	-0.0859**
(600-700m)	0.0319	0.056*	0.0291	0.0025	0.0005	0.1228***	-0.0752**
(700-800m)	-0.0382	0.0524*	-0.0021	0.0025	0.0139	0.1171***	-0.0434
(800-900m)	0.0146	0.0095	-0.0046	0.0199	-0.0383	0.0925***	-0.0218
(900-1000m)	-0.0275	0.0112	-0.0159	0.0096	0.0096	0.0508	-0.0253
Lakes	219	105	41	51	112	56	49
N	286,039	181,732	123,283	106,897	186,632	206,256	213,414

Notes: * *p*<0.1, ** *p*<0.05, ****p*<0.01. Column headings are climate regions define as Upper Midwest (UM), Northeast (NE), Northwest (NW), Ohio Valley (OV) South (S), Southeast, and West (W). Regressions are nominal sale price regressed on block group, state by year of sale, and month of sale fixed effects along with distance to shore dummy variables for 100m increments out to 1000m and interactions between these dummy variables and cyanoHAB occurrence percentage. Estimates for the interaction variables are shown in the Table 1.6. Properties sold farther than 1km from the nearest lake is the left out category. Samples include home sales within 10km of study lakes during the years 2008 to 2011 (years with cyanoHAB data). Standard errors clustered at the block group level.

1.10 Appendix: Identification assumption

Our main estimating equation is

$$\ln P_{it} = \alpha T_{it} + \beta T_{it} \times cyanoHAB_{jlt} + \theta_j + \tau_{ts} + \xi_m + \varepsilon_{it}. \tag{1.A1}$$

To understand conditions for identification, consider sales of homes *i* and *k* in census block group *j* during year *t*, where T_{it}=1 and T_{kt}=0, so that home *i* is 'treated' and home *k* is a 'control'. The conditional expectations for sales prices are

$$\begin{gathered} E\left(\ln P_{it} \mid T_{it}, cyanoHAB_{jlt}\right) = \theta_j + \tau_{ts} + \alpha + \beta cyanoHAB_{jlt} + E\left(\varepsilon_{it} \mid T_{it}, cyanoHAB_{jlt}\right) \\ E\left(\ln P_{kt} \mid T_{kt}\right) = \theta_j + \tau_{ts} + E\left(\varepsilon_{kt} \mid T_{kt}\right), \end{gathered} \tag{1.A2}$$

and the difference in expectations is

$$E\left(\ln P_{it} \mid T_{it}, cyanoHAB_{jlt}\right) - E\left(\ln P_{kt} \mid T_{kt}\right) = \alpha + \beta cyanoHAB_{jlt} + E\left(\varepsilon_{it} - \varepsilon_{kt} \mid T_{it}, T_{kt}, cyanoHAB_{jlt}\right). \tag{1.A3}$$

Note that the conditional expectation on the right-hand side of (A3) is unlikely to be zero. For example, the term *a* captures the average proximity premium for lakes in a climate region, as well as any systematic differences in properties nearer and farther from shore. If the unobserved lake-specific deviation from the average proximity premium is correlated with water quality then

$$E\left(\varepsilon_{it} - \varepsilon_{kt} \mid T_{it}, T_{kt}, cyanoHAB_{jlt}\right) = E\left(\underbrace{\Delta_{lt} + \tilde{\varepsilon}_{it}}_{\varepsilon_{it}} - \varepsilon_{kt} \mid T_{it}, T_{kt}, cyanoHAB_{jlt}\right) \neq 0, \tag{1.A4}$$

where D_{lt} denotes the deviation from the average proximity premium for homes near lake *l* during time *t*. This can occur if time-invariant lake characteristics such as surface area or lake depth are correlated with water quality and affect the size of the proximity premium.

To eliminate this source of bias via another difference, consider sales of homes *i′* and *k′* in census block group *j* during year *t′* and note that the difference in expectations for the two sales prices is

$$\begin{aligned} & E\left(\ln P_{i't'} \mid T_{i't'}, cyanoHAB_{jlt'}\right) - E\left(\ln P_{k't'} \mid T_{k't'}\right) \\ & = \alpha + \beta cyanoHAB_{jlt'} + E\left(\Delta_{lt'} \mid T_{i't'}, T_{k't'}, cyanoHAB_{jlt'}\right) + E\left(\tilde{\varepsilon}_{i't'} - \varepsilon_{k't'} \mid T_{i't'}, T_{k't'}, cyanoHAB_{jlt'}\right), \end{aligned} \tag{1.A5}$$

where we have substituted in the notation from (A4). Taking the difference between (A3) and (A5) we have a spatial difference in differences estimator:

$$\begin{aligned}
&\left[E\left(\ln P_{it} \mid T_{it}, cyanoHAB_{jlt}\right) - E\left(\ln P_{kt} \mid T_{kt}\right)\right] - \left[E\left(\ln P_{i't'} \mid T_{i't'}, cyanoHAB_{jlt'}\right) - E\left(\ln P_{k't'} \mid T_{k't'}\right)\right] \\
&= \beta\left(cyanoHAB_{jlt} - cyanoHAB_{jlt'}\right) \\
&+\left[E\left(\Delta_{lt} \mid T_{it}, T_{kt}, cyanoHAB_{jlt}\right) - E\left(\Delta_{lt'} \mid T_{i't'}, T_{k't'}, cyanoHAB_{jlt'}\right)\right] \\
&+\left[E\left(\tilde{\varepsilon}_{it} - \varepsilon_{kt} \mid T_{it}, T_{kt}, cyanoHAB_{jlt}\right) - E\left(\tilde{\varepsilon}_{i't'} - \varepsilon_{k't'} \mid T_{i't'}, T_{k't'}, cyanoHAB_{jlt'}\right)\right].
\end{aligned} \tag{1.A6}$$

Consistent estimation of *b* requires that the differences in the last two lines are zero. For this we need (a) the lake's expected deviation from its region-average proximity premium to be constant over time; and (b) the time varying unobserved drivers of price need to be uncorrelated with changes in cyanoHAB frequency over time. The first of these means that time invariant unobserved lake characteristics *can* be correlated with the lake-specific proximity premium without compromising the consistency of our estimate of *b*. The second of these relates to our inability to control for property characteristics. If near- and off-shore properties are systematically different in their (unobserved) attributes, then

$$E\left(\tilde{\varepsilon}_{it} - \varepsilon_{kt} \mid T_{it}, T_{kt}, cyanoHAB_{jlt}\right) \neq 0. \tag{1.A7}$$

For example, if homes near-shore are systematically larger than homes away from shore then the expectation will be positive. Indeed, using our limited data on property attributes, we do see modest differences in median lot size between near-shore and more distant homes for some regions. However, this will not bias our estimate of *b* so long as the systematic differences between homes sold near and far from shore is not changing over time in a way that is correlated with changes in cyanoHAB frequency. This condition describes the last line of equation (1.A6).

The results reported in Table 1.5 provide some support for this identification assumption. The table includes estimates of equation (1.4) using subsamples that (a) include only observations for which lot size is available; and (b) include only observations for which lot size, total rooms, and total bedrooms are available. We run models on these samples that do and do not control for the available attributes. The similar point estimates for *cyanoHAB* between the with and without attribute specifications provide some assurance against omitted variable bias.

Chapter 2

A residential sorting model of property values and water quality

With Daniel Phaneuf

2.1 Introduction

Inland lakes are a major source of freshwater in the United States. Water quality in lakes affects their ability to provide economic and ecological services for recreation activities, atheistic amenities, drinking water, and fisheries. Knowing the importance of improving water, it is still not clear either how water quality in lakes affects people's welfare or their willingness to pay for water quality improvement. Lakes provide amenity and recreation services for people in a wide space. People who live close to a lake obtain value through scenic views and convenient access. As for non-lakefront households who usually do not access the lake every day, they enjoy recreational visits to lakes that are within a reasonable travel distance. For both groups, water quality plays an important role in economic benefits from lakes. In this study, I examine the impact of water quality on household utility through the mechanisms of proximity amenities and recreational activities. I also want to estimate the marginal willingness to pay (MWTP) for water quality improvement for both lakefront and non-lakefront households.

Past studies find evidence on the amenity and recreational benefits of lakes, but in separate models with either housing transactions or recreational data. The hedonic model has been widely used to estimate the non-market value of water quality. Using housing transactions in one (Leggett and Bockstael 2000; Poor et al. 2007) or several counties (Gibbs et al. 2002; Walsh et al. 2017; Wolf and Klaiber 2017), researchers find a positive correlation between water quality and home prices. However, the overall benefits may be underestimated as most of the studies only focus on properties on or near the shoreline. Even though some studies cover a larger region around waterbodies, the impact of water quality on prices diminishes sharply or stops while moving away from the shoreline. The MWTP for water quality improvement is not estimated for non-waterfront households. In addition, the hedonic model usually assumes that the property price is affected only by water quality in its nearest lake, so that the MWTP is only for the improvement in local ambient water quality through the amenity channel. The recreational benefit associated with water quality in a larger region is not captured in the hedonic model. Studies focusing on the recreational benefit of water quality usually use the travel cost model to estimate the impact of water quality on recreational activities like fishing (Kaoru 1995) and boating (Lipton 2004). Some recent studies also explore alternative methods (Keiser 2019) or data (Keeler et al. 2015) to estimate the demand for clean water. But these studies fail to address amenity value of water quality improvement. In this paper, I develop a model to estimate the marginal benefits of water quality through both the amenity and recreational channels.

Before developing the model, it is important to understand the mechanisms providing benefits for lakefront and non-lakefront households. For lakefront households, there is both amenity value from the nearest lake where they locate and the recreational value from lakes in the larger region. For non-lakefront households, they only get the benefit from recreation. Water quality affects amenity value directly through the impact from the nearest lakes, while the recreational amenities are not only affected by water quality but also other lake characteristics like the size of the lake and the travel distance. The scenario described above brings about two main challenges to build the model: one is to distinguish the marginal benefits of water quality improvement for lakefront and non-lakefront households. The other is to capture the single dimensional amenity value in the nearest lake and the multidimensional recreation value from multiple lakes in a larger region at the same time. As the hedonic model and the travel cost model cannot individually measure both types of economic value, I am interested in a new modeling framework that can estimate the MWTP for water quality improvement from three perspectives: the amenity value for lakefront households, the recreational value for lakefront households, and the recreational value for non-lakefront households.

As an alternative to hedonic price analysis, the equilibrium sorting model has been widely used to value public goods like school quality, open space, air quality, and climate (Bayer and Timmins 2007; Klaiber and Phaneuf 2010; Hamilton and Phaneuf 2015; Sinha et al. 2018). But it hasn't been applied to estimate the value of water quality. The sorting model is derived from the discrete choice model that is used to estimate consumer demands for differentiated products. It assumes that households maximize their utility by choosing from a finite set of discrete alternatives. In the residential location, households pay for the bundle of housing type attributes and public goods while making a trade-off between the expenditure on houses and other goods given their budget constraints. I measure household preferences for public goods like water quality by estimating the residential sorting model. Different from the hedonic model, the unit of analysis in the sorting model is the housing type that includes a group of properties sharing the same housing attributes. In my study, the housing type is defined as the census tract of the property, the nearest lake of the property, and whether the property is in the lakefront or non-lakefront area. With this, three types of MWTP can be captured in different housing type attributes that are defined at one of the geographical levels. As water quality is measured at the lake level, all the housing types assigned to the same lake have the same water quality level regardless of their locations on the lake. To distinguish the lakefront and non-lakefront housing type, I include a dummy variable indicating the lakefront location which also measures the proximity premium of being close to a lake. The amenity value of water quality for lakefront households is captured by an interaction term of the water quality and the lakefront dummy variables. Meanwhile, I construct a recreation index at the census tract level. This index incorporates water quality, the size of the lake, and the distance from the census tract of a housing type to the lake into one single variable for all lakes in a large region around the census tract. With this, the MWTP for water quality through the recreation mechanism can be estimated from the recreation index. In addition, the residential sorting model allows heterogeneous preferences over housing and environmental attributes. Compared to the traditional hedonic model that assumes homogenous preferences, my model estimates how the demands for water quality and lake amenities change with the household race and annual income level.

My analysis uses data in Wisconsin. Wisconsin is one of the most lake-endowed state in the US. There are over 15,000 lakes and more than 2,000 of them are greater than 50 acres. When a household chooses a residential location in Wisconsin, I assume they consider water quality and other lake amenities around the neighborhood. My property transaction data comes from the ZTRAX dataset, which includes information on the location, the transaction time, and the price of residential home sales. The household characteristics of race and annual income level are obtained by merging the sales data with the Home Mortgage Disclosure Act (HMDA) dataset. Water quality data is from Wisconsin Department of Natural Resource. It records the Secchi depth values in Wisconsin lakes. Secchi depth measures the transparency of water, which is influenced by algae concentration and suspended sediments in water. Low Secchi depth indicates low water quality. As the Secchi depth is measured by human observations, it directly reflects what people observe and perceive about water quality in lakes.

The estimation results from my sorting model confirm there are benefits from an improvement in water quality at the nearest lake for lakefront households. For every 0.1m increase in Secchi depth, the value ranges from $1,442 to $3,195 for households with annual income from $22,000 to $960,100, respectively. This is a one-time payment based on the housing type price. In addition, I also develop a formula based on the recreation index that captures the recreational benefit of lakes. My exploratory analysis of the recreation index supports our hypothesis that water quality affects both lakefront and non-lakefront households though the recreational mechanism. The marginal value of water quality in a specific lake depends on the size of the lake and the distance to the lake.

This paper has five main contributions. First, this is the first paper using the equilibrium sorting model to estimate the demand for water quality and lake amenities. It contributes to the sorting literature and the literature on the non-market valuation of water quality. With different assumptions, it provides an alternative strategy to estimate the economic value of water quality. The hedonic model assumes that households choose the continuous quantity of structural characteristics and public goods to maximize their utility based on the budget constraint. (Kuminoff 2009). As a natural amenity, water quality can be considered as a continuous variable. However, the continuity is usually limited within waterbodies. Water quality and other lake characteristics usually vary across lakes in a discrete way. The equilibrium sorting model does not rely on continuity because households choose from a finite choice set of housing types to maximize their utility. Second, I estimate the amenity value and recreational value of water quality and lake amenities in one model, for both lakefront and non-lakefront households. My model provides an alternative way to understand how water quality affects household welfare. Third, I conduct an initial exploration of creating a recreation index to capture the recreational benefit of lakes in a large region. The illustrative modeling provides experiments on the different roles played by water quality and other lake amenities in generating recreational value. Fourth, I solved a common problem of missing data. In hedonic studies, researchers usually drop lakes with missing water quality data and link a property to its nearest lake with valid data. However, dropping a lake with missing data may lead to a false division of housing types across the landscape. In my study, I solve this problem with machine learning techniques to estimate the missing water quality data such that I can keep all the lakes in the dataset. Fifth, I explore the heterogeneity in household sorting behavior. Compared to the hedonic model which assumes homogeneous preferences for housing attributes and public goods, the sorting model allows MWTP to vary with household characteristics. I find that a white household with a higher annual income level is willing to pay more to live in a lakefront neighborhood. They also have a higher demand for water quality in the adjacent lake.

This paper proceeds in the following way. Section II provides background of past literature. Section III presents the models I used for our estimation. In Second IV, I introduce our main datasets and several techniques to process the data for the estimation. I also explain the machine learning approach to estimate the missing water quality data in this second. Section V explains our main results. Section VI discuss our findings and conclusions.

2.2 Background

The amenity and recreational value of water quality has been examined in two distinct literatures. The hedonic model measures the impact of water quality on home prices by running a regression of home prices on housing and neighborhood characteristics. The large hedonic property literature provides strong evidence on the positive impact of water quality improvement on property values. (Leggett and Bockstael 2000; Poor et al. 2007; Walsh et al. 2017; Gibbs et al. 2002; Wolf and Klaiber 2017, Zhang, Phaneuf, and Schaeffer 2022). However, most of the studies only find positive impact on home values in waterfront areas (Leggett and Bockstael 2000; Gibbs et al. 2002; Zhang and Boyle 2010), which corresponds to the amenity value in our study. Although researchers try to estimate the impact on non-waterfront properties by including transactions for both waterfront and non-waterfront properties or adding interactions of the water quality and distance dummy variables (Walsh et al. 2017; Wolf & Klaiber 2017), there is little property value evidence on the benefits of water quality for non-waterfront properties. While non-waterfront households may not receive amenity value of water quality, they can still receive the recreational value by visiting lakes nearby. Even for waterfront properties, past hedonic studies may underestimate the value of water quality as they only consider the ambient water quality in the nearest lake in most of the cases, which does not capture the benefit of water quality improvement in a larger region through the recreational mechanism. Thus, I conclude that the hedonic model only estimates the amenity value of water quality for lakefront households and fails to address its recreational value generally.

The recreational value of water quality is examined in different literature. The travel cost model is commonly used to analyze how water quality affects recreation amenities provided by lakes. They have shown positive effects from high water quality on recreation fishing (Yoshiaki 1995; Vesterinen et al. 2009), trip duration (Breen et al. 2017), boating (Lipton 2004), swimming (Vesterinen et al. 2009) and general aquatic activities (Keeler et al. 2015). In addition to the travel cost model, researchers also use the method of contingent valuation. They estimate the value of water quality improvement by directly asking visitors' MWTP to increase water quality to a higher level (Lipton 2004) or the additional trips they are willing to take when water quality improves (Lankia et al. 2017). For example, Lipton (2004) claims that the annual MWTP to increase water quality one-step forward for boat owners is $55-$93 based on an open survey question. Recently, people try to use alternative methods and data to measure the recreational benefit of water quality. Keiser (2019) develops a reduced-form model with instrument variables to measure the impact of water pollution in the participation in recreational activities. He claims that a 0.1 mg/l increase in phosphorus decreases water-based recreational trips at least 4 days per year. Also, Keeler et al. (2015) estimate the recreational demand for clean water using the geotagged photographs as a proxy for recreational visits to lakes. They find that the marginal willingness to pay for travel costs increases $22 for a 1m increase in Secchi depth. These studies estimate the demand for water quality through the recreational mechanism. However, it only considers the economic value of recreational activities for of visitors such as anglers, boaters, and swimmers. The amenity value for lakefront households who have easy access is not fully captured.

Considering that the hedonic model and the studies of recreation demand only partially estimate the benefit of water quality improvement, researchers have examined a model that combines both methods. Phaneuf et al. (2008) propose a revealed preference method to combine the long-run residential choice and the short-run recreational behaviors. They claim that the long-run decision of residential housing choice is affected by the availability and quality of local recreation sites. The recreational benefit of a property is affected by the quality of environmental condition and the location of the property. They first develop a random utility model to estimate the recreational benefits at a given residential location with data on trips to local recreation sites. Moving to the long-run residential decisions, they interpret the estimated benefit as an attribute of a property in a first-stage hedonic model. With the two-step analysis, they find a statistically significant impact on home values from the quality-adjusted recreation access index, which suggests the importance of estimating the recreational value of water quality. Kuwayama et al. (2022) also point out that the traditional hedonic model does not capture the recreational value of water in a large region. Following the approach in Phaneuf et al. (2008), they first estimate households' utility from recreational fishing trips. Then they include both local ambient water quality and the estimated recreational benefits in the hedonic model to estimate the MWTP for the amenity and recreational benefits of water quality improvement. They found a higher recreational benefit from the improvement in regional recreational waters than the benefit from the improvement in local ambient water quality. These two studies support the importance of measuring both the amenity and recreational value of water quality.

In this study, I use an alternative method of the residential sorting model to estimate both the amenity and recreational values. The residential sorting model is often used to estimate household preferences for neighborhood attributes like the crime rate and school quality (Bayer et al. 2007, Bayer and Timmins 2016). The sorting model assumes households choose from a finite set of housing types to maximize their utility. The housing types are usually defined by structural housing characteristics and the location of the house. (Klaiber and Phaneuf 2010, Bayer and Timmins 2007). The choice elements vary from metropolitan areas across the country (Bayer et al. 2012) to local communities like school districts and census tracts (Bayer et al. 2007, Kuminoff et al. 2013). More recent studies have used the residential sorting model to predict the demands for environmental amenities like open space (Klaiber and Phaneuf 2010), air quality (Hamilton and Phaneuf 2015), and tree cover level (Cao et al. 2018). However, the sorting model has not been applied to study water quality.

