7-2020

Using 'big data' to explain visits to lakes in 17 US states

Erik Nelson
Bowdoin College, enelson2@bowdoin.edu

Maggie Rogers
University of Minnesota

Spencer Wood
University of Washington

Jesse Chung
Bowdoin College

Bonnie Keeler
University of Minnesota

Using 'big data' to explain visits to lakes in 17 US states

Erik Nelson[1], Maggie Rogers[2], Spencer Wood[3], Jesse Chung[1], and Bonnie Keeler[2]

Abstract: We use large dataset on US lakes from 17 states to estimate the relationship between summertime visits to lakes as proxied by social media use and the lakes' water quality, amenities, and surrounding landscape features and socioeconomic conditions. Prior to estimating these relationships we worked on 1) selecting a parsimonious set of explanatory variables from a roster of more than 100 lake attributes and 2) accounting for the non-random pattern of missing water quality data. These steps 1) improved the interpretability of the estimated visit models and 2) widened our estimated models' scope of statistical inference. We used Machine Learning techniques to select parsimonious sets of explanatory variables and multiple imputation to estimate water quality at lakes missing this data. We found the following relationships between summertime visits to lake and their attributes across the 17-state region. First, we estimated that every additional meter of average summer-time Secchi depth between 1995 and 2014 was associated with at least 7.0% more summer-time visits to a lake between 2005 to 2014, all else equal. Second, we consistently found that lake amenities, such as beaches, boat launches, and public toilets, were more powerful predictors of visits than water quality. Third, we also found that visits to a lake were strongly influenced by the lake's accessibility and its distance to nearby lakes and the amenities the nearby lakes offered. Finally, our results highlight the biased results that "big data"-based research on recreation can generate if non-random missing observation patterns in the data are not corrected.

Keywords: Lake-based recreation; social media data; Flickr; big data; Secchi depth; LASSO; random forests; multiple imputation; Poisson count models

JEL Codes: Q26, C26, C55

[1] Department of Economics, Bowdoin College, Brunswick, ME. enelson2@bowdoin.edu.
[2] Hubert H Humphrey School of Public Affairs, University of Minnesota, Minneapolis, MN.
[3] EarthLab, College of the Environment, University of Washington, Seattle, WA.

1. Introduction

Measuring human interaction with nature is important for several reasons. First, such exercises help us understand the values that people ascribe to nature. Second, accurate descriptions of these interactions give policy-makers and nongovernmental organizations insights into the conservation and environmental policies that will increase human welfare the most.

In the past, estimating how we use and relate to nature may have been limited by data availability. Collecting data on human behavior in and around nature has traditionally been expensive and time-consuming (Wood et al. 2013, Richards and Friess 2015, Harari et al. 2016). But just as businesses and governments around the world have been inundated with "big data" that describe customer and constituent behavior (Hofacker et al. 2016, Matz and Netzer 2017, Ilieva and McPhearson 2018, Milne and Watling 2019), there are more and more large datasets that capture the state of the physical world and some of our interactions with nature (Roberge 2014, Di Minin et al. 2015, Hausmann et al. 2018, van Zanten et al. 2016). While this data revolution means that estimates of our interactions with nature can be more precise and robust, this big data also presents some unique analytical problems.

As data has become cheaper and easier to collect, the number of potential predictors of our behavior in and around nature has become very large and analysts may be tempted to use all of these predictors when estimating our interactions with nature. However, models that omit many potential predictors – parsimonious models – are desirous for several reasons. First, many of the potential predictors will have little to no relationship to the behavior of interest. The presence of many irrelevant variables in a model may make it hard to identify and isolate

the most essential correlates. Second, an over-parameterized model may be a poor predictor of human behavior when used with datasets that were not used to estimate the human behavior model. A model that hews too closely to the set of noise in the dataset used to "train" the model is likely to perform poorly when used with a separate dataset of the same variables but a different pattern of noise (Harrell et al. 1996). Therefore, a model parametrized with as little irrelevant noise (i.e., variables) as possible is likely to have less prediction error with novel datasets than an "overfit" model. Finally, a model that includes only the most relevant predictors is easier to explain and interpret. Policy makers are more likely to act on research that focuses on the most important correlates and can be easily explained to a broad audience (Pullin and Knight 2005, Karl et al. 2007, Ruckelshaus et al. 2015).

The ways in which "big data" are created can also create analytical difficulties for the data user. Unlike more consciously constructed research projects where data collection is carefully planned, "big data" are often an amalgam of various data streams that, when cobbled together, has missing observations. When this occurs one option is to drop every observation that does not have a complete set of attributes. However, this action can reduce the usefulness of an estimated model in two ways. First, it makes it the estimated model less precise. Second, this action will bias the estimated model if there is a pattern to the observations that are missing data. Therefore, correcting for patterns of missing data is likely to be important when using big data to estimate how humans use and relate to nature.

2. Our study

We estimate how lake water quality, lake amenities, and landscape features around lakes affect summer-time visits to a large number of US lakes. Understanding how lake water quality effects visits to lakes relative to other recreational drivers such as lake amenities and lake access can lead to more effective and coherent lake-based environmental and recreational policy.

The data we use to explain summer-time lake visits are "big" in two ways. First, *all* lakes 4 hectares or larger across 17 contiguous states have a measure of total summertime visits between the summers of 2005 to 2014 (N = 51,107). Second, there are more than 100 variables that describe each lake's water quality, set of amenities, and surrounding landscape over this same time period (SI Text 1). Having such expansive data on lake-based recreation is unique; typical park or lake visit studies may include at most 100 natural features and have data on 20 to 30 visit predictors (Phaneuf 2002, Hunt and Dyck 2011, Smirnov and Egan 2012, Donahue et al. 2018). Further, many past studies have focused on parks or lakes from the same landscape or region where differences in site attributes may be too slight to identify what lake features drive visitation rates (Yi and Herriges 2017). Therefore, our estimates on the relationships between summertime lake visits and lake attributes are 1) likely to be more valid in more external cases than usual and 2) more precise than past estimates due to the geographic scope and attribute heterogeneity in our data. However, while the bigness of our data gives our analysis heft and unprecedented scope, it also presents analytical complications.

One challenge we faced was explanatory variable selection. As noted in the Introduction, using all 100 plus explanatory variables at our disposal to explain lake visits would have create a model poorer at predicting visits to out-of-sample US lakes than more

parsimonious models. In addition, policy-makers and concerned citizens interested in crafting better lake recreation and conservation policy would likely ignore visitation model results if the analysis was too large and unwieldy for quick comprehension. Therefore, we used Machine Learning (ML) algorithms to systematically build parsimonious lake visitation models. The objective of ML algorithms we used is to find a parsimonious set of explanatory variables that predict out-of-sample visit rates better than alternative model specifications (e.g., Deryugina et al. 2019). The limited size of ML-informed models also aided us in our efforts to generate easily interpretable models.

Missing water quality and lake depth measures for most lakes in our dataset is another analytical challenge we faced. For example, Secchi depth, the most prevalent lake water quality measure in our dataset, and lake depth were measured for less than a fifth of the lakes. Further, we found strong evidence that missing Secchi and lake depth observations were not randomly distributed throughout our dataset (see below). Therefore, if we naïvely included Secchi and lake depth in our explanatory variable selection algorithms or visitation models we would have created two analytical difficulties. First, we would have lost estimation precision due to the omission of most lakes from the model. Second, model estimates would likely be biased given the non-random pattern in missing Secchi and lake depth observations. To avoid these outcomes, we experimented with imputing Secchi and lake depth observations for lakes missing this data.

Finally, we used the ML-generated models with and without imputed Secchi and lake depth data to estimate the relationships between lake visits and lake attributes across the 17-state region our dataset covers. We also compared the performance of the ML-constructed

models to the performance of a lake visitation model made up of variables suggested by the recreational demand literature. In addition, we determined how robust our default estimated relationships were to several modeling and data structure assumptions we made.

We found the following. First, on average, every additional meter of average Secchi depth during the summer months between 1995 and 2014 was associated with at least 7.0% more summer-time visits to a lake from 2005 to 2014, all else equal. However, while higher lake water quality was associated with more visits to a lake, we consistently found that lake amenities, such as beaches, boat launches, and public toilets, were more powerful predictors of visits than water quality. We reached this conclusion using two bits of information. First, the ML algorithms often selected lake amenities as strong predictors of visits while almost never identifying Secchi depth as a strong predictor of visits. Second, the estimated coefficients on the amenity variables in the lake visitation model were larger than Secchi depth's estimated coefficient. We also found that the rate of visits to a lake were strongly influenced by lake accessibility and the distance to nearby lakes and the amenities the nearby lakes offered. Finally, if we had only considered lakes with measured Secchi and lake depths our conclusions on the relationships between lake visits and lake attributed would have been biased. Across the non-random set of lakes with measured Secchi and lake depths the importance of lake amenities and the location of and attributes at nearby lakes are minimized relative to their explanatory power across the full set of lakes where Secchi and lake depth measures have been imputed when missing. The potential bias created by non-random missing data patterns in "big data" must be accounted for by researchers and policy-makers working with these data to accurately explain human behavior in nature.

3. Data

We assume that there is a quantifiable relationship between summer visits to a lake and its characteristics, including its water quality, amenities, and accessibility, of the form $V = f(\mathbf{Z})$ where V indicates the count of summer visits to the lake over time period T and $\mathbf{Z}$ is a vector that describes the lake's characteristics during T.

We do not observe lake visits. Instead, we observe Y, the number of photo-user days (PUDs) generated during the summers (June 15 to September 15) of 2005 through 2014 within the boundaries of each of the 51,107 lakes in our dataset (Fig. 1). The photos we used to generate summer PUD counts were found on the photo-sharing site Flickr (Wood et al. 2013; Figs. 2 and 3). A lake's summer PUD count increases in every one of its unique Flickr user-days during the 2005 to 2014 summers. For example, if Jack posted 5 photos and Jill posted 10 photo taken within lake *j*'s boundaries on August 1, 2008 to Flickr then *j*'s summer PUD count increased by 2. If Jill subsequently posted 3 photos taken within lake *j*'s boundaries on August 4 of 2008 then *j*'s summer PUD count increased by another unit.

Wood et al. (2013), Keeler et al (2015), Sonter et al. (2016), Sessions et al. (2016), Levin et al. (2017), and Tenkanen et al. (2017) have shown that the relationship between Flickr-based PUD counts (Y) and observed visits (V) to recreational sites can be represented by the linear relationship $V = \theta + \alpha Y$ where $\theta, \alpha > 0$.[1] Suppose our estimate of $\mathbf{Y} = f(\mathbf{Z})$ measures the

[1] Flickr is now a much less popular photo-sharing site than Instagram (it is not clear this was the case during our 2005 to 2014 timeframe). Further, Instagram tends to have younger users than Flickr. However, using data from Europe, van Zanten et al. (2016) found that Flickr and Instagram users tend to post photos of similar landscape features. When Tenkanen et al. (2017) compared Flickr to Instagram and Twitter data to manually counted monthly visits to South Africa parks they found that Instagram and Twitter predicted visits better. However, there was less discordance between the social media sites' predictive powers across Finland parks.

percentage change in summer 2005 to 2014 PUD counts at a representative lake given a one-unit increase in Z_i, all else equal. Assume $(Y'-Y)/Y = 0.2$ and $(Y'-Y)/Y = 0.3$ for one-unit increases in Z_j and Z_k, respectively, where Y' and Y are the initial and subsequent values of Y. Therefore, we can conclude that a one-unit change in Z_j was expected to increase the summer 2005 to 2014 PUD count at the lake by 20%, all else equal (and a one-change in Z_k causes a 30% increase, all else equal). However, despite not knowing θ and α, we can also use this estimate to say something useful about expected visits to the lake. If $V = \theta + \alpha Y$ then $(V'-V)/(V-\theta) = 0.2$ and $(V'-V)/(V-\theta) = 0.3$ given the one-unit increases in Z_j and Z_k, all else equal. Therefore, given it is reasonable to expect $\theta > 0$, we can interpret the estimate of $\mathbf{Y} = f(\mathbf{Z})$ to mean that a one-unit increase in Z_k increases $(V'-V)/V$ by *at least* 30 percentage points (lower bound). Further, a one-unit increase in Z_k increases V by at least 10 percentage points relative to the impact of a one-unit increase in Z_j.