My study is the first to estimate the value of water quality using the residential sorting model. To estimate the MWTP of water quality improvement for lakefront and non-lakefront households, I use a lakefront dummy variable to define the housing types along with the census tract and the nearest lake of the property. I include the water quality variable in two components of the utility function to separately estimate the amenity and recreational values. The amenity value of water quality improvement for lakefront households is captured in the interaction term of the lakefront dummy variable and the Secchi depth value in the nearest lake. The recreational value is captured in a recreation index that depends on water quality and other characteristics of lakes in the surrounding area of the housing type. I assume the utility of both lakefront and non-lakefront households is affected by water quality through the recreational mechanism. The specific MWTP depends on the size of the lake and the distance from the census tract of the housing type to the lake. The details of the model will be discussed in the next section.

2.3 Method

I develop my sorting model following the framework from Bayer et al. (2007) and Bayer and Timmins (2007). I assume that a household chooses the housing type to maximize utility. Water quality, as an environmental good, affects household utility along with other housing type attributes. A housing type is defined by three geographical variables: the census tract of the property, the nearest lake to the property, and a dummy variable indicating the lakefront area within 100m of the shoreline. The census tract determines socioeconomic and demographic features of the neighborhood, as well as local public goods. The other two variables jointly assign a property to either the lakefront or non-lakefront area of its nearest lake. Based on these three variables, the study area can be divided into a set of exhaustive and mutually exclusive choice alternatives with varying attributes. For example, a housing type in the choice set could be the lakefront area of a lake in a given census tract. For large lakes, properties around the same lake may locate in different census tracts.

Consider the setting where a household *i* chooses the housing type defined as the neighborhood in the area b of lake *l* in the census tract *c* to maximize its utility.[18] Letting h = *{b, l, c}*, we can write the utility function as:

$$U_h^i = U_{blc}^i = V\left(q_l, d_b, r_c, X_c, p_h, Cnty_j, \xi_h, \varepsilon_h^i\right) \, , \tag{2.1}$$

where the indirect utility is determined by the average water quality over six years at lake l q_l,[19] a dummy variable indicating a lakefront area within 100m of lakes d_b(=1, lakefront area), the recreation index for

[18] There are two areas: the lakefront and non-lakefront areas.
[19] In the machine learning model, I divide the six years into two periods. So that the average Secchi depth value was calculated by the mean value over two periods.

the census tract r_c, other characteristics of census tract X_c, the price of the housing type p_h, a county fixed effects $Cnty_j$, and the unobserved characteristics of housing type h defined by ξ_h.

For estimation the indirect utility function is written as:

$$V_h^i = \alpha_p p_h + \alpha_q^i q_{lc} \cdot d_b + \alpha_d^i d_b + \alpha_r r_c + \alpha_X X_c + Cnty_j + \xi_h + \varepsilon_h^i \ , \tag{2.2}$$

where the parameters α_A^i, A ∈ $\{q, d\}$ consists of a mean parameter that captures the average preference over all households and a household-specific component that accommodates heterogeneity in marginal utilities based on individual characteristics. We can expand the parameters α_A^i as:

$$\alpha_A^i = \alpha_A + \sum_k^K \alpha_{kA} \, I_k^i \ , \tag{2.3}$$

where I_k^i denotes the household characteristic k.

I assume that water quality affects the household utility through the amenity and recreational mechanisms. The corresponding marginal impact is estimated through different components of the model. The amenity value of water quality for lakefront households is captured in the interaction term of the lakefront dummy variable and the water quality variable. Meanwhile, I construct a recreation index r_c at the census tract level to capture the recreational value of water quality and lake amenities. The recreation index r_c considers multiple lakes in a region around the census tract. I do not differentiate the impact of water quality from the recreational channel for lakefront and non-lakefront households, so that all housing types in the same census tract have the same recreation index. The index is defined below.

The two components of water quality in the model allow me to estimate the impact for different groups through different mechanisms. Lakefront households gets both amenity value and recreational value from the nearest lake, which are separately captured in the interaction term and the recreation index. In addition to ambient water quality in the nearest lake, their utility is also affected by other lakes in the region. As for the non-lakefront households, water quality only affects their utility through the recreation index.

To estimate the model, we specify equation (2.2) into two parts:

$$V_h^i = \Theta_h + \Gamma_h^i + \varepsilon_h^i \ , \tag{2.4}$$

where

$$\mathit{\Theta}_h = \alpha_p p_h + \alpha_q q_l \cdot d_b + \alpha_d d_b + \alpha_d r_c + \alpha_X X_c + Cnty_j + \xi_h \ , \tag{2.5}$$

and

$$\mathit{\Gamma}_h^i = (\textstyle\sum_k^K \alpha_{kq}\, I_k^i) q_l \cdot d_b + (\sum_k^K \alpha_{kd}\, I_k^i) d_b \ . \tag{2.6}$$

In equations (2.5) and (2.6), $\mathit{\Theta}_h$ denotes the average utility of housing type h and $\mathit{\Gamma}_h^i$ denotes the heterogeneous part of the utility that varies with individual characteristics. I follow the two-stage estimation used in Klaiber & Phaneuf (2010) to estimate the model. At the first stage, I recover the heterogeneity parameters α_{kA} and the mean utility parameters Θ_h using the maximum likelihood method. I assume that ε_{blc}^i is independently and identically extreme value distributed. With this assumption, the probability of a household choosing housing type h is:

$$Pr_h^i = \frac{e^{\mathit{\Theta}_h + \mathit{\Gamma}_h^i}}{\sum_{h'} e^{\mathit{\Theta}_{h'} + \mathit{\Gamma}_{h'}^i}} \ . \tag{2.7}$$

The log likelihood function is

$$ll \ = \ \textstyle\sum_i^I \sum_h^H Y_h^i ln(\, Pr_h^i) \ , \tag{2.8}$$

where $Y_h^i \ = \ 1$ if household *i* chooses housing type h and $Y_h^i \ = \ 0$ otherwise. I use the contraction mapping algorithm proposed by Berry (1994) to estimate the first stage. At each iteration of the maximum likelihood search algorithm, the first order conditions of equation (2.8) implies that the predicted market share equals the true observed market share from the data:

$$\widetilde{S_h} = \textstyle\sum_i^I Pr_h^i = S_h \ , \tag{2.9}$$

I take as given the value of α_{kA} and solve for Θ_h from the equality condition in formula (2.9). Then, I plug all the parameters back into the objective function (2.8) to update the heterogeneity parameters through the gradient-based step until the objective function converges to its maximum. The second stage estimates parameters determining the average impacts, which decomposes the estimated average utility into housing type attributes in equation (2.5) using a linear regression. We can estimate the average marginal utilities of housing type attributes with OLS or IV methods.

2.3.1 Recreation Index

The recreation benefit from a lake depends on lake characteristics. In addition to water quality, I assume that the size of a lake and its proximity to the property affect its level of recreational services. To create the recreation index, I first define a recreational area by drawing a 30km buffer around the boundary of each census tract. Then, I select lakes with one of the recreational amenities of public lands, public parks, and public beaches. I assume that people are willing to travel a reasonable distance to a lake for recreational purpose, even though they may have different preferences for lakes with different characteristics. Every recreational lake within the 30km buffer contributes to the recreation index at the census tract level with different weights. The recreation index is defined in the following equation:

$$r_c = \sum_l^{L_c} \left(\frac{s_l}{s_{max}}\right)^{\gamma_1} \left(\frac{q_l}{q_{max}}\right)^{\gamma_2} \left(\frac{d_{max}-d_{cl}}{d_{max}}\right)^{\gamma_3}, \tag{2.10}$$

where L_c denotes the number of lakes within 30km of census tract c, q_l denotes the average water quality value at lake l, s_l is the surface area, and d_{cl} is the distance from census tract c to lake l. The recreation index for census tract c is the summation of the function of every lake in the buffer area. For each lake, the function is determined by the size of the lake, water quality, and the travel distance from the center of the census tract to the lake. Because the three variables are measured in different units, I rescale each variable as the fraction of its maximum value statewide. Since the travel distance usually negatively influences household utility, I rescale it so that a smaller distance positively contributes to the index. To allow the three lake features to differentially affect the index, I add a parameter for each rescaled-variable. The parameters are the elasticities of the recreation index with respect to individual lake characteristics, measuring the sensitivity of the recreation index to changes in the size of the lake, water quality, and the travel distance. For example, a 1 percent increase in water quality at a lake increases the recreation index by γ_2 percent. The parameters affect the pattern of how the recreation index changes with lake features. If the index parameters equal 1, it means the size of the lake, the distance to lake, or water quality contribute to the recreation index proportionately. If a lake feature has an elasticity greater than 1 and increases at a constant rate, its contribution to the recreation index will first increase in a lower rate and gradually change to a higher rate after the median point. An index parameter less than 1 indicates the opposite direction from a high rate to a low rate. The value of the index parameter determines the turning point and how sharp this transition will be. In addition, we can understand the tradeoff between any two of the variables required to keep the recreation index and the utility at the same level. Taking the size of the lake and water quality as an example, the marginal rate of substitution between quality and size:

$$\frac{MU_q}{MU_s} = \frac{\partial s}{\partial q} = \frac{\gamma_2}{\gamma_1}\frac{s}{q}\,. \tag{2.11}$$

As shown in formula (2.11), the marginal rate of substitution depends on reference values. We can either evaluate it at a specific point or the average level.

2.3.2 Macro Model

To explore the structure of the recreation index with lower computational burden and a richer dataset. I also develop an aggregate model by dropping the components with household characteristics from the utility function. The aggregate model assumes homogeneous preferences over housing type attributes, so that the parameters α_{kA} all equal to 0, which makes Γ_h^i 0 as well. With this assumption, the indirect utility function can be rewrite as:

$$V_h^i = \Theta_h + \varepsilon_h^i\ . \tag{2.12}$$

By maintaining the extreme value distribution assumption, the equality between the predicted market share and the true observed market share still holds:

$$\widetilde{S_h} = \frac{e^{\Theta_h}}{\sum_{h'} e^{\Theta_{h'}}} = S_h\ . \tag{2.13}$$

Taking log of both sides, I obtain a linear equation with the average utility Θ_h as the unknown parameter for each housing type h:

$$logS_h = \ \Theta_h - \log\left(\sum_{h'} e^{\Theta_{h'}}\right)\,. \tag{2.14}$$

Normalizing Θ_1 as 0 and subtracting housing type *1* from housing type h for all other h I obtain

$$\Theta_h = \ logS_h - \ logS_1. \tag{2.15}$$

Then, I plug the specification for average utility Θ_h from equation (2.5) back in and estimate the average impact of housing attributes with the linear regression.

2.3.3 Instrument variables

The average impacts of housing attributes are estimated from a linear regression in both the micro and aggregate models. However, the OLS estimation may be biased due to price endogeneity. Properties with preferred attributes have a higher price. It is likely that the unobserved housing attributes are correlated with the housing price. I need an instrumental variable to deal with the endogeneity problem. In the industrial organization literature for differentiated products, a common approach proposed by Berry, Levinsohn, &Pakes (1995) is to use the same characteristics of other products from the same firm and the sums of the values of the same characteristics of products from competitor firms to build the instrument variables. In addition, Nevo (2000) develops an approach of using the prices of the brand in other cities and quarters as instrumental variables. One important assumption with this method is that the utility of consuming a product depends only on the characteristics of that product, and the instrumental variables affect the product price either through the market equilibrium (Berry, Levinsohn, &Pakes 1995) or the common marginal costs (Nevo 2000), but do not influence the market-specific valuation. The same logic can be applied in the housing market. The utility of a housing type is only determined by its own attributes and those in nearby neighborhoods. But its price is correlated with attributes of distant neighborhoods through the common equilibrium structure. With these assumptions, I build my instrumental variable from attributes of distant housing types in the same market. Following Bayer & Timmons (2007) and Klaiber and Phaneuf (2010), I first move the price term to the left side and rewrite equation (2.5) as:

$$\Theta_h - \alpha_p^* p_h = \alpha_q q_l \cdot d_b + \alpha_d d_b + \alpha_d r_c + \alpha_X \widetilde{X_h} + Cnty_j + \xi_h \tag{2.16}$$

where I start with a plausible guess of the value of α_p and label it as α_p^*. In practice, I use housing price parameter from a second stage estimation without the instrument variables. I also add additional attributes of housing types within 0-1km, 1-3km, and 3-5km rings around the location of housing type h along with its own attributes and denote them as $\widetilde{X_h}$. Then I estimate equation (2.16) using OLS to obtain all the parameters on the right-hand side. After controlling for the attributes of in-location and nearby housing types, the residuals are assumed to be exogenous. Next, I set $\xi_h = 0$ and set up the following equations that satisfy the market clearing conditions:

$$S_h = \frac{1}{N}\sum_i^N \frac{e^{\widehat{\Theta_h}+\widehat{\Gamma_h^l}}}{\sum_{h'} e^{\widehat{\Theta_{h'}}+\widehat{\Gamma_{h'}^l}}} \qquad \forall h = 1, \dots, H \, , \tag{2.17}$$

$$\widehat{\Theta_h} = \widehat{\alpha_p} p_h^{iv} + \widehat{\alpha_q} q_h \cdot d_h + \widehat{\alpha_d} d_h + \widehat{\alpha_r} r_h + \widehat{\alpha_X}\widetilde{X_h} + Cnty_j \, , \tag{2.18}$$

where the $\widehat{\Theta_h}$ is the predicted value from equation (2.16), $\widehat{\Gamma_h^l}$ is calculated from the first-stage estimates, and p_h^{iv} is the unknown variables to be solved. As I include the in-location and nearby attributes in equation (2.18), the predicted price p_h^{iv} captures all the residual variation from distant housing type attributes beyond the 5km buffer. I assume the indirect utility are not affected by attributes from distant neighborhoods, so that the constructed instrument variables p_h^{iv} will only affect Θ_h through its correlation with the price index variables. Once I estimate equation (2.18) with the instrument variables p_h^{iv}, I update the coefficient of price index α_p^* and repeat the whole process until α_p^* converges.

2.4 Data

My main data is collected from three datasets. The property transaction data is obtained from the Zillow's Transaction and Assessment Database (ZTRAX). ZTRAX is the largest nationwide real estate database created by the Zillow company. It includes more than 400 million records of individual property transaction in more than 2750 counties.[20] This dataset includes two subsets. The assessment data contains housing structural variables like the lot size, the number of rooms, and the number of bedrooms. It also has the coordinates of properties that can be used for geolocation and spatial analysis. The transaction

[20] For more information on the ZTRAX dataset, please visit: https://www.zillow.com/research/ztrax/

data has information on the sale prices, the sale dates, and mortgages. A property with repeated sales may appear multiple times in the transaction data but only once in the assessment data. There is a variable that can uniquely identify each property in both subsets. With this unique identifier, I merge the two subsets and obtain a cross-sectional transaction dataset with repeated sales.

The household characteristics are collected from the Home Mortgage Disclosure Act dataset. The HMDA data was enacted by Congress in 1975 and was implemented by the Federal Reserve Board's Regulation C, which requires lending institutions to report public loan data.[21] This provides loan-level information for home mortgages lending activities starting in 2007 in the US. Each observation includes the annual income, race, and gender of the mortgage applications who are considering a home purchase. Also, it provides information on the mortgage loan amount, the lender's name, and the census tract of the property.

For lake water quality I use Secchi depth data obtained from the Wisconsin Department of Natural Resource (DNR). It has more than 300,000 records of Secchi depth in Wisconsin lakes from 2006 to 2019. The Secchi depth data comes from two main resources. One is the Citizen Lake Monitoring Network for which volunteers across the state measure water quality using the Secchi disk.[22] The other is a satellite monitoring program that collects water clarity data for lakes across the state.[23] Using a model developed by the University of Wisconsin-Madison Environmental Remote Sensing Center (ERSC), scientists estimate the Secchi depth value from satellite images in lakes greater than 5 acres. A lake may have multiple records at different times and monitor sites from both programs.

Wisconsin has over 15,000 lakes statewide. The characteristics of lakes are also obtained from Wisconsin DNR. The DNR data has information on the size and location of lakes, as well as whether a lake has public lands, public parks, and public beach around it. The shapefile of lakes is from the 24K

[21] For more information on the HMDA dataset, please visit: https://www.ffiec.gov/hmda/

[22] For more information on the Citizen Lake Monitoring Network, please visit: https://dnr.wisconsin.gov/topic/lakes/clmn

[23] For more information on the satellite monitoring program, please visit: https://dnr.wisconsin.gov/topic/lakes/satellitemonitoring.html

DNR hydrography database, which includes information about surface water features represented on the United State Geological Survey (USGS) 1:24,000-scale topographic map series. I merge the lake characteristic data with the spatial data using a unique waterbody identification code that is included in both datasets. I exclude lakes less than 50 acres to make sure the lakes are large enough to provide amenity and recreational values. This reduces the number of lakes to 2113.

2.4.1 Machine learning for missing water quality data

My study period is from 2011 to 2016, during which the Secchi depth data has records in 2029 lakes greater than 50 acres. However, the number of water quality records per lake varies from 1 to 988. Limited records may not be sufficient to represent water quality in a lake. Even for lakes with several records, if all the records cluster around 1 year during our 6-year study period, they may also falsely represent the average water quality during the whole study period. Since water quality usually changes gradually overtime, I do not require Secchi depth records for every year of my analysis. Instead, I considered all records in 2011-2013 as *period1* and records in 2014-2016 as *period2*. I select valid water quality data at the lake-by-period level by excluding lake-by-period units with less than 3 observations. If all the units have valid water quality data, I would have 4226 indicators of water quality in 2113 lakes over the 2 aggregate periods, after combining Secchi depths at the lake by period level. However, I only end up with 3585 lake-period units with valid water quality data in 1993 lakes. Among the 2113 lakes, some only have valid water quality data in one period and a small proportion of them do not have any valid data in both periods.

Missing values are a common problem in environmental datasets. I cannot drop lakes with missing data because in our sorting model, the choice set is constructed based on the location of lakes. The distribution of available housing types in the choice set will change if I drop lakes from the original list. To solve the problem, I use machine learning to predict missing values for the water quality data. In my data, the unit of observations is lake by period. The dependent variable at each row is the average value of Secchi depth per period-by-lake. The attributes of the model cover the location and characteristics of lakes. First, I created a series of dummy variables to indicate the period of the record and the county of the lakes. Then, I include the size of the lake, and three additional dummy variables to indicate whether the lake has public parks, beach, or lands nearby. In addition, water quality in a lake is affected by the surrounding landscapes. I collect the level-1 land cover division from the Wisconsin DNR and merge it with lakes in our dataset. I compute the percentage of eight land use types in a 10km buffer area around each lake and use them as additional attributes.[24]

Next, I need to choose a specific machine learning model for the estimation. Logistic regression, regression trees, random forest regression, and neural networks are common methods to estimate continuous dependent variables. Due to the limited sample amount, I cross validate the models mentioned above on the subset of the lake-by-period units with valid Secchi depth values. Cross-validation is a resampling procedure used to evaluate the performance of a machine learning model.[25] In my case, I am using the 10-fold cross validation approach. It splits the full dataset into 10 groups. I alternatively take one group as the test set and the rest of data as the training set. For each iteration, I train the model on each training set and evaluate it on the test set. I use three performance metrics to evaluate the performance of models: the minimum squared error, the mean absolute percentage error, and the percentage of observations whose difference between the predicted and the real value is less than 2 feet. I compute the average of the performance metrics over 10 integrations and compare them among the models listed above.

The random forest regression model performs the best when evaluated with different metrics. So that I train the model with the subset of lake-by-period units with Secchi depth data and predict for the units do not have valid water quality data. Random forest regression is an ensemble algorithm that fits several regression trees on different sub-samples of the dataset.[26] The training from multiple regression

[24] The eight land use types are urban, agriculture, grassland, forest, open water, wetland, barren, and shrubland.

[25] For more information on cross validation, please visit: https://scikit-learn.org/stable/modules/cross_validation.html

[26] For more information on the random forest regression, please visit:

trees with different samples gives stability to the model. A regression trees is a decision tree that is used to predict continuous variables.[27] I use the machine learning algorithm as a preferred alternative to using sample averages. For sample averages, it is unknow at which level to take the average. For example, when a lake does not have valid water quality data in period 2, I have several options to replace the missing value such as the average value in period 1, the average Secchi depth at the county level in period2, or the average Secchi depth of lakes with similar land covers. The decision tree regression uses the tree structure that starts with a root node and ends with a decision made by leaves. Starting from the root, it breaks down the training dataset into smaller subsets based on the attributes mentioned above. The rules of splitting the data are determined in the training process to minimize the variation of the dependent variable in the subsets. Then the model predicts the value in each subset by taking the average of dependent variables in the training dataset. The idea is like the approach of using the average in a larger spatial or temporal scale to replace the missing data, but with a higher accuracy level. If I compare the predicted value from the random forest model with the average value at the county-by-period level, the mean absolute percentage error decreases by 24%.[28]

I aggregate the Secchi depth data at the lake level. The average Secchi depth over the 6-year study period and the size of lakes are summarized in Table 2.1. I observe large variation in the size of lakes. The largest lake among the 2,113 lakes is more than 130,000 acres while the size smallest lake is 50 acres, which I defined as the threshold for a lake to provide economic benefits. The average water quality is a Secchi depth value of 2.53m. In addition to the full set of lakes, I also summarize the same variable for recreation lakes, which are defined as those with at least one lake amenity like public park, public beach, and public land. The average lake size is 681 acres among recreation lakes, larger than the average value of 427 among all lakes. The average Secchi depth is slightly smaller, at 2.47m.

https://scikit-learn.org/0.16/modules/generated/sklearn.ensemble.RandomForestRegressor.html

[27] For more information on decision trees and regression trees, please visit: https://www.analyticsvidhya.com/blog/2021/08/decision-tree-algorithm/

[28] The ratio of splitting the training and test data is 6:4 for this calculation.

2.4.2 Estimation of housing type price

The sale prices for individual transactions from the ZTRAX data cannot be used directly in our sorting model. I need to estimate the housing price for each housing type in the choice set. To estimate these prices, I follow Klaiber & Phaneuf (2010) and use a log-linear regression:

$$LnP_h^i = \rho_h H_h^i + y_t^i + \epsilon_h^i \tag{2.19}$$

where P_h^i is the sale price of a property in the housing type h that is defined by the distance bin b of lake l in census tract c, H_h^i is a series of dummy variables indicating each housing type, and ϵ_h^i is the regression error. I also include year fixed effects y_t^i to control for systematic changes in home prices over time. The price estimator is calculated from the estimated parameters of housing type dummy variables following the equation:

$$p_{blc} = e^{\rho_{blc}} . \tag{2.20}$$

The unique price estimate for each housing type is used in the sorting model as the price variable in the utility function.

2.4.3 Merge ZTRAX data with HMDA data

To estimate heterogeneity in household preferences, I need to merge the ZTRAX data and the HMDA data to get the sale and household information for each transaction. The HMDA data is organized by year. I merge the HMDA data with the ZTRAX data individually for every year in 2011-2016. There are two challenges in the merging process. First, there is not an identification code in both datasets that can uniquely identify a property. Following the strategy from past studies (Bayer et al. 2016; Bishop & Timmins 2018), I use four common variables in both datasets to merge data. These include county identifiers, census tract numbers, loan amounts, and the mortgage lender’s names. The second challenge is that the ZTRAX data does not have a respondent identifier for each transaction, but rather a lender’s name that corresponds to each respondent identifier. Bernstein et al. (2021) proposes matching based on unique occurrences of loan amounts and census tracts to create a correspondence between respondent id and lender name based on loan amounts, census tracts, and lender names. However, the number of transactions that can be uniquely identified with the three variables is limited, and we may not be able to get all the matches between the lender's name and correspondent identifier from it. Instead, I collect the financial institution dataset from the Consumer Financial Protection Bureau, which includes information on the respondent identifier and the lender's name. One caveat of using the institution data is that lenders' names in ZTRAX are not fully matched with CFPB data. Pattern changes like replacing 'Credit Union' with 'CU' is necessary.

My merging algorithm is derived from a national project of merging the ZTRAX dataset and the HMDA dataset completed by Stephen Billings.[29] I have tried different algorithms, and the following strategy gives the best outcome. First, I clean both datasets and only keep arm-length transactions with non-zero mortgage values and define a unique identifier *unique_id* for each transaction in the ZTRAX data. Second, I combine the county identifier, census tract number and the loan amount as a single variable named *mergeid1* in both the ZTRAX and HMDA datasets. Then, I merge these two datasets with *mergeid1,* and select the subset with uniquely merged observations as the first part of the output, which I denote as *merged_part1*. The rest of the merged data excluding the uniquely merged subset is used for the next step, which is denoted as *merge_part2*. Third, I merge the dataset *merge_part2* with the institution data with the lender's name and excluding the observations with different values of the respondent identifier from the HMDA and the institution data. In this way, I add additional information to the current *mergeid1* to further merge observations that cannot be uniquely identified by *mergeid1*. Fourth, I remove observations in *merge_part2* that have duplicate *unique_id* to guarantee every observation in the second part of the merged data is uniquely identified. The last step is to combine the datasets *merge_part1 and merge_part2* so that we have a transaction dataset with household characteristics.

The merging rates of my data for each year are summarized in Table 2.2. Since the loan amount is one of the variables to identify the transactions, I am only able to merge transactions with positive loan

[29] The code and report are posted on his website: https://sites.google.com/a/colorado.edu/stephen-billings/code

amounts. I summarize both merging rates with and without counting transactions with zero loan amount in the ZTRAX data. When I only count transactions with mortgages in the denominator, the merging rates are higher (and similar) in recent years from 2014 to 2016. Also, the merging rate with the entire sample increases over years. This is consistent with the increasing mortgage rate as shown in the fifth row, which records the percentage of transactions with mortgages. Past studies have successfully merged the HMDA data with multiple property transaction datasets. The merging rate ranges from 40% to 70% (Bayer et al. 2016, Tra et al. 2013, Bishop and Timmins 2018, Bernstein et al. 2021). This is consistent with my results. To compare the entire transaction datasets with the subsets that have household characteristics, I plot the distribution of the home price and the lot size before and after the merge in Figure 2.1 and Figure 2.2.[30] Both variables have higher median values in the merged data than the whole data. Also, the percentages of transactions on the lower side of the values of two variables are higher in the whole data. It suggests that houses in the merged dataset tend to have a higher quality. This is under my expectation because purchasing a house with a higher price is more likely to require a mortgage. However, I think the discrepancy between the two datasets is acceptable as the shapes of the distributions are still similar. Therefore, the merged subset provides a reasonable representative of housing transactions happened during our study period.

Recall that I do not need household information for the aggregate model. Therefore, I aggregate the full transaction dataset at the housing type level depending on the census tract of the property, the nearest lake to the property, and whether the property is in the lakefront or non-lakefront area. After merging the aggregate data with data on water quality and the housing type prices, I generate a choice set of 4,849 housing types. The summary statistics of the housing type attributes are reported in Table 2.3. All the variables are rescaled for computational purpose. I am interested in three main variables: the lakefront dummy variable, the interaction of the lakefront dummy variable and the Secchi depth, and the recreation index that is constructed from equation (2.10). In addition, I also include other attributes at the

[30] I remove the upper and lower 5 percent of the observations to make the plots clear.

census tract level, which includes the ratio of white household, the unemployment rate, the number of families with kids, the medium income level, the percentage of households below the poverty line, and the medium home value.

Moving to the micro model, I generate a smaller choice set by aggregating the subset of the transactions data that is merged with the household data, which gives me 2,837 housing types. In addition, I also extract information on household characteristics and the housing type chosen for each transaction from the merged dataset. The housing type attributes, and household characteristics are summarized in Table 2.4. As shown in the panel B of Table 2.4, the household income varies largely from $22,000 to $961,000, while the race of the household does not have much variation with 95 percent of them as white. Comparing Table 2.3 and Table 2.4, I do not find evidence that the housing type in one dataset is better than the other. However, I see differences in some variables. For example, the average housing price is higher in the micro-data, but the percentage of lakefront housing type is higher in the macro-data. The variation in the distribution of housing attributes in these two datasets may affect the estimation results of the sorting model.

2.5 Results

2.5.1 Macro model

I first estimate the aggregate model with the dataset of 4,849 housing types. My main interest focuses on three variables: the lakefront dummy variable, the interaction of the lakefront dummy with Secchi depth, and the recreation index. I also include other housing type attributes as shown in Table 2.3 as additional control variables. Before estimating the model, I need to create the recreation index by defining its parameters. I make two assumptions on how the features of a lake affect its ability to provide recreational amenities: First, a higher value of a lake feature (ex: larger surface area, shorter travel distance to the lake, higher water quality) positively contributes to the recreation index up to certain threshold. For example, people are likely to prefer a 5000-acre lake over 500-acre lake. However, when the lake is large enough, additional size may not be an important factor in recreational or residential decisions. Second, I assume the three quality dimensions differentially affect the recreation index, which is reflected in the different values of the index parameters $\gamma_1, \gamma_2, \gamma_3$ in equation (2.10). To implement these two assumptions, I first set a threshold for each lake feature according to its distribution.[31] Then, I replace any value beyond the threshold with the same value of the threshold. Next, I examine different structures of the recreation index by setting the distance parameter as 1 and varing the other two parameters. Recall that the parameters represent the elasticity of each lake characteristic for the recreation index. Relative to distance and water quality, the size of lake is more likely to affect the recreation index in a non-linear way. I assume that the elasticity of size for the recreation index is larger than the elasticities of distance and water quality. Some aquatic activities require more surface areas, so small lakes may have less recreational value. As the size of the lake remains in the lower range, a slight increase in lake size may still contribute to its recreational value. I expect to see a significant impact from the increase of the lake size on the recreation index when the size is large enough. As for the water quality and the distance to lake, I assume their impacts on the recreation index are more likely to be linear with an elasticity close to 1. Therefore, I start by assigning a larger index parameter on the size of the lake while keeping the other two parameters low. I examine the macro model with different combinations of $\gamma_1, \gamma_2, \gamma_3$, with selected results summarized in Table 2.5.

All the models are estimated by IV. The coefficients of the main parameters except for the recreation index are robust across different values of the recreation index. As expected, the coefficient on price is negative and significant, suggesting a negative impact of the home price on utility. The coefficients on both the lakefront dummy and the interaction of the lakefront dummy with the Secchi depth in the nearest lake are positive and significant. The results show that, controlling for price, residents prefer to live close to lakes and their utility increases with the water quality level in the nearest lake. Dividing the coefficients on the lakefront dummy variable and the interaction variable by the coefficient

[31] The thresholds for the size of the lake, the distance to lake, and the Secchi depth value are 10,000 acres, 0.1km, and 6m, respectively.

on the price index, I estimate the proximity premium and the marginal value of water quality. On average, people are willing to pay ~$50,000 more to purchase the type of houses within 100m of a lake in which the water quality is at the average level of 2.16m.[32] For lakefront households, the MWTP for a 0.1m increase in Secchi depth of the nearest lake ranges from $1,799 to $1,812. Since we are evaluating the MWTP based on the housing price at the time of purchase, the additional price they are willing to pay is a one-time payment for a perpetual improvement in water quality.

The hedonic literature has provided strong evidence on the amenity value of water quality for lakefront households. However, this is the first study to estimate the amenity value of water quality using the residential sorting model. For comparison, a recent hedonic study by Moore et al. (2020) estimates a change of $3971 in the average home price with a 0.1m change in Secchi depth. This impact is on properties within 0.1 mile of their nearest lake. A meta-analysis of property values and water quality (Guignet et al. 2021) suggests that a one-percent increase in water clarity leads to 0.19 percent increase in waterfront home prices. In our study, the average Secchi depth value in 2,113 lakes is 2.54m, and the average housing price for properties within 100m of their nearest lakes is $313,760. The elasticity from Guignet et al. (2021), using our data suggests a home price increase of $2,347 for a 0.1m improvement in Secchi depth. Compared to the hedonic model that considers water quality as individual housing characteristics, the sorting model addresses water quality as a housing type attribute. Even with different settings and model structures, my findings are consistent with the lower end of the value found in the hedonic literature. The comparison results from the sorting model with results in the hedonic literature also provides some evidence of convergent validity on the amenity value of water quality.

Moving to the estimates of the recreation index, the coefficient varies with different index parameters. As the initial exploration of the recreation index, I hope to understand how these three variables jointly affect the recreation value generated from lakes. Instead of estimating the parameters from the data, I simplify the problem by assigning different values to the index parameters. Column (1)

[32] This average is calculated based on the Secchi depth variable for each housing type.

shows a positive and significant coefficient while keeping the parameter of the lake size at a high level of 6.[33] When I decrease the parameter for the lake size in columns (2) – (4), the coefficient of the recreation index is still positive but not statistically significant. The magnitude also decreases when I lower the parameter. When I assign high values for the size and distance parameters, I find no significant impact as shown in columns (5), (6), and (7). In addition, I have also tried to assign a parameter less than 1 on the size and water quality variables, and the coefficients are also statistically insignificant. Among these functional forms for the recreation index, the ones that put a high weight on the size of lake are more likely to provide significant and intuitive estimates. For the recreation index in column (1), a 1 percent increase in lake size leads to a 6 percent increase in the recreation index, while the elasticity of distance and water quality is much smaller at 1. In addition, the ratio of parameters also reflects the trade off a household makes to maintain the same utility level. The high ratio of the size parameter and water quality parameter suggests that households are more sensitive to the changes in water quality, but the travel distance affect utility in the same way as water quality.

Some intuition on what these parameters represent is available from past studies. Egan et al. (2009) estimates the recreation demand for water quality with a repeated mixed logit model using survey data on visits to lakes in Iowa. They define the recreation trip utility as a function of features of lakes, and multiple water quality measurements, which includes Secchi depth. In their model, the marginal utility of Secchi depth is 2.4, which is the coefficient on the variable. Meanwhile, the marginal utility of the size of the lake is determined by both the coefficient and the value of the size because it is in log form. The average size of the lake in the study area is 672 acres, which leads to a marginal utility of 0.0070 on average. Taking the fraction of the marginal utility of two variables, I obtain the marginal rate of substitution between Secchi depth and the size of the lake, which is 343. This means that for every 1m decrease in Secchi depth, the size of the lake needs to increase by 343 acres to maintain the same utility level. Recall that in my data the average size of a lake is 681 acres, and the average Secchi depth is

[33] I also try higher values (>6) of the index parameter for the lake size, the coefficients are also significant and positive.

2.47m. Following equation (2.11), the marginal rate of substitution in my study is 46. To keep utility at the same level, for the same change (1m) in water quality, my result requires a smaller change of 46 acres in lake size. Note that the average Secchi depth in my study is more than twice what is in Egan et al (2009). With lower water quality in their case, it makes sense that a larger decrease in water quality needs a higher trade-off with the size of the lake. If I define the recreation index to have the same marginal rate of substitution as Egan et al (2009), the parameters for lake size and Secchi depth should be similar. In this case, the coefficient on the recreation index loses its significance and becomes smaller, even negative, as is shown in columns (4)-(6) of Table 2.5. It is not possible to fully compare my results with theirs because of different model structures and datasets. But both results support the importance of the size of the lake while estimating the recreation demand of water quality. Meanwhile, the difference in the results of two studies illustrates the flexibility and complexity when modeling the recreation index. For example, if I simply borrow the results from Egan et al. (2009) to define the formula for my recreation index, I end up with zero impact of the recreation index on household utility. With my current analysis, although I cannot claim the formula used in column (1) of Table 2.5 provides the best estimate of the recreation index because the parameter does not come from causal analysis, I gain insights on possible ways of modeling the recreation index – by considering the tradeoff between water quality and other lake features. As for specific formula, the robustness of my findings needs to be examined in different scenarios.

Figure 2.3 shows the spatial distribution of the recreation index from column (1), which suggests a large variation across census tracts. In addition, I also plot the location and size of all the recreational lakes in Figure 2.4 and the distribution of water quality in lakes in Figure 2.5. Census tracts with a large recreation index cluster in the northwest, south, and central regions. We can see how different lake characteristics and the richness of lakes affect the recreation index. The northwest region has many medium sized lakes on both the west and east sides, and a fair number of lakes have high water quality. As we move to the northeast side where the number and size of lakes decrease, the recreation index becomes smaller even though there are more lakes with high water quality. Meanwhile, the high recreation index in the south and central Wisconsin results from multiple extremely large lakes in these regions. The southeast region has some large lakes but most of them has low water quality, which results in a smaller recreation index compared to the south region.

The recreation index of a housing type is the sum of recreation index for every lake in the recreational area around the census tract of this housing type. With the formula and model, I can calculate the marginal recreational value of water quality improvement for an individual lake. Using the estimates from column (1) as an example, if there is a median-size lake with a median travel distance from the census tract of a housing type, the one-time MWTP at the purchase time for a 0.1m increase in the Secchi depth is $6.[34] Meanwhile, if the size of the lake is at the 0.9 quantile, the MWTP increases to $200.[35] To better present the marginal recreational value of water quality, I plot the MWTP for a 0.1m increase in the Secchi depth with different values of the lake size and the distance in Figure 2.6. The size of the bubble represents the value of the MWTP, which ranges from a minimum value close to $0 to the largest value of $752 where the largest lake has the closest distance to a housing type.[36] The figure also shows a sharp increase when the size of the lake is beyond the medium value, which is expected from the large elasticity of lake size.

2.5.2 Micro model

The macro model only estimates the average impact of housing type attributes on utility, without differentiating the impact across households. Next, I estimate the residential sorting model assuming heterogeneous preferences across households with different races and income levels. Compared to the utility function in the macro model, the heterogeneity is addressed by the additional interaction terms of household characteristics with two housing type attributes -- the lakefront location and the Secchi depth value in the nearest lake. I follow the same formula in the column (1) of Table 2.5 to construct the

[34] This result is calculated by (5.11*0.5^6*0.5^1*(1/6)^1)/11.32*100000*0.1

[35] This result is calculated by (5.11*0.9^6*0.5^1*(1/6)^1)/11.32*100000*0.1

[36] This result is calculated by (5.11*1^6*1^1*(1/6)^1)/11.32*100000*0.1

recreation index.[37] The results from the first-stage estimation are summarized in the Panel A of Table 2.6. All the coefficients are statistically significant and consistent with our assumptions. The positive coefficients on the interactions with household income suggests wealthier people are willing to pay more to buy a house in a lakefront neighborhood and for a higher Secchi depth value in the lake where the neighborhood is located. Similarly, the interactions with the dummy variable indicating race shows that white households have a higher preference for the lakefront neighborhoods and for better water quality. In addition to the interaction parameters, I also estimate the mean utility level for each housing type and use it for the second stage estimation.

The results from the second stage estimation are summarized in Panel B. I start with the OLS regression. As shown in the first column, many parameters of the housing type attributes are either not statistically significant or counterintuitive. For example, the negative and significant coefficient of the Secchi depth suggests lakefront residents prefer to living on lakes with lower water quality, which contradicts intuition. Following equations (2.16) – (2.18), I estimate the IV model using housing type attributes in the distance neighborhoods. I see some changes in the counterintuitive parameters from the OLS regression. The coefficient on Secchi depth is positive and significant, suggesting a positive impact on utility. For households in the lakefront neighborhoods, the average part of the one-time MWTP at the time of purchase for 0.1m perpetual increase in the Secchi depth value is $2,413. The coefficient of the lakefront dummy variable is still negative but loses its significancy. The coefficient on the recreation index is negative and insignificant in both the OLS and IV models. Recall that the total number of housing types decreases from 4,948 to 2,837 after merging the transaction data with household characteristics. Therefore, the limited data may fail to identify the proximity premium for lakefront neighborhoods. In addition to the main variables, the coefficients of other housing type attributes also change when I apply the IV estimator. For example, the coefficient on the median home value at the census tract level becomes positive with a larger magnitude. Because a higher medium home at the census

[37] For the baseline model, I set $\gamma_1 = 6,\ \gamma_2 = 1,\ \gamma_3 = 1$. I have also examined other combinations of the parameters. The estimated coefficients of other variables remain consistent.

tract level usually means better amenities in this census tract, the neighborhoods in this census tract may also benefit from it. The difference between the OLS and the IV model supports our hypothesis of price endogeneity. The rational results from the IV model also show the usefulness of our IV approach.

With both the results from the first and second stages, I can estimate the MWTP for a specific group of households. For example, a white household with $50,000 annual income that lives in the lakefront area is willingness to pay $2,490 for a 0.1m increase in the Secchi depth in the lake.[38] If annual income increases to $200,000, the MWTP increases to $2,587.[39] A non-white household is willing to pay less even with the same income level. For a non-white household that earns $200,000 per year, the marginal value of a 0.1m increase in Secchi depth is $2,541[40]. The results do not show much heterogeneity across race and income level. Recall that more than 95 percent of the households in our data are white. In addition, the lakefront average household income ($175,633) is twice of the non-lakefront average income ($87,625), and the annual income level among lakefront households clusters more on the higher end. The distribution of the housing characteristics limits the variation in the household income.