Unlike lake visits, lake characteristics – variables that could be part of **Z** – are directly observed. Most of the lake data come from LAGOS-NE-LIMNO v1.087.3 (Soranno et al. 2019) and LAGOS-NE-GEO v1.05 (Soranno and Cheruvelil 2017). LAGOS (Soranno et al. 2017) is a set of data products that contains water quality (LAGOS-NE-LIMNO) and ecological and landscape context (LAGOS-NE-GEO) data for lakes found in 17 US states. In our study we limit ourselves to the 51,107 lakes in LAGOS that are 4 hectares or greater. Of all the lake water quality measures included in LAGOS-NE-LIMNO, Secchi depth is the most prevalently recorded measure across the 51,107 lakes (Fig. 4). The greater a lake's Secchi depth, the greater the lake water's clarity. Measures of a lake's total phosphorous (TP), chlorophyll-a (Chlor), and nitrate and nitrite (NO_2NO_3) are less common in LAGOS. In this study we will rely on the mean of Secchi

measurements taken from June 15 to September 15 in the years 1995 to 2013 as a representative measure of a lake's quality during the summers of 2005 through 2014. Only 9,005 of the 51,107 lakes in our dataset (17.62%) have at least one Secchi measure taken from June 15 to September 15 in the years 1995 to 2013 (the percentages are 10.66, 13.32, and 6.58, respectively, for TP, Chlor, and NO_2NO_3).[2]

Besides water quality measures, LAGOS records include many other lake characteristic variables. Each lake's location, size, depth, and home subwatershed and county are given. Further, land use distributions at the 500-meter buffer and at the subwatershed-level (12-unit hydrological units) in 2001, 2006, and 2011 are given for each lake. Subwatershed-level data on thirty-year climate averages and stream, wetland, and lake density are also provided in LAGOS. Unfortunately, mean and maximum lake depth information is only given for a minority of the lakes in the dataset. For example, maximum depth is provided for 9,371 or 18.3% of the 51,107 lakes in the dataset.

We generated several other lake and subwatershed-level variables that could potentially explain summer visits to a lake. We used OpenStreetMap to count the number of features tagged as a marina, boat launch, beach, hotel, shelter, toilet, picnic area, or BBQ facility (OpenStreetMap contributors 2015) at each lake as of 2016. For example, a lake with a '5' for the picnic amenity had 5 features surrounding the lake tagged as a picnic area as of 2016 (Figure 5). Boat launches are the most prevalent amenity across the 51,107 lakes in our

[2] LAGOS does not have 2014 Secchi measurement data. Even though Secchi measurement data from 1995 to 2004 does not overlap with the PUD dates, we reach that far back in time to make the roster of lakes with summer Secchi measures as large as possible. For example, if we limited our data to summer Secchi measures from 2005 to 2013 only 5,817 lakes would have a summer Secchi measure.

dataset: 1,357 or 2.7% of the lakes have at least one boat launch feature as of 2016. Beach features are found at 1.1% of the 51,107 lakes. Only 1,919 lakes or 3.8% of the 51,107 lakes have at least one amenity of any type. We also found each lake's distance to the nearest core-based statistical area (CBSA; in kilometers; US Census 2017A).

In addition, we created a suite of subwatershed-level socioeconomics variables. First, we found the 2010 population and population density (people per square kilometer) in each subwatershed with USEPA's 30m dasymatric population raster (USEPA 2013). Then, using 2011-2015 American Community Survey, 5-year average data at the block group level in conjunction with the dasymatric population raster, we created a series of 2011-2015 socio-economic population maps at the 30-meter grid cell level, including number of people with a Bachelor's degree, number of Non-Hispanic whites, etc.[3] Finally, these maps were used to summarize socio-economic conditions circa 2011-2015 in each subwatershed. The socioeconomic variables in our dataset include the 1) percentage of the subwatershed population that is Hispanic or Latino of any race, 2) percentage of the subwatershed population that is Non-Hispanic white, 3) percentage of the subwatershed population that is black, 4) percentage of the subwatershed population with a Bachelor's degree or more, 5) percentage of the subwatershed population that is living below the Federal poverty, 6) the subwatershed's median household income, and 7) the subwatershed's median age.[4] Each lake *j* was assigned the socioeconomic values of its parent subwatershed.

[3] For example, if raster cell j is in block group k then the population in cell j according to the dasymatric population raster was multiplied by the percentage of Non-Hispanic whites in block group k to generate the number of Non-Hispanic whites in cell j.

[4] For median and mean household (HH) income we used a gridded map of household population rather than a gridded population map. If raster cell j is in block group k then the number of households (HHs) in cell j was multiplied by the median or mean HH income in block group k to generate the total HH income in cell j. Then the

Further, we also generated a set of spatially lagged variables. These variables summarize lake characteristics at *j*'s nearest neighbors. For example, lake *j*'s spatial lag of summer PUDs is given by $\mathbf{w}_j\mathbf{Y}$ where $\mathbf{w}_i$ is a [1 x 51107] vector of distance weights between lake *j* and all other lakes in our dataset and **Y** is the [51107 x 1] vector of summer PUD counts. In our case, $w_{ji} = \frac{1/d(j,i)}{\sum_{i=1}^{51107} 1/d(j,i)} \in [0,1]$ where $d(j,i)$ is the Euclidean distance between lakes *j* and lake *i*. Furthermore, we also created eight lake amenity spatial lag variables for each lake *j* using $\mathbf{w}_j\mathbf{A}_k$ where $\mathbf{A}_k$ is the [51107 x 1] amenity *k* feature count vector across all lakes (there are eight $\mathbf{A}_k$, one for each amenity type). In addition, we made the variable *lag Secchi*$_j$ equal to $\mathbf{u}_j\mathbf{S}$ where $\mathbf{u}_j$ is a [1 x 9005] vector of inverse distance weights between lake *j* (a lake that may or may not have a summer Secchi measure) and all lakes with a summer Secchi measurement in our dataset and **S** is the [9005 x 1] vector of observed average summer Secchi depths across lakes with observed summer Secchi depths. The spatial lags of summer PUD count, amenity counts, and average summer Secchi for lake *j* are higher if *j*'s nearest neighbors (or at least those neighbors with average summer Secchi measures in the latter case) have higher summer PUD counts, amenity counts, and average summer Secchi depths, respectively.

While the lag variables summarize characteristics of the lake nearest to *j*, they do not indicate if lake *j* has several lake neighbors nearby or if the lake is spatially isolated on the landscape (the row normalization of the inverse distances between lake *j* and all other lakes

median or mean HH in a subwatershed was found by summing the cell values across all cells in a watershed and dividing total HH income (median or mean-based) in the watershed by the total HHs in in the watershed. For median age we used the gridded population map. If raster cell j is in block group k then the number of people in cell j was multiplied by the median age in block group k to generate aggregate age in cell j. Then the median age in a subwatershed was found by summing the cell values across all cells in a watershed and dividing aggregate age in the watershed by the watershed's population.

obviate any measure of absolute distances between lakes). To summarize *j*'s spatial position on the landscape relative to other lakes we measured 1) the average distance between lake *j* and its five nearest neighbors; 2) the average distance between lake *j* and its five nearest neighbors that have summer Secchi measures; 3) the average distance between a lake and its closest CBSA of lake *j*'s five nearest neighbors; and 4) the average size of lake *j*'s five nearest neighbors.

3.a. The set of lakes with summer Secchi and maximum lake depth measurements is not a random sample from the population of lakes.

As we noted above, data on summer Secchi depth, our primary water quality variable, and maximum lake depth, another indicator of lake water quality, are missing for most of the lakes in our dataset. (Deeper lakes are less likely to transition to the eutrophic state than shallower lakes, all else equal (Qin et al. 2020).) Therefore, an estimate of **Y** = f (**Z**) where **Z** includes average summer Secchi and/or maximum lake depth would be based on a limited number of lakes (whereas data on all other variables we considered for **Z** are consistently observed). While omitting lakes without Secchi and/or maximum lake depth from an estimate of **Y** = f (**Z**) would mean a loss in statistical power, it would not affect model inference if the dropped subset of lakes was a random draw from the population of lakes (Jakobsen et al. 2017). On the other hand, if the lakes with the depth measures was not a random draw from the entire population of lakes then any estimate of **Y** = f(**Z**) where **Z** includes Secchi and/or maximum lake depth would produce results that could not be used to infer the relationships between **Y** and **Z** across the population of lakes.

O'Sullivan and Unwin (2010) suggest two methods for testing whether a sample of spatial points could reasonably be considered a random sample of the population of spatial points. To conduct the first suggested analysis, we began by randomly selecting (without replacement) 9,005 lakes from the dataset of 51,107 lakes 1,000 times (recall there are 9,005 lakes with average summer Secchi depth based on 1995 to 2013 measurements). Next, for each of the 1,000 samples, indexed by *s*, we calculated the mean nearest lake distance, given by $\bar{d}_{min,s}$.

$$\bar{d}_{min,s} = \frac{\sum_{j=1}^{9005} d(l_j)}{9005} \quad (1)$$

where $d(l_j) = \min\{d(l_j, l_1), \dots, d(l_j, l_{i-1}), d(l_i, l_{i+1}), \dots, d(l_i, l_{9005})\}$ and $d(l_j, l_i)$ is the Euclidean distance between sampled lakes *j* and *i*.

The mean and standard deviation of $\bar{d}_{min}$ across the 1,000 samples are 5,133.9 and 45.0 meters, respectively. Given $\bar{d}_{min}$= 4,462.7 meters for the 9,005 lakes with average summer Secchi depth measurements based on 1995 to 2013 measurements (the "Secchi lakes") there is an infinitesimally small probability that the Secchi lakes' spatial pattern was a random draw from the 51,107 population of lakes (Fig. 5A). A similar analysis of the 9,371 lakes with measured maximum depth also indicates there is an infinitesimally small probability that the spatial pattern of lakes with observed maximum depth measures was a random draw from the 51,107 population of lakes (Fig. 6A). Instead, some spatial process likely explains which lakes have and have not been measured for Secchi depth and maximum depth.

One criticism of the $\bar{d}_{min}$ test it only considers each lake's nearest neighbor. If nearest-neighbor distances are short relative to distances to all other lakes then the $\bar{d}_{min}$ test could be spurious (O'Sullivan and Unwin 2010). Therefore, we used an alternative test that uses all lake

distances to verify that the subsets of lakes with average summer Secchi and maximum lake depth measures are not representative of a random sample of the population of lakes. We first found the *K* function for each of the 1,000 random draws of 9,005 lakes (without replacement). The *K* function for a sample was generated by drawing a series of concentric circles with radius *d* around *each* lake *j* in the sample, counting the number of the other 9,004 sampled lakes in each circle of radius *d* around lake *j*, and then estimating the mean density of sampled lakes in each circle of radius *d* across all *j*,

$$K(d) = \frac{\sum_{j=1}^{9005} \#[S \in C(j,d)]}{9005 \times \left(\frac{9005}{A}\right)} \quad (2)$$

where *S* is set of sampled lakes, $\#[S \in C(j,d)]$ indicates the number of sampled lakes in the circle *C* centered on *j* with radius *d*, *A* is the area of the LAGOS region, and (9005/A) measures the intensity of sampled lakes on the landscape. We also generated the *K* function for the set of 9,005 Secchi lakes.

Finally, we plotted the 1,000 normalized *K* functions (called *L* functions) of the 1,000 randomly drawn samples of 9,005 lakes and the normalized *K* function of the 9,005 Secchi lakes for diameters of 100, 200, 300, …, 70000 meters on the same graph (Fig. 5B). We repeated the normalized *K* function analysis for maximum lake depth as well (see Fig. 6B). Evidence that the Secchi lakes and lakes measured for maximum depth do not represent a random draw from the population of lakes is indicated by the normalized *K* functions lying outside the ranges formed by the 1000 random sample normalized *K* functions. Instead some unknown spatial process determined which lakes were measured for Secchi and maximum lake depths.

3.b. Imputing missing summer Secchi and maximum lake depth

Because we found that lakes with average summer Secchi and maximum lake depth measurements are likely not a random sample of all lakes, an estimate of **Y** = f(**Z**) when **Z** includes average summer Secchi and/or maximum lake depth and does not correct for measurement selection bias could only be used to infer relationships between **Y** and **Z** across lakes measured for this data and not the population of lakes. Econometricians have developed several approaches to correct for selection bias in data and make models estimated with the biased data appropriate for population-level inference. However, we eschewed these approaches in this research because we have not identified the spatial processes which determine whether lakes were sampled or not for Secchi and maximum lake depth. Instead we use multiple imputation (MI) to estimate Secchi and maximum lake depth at the lakes where these measures are missing.