2.5.3 Micro-Macro Model

In the macro model, I estimate the average effect of housing type attributes using the complete housing transaction data from ZTRAX. After merging with the HMDA data to obtain household characteristics, I lose more than 40% of the transactions data in the micro model. Following Berry, Levinsohn, &Pakes (2004), I now estimate a micro-macro model which allows me to use the full transaction data and the household characteristics. The procedure of the estimation is the same as the micro model. The only difference is that I use both the full (aggregate data) dataset and merged dataset (micro data) at different stages of the estimation. I identify the housing types and their attributes from the

[38] The result is calculated by (0.007*5+0.0486+2.6196)/10.8570*100000*0.1
[39] The result is calculated by (0.007*20+0.0486+2.6196)/10.8570*100000*0.1
[40] The result is calculated by (0.007*20+2.6196)/10.8570*100000*0.1

aggregate data, such that I still have 4,849 housing types in the aggregate data. To estimate the heterogeneity parameters α_{kA}, I use the micro transaction data that is merged with household characteristics. Using this approach, I can include more housing types in the aggregate data therefore have more variation to estimate the average effect of housing attributes. Meanwhile, I can still uncover the heterogeneity in household preference through the micro dataset.

In Table 2.7, I report the estimation results from both stages of the micro-macro model. As shown in Panel A, the coefficients of the interactions with the income level are still positive and significant, suggesting the existence of heterogeneity among households with different income levels. However, I do not find any correlation between the race of households and their preferences for both the lakefront location and high-water quality.

Moving to the second stage, the results estimated with the OLS regression still generate many counterintuitive results as with the micro model. After instrumenting for price, I see large changes in both the signs and magnitudes towards the expected way. The coefficient on Secchi depth changes from negative to positive and remains significant. It shows a positive preference for high water quality through the amenity mechanism for households in the lakefront area. In addition to Secchi depth, I also find evidence on the proximity premium for lakefront neighborhoods, which is different from the micro model but consistent with the macro model. To obtain the MWTP, I combine the results from the first and second stage estimation. Using the same example as in the micro model, a household with the annual income of $50,000 is willing to pay extra $52,313 to live in a neighborhood within 100m of a lake regardless of the race of the household.[41] A higher income level of $200,000 will increase the MWTP to $59,403.[42] For households living in the lakefront neighborhood, the MWTP for a 0.1m increases in the Secchi depth in the closest lake varies from $1,494 for a household that earns $50,000 a year to $1,774 for a household earns $200,000 a year.[43] Comparing to the results from the micro model, these results suggest

[41] This result is calculated by (0.007*5+2.0022+(0.019*5+1.4232)*2.16)/10.1629*100000

[42] This result is calculated by ((0.007*20+2.0022+(0.019*20+1.4232)*2.16)/10.1629*100000

[43] These results are calculated by (0.019*5+1.4232)/10.1629*100000*0.1 and (0.019*20+1.4232)/10.1629*100000*0.1

more heterogeneity in preference for water quality across households with different income levels. The coefficient of the recreation index is similar to the macro model with a slight change in the magnitude.

2.6 Discussion and conclusion

Comparing three models, the micro-macro model uses more information from both the transaction data and the household data than the other two models. The results also meet our hypothesis with the IV estimation. The average impacts from housing type attributes are similar in the macro and the micro-macro models as they both use the full version of the transaction data. While comparing with the micro model, some coefficients are not consistent. As the micro model uses the merged dataset with limited housing types, the estimation results from the micro and the micro-macro model are more convincing.

I find heterogeneity in household preferences towards lakefront neighborhoods and a higher Secchi depth level in their nearest lakes in both the micro and the micro-macro models. The micro model provides evidence on the heterogeneity over both income level and race. The micro-macro model, however, suggests that only the income level affects people's MWTP for living close to a lake with higher water quality. In the micro model, the coefficient of the lakefront dummy is insignificant. However, the coefficient of the interaction between dummy variables indicating the lakefront location and the race of white is high, suggesting a white household is willing to pay $13,376 more for a lakefront neighborhood than a non-white household.[44] Let's consider a white household with the annual income level of $100,000. The micro model estimates its MWTP for a lakefront neighborhood is $56,022.[45] Meanwhile, the micro-macro model estimates a MWTP of $54,682 for the lakefront premium[46]. Note that more than 95% of households in our data are white. The ratio of white households is even higher in the lakefront

[44] This result is calculated by (1.3658+0.0486*1.78)/10.8570*100000

[45] This result is calculated by (0.0548*10+1.3658-0.7055+(0.007*10+0.0486+2.6196)*1.78)/10.8570*100000.

[46] This result is calculated by (0.007*10+0.0006+2.0022+(0.019*10+1.4232)*2.16)/10.1629*100000.

neighborhood. Such that the estimation of the parameters for the race heterogeneity maybe biased due to insufficient variation. On the other hand, the additional MWTP of white household can represent the average effect in the micro model. For both the micro and micro-macro models, I find significant coefficient of the interaction variables with the income level. The annual household income ranges from $22,000 to $961,000. Calculating with the results from micro-macro model, the MWTP for the amenity value from a 0.1m increase in Secchi depth varies from $1,442 to $3,195. In other words, for every extra $10,000 a household earn each year, they would like to take $19 more for a 0.1m improvement in Secchi depth.

The interaction variable in the model helps me understand the amenity value of water quality for lakefront households. In addition, I also find evidence on the recreational value of water quality through the exploration of the recreation index. I need more analysis and evidence to provide a specific formula of the recreation index. But I can at least claim that water quality in a large area affects household utility in a different way from the ambient water quality in the nearest lake. A lakefront household may have different MWTPs for the amenity value and the recreational value from water quality improvement even in the same lake. For example, if a housing type is within the lakefront area of a large lake the size of which is at the 0.9 percentile, and the distance from its census tract to the lake is 5km. The MWTP for 0.1m increase in the Secchi depth value that contributes to the amenity benefit is $1,811. While the MWTP for the same change in water quality that contributes to the recreational benefit is $368. Although non-lakefront households are not treated by amenity benefit from the nearest lake, their utility is also affected by water quality through the same recreational mechanism as the lakefront households. Knowing the size of the lake and the distance to the census tract where a housing type locates, I can estimate the marginal recreational value of water quality for every lake in the 30km buffer area.

Our study measures the amenity and recreational values of water quality for different groups of households. I find strong evidence on the marginal willingness to pay from lakefront households for a higher amenity value associated with water quality improvement. Also, I find household preferences for lakefront proximity and water quality in the nearest lake of the lakefront neighborhood change with the household income level. Our initial exploration of the recreation index provides insights on the estimation of the recreational value of water quality. I develop a formula of the recreation index that allows me to estimate the MWTP for water quality improvement that contributes to the recreational value. However, further studies are needed to examine the robustness of the recreation index in other regions.

2.7 Bibliography

Bayer, Patrick, Fernando Ferreira, and Robert McMillan. "A unified framework for measuring preferences for schools and neighborhoods." *Journal of political economy* 115, no. 4 (2007): 588-638.

Bayer, Patrick, and Robert McMillan. "Tiebout sorting and neighborhood stratification." *Journal of Public Economics* 96, no. 11-12 (2012): 1129-1143.

Bayer, Patrick, and Christopher Timmins. "Estimating equilibrium models of sorting across locations." *The Economic Journal* 117, no. 518 (2007): 353-374.

Bayer, Patrick, Robert McMillan, Alvin Murphy, and Christopher Timmins. "A dynamic model of demand for houses and neighborhoods." *Econometrica* 84, no. 3 (2016): 893-942.

Bernstein, A., Billings, S. B., Gustafson, M., & Lewis, R. “Partisan residential sorting on climate change risk”. *Available at SSRN 3712665*,(2021).

Berry, Steven T. "Estimating discrete-choice models of product differentiation." *The RAND Journal of Economics* (1994): 242-262.

Berry, Steven, James Levinsohn, and Ariel Pakes. "Automobile prices in market equilibrium." *Econometrica: Journal of the Econometric Society* (1995): 841-890.

Berry, Steven, James Levinsohn, and Ariel Pakes. "Differentiated products demand systems from a combination of micro and macro data: The new car market." *Journal of political Economy* 112, no. 1 (2004): 68-105.

Bishop, Kelly C., and Christopher Timmins. "Using panel data to easily estimate hedonic demand functions." *Journal of the Association of Environmental and Resource Economists* 5, no. 3 (2018): 517-543.

Cao, Xiang, Kevin J. Boyle, Shyamani D. Siriwardena, and Thomas P. Holmes. "Estimating Demand for Urban Tree Cover Using a Residential Sorting Model." (2018).

Egan, Kevin J., Joseph A. Herriges, Catherine L. Kling, and John A. Downing. "Valuing water quality as a function of water quality measures." *American Journal of Agricultural Economics* 91, no. 1 (2009): 106-123.

Gibbs, Julie P., John M. Halstead, Kevin J. Boyle, and Ju-Chin Huang. "An hedonic analysis of the effects of lake water clarity on New Hampshire lakefront properties." *Agricultural and Resource Economics Review* 31, no. 1 (2002): 39-46.

Guignet, Dennis, Matthew T. Heberling, Michael Papenfus, and Olivia Griot. "Property values, water quality, and benefit transfer: A nationwide meta-analysis." *Land Economics* (2021): 050120-0062R1.

Hamilton, Timothy L., and Daniel J. Phaneuf. "An integrated model of regional and local residential sorting with application to air quality." *Journal of Environmental Economics and Management* 74 (2015): 71-93.

Kaoru, Yoshiaki. "Measuring marine recreation benefits of water quality improvements by the nested random utility model." *Resource and Energy Economics* 17, no. 2 (1995): 119-136.

Keeler, Bonnie L., Spencer A. Wood, Stephen Polasky, Catherine Kling, Christopher T. Filstrup, and John A. Downing. "Recreational demand for clean water: evidence from geotagged photographs by visitors to lakes." *Frontiers in Ecology and the Environment* 13, no. 2 (2015): 76-81.

Keiser, David A. "The missing benefits of clean water and the role of mismeasured pollution." *Journal of the Association of Environmental and Resource Economists* 6, no. 4 (2019): 669-707.

Klaiber, H. Allen, and Daniel J. Phaneuf. "Valuing open space in a residential sorting model of the Twin Cities." *Journal of environmental economics and management* 60, no. 2 (2010): 57-77.

Kuminoff, Nicolai V. "Decomposing the structural identification of non-market values." *Journal of Environmental Economics and Management* 57, no. 2 (2009): 123-139.

Kuminoff, Nicolai V., V. Kerry Smith, and Christopher Timmins. "The new economics of equilibrium sorting and policy evaluation using housing markets." *Journal of economic literature* 51, no. 4 (2013): 1007-62.

Kuwayama, Yusuke, Sheila Olmstead, and Jiameng Zheng. "A more comprehensive estimate of the value of water quality." *Journal of Public Economics* 207 (2022): 104600.

Leggett, Christopher G., and Nancy E. Bockstael. "Evidence of the effects of water quality on residential land prices." *Journal of Environmental Economics and Management* 39, no. 2 (2000): 121-144.

Lipton, Douglas. "The value of improved water quality to Chesapeake Bay boaters." *Marine Resource Economics* 19, no. 2 (2004): 265-270.

Moore, Michael R., Jonathan P. Doubek, Hui Xu, and Bradley J. Cardinale. "Hedonic price estimates of lake water quality: Valued attribute, instrumental variables, and ecological-economic benefits." *Ecological Economics* 176 (2020): 106692.

Nevo, Aviv. "A practitioner's guide to estimation of random-coefficients logit models of demand." *Journal of economics & management strategy* 9, no. 4 (2000): 513-548.

Phaneuf, Daniel J., V. Kerry Smith, Raymond B. Palmquist, and Jaren C. Pope. "Integrating property value and local recreation models to value ecosystem services in urban watersheds." *Land Economics* 84, no. 3 (2008): 361-381.

Poor, P. Joan, Keri L. Pessagno, and Robert W. Paul. "Exploring the hedonic value of ambient water quality: a local watershed-based study." *Ecological Economics* 60, no. 4 (2007): 797-806.

Sinha, Paramita, Martha L. Caulkins, and Maureen L. Cropper. "Household location decisions and the value of climate amenities." *Journal of Environmental Economics and Management* 92 (2018): 608-637.

Tra, Constant I., Anna Lukemeyer, and Helen Neill. "Evaluating the welfare effects of school quality improvements: A residential sorting approach." *Journal of Regional Science* 53, no. 4 (2013): 607-630.

Vesterinen, Janne, Eija Pouta, Anni Huhtala, and Marjo Neuvonen. "Impacts of changes in water quality on recreation behavior and benefits in Finland." *Journal of Environmental Management* 91, no. 4 (2010): 984-994.

Walsh, Patrick, Charles Griffiths, Dennis Guignet, and Heather Klemick. "Modeling the property price impact of water quality in 14 Chesapeake Bay Counties." *Ecological economics* 135 (2017): 103-113.

Wolf, David, and H. Allen Klaiber. "Bloom and bust: Toxic algae's impact on nearby property values." *Ecological economics* 135 (2017): 209-221.

Zhang, Congwen, and Kevin J. Boyle. "The effect of an aquatic invasive species (Eurasian watermilfoil) on lakefront property values." *Ecological Economics* 70, no. 2 (2010): 394-404.

Zhang, Jiarui, Daniel J. Phaneuf, and Blake A. Schaeffer. (2022). “Property values and cyanobacterial algal blooms: evidence from satellite monitoring of inland lakes.” *Ecological Economics*, forthcoming.

2.8 Figures and Tables

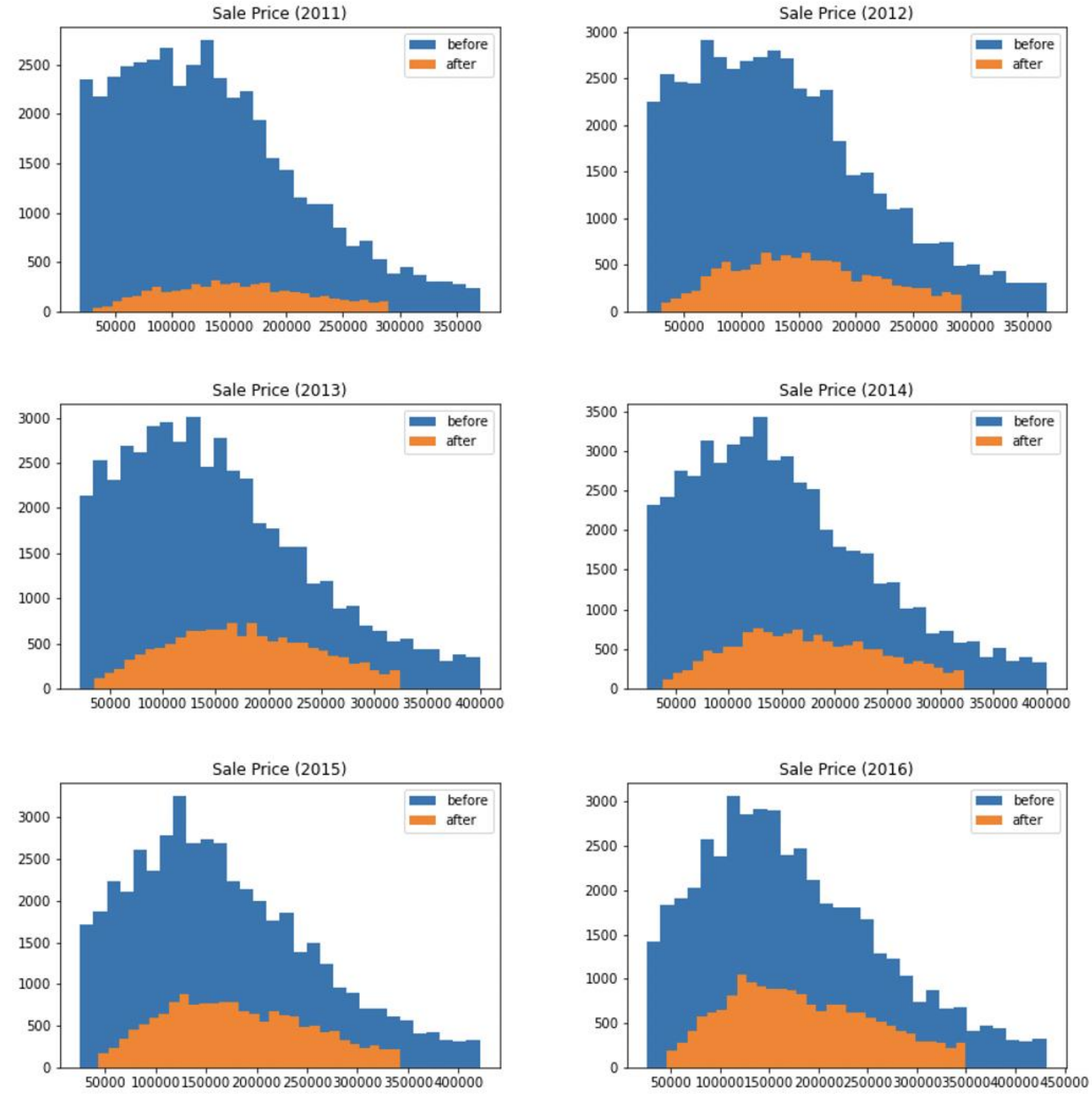

Figure 2.1: The distribution of sale prices ($) before and after merging. The sale price distribution of all the transactions in the ZTRAX dataset is plotted in blue and marked as ‘before’. The sale price distribution of transactions in the ZTRAX dataset that are merged with the HMDA dataset is plotted in orange and marked as ‘after’. The upper and lower 5 percent of the observations are removed for better visualization.

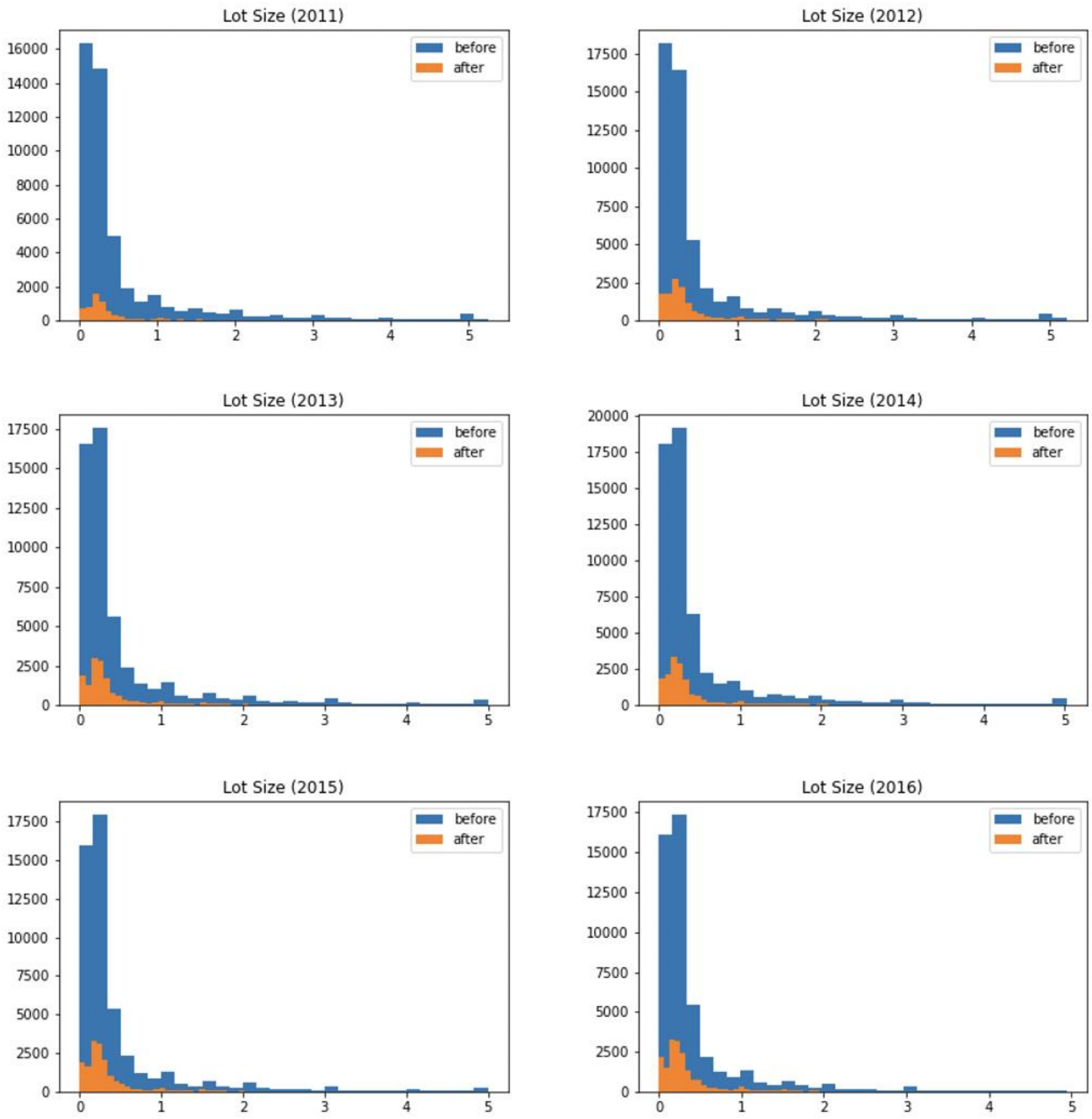

Figure 2.2: The distribution of the lot size (Acres) before and after merging. The lot size is measured in square feet. The sale price distribution of all the transactions in the ZTRAX dataset is plotted in blue and marked as ‘before’. The sale price distribution of transactions in the ZTRAX dataset that are merged with the HMDA dataset is plotted in orange and marked as ‘after’. The upper and lower 5 percent of the observations are removed for better visualization.