MI will generate estimates of missing average summer Secchi and maximum lake depth measures that allow population-level valid inference if average summer Secchi and maximum lake depth are missing at random (MAR). These data are MAR if the probability of their "missing"-ness does not depend on unobserved data but rather can be explained by a combination of **Y** and other observed data (Jakobsen et al. 2017).

Unfortunately, we cannot prove that the "missing"-ness pattern is MAR given unobserved data is unobserved (Jakobsen et al. 2017)! However, we generated a random forest over the observed lake and subwatershed data that accurately predicts whether a lake has an average summer Secchi or maximum lake depth measure 95% of the time (Table 2). We believe that this analysis allowed us to conclude that MAR for is a reasonable assumption for missing average summer Secchi or maximum lake depth data.

For a continuous variable with a restricted range, such as average summer Secchi and maximum lake depth, nearest matching or 'predictive mean matching' (PMM) is a recommended MI method (Raghunathan et al. 2001). Our PMM specification sets average summer Secchi or maximum lake depth at a lake missing one or both measures equal to the average summer Secchi or maximum lake depth at one of the lake's 'nearest' 5 lakes (a lake in the near set will always have an observed average summer Secchi or maximum lake depth). Nearest here refers to average summer Secchi predictive distance, not Euclidean distance.[5] In each MI iteration the lake that 'donates' its average summer Secchi or maximum lake depth to lake *j* is randomly chosen from the set of *j*'s nearest 5 neighbors. We created 20 sets of imputed summer Secchi and 20 sets of maximum lake depths. [6]

Finally, we merged each imputed vector of average summer Secchi depth data with the observed average summer Secchi to create 20 vectors of average summer Secchi depth that had a value for each lake our dataset, either the observed depth or, if this was missing, an

[5] For example, suppose we are imputing average summer Secchi depth. First, a linear model with average summer Secchi depth as the dependent variable and a covariate vector comprised of **Y** and all of the potential explanatory variables is estimated using least squares across all lakes with observed average summer Secchi depth measures. Second, new model parameters are simulated from their joint posterior distribution under the conventional noninformative improper prior $Pr(\boldsymbol{\beta},\boldsymbol{\sigma}^2) \propto 1/\boldsymbol{\sigma}^2$ where $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}^2$ are the least square estimates of the average summer Secchi depth model. Using the simulated model, average summer Secchi depth is then predicted for all lakes missing this data. Suppose lake *j* is missing average summer Secchi depth. Suppose its predicted average summer Secchi depth is *X*. The 5 lakes that have a measured average measured Secchi depth closest to *X* are matched to *j*. In imputation iteration 1 one of the nearest 5 lake's average summer Secchi depth measures is randomly assigned to lake *j*, in imputation iteration 2 one of the nearest 5 lake's average summer Secchi depth measures is randomly assigned to lake *j*, etc.

[6] We simultaneously impute summer Secchi and maximum lake depth, in both cases using predictive mean matching (PMM). To implement this simultaneous imputation we used the multivariate MI method know as MICE. Under this method summer Secchi and maximum lake depth are imputed iteratively with summer Secchi depth imputed first, if need be, conditional on maximum lake depth and all potential explanatory variables (using PMM) and then maximum lake depth is imputed, if need be, conditional on summer Secchi depth and all potential explanatory variables (using PMM).

imputed value. We similarly created 20 vectors of maximum lake depth that had an observed or imputed value for each lake in the dataset.[7]

3.c. Measuring the accuracy of Secchi and maximum lake depth imputation

We tested MI accuracy with a 10-fold cross validation analysis. First, we retained the lakes with an observation for every variable in our dataset, including observed average summer Secchi and maximum lake depths ("complete case" lakes, N = 6,755). Second, we randomly divided the complete case dataset into 10 folds (approximately 675 lakes each). Third, we deleted the observed average summer Secchi and maximum lake depths in the first fold and then, using the same MI algorithm described above, we imputed 20 values of average summer Secchi and maximum lake depth for each lake in the first fold based on the remaining nine folds of complete data. We repeated this process nine more times, each time deleting the observed average summer Secchi and maximum lake depths in the i^{th} fold of lakes and using the other nine folds to impute the missing depth values. Therefore, for each lake k in the complete case dataset we have an observed measure of average summer Secchi depth (s_k), an observed measure of maximum lake depth (d_k), the mean ($\hat{s}_k$) and variance of 20 imputed average summer Secchi depths, and the mean ($\hat{d}_k$) and variance of 20 imputed maximum lake depths.

The correlation coefficient between the vectors $(s, \hat{s})$ is 0.62 and the root mean square error between the predicted and observed average summer Secchi depth is 1.39 meters. The

[7] Technically not every lake in the 51,107-lake dataset was assigned an average summer Secchi depth measure if it was missing an observation. If a lake was missing data on a variable used in the MI process than it was not assigned an imputed average summer Secchi depth. The same issue applied to the assignment of an imputed maximum lake depth to lakes missing this data. Therefore, we created 20 vectors of average summer Secchi depth that had a value for almost every lake in our 51,107-lake dataset

correlation coefficient between the vectors $(d, \hat{d})$ is 0.34 and the root mean square error between the predicted and observed maximum lake depth is 8.63 meters. Therefore, predictions of maximum lake depth are less accurate than predictions of average summer Secchi depth. As can be seen in Fig. 7, the absolute deviations between the observed and predicted average summer Secchi depth and observed and predicted maximum lake depth increase in depth.

4. Methods

In this section we first use recreational demand theory to explain what type of variables should be part of the explanatory variable **Z** vector. Second, we describe how our dataset can be used to construct the variables consistent with theory. Third, we describe several methods for selecting the parsimonious set of variables that can be used to explain lake visits (or more appropriately, our proxy for lake visits). Finally, we describe the functional form we assumed when estimating our lake visitation models (hereinafter we refer **Y** = f(**Z**) as a lake visitation model despite **Y** being a proxy for lake visitation).

4.a. Theoretical foundation for an aggregate visit site model

Assume two recreation locations (e.g., lakes) on a landscape, indexed by *j*. Assume a set of households on the same landscape, *i* = 1,…,*I*. Household *i*'s utility is defined over the number of visits to each recreation site and the consumption of *z* units of the numeraire good. We assume the household's objective is to,

$$\max_{\mathbf{X},z} u_i(\mathbf{X}_i, \mathbf{Q}, z_i, \boldsymbol{\varepsilon}_i) \qquad \text{St: } \mathbf{P}_i\mathbf{X}_i + z_i \leq M_i \qquad (3)$$

where $\mathbf{X}_i = (x_{1i}, x_{2i})$ indicates the number of times *i* visits each site *j*, $\mathbf{Q} = (q_1, q_2)$ indicates the level or density of attributes at each site (e.g., water quality, recreational amenities, quality of vista, etc.), $\mathbf{P}_i = (p_{1i}, p_{2i})$ indicates the household's cost to visit each site in terms of *z* units (cost includes actual outlays and any opportunity costs), $\boldsymbol{\varepsilon}_i = (\varepsilon_{1i}, \varepsilon_{2i})$ captures unobserved factors that influence household location preferences, and M_i is the household's income. $\boldsymbol{\varepsilon}$ is known to the household but unknown to the modeler of demand (this theory section follows von Haefen and Phaneuf 2005).

Assuming consumption of the numeraire good is essential to the household (*z* must be greater than 0), the first-order Kuhn-Tucker conditions for utility maximization are,

$$u_{zi}(\mathbf{X}_i, \mathbf{Q}, \boldsymbol{\varepsilon}_i, M_i - \mathbf{P}_i\mathbf{X}_i) = \delta_i \tag{4}$$

$$\underbrace{\frac{u_{ji}(\mathbf{X}_i, \mathbf{Q}, \boldsymbol{\varepsilon}_i, M_i - \mathbf{P}_i\mathbf{X}_i)}{u_{zi}(\mathbf{X}_i, \mathbf{Q}, \boldsymbol{\varepsilon}_i, M_i - \mathbf{P}_i\mathbf{X}_i)}}_{\substack{\text{MRS between} \\ \text{a trip to } j \text{ and } z}} \leq \underbrace{p_{ji}}_{\substack{\text{Cost of visiting } j \\ \text{an additional time}}} \quad \text{for } j = 1,2 \tag{5}$$

$$x_{ji} \geq 0 \quad \text{for } j = 1,2 \tag{6}$$

$$x_{ji}\left[\frac{u_{ji}(\mathbf{X}_i, \mathbf{Q}, \boldsymbol{\varepsilon}_i, M_i - \mathbf{P}_i\mathbf{X}_i)}{u_{zi}(\mathbf{X}_i, \mathbf{Q}, \boldsymbol{\varepsilon}_i, M_i - \mathbf{P}_i\mathbf{X}_i)} - p_{ji}\right] = 0 \ \text{ for } j = 1,2 \tag{7}$$

where δ_i is *i*'s marginal utility of money. Therefore, *i*'s number of visits to each recreation location is given by the vector of **X** that satisfies (4) – (7),

$$\mathbf{X}_i^* = \left(x_1^*(\mathbf{Q}, \mathbf{P}_i, \boldsymbol{\varepsilon}_i, M_i), x_2^*(\mathbf{Q}, \mathbf{P}_i, \boldsymbol{\varepsilon}_i, M_i)\right) \tag{8}$$

The household comes up with $\mathbf{X}_i^*$ by comparing their willingness to pay (WTP) for a visit to each site to the cost of a visit to a site. The left-hand side of eq. (4) indicates the price the household (HH) is willing to pay (in terms of sacrificed *z* units) for an additional visit to site *j* and the right-hand side of the equation indicates the actual price of a visit (again, in terms of sacrificed *z* units). Assume $x_{1i}^* = 0$. In this case, *i*'s WTP for even 1 visit to site 1 is strictly less

that what it costs to visit. Further, assume $x_{2i}^* > 0$. In this case, *i*'s WTP for the last visit is equal to p_{j2} and WTP values for previous visits to site 2 were greater than p_{j2}.

Now assume $p_{1i} = p_{2i}$ and both x_{1i}^* and x_{2i}^* are greater than 0. In this case, the household will allocate its visits across *j* such that the marginal utilities from the last visit to each *j* are equal. Because *i*'s utility increases in site quality q_j, price equality means that *i* will visit the site with the highest quality the most. Now assume $q_1 = q_2$ and p_{ji} varies across *j*. Now the household will maximize utility by visiting the least costly site more than the costlier site. Simultaneous variation in q_j and p_{ji} across *j* creates a more varied set of choices but the general pattern holds: sites with better quality and less costly to visit will garner the most visits from *i*, all else equal

Suppose the site visit data we observe is not defined at the household level. Rather assume we only observe the welfare maximizing total number of visits to site *j* made across all households,

$$\underbrace{V_j^*}_{\text{Observed}} = \sum_{i=1}^{I} x_{ji}^*\left(\mathbf{Q}, \mathbf{P}_i, \boldsymbol{\varepsilon}_{i,} M_i\right) \geq 0 \tag{9}$$

Therefore, aggregate demand for *j* is also a function of ***Q***, vectors of $\mathbf{P}_i$, vectors of $\boldsymbol{\varepsilon}_i$, and the vector **M**. Just as we do not observe individual household visits to each site, we do not observe vectors of $\mathbf{P}_i$ and **M**. However, if we can find data that allows us to approximate whether visiting site *j* is expensive or not for a representative household (a representation of the vectors $\mathbf{P}_i$) and the income of the representative household that may or may not visit *j* (a representation of **M**) then the aggregate site visit model,

$$V_j^* = f\left(q_j, \mathbf{Q}_{-j}, p_j, \mathbf{P}_{-j}, M_j\right) + \epsilon_j \tag{10}$$

is consistent with assumption of utility-maximizing households. In equation (10) the quality of *j* and the representative cost to visit *j* (q_j and p_j, respectively) have explicitly been separated from the quality of all other *j* and the representative cost to visit all other *j* (Q_{-j} and P_{-j}, respectively). Finally, M_j is the representative income of households that may or may not visit *j* and ϵ_j are the unobservable factors that influence household location preferences that may or may not visit *j*.