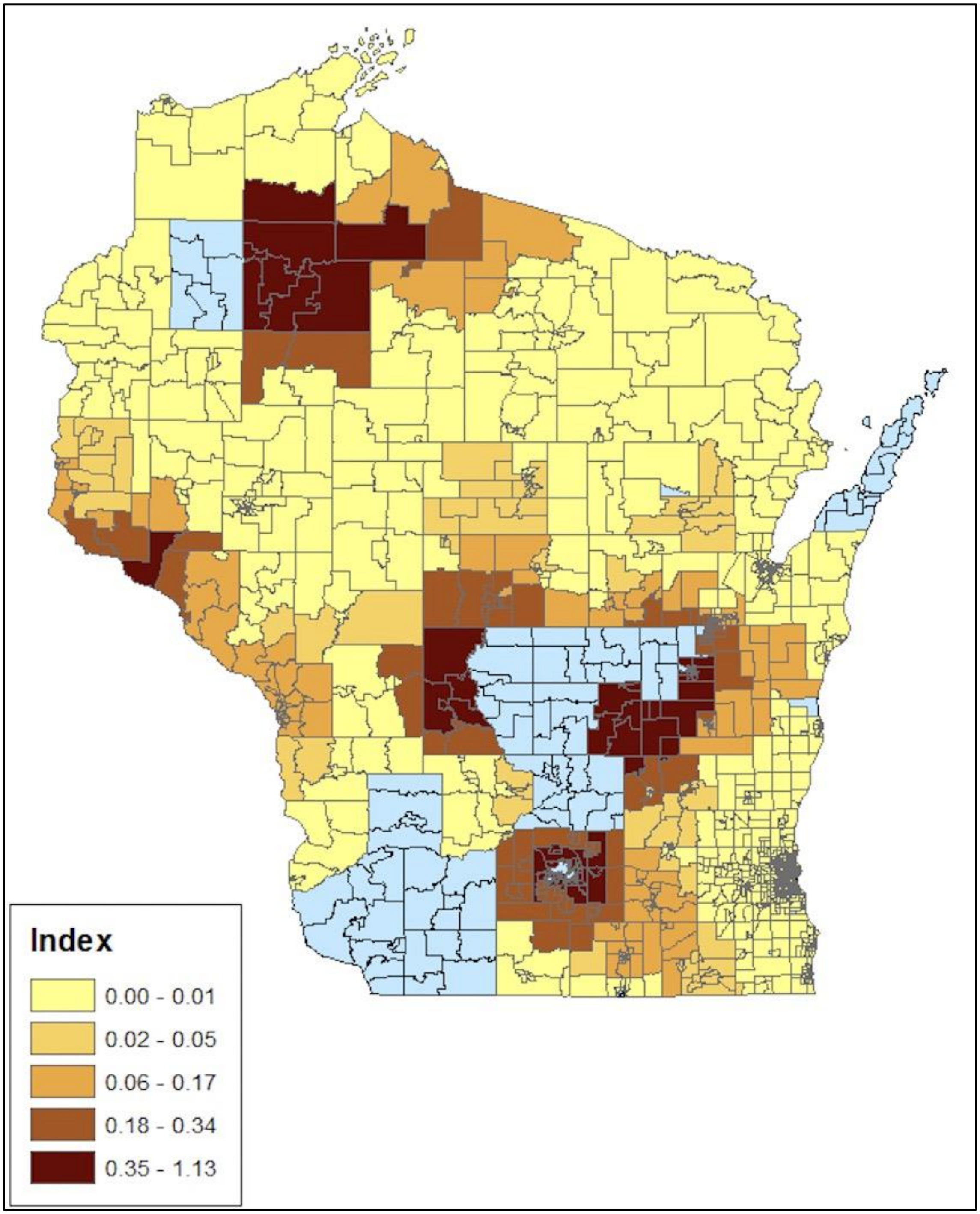

Figure 2.3: Distribution of the recreation index. The recreation index was defined with γ_1=*6*, γ_2=*1*, γ_3=*1*. Census tracts in blue do not have transaction data in the ZTRAX dataset. The recreation index is calculated and plotted at the census tract level.

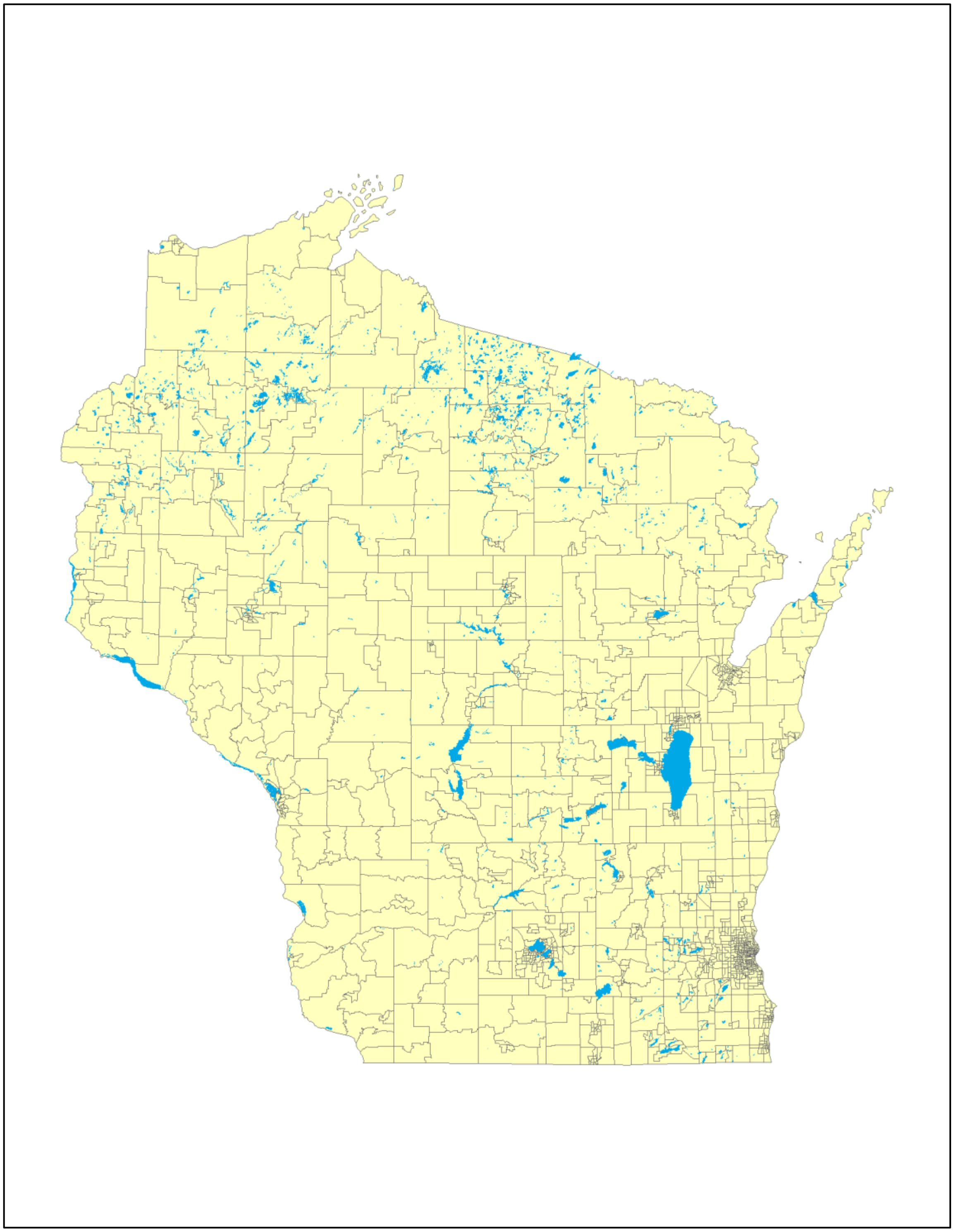

Figure 2.4: Distribution of recreational lakes (N = 1599) in WI. The shape and size of the lake are plotted in blue. The recreation lakes are defined as lakes >=50 acres and have at least one of public parks, public beaches, and public lands.

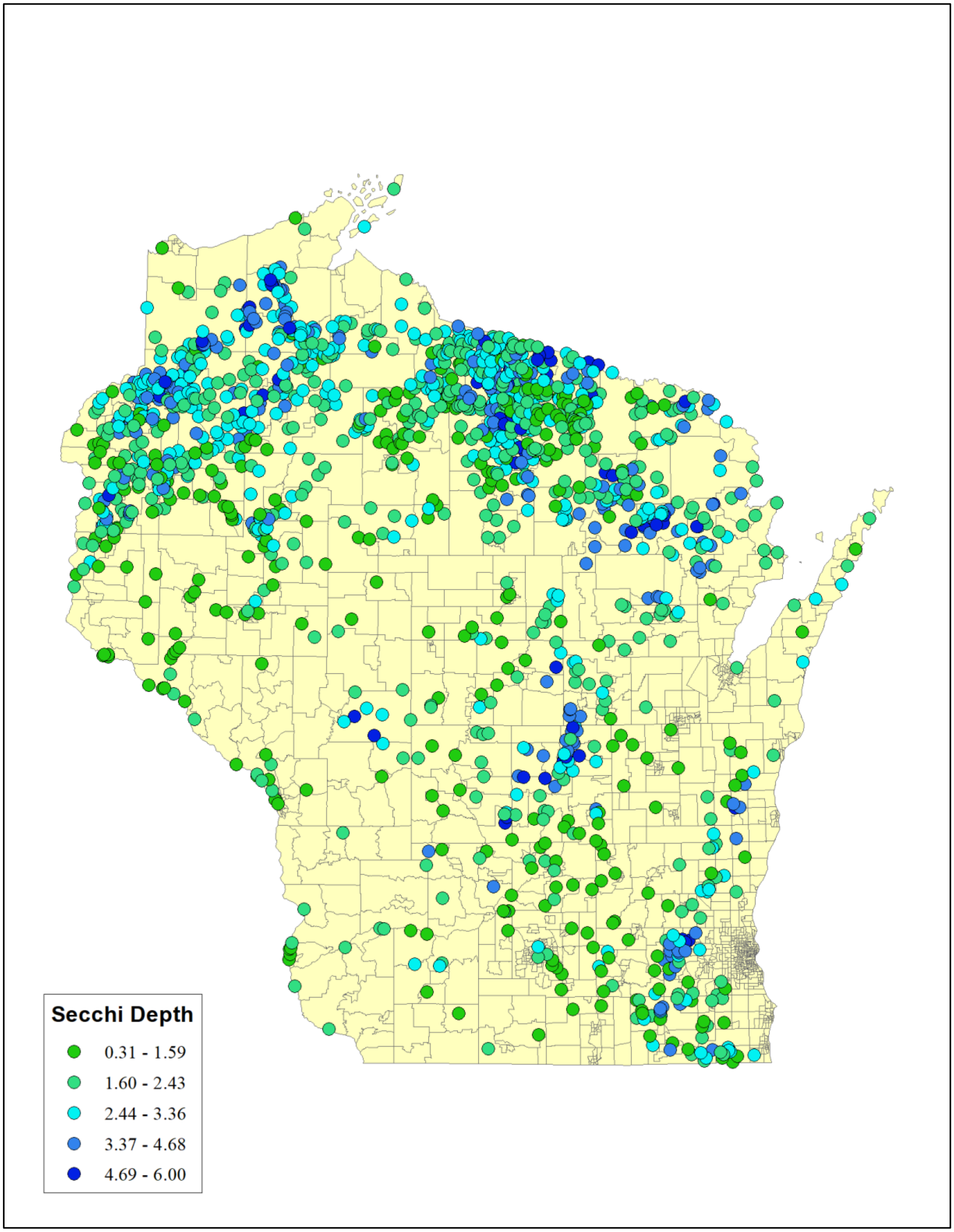

Figure 2.5: Water quality of recreational Lakes (N – 1,599). Secchi depth is measured in meters. The recreation lakes are defined as lakes >=50 acres and have at least one of public parks, public beaches, and public lands.

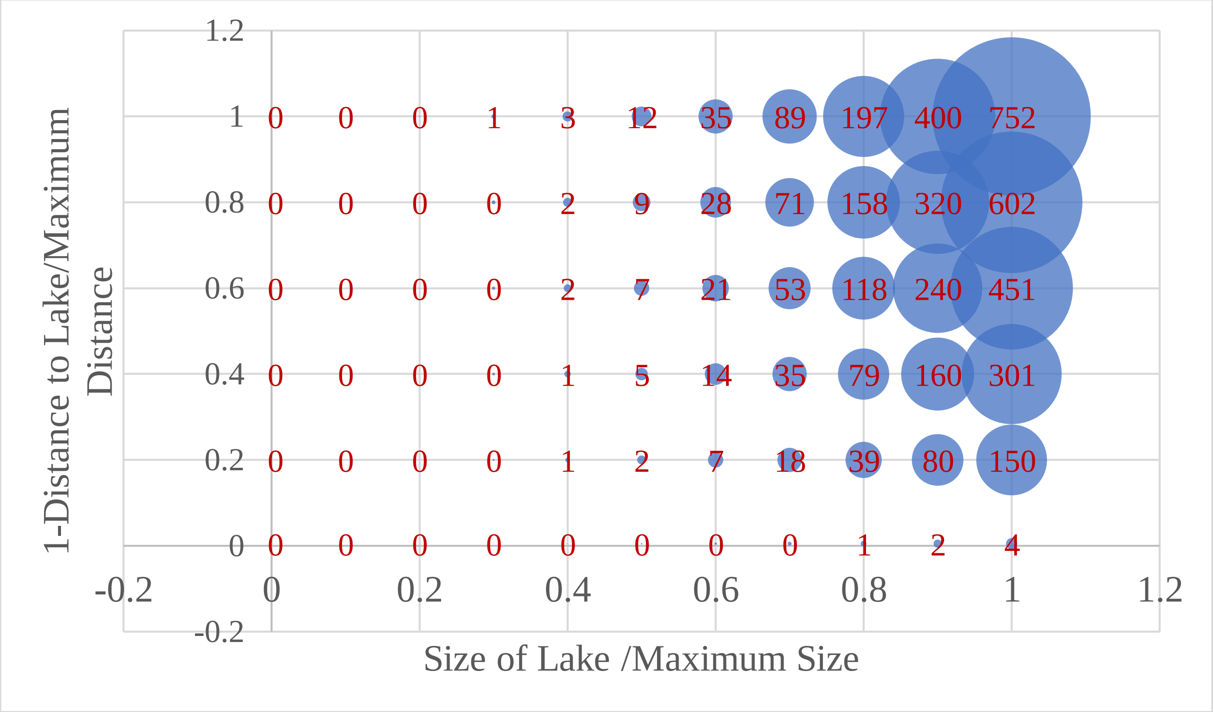

Figure 2.6: MWTP ($) for additional recreational value from water quality improvement. The red number in each bubble is the amount of MWTP for a 0.1m increase in Secchi depth. MWTP is calculated with the estimate results from the aggregate model in column (1) of Table 2.5. The horizontal axis is fraction of the size of lake the largest lake size across the state. The vertical axis is 1 minus the fraction of the travel distance from the census tract to the lake over the largest travel distance across the state.

Table 2.1: Summary Statistics of Lake Characteristics

Panel A: All lakes (N = 2,113)				
	Mean	Std	Max	Min
Size of lake (acre)	427	3102	131039	50
Secchi (meter)	2.53	1.13	8.43	0.31
Panel B: Recreation lakes (N = 1,559)				
Size of lake (acre)	681	3775	131939	50
Secchi (meter)	2.47	1.17	8.43	0.31

Note: Lakes >=50 acres are summarized in Panel A. Lakes >= 50 acres and have at least one of public parks, public beaches, and public lands are summarized in penal B.

Table2.2: Merging Rate of ZTRAX and HMDA Data

	2,016	2,015	2,014	2,013	2,012	2,011
N_Before	51,014	51,672	58,237	53,645	53,830	49,099
N_After	20,076	17,984	16,903	15,588	13,531	6,625
Merging Rate (All Data)	39%	35%	29%	29%	25%	13%
Merging Rage (Mortgage Only)	69%	68%	72%	62%	58%	37%
Mortgage Rate	57%	51%	40%	47%	43%	36%

Note: The third row summarizes the merging rate with all the transactions in the denominator. The fourth row summarizes the merging rate with only transactions with mortgage in the denominator. The mortgage rate is calculated as the number of transactions with mortgage over the number of transactions in the ZTRAX dataset.

Table 2.3: Summary Statistics of housing types for the aggregate model (H = 4849)

Variables	Mean	Std	Max	Min
Price Index (in $100,000)	1.48	0.97	56.78	0.08
Secchi_Lakefront (meter)	0.58	1.09	8.26	0
Lakefront (= 1 if within 100m of a lake)	0.23	0.39	1	0
Rec Index	0.08	0.13	1.13	0
White Ratio (%)	0.91	0.17	1	0
Unemployed (%)	0.06	0.03	0.36	0
Family with Kids (in 1,000)	0.77	0.46	4.63	0
Med Income (in $10,000)	5.54	1.91	15.11	1.25
Poverty Line (%)	0.08	0.08	0.64	0
Med Home Value (in $100,000)	1.76	0.66	6.14	0.42

Table 2.4: Summary Statistics of data for the micro model

	Mean	Std	Max	Min
Panel A: Housing type attributes (H = 2837)				
Price Index (in $100,000)	1.46	0.97	12.12	0.19
Secchi_Lakefront (meter)	0.44	1.09	8.26	0
Lakefront (= 1 if within 100m of a lake)	0.18	0.39	1	0
Rec Index	0.07	0.13	1.13	0
White Ratio (%)	0.90	0.17	1	0
Unemployed (%)	0.06	0.03	0.31	0
Family with Kids (in 1,000)	0.89	0.46	4.63	0
Med Income (in $10,000)	5.86	1.91	15.00	1.25
Poverty Line (%)	0.08	0.08	0.60	0
Med Home Value (in $100,000)	1.85	0.67	6.14	0.42
Panel B: Household characteristics (N = 83349)				
Household Income Level (in $ 10000)	9.01	6.95	96.1	2.2
White (= 1 if white)	0.95	0.22	1	0

Table 2.5: Results for the aggregate model

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Price Index	-11.32^{***}	-11.29^{***}	-11.27^{***}	-11.15^{***}	-11.14^{***}	-11.23^{***}	-11.27^{***}	-10.98^{***}
Secchi_Lakefront	2.05^{***}	2.04^{***}	2.03^{***}	2.01^{***}	2.01^{***}	2.02^{***}	2.04^{***}	1.99^{***}
Lakefront	1.94^{**}	1.94^{**}	1.95^{**}	1.93^{**}	1.92^{**}	1.95^{**}	1.95^{**}	1.86^{*}
Rec Index	5.11^{*}	3.69	1.72	0.09	-1.08	2.01	8.90	-0.08
γ_1	6	4	2	1	4	2	6	0.5
γ_2	1	1	1	1	4	2	2	0.5
γ_3	1	1	1	1	1	1	1	1

Notes: * $p<0.1$, ** $p<0.05$, $^{***}p<0.01$. γ_1 is the index parameter of lake size, γ_2 is the index parameter of water quality, γ_3 is the index parameter of distance. These models are estimated with the full transaction data from ZTRAX.

Table 2.6: Results for the micro model

Panel A: First Stage Estimation		
Secchi_Lakefront -X-Income	0.0070^{***}	
Secchi_Lakefront -X-White	0.0486^{**}	
Lakefront-X-Income	0.0548^{***}	
Lakefront-X-White	1.3658^{***}	
Panel B: Second Stage Estimation		
	OLS	IV
Price Index	-0.1412^{***}	-10.8570^{***}
Secchi_Lakefront	-0.0074	2.6196^{***}
Lakefront	-2.9680^{***}	-0.7055
Rec Index	-0.3566	-0.5441
White Ratio	0.2649	1.5258
Unemployed	-3.0444^{**}	6.4412
Family with Kids	0.3658^{***}	0.5121
Med Income	-0.0305	-0.3161^{*}
Poverty Line	-2.8335^{***}	-6.5397^{*}
Med Home Value	-0.3196^{***}	8.5879^{***}

Notes: * p<0.1, ** p<0.05, ***p<0.01. The recreation index was defined with $\gamma_1=6$, $\gamma_2=1$, $\gamma_3=1$. This model is estimated with the merged data from ZTRAX and HMDA.