4.b. The set of variables that could explain visits to a lake

We can use published lake and park visitation studies to guide us in the selection of variables to represent q_j, $\mathbf{Q}_{-j}$, p_j, and $\mathbf{P}_{-j}$. Fortunately, our lake-level dataset is rich with the suggested variables. In many lake and park visitation studies q_j includes a suite of variables that quantify the recreational activities and related infrastructure available at site *j* and *j*'s water quality, depth, and size (e.g., Parsons et al. 2003, Fleming et al. 2008, Egan et al. 2009, Vesterinen et al. 2010, Kasul et al. 2010, Keeler et al. 2015, Schneider et al. 2005, Allan et. al. 2015, Ziv et al. 2016). While a positive relationship between lake use and its water quality is usually found (e.g., Egan et al. 2009, Vesterinen et al. 2010, Keeler et al. 2015), there are exceptions (e.g., Ziv et al. 2016). Many park visitation models (e.g., Donahue et al. 2018, Zhang and Zhou 2018, Hale et al. 2019) also include measures of land cover (e.g., percent of site with tree canopy, percent of site that is vegetated) and climate at the site to characterize the quality of the site's vista and recreation experience. Therefore, we consider land cover around the lake, the lake's climate, and the hydrological system around the lake (e.g., the extent and density of wetlands and streams around the lake) as descriptors of q_j as well.

Given that summer PUD counts at lake *j* (our proxy for lake visits) cannot be connected to specific households we cannot explicitly estimate p_j, the representative cost of visiting *j*. When faced with a similar difficulty, previous literature has used population density around lakes (in our case, the lake's subwatershed) (Keller et al. 2015, Zhang and Zhou 2018, Hamstead et al. 2018, Donahue et al. 2018) and lake's distance to the nearest population center (Zhang and Zhou 2018) as proxies for p_j. Lakes in or near areas of greater population density will contain many households with low p_j due to close proximity to the lake. This should mean higher than average summer visits (or more precisely, summer-time PUD counts) for these lakes, all else equal, relative to lakes in less dense landscapes and far from population centers (Keeler et al. 2015, Ziv et al. 2016). Such isolated lakes present low-cost recreation opportunities for relatively few households. In addition, past lake visitation studies have found that households occasionally visit multiple lakes on a recreation trip (e.g., Keller et al. 2015). Households displaying such behavior will find visiting *j* costlier if it is isolated from other lakes on the landscape. Therefore, we can use average distance between *j* and its nearest 5 neighbors as an additional indicator of the typical cost of visiting *j*. Finally, the representative household's cost of visiting a lake will be affected by the lake's accessibility. Lakes with parking lots, several access roads, and nearby stores and other built infrastructure will be less costly and easier to visit (i.e., have lower p_j for the representative household). Past research has measured site accessibility with the road density around the site or proximity to public transportation (Zhang and Zhou 2018, Hamstead et al. 2018, Donahue et al. 2018). We can use the percentage of a lake's 500-meter buffer area in developed land covers (e.g., roads, buildings, etc.) as a proxy for lake accessibility.

We can use the same variables that define quality at lake *j* to define the quality of *j*'s set of potential visit substitutes. As discussed in the theory section, households choose among a set of recreation sites to visit; in this case, the choice is among a set of lakes. Presumably, a household that only allocates a few hours for a visit to a lake will choose among the set of lakes within the household's local landscape. On rarer occasions, the household may plan a multiday trip that could include a visit to a faraway lake or two (Vesterinen et al. 2010). In these cases, the set of lakes in the choice set is defined over a much larger region. Therefore, *j*'s competition is generally its neighboring lakes and rarely lakes from very far away. Accordingly, the quality of *j*'s substitutes, given by $\mathbf{Q}_{-j}$, should mostly be defined by the quality of lakes near *j* and less so by the quality at lakes far from *j*. Therefore, we can include the spatially lagged Secchi and amenity feature count variables are part of $\mathbf{Q}_{-j}$.[8]

Above we suggested that the representative household's p_j could be indicated by *j*'s relative proximity to population centers. Therefore, we can use the isolation of *j* relative to *j*'s substitutes as a measure of $\mathbf{P}_{-j}$. For example, suppose lake *j* is 100 miles from the nearest metropolitan area but its 5 nearest lake neighbors are, on average, only 50 miles from the nearest metropolitan area. In this case, assuming distance to a lake is the strongest indicator of visit cost, [9] $\mathbf{p}_j > \mathbf{P}_{-j}$ for a large swath of households on the landscape. Presumably this will mean lest visits to *j*, all else equal. Therefore, the average distance between a lake and its closest CBSA of lake *j*'s five nearest neighbors will be part of $\mathbf{P}_{-j}$.

[8] Interestingly, other than Egen et al. (2009), none of the cited lake and park visitation studies explicitly control for the impact of site substitutes on visitation rates to site *j*.

[9] Vesterinen et al. (2010) modeled lake visitation using data from a survey. They used distance between *i*'s home and the nearest lake site as a measure of p_{ij}.

We used *j*'s subwatershed-level household median income to represent M_j, the income of the representative household that may or may not visit *j*. While a lake can attract visitors from across the US, we assume that most visits to a lake are from local people (e.g., Kasul et al. 2010, Vesterinen et al. 2010). Therefore, we believe median household income in j's subwatershed is the best estimate of representative visitor income we have. Finally, to control for taste preferences of a recreation site's most likely visitors, past literature has also included the racial identities, poverty status educational status, and ages of people living in the areas immediately around site *j* (Zhang and Zhou 2018, Hamstead et al. 2018, Kasul et al. 2010, Vesterinen et al. 2010). To replicate this practice owe can include *j*'s subwatershed-level measures of educational attainment, poverty, racial mix, and age in our lake visitation model. Let the set of taste preferences that affect lake visits controls be collected in vector **S**.

Therefore, based on recreation demand theory and a review of the lake and park visitation modeling literature, our lake visitation model has the form,

$$Y_j = f\left(\underbrace{\mathbf{q}_j, \mathbf{Q}_{-j}, \mathbf{p}_j, \mathbf{P}_{-j}, M_j, \mathbf{S}_j, lagY_j}_{\mathbf{Z}}\right) + \epsilon_j \tag{11}$$

where q_j, $\mathbf{Q}_{-j}$, p_j, $\mathbf{P}_{-j}$, M_j, and $\mathbf{S}_j$ can be made up of the variables discussed above and *lag* Y_j is *j*'s spatial lag of *Y*. See Table 3 for a complete list of variables that, according to our review of the literature, can be used to describe q_j, $\mathbf{Q}_{-j}$, p_j, $\mathbf{P}_{-j}$, M_j, and $\mathbf{S}_j$. Let this collection of variables be called the literature review-informed **Z** or $\mathbf{Z}_{\text{Lit}}$.