Table 2.7: Results for the micro model with full transaction data

Panel A: First Stage Estimation		
Secchi_Lakefront -X-Income	0.0190^{***}	
Secchi_Lakefront -X-White	-0.0000	
Lakefront-X-Income	0.0070^{***}	
Lakefront-X-White	0.0006	
Panel B: Second Stage Estimation		
	OLS	IV
Price Index	-0.0685^{***}	-10.1629^{***}
Secchi_Lakefront	-0.2479^{***}	1.4232^{***}
Lakefront	-0.6900^{***}	2.0022^{*}
Rec Index	-0.4090	4.0584^{*}
White Ratio	-0.2852	-0.4228
Unemployed	-1.8222^{*}	1.4750
Family with Kids	0.4102^{***}	0.4916
Med Income	-0.1608	-1.7393
Poverty Line	-1.3492^{**}	-4.6669
Med Home Value	-0.4649^{***}	7.5039^{***}

Notes: * p<0.1, ** p<0.05, ***p<0.01. The recreation index was defined with $\gamma_1=6$, $\gamma_2=1$, $\gamma_3=1$. This model is estimated with both the full transaction data from ZTRAX and the merged data from ZTRAX and HMDA.

Chapter 3

Valuing water quality in the US using a national data set on property values

With Saleh Mamun, Adriana Castillo Castillo, Kristen Swedberg, Tihitina Andarge, Kevin J. Boyle, Diego Cardoso, Catherine L. Kling, Christoph Nolte, Michael Papenfus, Daniel Phaneuf, Stephen Polasky

3.1 Introduction

The continental United States enjoys a rich tapestry of freshwater lakes that provide multiple ecosystem services including water-based recreation, habitat, biodiversity, aesthetic values, cultural values, and drinking water. Yet the supply of clean water is scarce. Keiser and Shapiro (2019) note that "over half of rivers and substantial shares of drinking water systems violate (water quality) standards." Under the auspices of the Clean Water Act, the US EPA (U.S. Environmental Protection Agency), USDA (U.S. Department of Agriculture), and other federal and state agencies take actions to protect the quality of these waters through regulations (e.g., Navigable Waters Protection Rule), fines for violations (e.g., Comprehensive Environmental Response, Compensation and Liability Act) and incentives (e.g., USDA's Conservation Reserve Enhancement Program). These actions can be divisive, in part because the extent of benefits from improvements have been inadequately quantified (Keiser, Kling, and Shapiro, 2019). This is illustrated by the recent debate over the Waters of the US (WOTUS) rule (e.g., Boyle, Kotchen and Smith 2017; Keiser et al. 2021). At the core of the WOTUS debate, and a challenge for many national water quality policy analyses, is the quantification of the monetary benefits of surface water quality improvements. In the absence of credible estimates of specific economic benefits that are salient to the public, policy costs that are concentrated in readily identified industries receive more emphasis than the often diffuse and abstract economic benefits.

The evidence based on the value of water quality, for agencies like the US EPA and USDA at the national scale empirical literature is quite thin. Most studies estimating public benefits from improving surface water quality have been conducted at the local level (Wolf and Klaiber 2017; Walsh et al. 2011). Policy applications focused on other environmental media such as air quality, however, have benefited from influential national scale studies (Chay and Greenstone 2005). We know of only four studies that estimate national values for surface water quality: a stated preference study from 1983 by Carson and Mitchell (1993) valuing surface water quality; a recreation study by Keiser (2019) that examines the USDA's Conservation Reserve Program (CRP); and a recent hedonic study by Moore et al. (2020) valuing lake water quality using property sales data from 2010 through 2013 and 2007 water quality data. In addition, Keiser and Shapiro (2019) conducted a nationwide hedonic study on the downstream benefits of wastewater treatment grants. However, each of these studies has features, such sample size, spatial distribution, and scope, that complicate its use for national water quality policy analysis.

Given the lack of suitable national-level studies of the benefits of improving or protecting surface water quality, analysts at the US EPA rely on "benefit transfers" for policy analysis. This is a technique whereby the results from multiple case studies from around the US are aggregated and transferred to other contexts to develop national benefits estimates (Corona et al. 2020; Newbold et al. 2018). Benefit transfer is acknowledged as a second choice relative to primary research for several reasons (Rolfe et al. 2015); among these is the thinness of the empirical literature in terms of the number of well-executed studies, their limited spatial distribution across the US, and concerns about the validity of predictions (Guignet et al. 2020; Johnston et al. 2017; Moeltner et al. 2019). These limitations matter for several reasons, including the often-narrow extent to which water quality regulations pass cost-benefit comparison tests. Unlike air quality regulations, which generate large economic benefits through reduced mortality, there is no existing literature linking surface water quality improvements to widespread human health impacts. The degree to which water quality regulations result in positive net benefits therefore hinges on the size and spatial extent of improvements to recreation and residential amenities – and the 'measured' benefits from these in past studies can be quite small (Keiser et al. 2019). Relative to air quality regulation, where benefits are thought to have greatly exceeded costs (Currie & Walker 2019), the close-call nature of water quality regulations places a still higher premium on accurate, policy-relevant, and national-scale estimates of improving and protecting surface water quality.

Here, we focus on a national hedonic property value study of lake water quality. The hedonic model assumes that the price of a house is determined by a bundle of housing and neighborhood characteristics. It is commonly used to estimate environmental amenities. Lakes have an important story to tell as the recipients of upstream pollution loads and the ecological beneficiaries of upstream watershed protection efforts. We estimate the impact of water quality on home prices by comparing the sale prices of properties with different water quality levels. In building out the national benefits of protecting and improving surface water quality, a national hedonic model provides important insights. First, the model shows that people pay for higher water quality through price premiums on properties located on and near lakes. Second, owners of these properties provide the last line of protection of lake water quality through the landscaping on their properties, particularly in the riparian zone, and the empirical results can be used to show these property owners that it is in their own interest to take actions to protect lake water quality and thereby their property values. Third, property taxes provide an important source of revenue, particularly to support primary and secondary education, in communities and this link to hedonic model results can be used to provide a financial incentive to support water quality protection upstream in watersheds. Collectively, these insights demonstrate that hedonic models of lake water quality can be a powerful tool in policy analyses and public education efforts.

In this paper, we leverage newly available spatially extensive data on water quality and property sales near lakes to estimate the price premium for homes near lakes with clear waters. Specifically, we combine water quality data from LAGOS-NE (Lake Multi-Scaled Geospatial and Temporal Database)

(Soranno et al. 2017) and US EPA's Water Quality Portal Data (National Water Quality Monitoring Council 2021) with property sales data from Zillow to develop a sample of nearly 746,424 property transactions near 1,632 lakes across the US. We use two common metrics of water quality: Secchi depth and Chlorophyll-a (chl-a). Secchi depth is a direct measure of water clarity, and is what buyers see when viewing a lake. Chlorophyll-a is a measure used by limnologists to monitor the effects of nutrient pollution on lake water quality, which is one of the factors affecting water clarity. By exploring directly visible measures vis-à-vis a measure of a lake's overall ecological condition, we advance understanding of how water quality improvements translate into perceived services.

With this effort we build on and enhance the existing hedonic property value analyses at the local level (e.g., Michael et al. 2000; Gibbs et al. 2002, Walsh et al. 2011; Weng et al. 2020; Liu et al. 2017; Walsh & Milon, 2016; Weng et al., 2020) and recent national level analyses conducted by Moore et al. (2020) and Keiser & Shapiro (2019). Prior case studies indicate that a 0.1m increase in Secchi depth generates property price increases between $850 (Michael et al. 2000; 2019 $s) and $5,400 (Walsh et al. 2011; 2019 $s), whereas the Moore et al. (2020; 2019 $s) estimate was on the upper end at $4,356. In spatial scale our study is most comparable to Moore et al. (2020) who value lake water quality using property sales data from 2010 through 2013 and 2007 water quality data. However, this study is based on Secchi depth measurements of water quality from 113 lakes using 1,462 property sale observations, and the paucity of data (an average of 13 sale observations per lake) limits the authors ability to test model assumptions. Our enhanced spatial coverage allows us to move beyond implicit price estimates, calculating the national-scale economic benefits that would accrue from improving water quality to two policy relevant levels: restoring lakes to water quality levels typical of original natural conditions, as well as less ambitious, and likely more achievable, improvements of 10% from current baselines in all lakes across the US. Together these two scenarios demonstrate the range of potential benefits of potential policy interventions to improve water quality in lakes and their surrounding watersheds.

Our results provide compelling nationwide evidence that homeowners value lake water quality substantially. On average, homeowners are willing to pay $3,681 for a home near a lake with as little as a

0.1m greater Secchi depth and $4,359 for a 1.0 microgram/liter lower level of Chlorophyll-a. When extrapolated nationally, our estimates indicate that a 10% improvement in Secchi depth generates $9.22 billion in economic benefits and restoring lakes to pristine ecological condition results in $26.67 billion in economic benefits.

3.2 Materials and methods

3.2.1 Property data

We assemble several datasets for this study. The property dataset used in this analysis is constructed by linking parcel boundary data with Zillow’s Transaction and Assessment Database (ZTRAX, version: Oct 09, 2019). ZTRAX provides sales related information (sale dates, prices, inter-family transfer), parcel and building characteristics such as building area, lot size, and built year. We link parcel boundaries with ZTRAX using assessor's parcel number and customized pattern matching algorithm (Nolte et al. 2020). If parcel subdivision and consolidation results in unsuccessful or partial linking of parcels with ZTRAX dataset, we ignore those transactions. To clean up ZTRAX transaction values and coordinates, we follow the best practice as described by Nolte et al. (2021). Fair market value is ensured by (1) removing transactions involving inter-family transfer based on name similarity index; (2) removing transactions involving public buyers and public sellers. In addition, we remove transactions if the sale price is less than $10,000 (Gindelsky et al. 2018) and the top 1 percentile of sale price per hectare to address remaining outliers. We identify residential developed properties if there is any building area and Zillow reported a building code of ‘RR’ (residential).

3.2.2 Water quality data

Water quality data is collected from two national level data repositories: water quality portal (WQP47) (National Water Quality Monitoring Council 2021) from the United States Environmental Protection Agency (USEPA) and the LAGOS-NE. We merge LAGOS-NE and WQP datasets to create a comprehensive dataset. As they do not have a common identifier, we used the National Hydrography dataset (NHD) to combine them using spatial join technique. The two most common water quality measurements are Secchi depth and Chlorophyll-a concentration. Secchi depth is measured by inserting a circular disk into a water column and determining the distance from the surface at which the disk is no longer visible. Hence, it measures how deep sunlight can penetrate water. As water clarity is a salient feature of water quality to homeowners, and Secchi depth captures water clarity, we use Secchi depth as a measure of water quality in our hedonic pricing model. Chlorophyll-a is a measure of the number of algae in a waterbody.

3.2.3 Merging property data with water quality data

The property dataset contains transaction records of property within 2000m from the shoreline of lakes between 2000-2019. The lake water quality dataset also consists of water quality measures for lakes between 1900-2020. However, the longitudinal dimension of lake water quality is limited partially due to the high cost of collecting water quality data (Sprague et al. 2017). We used a fuzzy date matching algorithm to link properties with water quality. The algorithm searches the nearest lake to a property by NHD lake ID and then finds the closest water quality sample year. If the difference between the water quality sample year and property sale year is more than 5 years, we omit those transactions. We also check robustness based on alternative cutoff values of this difference (i.e., 1 year, 3 year).

[47] The dataset contains more than 50 million observations of 3,393 different water quality parameters by activity date and monitoring location. However, the frequency of parameters varies substantially.

After merging the water quality and property transaction data, our sample consists of 746,434 transactions around 1,632 lakes greater than 4 ha (Figure 3.1). The lakes in our sample are clustered in lake-rich regions like the Midwest and East Coast. However, adding restrictions to the sample based on the number of sale observations within 100m of the lake greatly reduces the total number of lakes included in the sample and overall spatial coverage (Figure 3.1B). Thus, we select the widest sample as the baseline for comparison.

We observe wide variation in both Secchi depth and Chlorophyll-a among census tracts throughout the sample, with generally poor water quality conditions in the upper Midwest (Figure 3.1C and D). Table 3.1 reports the summary statistics of our main variables after merging property transactions with water quality values. It also shows the wide range of the Secchi depths from a few centimeters to 10m with a median value of 1.52m. Lakes with Secchi depths less than 1 - 2m would be classified with poor or bad water quality (Sondergaard, et al. 2005, Moss et al. 2003), suggesting more than half of properties in our data may suffer from water pollution in their nearest lake. A detailed description of data cleaning and combining is provided in Appendix.

3.3 Methods

In this study, we use a hedonic model to estimate the impact of lake water quality on housing prices at a nationwide scale. Assume home buyers are consumers and home sellers are producers. Then, the housing market can be characterized as a differentiated goods market. Consumers choose a house and other goods to maximize their utility given a budget constraint based on their income levels. Consumers' utility from a house depends on the characteristics of the house, which also makes it differentiated from other houses in the market. The hedonic model defines the price of a house as a function of structural, neighborhood, environmental characteristics. To measure the implicit price of water quality, which is an environmental amenity affecting consumers' housing preferences, we estimate the following hedonic model:

$$lnP_{it} = \alpha lnQ_{it} + \beta_s S_{it} + \beta_f F_{it} + \theta_s lnQ_{it} \times S_{it} + \theta_f lnQ_{it} \times F_{it} + \gamma \boldsymbol{X}_{it} + \sigma Y_l + \tau_{ct} + \epsilon_{it} \qquad (3.1)$$

Where P_{it} is the sale price of property *i* in transaction year *t* , Q_{it} is the water quality assigned to property *i* sold in year *t*, S_{it} and F_{it} are dummy variables indicating that a property is within 100m and 100-300m of its nearest lake, X_{it} is the vector of housing characteristics, Y_l denotes lake characteristics, and τ_{ct} denotes census tract by year fixed effects. To address the potential bias with time and regions, the home price is adjusted using the housing price index (HPI).

Existing literature finds that lakefront homeowners have a higher marginal willingness to pay for water quality improvements (Walsh et al. 2011; Walsh et al. 2015; Wolf & Klaiber 2017). Therefore, the distance to the lake is an important factor affecting residential decisions. However, the cutoff distance that defines the lakefront area varies across studies in past literature. Here we divide our properties into three distance bins according to the proximity from the property to its nearest lakes: 0-100m, 100m – 300m, and the area beyond 300m.

In our baseline model, we use Secchi depth to capture water quality as it reflects both clarity and quality of water. We assume home buyers are sensitive to this measurement because the aesthetic and recreational amenities are highly associated with the value of Secchi depths. We also estimate the same model using Chlorophyll-a as the water quality variable. We interact the distance dummy variables with water quality to capture the heterogeneity in the impact of water quality. Our primary interest focuses on three parameters. Given the log-log functional form, our first parameter of interest, α, measures the percentage change in home prices from a one percent change in Secchi depth for properties within 2000m of their nearest lakes. The parameters θ_s and θ_f capture the impact of water quality on the price premium for properties within 100m and 100-300m of lakes, respectively. Given that homeowners living within 300m of lakes have easier access to lake amenities, we expect positive values of θ_s and θ_f to represent additional impact from water quality on properties close to lakes.

In addition to the key variables mentioned above, we also include variables of housing characteristics. After removing variables that are highly correlated, we end up with six variables that address the lot size, the average slope, the elevation of the parcel, the proximity to highway, the age of the building, and the area of the building. As lakes with larger sizes usually provide more recreational amenities, we also include the size of the nearest lake to represent all lake features. In addition, the census tract by year fixed effects capture neighborhood characteristics on an annual basis. Properties in the same census tract are assigned to different lakes that have different water quality, even with the inclusion of fine-scale spatial by temporal fixed effects, so that there are still enough variations in water quality and home prices for the estimation. As a robustness check, we examine the sensitivity of our results to the inclusion of spatial fixed effects at different levels. We cluster standard error at the census tract level to account for unobserved correlation within the census tract. In addition, past literature adds restrictions on the number of lakeshore observations and minimum number of water quality records per lake to control for the bias from the sample (Zhang et al. 2022). However, we will lose a large amount of data if we impose strict restrictions. It is important to include a wide coverage of lakes and property transactions that can represent our study area at a nationwide scale in the baseline model. As an alternative, we estimate models with different restrictions on water quality and property transaction data as robustness checks.

3.4 Results

We evaluate two water quality measurements: Secchi depth and Chlorophyll-a. Table 3.2 summarizes the estimation results for the main model. Using transactions within 2km of lakes, we found a positive and significant impact of water quality on home prices for properties in the lakefront areas with larger impact for properties closer to the shoreline. As shown in the first column, the estimate results suggest that the average home prices of properties within 100m of lakes are 56 percent higher than properties beyond the 300m boundary.[48] Lake amenities contribute to this price premium, which depends on water quality in the lake. So that water quality is most salient to residents in this area. The coefficient

[48] This result is calculated by 100*(e^0.4473-1).

of the interaction term shows that a 1 percent increase in the value of Secchi depths increases the proximity premium by 0.1673 percent for properties within 100m of lakes. Additionally, the joint F-test of both the Secchi depth variable and the interaction term is also significant, suggesting a 1 percent increase in the value of Secchi depth increases home prices by 0.1749 percent within 100m of lakes. In the same sense, both the price premium of being close to the lake and the impact of water quality on the proximity premium are smaller when moving farther from the lakeshore: for properties in the 100-300m area of lakes, the average price is 5 percent higher than properties 300m away. And a 1 percent increase in the value of Secchi depths leads to a 0.0481 percent increase in home prices. These results are consistent with other spatial fixed effect levels (e.g. census block group and county). For properties located beyond the 300m threshold, water quality has a positive but statistically insignificant effect on home values. Note that properties far from lakes are less likely to be affected by lake amenities. It makes sense that water quality has little impact on home values. Similar insignificant overall impact has been found in other studies (Wolf & Klaiber 2017, Walsh et al. 2017).

The average value of Secchi depths in our sample is 2.61 meters and the average home price within 100m of lakes is $549,245. Calculating the implicit price of water quality at the average values, a 0.1m improvement, which is 4 percent increase in the value of Secchi depth increases the home price by $3,681 on average. Correspondingly, the implicit price for a 0.1m increase in the Secchi depth value is $764 for properties in the 100-300m area of lakes.[49] We examine the robustness of our results regarding the structure of the model, data selection, and the parameter definition. The coefficients on the three main variables are summarized in Figure 3.2. The violin plot suggests that the estimate results from the robustness checks are general consistent with our baseline model.

In addition to Secchi depths, Chlorophyll-a is also commonly used to measure water quality (Weng et al. 2020, Liu et al. 2017). Column 2 of Table 3.2 shows the results of the baseline model using Chlorophyll-a as the water quality variable. The results suggest that Chlorophyll-a also only affects home

[49] The average Secchi depth within 100-300m of lakes is 2.20m. The average housing price is $349,657.

value of properties within 300m of lakes. As shown in Table 3.2, the marginal effect of Chlorophyll-a is smaller than the Secchi depths: a 1 percent decrease in Chlorophyll-a leads to a 0.1162 percent increase in home prices for properties within 100m of lakes and 0.0391 percent increase in home prices for properties within 100-300m of lakes. Knowing that the average concentration of Chlorophyll-a is 14.52 microgram/liter for properties within 100m of lakes, people are willing to pay $4,395 for a 1 μg/L (7%) decrease in the level of Chlorophyll-a. Similarly, residents living in the 100-300m buffer area of lakes are willing to pay $783 for the same reduction in the level of Chlorophyll-a.

3.4.1 Capitalization effects

To understand the full benefits of water quality capitalized in lake housing markets, economists consider the combined effects of water quality improvements for all properties throughout the sample or study area. Given the broad spatial coverage of our dataset, we evaluate capitalization effects of lake water quality at a national scale. To do so we consider all properties surrounding lakes larger than 4 ha within the continental United States. These properties are divided into two buffer groups: those within 100m and those between 100-300m from the lakeshore indicated by subscripts *s* and *f* respectively. The total combined capitalization effects are

$$[(\alpha + \theta_s)\bar{P}_s n_s + (\alpha + \theta_f)\bar{P}_f n_f] \times \%Change \ , \tag{3.2}$$

where $\alpha + \theta_{s(f)}$ is the combined elasticity estimate from the baseline model, $\bar{P}_{s(f)}$ represents last sale prices or assessed values averaged for each buffer group at the lake level, *n* is the number of properties within the buffer group for each lake, and *%Change* is a percent change in water quality. For lakes that span multiple counties, $\bar{P}_{s(f)}$ is averaged for properties adjacent to the lake in each county.

For a marginal change in water quality, the shape of the hedonic price function remains unchanged and, under the assumption that preference, income, and technology remain constant, we can interpret the capitalization effects as the marginal willingness to pay (Kuminoff & Pope 2014). However, what constitutes a marginal change in water quality is not well-defined. For example, both 1% and 10%

changes in Secchi depth for a lake with 1m of clarity (1cm and 10cm respectively) may not be distinguishable to a homebuyer and thus could be considered marginal. In fact, 1cm change in water clarity may result from typical fluctuations in water quality throughout the summer. Considering that the median Secchi depth value for our sample is 1.52m, we want to make sure the change exceeds the seasonal fluctuation but is not large enough to affect the housing market equilibrium. We therefore evaluate the effects of 10% improvements of both Secchi and Chlorophyll-a using average assessed values and average last sale prices. Our results indicate capitalization effects for marginal changes in Secchi range from \$9.22 to \$11.09 billion (Table 3.3). Most of these benefits are from properties within 100m of the lake, \$8.09 billion compared to \$2.99 billion for properties 100-300m from the lakeshore considering last sale prices. For Chlorophyll-a we find smaller capitalization effects that range from \$4.33 to \$5.21 billion.

A limitation of analyzing marginal improvements in water quality is that the resulting capitalization effects only serve as a lower bound for the benefits of water quality policies, as intended policy outcomes are often nonmarginal in nature. Thus, to provide an upper bound, we consider the capitalization effects if all lakes were restored to pristine ecological conditions. We define pristine as the water quality conditions on “reference lakes” from the National Lakes Assessment (NLA), which provides water quality samples for an ecologically representative national sample of lakes. These “reference lakes” refer to the least disturbed lakes for each ecoregion and serve as the baseline by which the NLA gauges lake health within an ecoregion (USEPA 2016; USEPA 2012). Using the NLA 2012 and 2017, we calculate the percent changes in Secchi depth required to shift the average water clarity in each ecoregion to its respective reference lake levels by

$$\%Change = (\bar{Q}_{eco,ref} - \bar{Q}_{eco}) / \bar{Q}_{eco} , \tag{3.3}$$

where $\bar{Q}_{eco}$ is the average Secchi depth for lakes in each ecoregion and $\bar{Q}_{eco,ref}$ is the average Secchi depth for the reference lakes in each ecoregion. We obtain the capitalization effect by multiplying the percentage change of Secchi depth with the elasticity estimated from the baseline model. In this scenario. we find capitalization effects ranging from $25.18 to $26.67 billion in assessed values using the 2012 and 2017 NLA respectively. Because the changes in water quality in equation (3.3) can no longer be considered as marginal, we cannot interpret these results in willingness to pay terms. However, the value of water quality capitalized in housing markets still provides important economic insights. Property taxes are a large source of government revenue that support local school systems, public safety, and other government programs. Thus, the capitalized benefits of improving water quality at the national level extend beyond gains to individual property owners and are relevant to policy makers.

3.5 Discussion and conclusion

The results of this study substantiate the work of prior hedonic studies focused on small lake samples, and through expanded spatial scale, provide strong evidence in support of nationwide water quality policy. We find a 0.1m increase in Secchi depth leads to a $3,681 price premium for an average lakefront home located within 100m of lakes. This estimate falls between the bounding of the regional literature, which estimates implicit prices as low as $850 in Northern Maine (Michael et al. 2000) and $5,540 in Orange County Florida (Walsh et al. 2011). While Moore et al. (2020) report a nationwide implicit price estimate of $4,354, their sample included only 113 lakes that may be more representative popular lakes with more frequent and abundant lakefront sales. In contrast we consider 1,632 lakes, and our results are robust to multiple sample requirements and model specifications. We construct different samples of lakes by restricting the number of water quality records and the number of housing transactions per lake and find that the results across our broadest sample of lakes is the most robust. Only in our most restrictive models that drastically reduce the number of lakes below 100 do we observe differences in the estimated effects, providing compelling evidence of sample selection bias in broad scale hedonic analysis with limited numbers of observations. Given the strength of our results and broad spatial coverage (43 states), we conclude water quality, both water clarity and algal concentration, is capitalized in lakefront property values at a national scale.

These results not only allow us to identify nationwide capitalization effects for lake water quality but reveal which property values will benefit the most from improved water quality. We find the largest effects on properties within 100m from the lakeshore and no significant effects beyond 300m. Prior studies have found similar decay in the effects of water quality moving away from the waterfront (Walsh et al. 2017). Understanding the effects of water quality on lakefront property values is important for policy considerations, as lakefront homeowners will bear the highest costs, both implicit and explicit, in protecting and improving water quality. The marginal benefits of water quality improvements can influence individual homeowners to take actions, such as reducing lawn fertilizer or maintaining riparian buffers. In aggregate, these benefits can lay the groundwork for national policies designed to protect and restore lakes. Our capitalization results show that even modest policies aimed at 10 percent improvement in lake water quality have large welfare implications, ranging from over 6 billion dollars in consumer surplus for Chlorophyll-a and over 9 billion dollars for Secchi depth.

The potential benefits increase in relation to large, policy relevant, changes in water quality. We find property values could differ by more than 25 billion dollars if all lakes return to pristine ecological conditions. However, eutrophication in lakes is a difficult and costly process to reverse, so losses in property values and the corresponding tax revenue are unlikely to be restored. However, policies aimed at protecting existing lakes from further nutrient pollution can prevent future losses in property values.

3.6 Bibliography

Boyle, Kevin J., Matthew J. Kotchen, and V. Kerry Smith. "Deciphering dueling analyses of clean water regulations." *Science* 358, no. 6359 (2017): 49-50.

Carson, Richard T., and Robert Cameron Mitchell. "The value of clean water: the public's willingness to pay for boatable, fishable, and swimmable quality water." *Water resources research* 29, no. 7 (1993): 2445-2454.

Chay, Kenneth Y., and Michael Greenstone. "Does air quality matter? Evidence from the housing market." *Journal of political Economy* 113, no. 2 (2005): 376-424.

Corona, Joel, Todd Doley, Charles Griffiths, Matthew Massey, Chris Moore, Stephen Muela, Brenda Rashleigh, William Wheeler, Stephen D. Whitlock, and Julie Hewitt. "An integrated assessment model for valuing water quality changes in the United States." *Land Economics* 96, no. 4 (2020): 478-492.

Currie, Janet, and Reed Walker. "What do economists have to say about the Clean Air Act 50 years after the establishment of the Environmental Protection Agency?." *Journal of Economic Perspectives* 33, no. 4 (2019): 3-26.

Gibbs, Julie P., John M. Halstead, Kevin J. Boyle, and Ju-Chin Huang. "An hedonic analysis of the effects of lake water clarity on New Hampshire lakefront properties." *Agricultural and Resource Economics Review* 31, no. 1 (2002): 39-46.

Gindelsky, Marina, Jeremy Moulton, and Scott A. Wentland. "Valuing housing services in the era of big data: A user cost approach leveraging Zillow microdata." In *Big Data for 21st Century Economic Statistics*. University of Chicago Press, 2019.

Guignet, Dennis, Matthew T. Heberling, Michael Papenfus, and Olivia Griot. "Property values, water quality, and benefit transfer: A nationwide meta-analysis." *Land Economics* (2021): 050120-0062R1.

Johnston, Robert J., Elena Y. Besedin, and Ryan Stapler. "Enhanced geospatial validity for meta-analysis and environmental benefit transfer: an application to water quality improvements." *Environmental and Resource Economics* 68, no. 2 (2017): 343-375.

Keiser, David A. "The missing benefits of clean water and the role of mismeasured pollution." *Journal of the Association of Environmental and Resource Economists* 6, no. 4 (2019): 669-707.

Keiser, David A., Catherine L. Kling, and Joseph S. Shapiro. "The low but uncertain measured benefits of US water quality policy." *Proceedings of the National Academy of Sciences* 116, no. 12 (2019): 5262-5269.

Keiser, David A., Sheila M. Olmstead, Kevin J. Boyle, Victor B. Flatt, Bonnie L. Keeler, Catherine L. Kling, Daniel J. Phaneuf, Joseph S. Shapiro, and Jay P. Shimshack. "A water rule that turns a blind eye to transboundary pollution." *Science* 372, no. 6539 (2021): 241-243.

Keiser, David A., and Joseph S. Shapiro. "Consequences of the Clean Water Act and the demand for water quality." *The Quarterly Journal of Economics* 134, no. 1 (2019): 349-396.

Kuminoff, Nicolai V., and Jaren C. Pope. "Do “capitalization effects” for public goods reveal the public's willingness to pay?." *International Economic Review* 55, no. 4 (2014): 1227-1250.

Liu, Tingting, James J. Opaluch, and Emi Uchida. "The impact of water quality in Narragansett Bay on housing prices." *Water Resources Research* 53, no. 8 (2017): 6454-6471.

Michael, Holly J., Kevin J. Boyle, and Roy Bouchard. "Does the measurement of environmental quality affect implicit prices estimated from hedonic models?." *Land Economics* (2000): 283-298.

Moeltner, Klaus, Jessica A. Balukas, Elena Besedin, and Ben Holland. "Waters of the United States: Upgrading wetland valuation via benefit transfer." *Ecological Economics* 164 (2019): 106336.

Moore, Michael R., Jonathan P. Doubek, Hui Xu, and Bradley J. Cardinale. "Hedonic price estimates of lake water quality: Valued attribute, instrumental variables, and ecological-economic benefits." *Ecological Economics* 176 (2020): 106692.

Moss, Brian, Deborah Stephen, Cristina Alvarez, Eloy Becares, Wouter Van De Bund, S. E. Collings, Ellen Van Donk et al. "The determination of ecological status in shallow lakes—a tested system

(ECOFRAME) for implementation of the European Water Framework Directive." *Aquatic Conservation: Marine and Freshwater Ecosystems* 13, no. 6 (2003): 507-549.

Newbold, Stephen, R. David Simpson, D. Matthew Massey, Matthew T. Heberling, William Wheeler, Joel Corona, and Julie Hewitt. "Benefit transfer challenges: perspectives from US practitioners." *Environmental and Resource Economics* 69, no. 3 (2018): 467-481.

Nolte, Christoph. "High-resolution land value maps reveal underestimation of conservation costs in the United States." *Proceedings of the National Academy of Sciences* 117, no. 47 (2020): 29577-29583.

Nolte, Christoph, Kevin J. Boyle, Anita M. Chaudhry, Christopher M. Clapp, Dennis Guignet, Hannah Hennighausen, Ido Kushner et al. "Studying the impacts of environmental amenities and hazards with nationwide property data: best data practices for interpretable and reproducible analyses." *Available at SSRN* (2021).

Rolfe, John, Jill Windle, and Robert J. Johnston. "Applying benefit transfer with limited data: unit value transfers in practice." In *Benefit Transfer of Environmental and Resource Values*, pp. 141-162. Springer, Dordrecht, 2015.

Ross, Matthew RV, Simon N. Topp, Alison P. Appling, Xiao Yang, Catherine Kuhn, David Butman, Marc Simard, and Tamlin M. Pavelsky. "AquaSat: A data set to enable remote sensing of water quality for inland waters." *Water Resources Research* 55, no. 11 (2019): 10012-10025.

Søndergaard, Martin, Erik Jeppesen, Jens Peder Jensen, and Susanne Lildal Amsinck. "Water Framework Directive: ecological classification of Danish lakes." *Journal of Applied Ecology* 42, no. 4 (2005): 616-629.

Soranno, Patricia A., Linda C. Bacon, Michael Beauchene, Karen E. Bednar, Edward G. Bissell, Claire K. Boudreau, Marvin G. Boyer et al. "LAGOS-NE: a multi-scaled geospatial and temporal database of lake ecological context and water quality for thousands of US lakes." *GigaScience* 6, no. 12 (2017): gix101.

Sprague, Lori A., Gretchen P. Oelsner, and Denise M. Argue. "Challenges with secondary use of multi-source water-quality data in the United States." *Water research* 110 (2017): 252-261.

Stachelek, Joseph, Samantha Oliver, Farzan Masrour. LAGOSNE: Interface to the Lake Multi-Scaled Geospatial and Temporal Database. (2020). R package version 2.0.2, https://CRAN.R-project.org/package=lagosne.

United States Environmental Protection Agency (USEPA), 2003. Ambient water quality criteria for dissolved oxygen, water clarity and Chlorophyll-a for the Chesapeake Bay and its tidal tributaries. Office of Water. EPA, Annapolis MD (903-R-03-002). https://cdn.ioos.noaa.gov/media/2017/12/ambient_water_quality_criteria.pdf

United States Environmental Protection Agency (USEPA), National Lakes Assessment 2012: A Collaborative Survey of Lakes in the United States. Available at https://www.epa.gov/sites/default/files/2016-12/documents/nla_report_dec_2016.pdf (2016).

United States Environmental Protection Agency (USEPA), National Lakes Assessment 2017: A Collaborative Survey of the Lakes in the United States Available at https://www.epa.gov/national-aquatic-resource-surveys/data-national-aquatic-resource-surveys (2021).

Walsh, Patrick, Charles Griffiths, Dennis Guignet, and Heather Klemick. "Modeling the property price impact of water quality in 14 Chesapeake Bay Counties." *Ecological economics* 135 (2017): 103-113.

Walsh, Patrick J., and J. Walter Milon. "Nutrient standards, water quality indicators, and economic benefits from water quality regulations." *Environmental and Resource Economics* 64, no. 4 (2016): 643-661.

Walsh, Patrick J., J. Walter Milon, and David O. Scrogin. "The spatial extent of water quality benefits in urban housing markets." *Land Economics* 87, no. 4 (2011): 628-644.

Weng, Weizhe, Kevin J. Boyle, Kaitlin J. Farrell, Cayelan C. Carey, Kelly M. Cobourn, Hilary A. Dugan, Paul C. Hanson, Nicole K. Ward, and Kathleen C. Weathers. "Coupling Natural and Human

Models in the Context of a Lake Ecosystem: Lake Mendota, Wisconsin, USA." *Ecological Economics* 169 (2020): 106556.

Wolf, David, and H. Allen Klaiber. "Bloom and bust: Toxic algae's impact on nearby property values." *Ecological economics* 135 (2017): 209-221.

Zhang, Jiarui, Daniel J. Phaneuf, and Blake A. Schaeffer. (2022). “Property values and cyanobacterial algal blooms: evidence from satellite monitoring of inland lakes.” *Ecological Economics*, forthcoming.

3.7 Figures & Tables

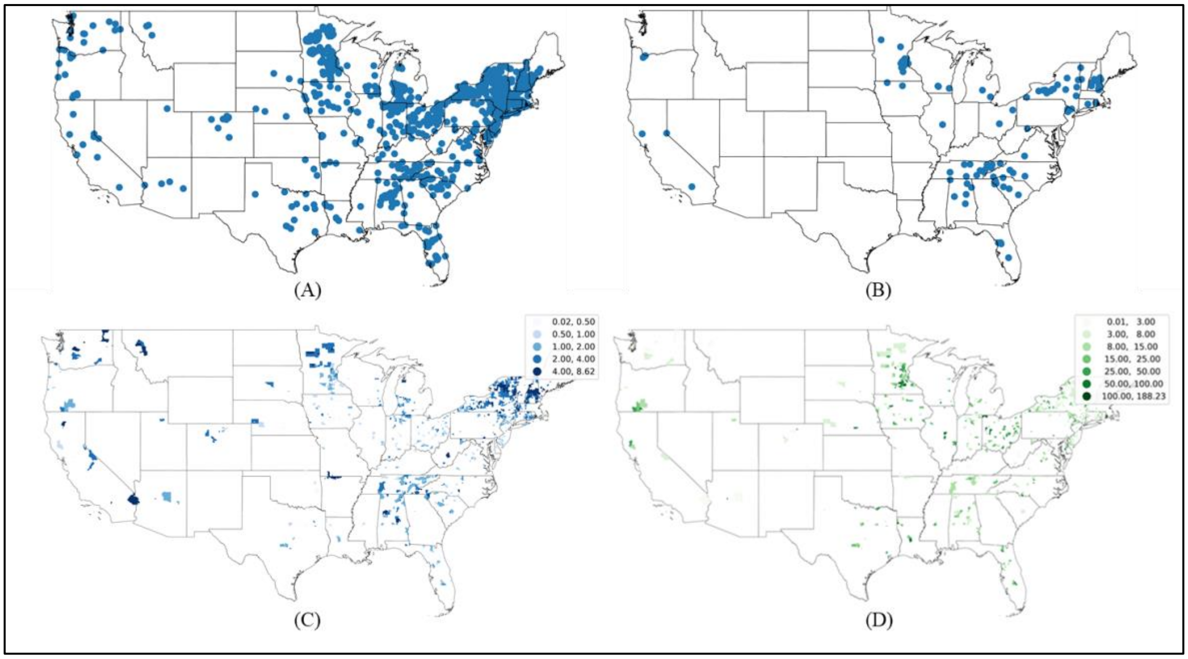

Figure 3.1: (A) Distribution of study lakes with at least one water quality sample and one property sale within 100m from lakefront (N = 1,632) (B) Distribution of study lakes with at least one water quality sample and 100 property sales within 100m from lakefront (N = 109) (C) Most recent Secchi depth (m) samples for lakes averaged over census tracts (D) Most recent chlorophyll-A (ug/L) samples for lakes averaged over census tracts

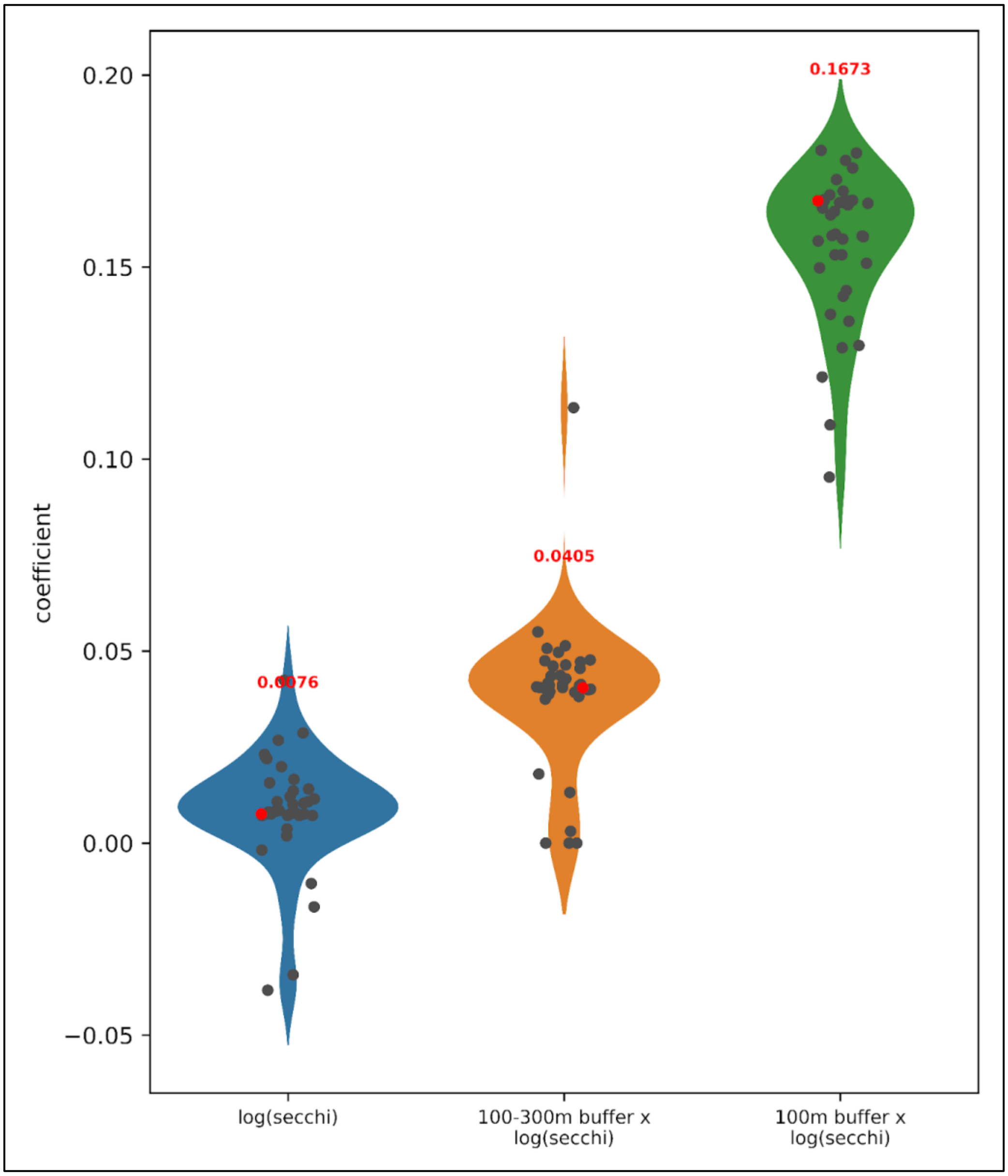

Figure 3.2: Violin plot of Secchi depth results across all model specifications and robustness tests. The red dots and labels indicate results from the baseline model.

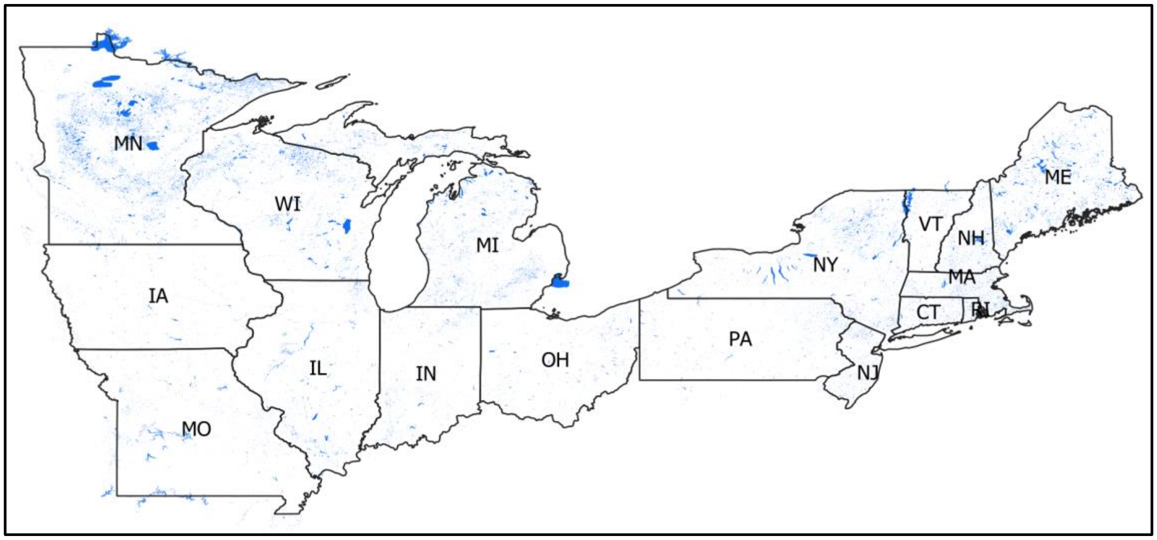

Figure 3.3: LAGOS-NE States and location of lakes considered.

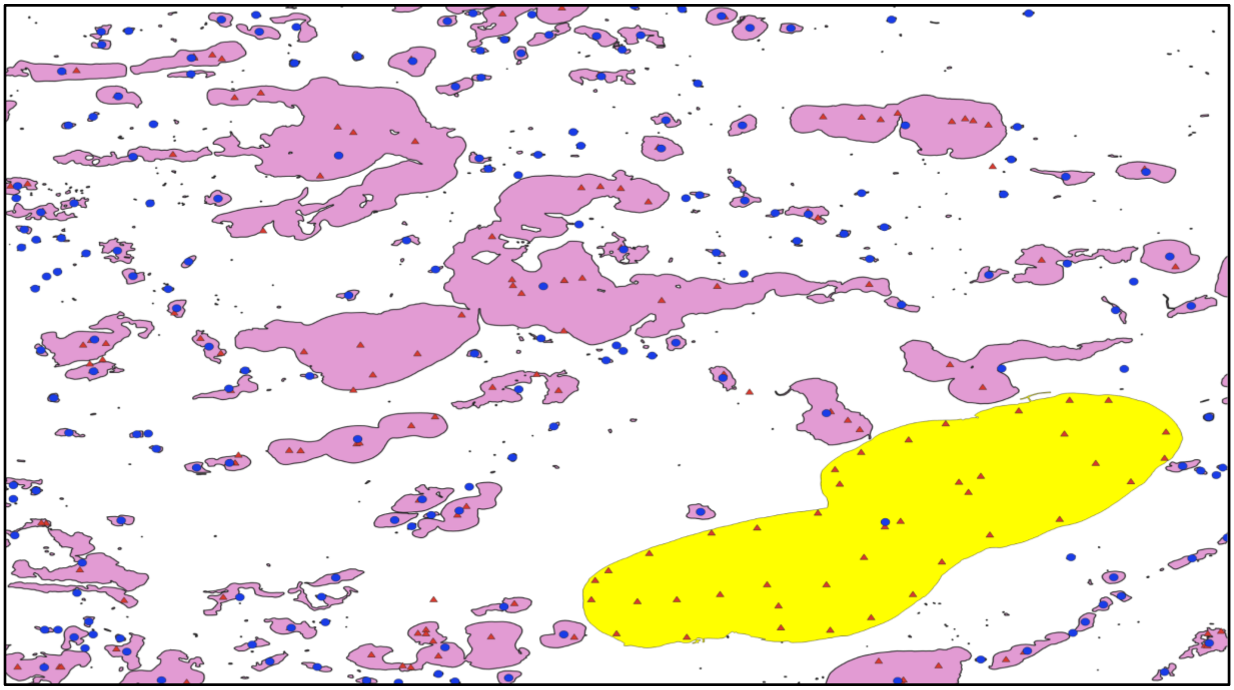

Figure 3.4: WQP monitor locations vs LAGOS-NE Lake Location. Otter Tail Lake is highlighted in yellow. The red triangles are WQP monitoring station locations while blue dots are lake central points from LAGOS-NE.

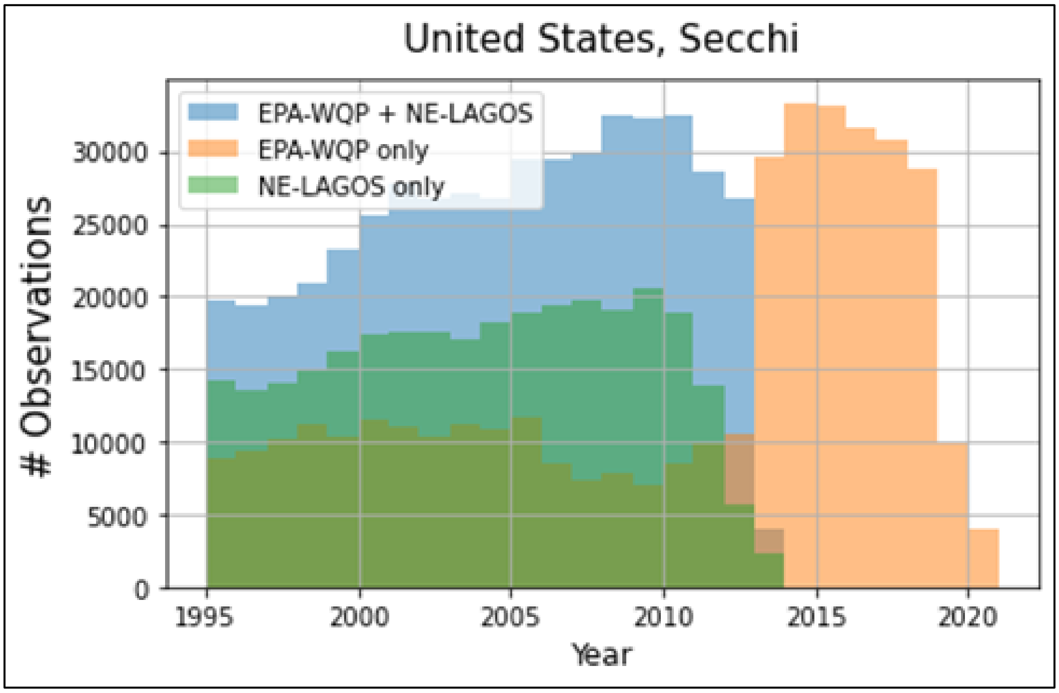

Figure 3.5: The complementarity of water quality data between LAGOS-NE and WQP. Only Secchi depth is shown in the figure.

Table 3.1: Summary statistics for baseline model observations

	Min	Max	Median	Mean	Std
0-100m buffer	0	1	0	0.09	0.28
100-300m buffer	0	1	0	0.15	0.35
Secchi (m)	0.02	9.70	1.52	2.01	1.59
Chl-a (ug/L)[1]	0.01	200.00	11.60	19.01	21.54
Lot area (m^2)	100	5,865,905	1,086	3,035	19,898
Lake size (km^2)	0.04	1,650.28	1.12	39.82	122.67
Price (2019 $)	10,000	39,524,515	276,544	344,307	348,070

Notes: N = 746,424. Fewer Chl-a observations (N = 637,548)

Table 3.2: Estimation results for Secchi depth and chl-a

	Secchi	Chl-a
0-100m buffer	0.4473**	0.7600**
100-300m buffer	0.0496**	0.1283**
Log(Secchi)	0.0076	-
Log(Secchi) x 0-100m buffer	0.1673**	-
Log(Secchi) x 100-300m buffer	0.0405**	-
Log(chl-a)	-	-0.0115*
Log(chl-a) x 0-100m buffer	-	-0.1047**
Log(chl-a) x 100-300m buffer	-	-0.0276**
N	746,424	637,548

Notes: *p<0.05, **p<0.01 Standard errors clustered at census tract level. The first column shows the estimate results using Secchi depth as water quality measurement. The second column shows the estimate results using Chlorophyll-a as water quality measurement.

Table 3.3: Total capitalization effects for marginal improvements and lake restoration

	Market Value (billion \$)^2	Last Sale Price (billion \$)^2
Marginal Improvements		
10% increase in Secchi	9.22	11.09
10% decrease in chl-a	6.46	7.81
Pristine Ecological Condition		
2012 NLA standards	25.18	-
2017 NLA standards	26.67	-

Notes: Overall impacts for all properties within 300m of lakefront for all NHD lakes (n=76,131) larger than 4 ha identified using PLACES parcel data. Most recent market values and last sales prices from ZTRAX averaged at the lake level within a county for each buffer distance (0-100m and 100-300m)

Table 3.4: Aggregating water quality parameters[50]

Parameter	**WQP Characteristic Name**	**WQP measurement units**
Secchi	Depth, Secchi disk depth; Light attenuation at measurement depth; Light attenuation coefficient; Light attenuation, depth at 99%	m; ft; cm; in; %
chla	Chlorophyll-a; Chlorophyll-a - Phytoplankton (suspended); Chlorophyll-a (probe relative fluorescence); Chlorophyll-a (probe) Chlorophyll-a, uncorrected for pheophytin	µg/L; ug/m3; mg/L; mg/m3; ppb

[50] Ross et al. (2019) also used a similar approach to aggregate data.

Table 3.5: EPA-WQP VS LAGOS-NE epilimnion water quality data for Minnesota

	EPA-WQP	LAGOS-NE	Match	Match (%)
Latest data date	11/6/2019	11/18/2012	NA	NA
Number of lakes covered	5,018	4,565	4,349	95.30%
Number of activities	599,989	358,009	349,790	97.70%
Secchi depth result	582,100	350,404	345,842	98.70%

Table 3.6 Secchi depth results for varying spatial fixed effects

	Tract x year	Block group x year	County x year
0-100m buffer	0.4473**	0.4407**	0.4342**
100-300m buffer	0.0496**	0.0518**	0.0695**
Log(secchi)	0.0076*	0.0287*	0.0199*
Log(secchi) x 0-100m buffer	0.1673**	0.1581**	0.1778**
Log(secchi) x 100-300m buffer	0.0405**	0.0388**	0.0435**

Note: *p<0.05, **p<0.01 Standard errors clustered at census tract level. N = 746,434. The first column shows the same as the baseline model in Table 3.2. The second and third columns show the estimate results with block group by year and county by year fixed effects, respectively.

Table 3.7: Secchi depth results for varying minimum years of water quality samples per lake

	1-Year	2-Year	3-Year	5-Year	7-Year	10-Year
0-100m buffer	0.4473***	0.4624***	0.4723***	0.4723***	0.4644***	0.4539***
100-300m buffer	0.0496***	0.0505***	0.0541***	0.0459***	0.0393***	0.0431***
Log(Secchi)	0.0076	0.0085	0.0072	0.0136	0.0101	-0.0018
Log(Secchi) x 0-100m buffer	0.0405***	0.0414***	0.0401***	0.0455***	0.0550***	0.0497***
Log(Secchi) x 100-300m buffer	0.1673***	0.1573***	0.1532***	0.1586***	0.1666***	0.1377***
Number of lakes	1,632	1198	940	681	503	299
States	43	42	40	34	34	27
N	746,424	675,622	608,687	505,254	438,366	313,698

Note: *p<0.05, **p<0.01 Standard errors clustered at census tract level. This table summarizes the estimate results with different subsets of the transaction data. For example, the second column labeled as ‘2-Year’ summarize the results estimated with the subset that only contains lakes with Secchi depth records in at least 2 years.

Table 3.8: Secchi depth results for varying minimum number of observations within 100m

	1	10	30
0-100m buffer	0.4473^{***}	0.4785^{***}	0.4907^{***}
100-300m buffer	0.0496^{***}	0.0548^{***}	0.0538^{***}
Log(Secchi)	0.0076	0.0115	0.0268^{*}
Log(Secchi) x 0-100m buffer	0.0405^{***}	0.0395^{***}	0.0464^{***}
Log(Secchi) x 100-300m buffer	0.1673^{***}	0.1498^{***}	0.1424^{***}
Number of lakes	1632	571	312
States	43	34	30
N	746,424	515,025	401,025

Note: *p<0.05, **p<0.01 Standard errors clustered at census tract level. This table summarizes the estimate results with different subsets of the transaction data. For example, the second column labeled as ‘10’ summarizes the results estimated with the subset that only contains lakes with at least 10 transactions during our study period.

Table 3.9: Secchi depth results for repeated sales with varying fixed effects and sample restrictions

	(1)	(2)	(3)	(4)
Log(Secchi)	-0.0145	0.0240	-0.0094	0.0280
Fixed effect	property id	property id	property id + year	property id + year
Sample restriction	none	within 300m	none	within 300m
N	568,352	36,986	568,352	36,986

Note: *p<0.05, **p<0.01 Standard errors clustered at census tract level. There are no restrictions on the minimum water quality records or housing transactions per lake.

3.8 Appendix: Water quality data -- aggregation and harmonization

The water quality data comes from two sources: water quality portal (WQP) of the United States Environmental Protection Agency (USEPA) and the ‘Lake Multi-scaled Geospatial and Temporal Database (LAGOS-NE). Although there exists a complementary relationship between these two water quality datasets, LAGOS-NE also provides lake limnological context with validation but is limited in geospatial and temporal dimension. We aim to clean WQP data and then merge with LAGOS-NE data. The water quality data is then merged with the property dataset.

The WQP is a community-sourced repository that is updated bi-weekly. We downloaded the data on March 03, 2021, for all the lakes in the US. There are more than 50 million water quality observations of 3,393 water quality parameters. The WQP reports these by monitoring stations and by activity date. Out of all these parameters, Secchi depth and Chlorophyll-a are the most used water quality measures. However, in the database there are several measures that can lead to Secchi depth and Chlorophyll-a. They varied by names, measurement units, or measurement method altogether. Table 3.4 provides frequency and WQP raw characteristic names of Secchi depth and Chlorophyll-a measures that are used in our work. Note that the measurement units and/or methods may be different. In the case of measurement, units we used appropriate conversion factors to harmonize Secchi depth unit as meter (m) and Chlorophyll-a unit as microgram per liter (µg/L). We convert light attenuation coefficient measure to Secchi depth by dividing by 1.45 as described in Walsh et al. (2017) and USEPA (2003).

LAGOS-NE provides validated water quality data for 17 northeast and midwestern states in the US (Soranno et al. 2017). We used the latest publicly available LAGOS-NE dataset through an R package (lagosne v1.087) (Stachelek et al. 2020). A map showing LAGOS-NE lakes and states is in Figure 3.3. Epilimnion water quality data includes 17 parameters including Secchi depth, Chlorophyll-a, apparent color, true color, total kjeldahl nitrogen, and dissolved organic carbon. The sample date ranges from 1925 to 2016. There are altogether 816,095 activities of epilimnion water quality data collected for 14,657 lakes.

In addition to epilimnogical parameters LAGOS-NE also provides lake geographic location and corresponding National Hydrographic Dataset (NHD) Permanent Identifier (ID). To merge WQP with LAGOS-NE we use a geospatial join technique to match WQP monitor location with NHD data. WQP monitor spatial data also varied in terms of geo datum. We used the most common three ('NAD83', 'WGS84', 'NAD27') and 'UNKWN' is coded as 'WGS84'. A total of 97.6% of the monitoring station's location can be matched. Unlike LAGOS-NE, the WQP reports by monitoring location and by activity date. LAGOS-NE converts multiple monitoring locations into a centroid, regardless of monitoring locations (Ross et al. 2019). An illustrative example of lakes with multiple monitoring stations are shown in Figure 3.4. In this event, we take the average of water quality parameters by lake. We then match LAGOS-NE and WQP data by activity date and NHD lake ID. The WQP and LAGOS-NE water quality parameters are complementary to each other, yet distinct in geographic and temporal dimensions. Figure 3.5 shows their overlap and extensions. The WQP is both geographically and temporarily diverse as compared to LAGOS-NE. For some states and years where LAGOS-NE provides data, it has more coverage than the WQP. But LAGOS-NE does not have any data after the third quarter of 2016 and outside of 17 states. We prioritize LAGOS-NE data over WQP data as LAGOS-NE is compiled by harmonizing the metadata and comes with lake ecological context. See Table 3.5 for an example of comparison of matching between LAGOS-NE and WQP Secchi depth measures in Minnesota.

The merged dataset has many zero values and inf values along with some unusual high values coming from user uploaded WQP data. We deleted zero and inf values and used a 99th percentile cutoff for maximum value. The water quality data is then aggregated by year and by lake only for summer months (April to September), as lake clarity and Chlorophyll-a measurement are very unusual for winter months, especially for northern part of the country where lakes freeze during winter. We used the median water quality measures of the summer months.

3.9 Appendix: Robustness analysis

We first examine the model with different spatial fixed effects to address the concern that the spatial unit of the census tract in our baseline model may absorb too much variation or leave out endogenous variables. As shown in column 2 and 3 of Table 3.6, the results are robust when we control for the fixed effects at the year by census block group and year by county levels.

Considering the sparsity of our water quality data over the long study period, lakes with limited amounts of observations may have biased Secchi depth values and further that affect our estimation of the impact. Results in Table 3.7 are estimated from models with different restrictions on the minimum years having water quality records for each lake. All columns show robust positive marginal willingness to pay from residents within the 0-100m and 100-300m buffer areas of lakes.

In addition to the concern on water quality data, the small coverage of the lakeshore area may also restrict the number of lakeshore observations. If the amount of lakeshore transactions is limited, it may falsely represent the housing price and characteristics that may cause bias in the estimates. To assess this impact, we also estimated the baseline model with different samples by restricting the minimum number of lakeshore observations at 10 and 30. The results shown in Table 3.8 are consistent with minor changes in the magnitude of the coefficients.

To fully capture time-invariant housing characteristics, we estimated a series of repeated sale models with properties sold more than once. The main results are summarized in Table 3.9. In addition to the Secchi depth variable, column (1) and (2) include property fixed effects. We found a negative but insignificant coefficient of Secchi depth while estimating with all data. When we move to the model with only transactions within 300m of lakes, the coefficient becomes positive but still insignificant. This is inconsistent with the results of our baseline model that lakefront properties are more likely to be affected by water quality. Thus, we believe the positive coefficient of the Secchi depth variable also captures other features in addition to water quality for two main reasons: First, the price changes may be driven by the general trend of the housing market over years. Second, property fixed effects do not control for timevariant housing characteristics like the improvement of house conditions with remodeling. With these two considerations, the coefficients on the Secchi depth variable in column (1) and (2) of Table 3.9 may be biased. To account for this possible bias in estimations, we include year fixed effects to control for average changes over years. Columns (3) and (4) show insignificant results for models running with all transactions and transactions within 300m of lakes. Note that more than half of the properties were only sold twice during our study area. Both the property and year fixed effects may take up too much of the variation. With the concerns on overidentification and insufficient variation, we think our baseline model is a better choice than the repeated sale model.