Please note our visitation model cannot use the preferred V_j^*, as we do not observe it. Instead we use its observed proxy Y_j. However, as we discussed above, we can interpret (11)'s estimated marginal effects as lower bounds on the impact of changes in explanatory variables

on visits to a lake. Finally, we include *lag* Y_j in our visitation model because a Moran's I test of summer PUD count's spatial pattern indicates we can reject the null hypothesis that there is zero spatial autocorrelation in the dependent variable.[10] Therefore, including a spatial lag of Y in model (11) can reduce the bias in the estimate that the otherwise uncontrolled spatial autocorrelation could cause.[11]

4.c. Using machine learning rather than literature and theory to select variables to include in Z

We use two ML techniques that are designed to identify the parsimonious subset of explanatory variables that accurately predict out-of-sample responses to generate alternatives to $\mathbf{Z}_{\mathbf{Lit}}$ for several reasons. First, a more parsimonious **Z** than $\mathbf{Z}_{\mathbf{Lit}}$ is likely to improve the 1) policy-relevance and 2) out-of-sample predictive accuracy of estimated visitation model (11). With regard to the first point, we believe that research becomes more relevant to policy-makers and concerned citizens if the scope of the model is limited to the most important variables. A model with too many variables and too much detail is likely to be glossed over and disregarded due to limited attention spans in the policy world. As to the second point, an estimate of model (11) based on the literature-informed **Z** is likely to be less predictive of out-of-sample data summer PUDs than an estimate of ML-informed **Z** due to a greater extent of

[10] Global Moran's I Summary of summer PUD count from ArcGIS. The statistic ignores summer PUD counts at lakes 100,000 meters or more from lake j. Moran's Index: 0.001328; Expected Index: -0.000020; Variance: 0.000000; z-score: 4.900560; p-value: 0.000001. Therefore, we can reject the null hypothesis that there is zero spatial autocorrelation present in summer PUD count.

[11] We suspect that much of this spatial clustering in summer PUD counts can be explained by the spatial clustering of explanatory variables. However, even after accounting for the spatial clustering of explanatory variables, we suspect that popularity of lake *j* is influenced by the relative popularity of its neighboring lakes and vice-versa. For example, when people visit lake j they learn of nearby lake k and therefore, may be more likely to visit k in the future. Further, lake k may be an option for a visit when j is too congested (i.e., very popular) for a recreator's liking.

over-fitting in estimate of model (11) based on the literature-informed **Z** than in the estimate based on the ML-informed **Z** (James 2013).

A second reason to use ML-informed **Z**s is to overcome the model selection bias in previous lake and park visitation model literature. Most of these studies refer to each other, reinforcing the notion that a core set of explanatory variables, such as water quality and recreation attributes, are the appropriate variables to use in visitation studies. However, largely unexamined combinations of other lake and subwatershed-level attributes may just as important or even more important in explaining and predicting lake visits. Therefore, running the variable-selection ML algorithms over all the lake and subwatershed-level in variables in our dataset, not just the variables that agree with the literature, can mean models that better predict summer PUD counts than those suggested by the literature.

The least absolute shrinkage and selection operator (LASSO) and random forests are the two ML techniques we use to generate alternative parsimonious **Z** vectors.[12] The coefficients on **Z**, contained in the vector **β** of length *P*, that solves,

$$\min_{\boldsymbol{\beta}} \frac{1}{N}\sum_{i=1}^{N} \ell(Y_i, \boldsymbol{\beta}\mathbf{Z}_i') + \lambda \sum_{p=1}^{P} \left|\beta_p\right| \tag{12}$$

are the (single) LASSO coefficients where **Z'** includes *all* potential explanatory variables (**Z'** contains *P* variables). In this case, the function ℓ is the negative of the Poisson log-likelihood function because the dependent variable, summer PUDs, are count data. The term $\lambda \sum_{p=1}^{P} \left|\beta_p\right|$

[12] Egan et al. (2009) is the only lake visitation modeling exercise that we know of that systematically searched for a parsimonious set of explanatory variables. They estimated a mixed-logit model with a Iowa household survey of lake visits many times, experimenting with the form of 5 water quality measures (e.g., Secchi depth enters the model linearly in some specifications and logged in others) and the combination of water quality measures in the model (e.g., in one model Secchi depth and Chlorophyll levels are the water quality measures, in another model Secchi depth, Chlorophyll, and bacteria are the water quality measures). Their preferred model generates the loest log-likelihood value when estimated with maximum likelihood. They then interpret their preferred model.

is a penalty that increases in the sum of the absolute magnitudes of coefficients on **Z'** where the modeler chooses the value of the tuning parameter λ. Therefore, the minimization problem (12) exhibits a tradeoff: while the 'min' operator has incentive to choose the β's, even very large coefficients, that minimize ℓ, there is a competing incentive to keep the number of nonzero β's to a minimum; otherwise the penalty function's value can explode (assuming $\lambda > 0$). In other words, problem (12) selects the subset of potential explanatory variables that best explain **Y** – their coefficients are nonzero – given a constraint of parsimony. The greater the value λ, the tighter the parsimony constraint.

We set λ equal to the λ that minimizes the 10-fold mean cross validation error found when solving (12) over a subset of observations from the dataset (the "training" set). Then we solved (12) using λ_{min} with the subset of observations not used to find λ_{min} (the "test" set). The resulting variables with nonzero estimated coefficients form the parsimonious **Z** constructed with the single LASSO (the single LASSO-informed **Z** or $\mathbf{Z_{SL}}$).

One of the main purposes of this research is to determine to what extent lake water quality, as measured with summer Secchi depth, is related to lake visits in our study area. If the single LASSO-informed **Z** does not include average summer Secchi depth then we have gained information that the relationship is relatively weak. However, for policy analysis purposes, assume we want to force the inclusion of summer Secchi depth in our LASSO-informed **Z**.

We could do this by adding average summer Secchi depth to the single LASSO informed **Z** *ex post*. However, this process could lead to omitted variable bias when we estimate model (11) (Urminsky et al. 2015). Instead we used the double LASSO method to create a LASSO informed-**Z** that ensures the inclusion of average summer Secchi depth but avoids omitted

variable bias (Urminsky et al. 2015). In the first step of the double LASSO method we repeated the single LASSO algorithm without including average summer Secchi depth in **Z'**. In we repeated the single LASSO algorithm where summer Secchi depth is the dependent variable and all other potential explanatory variables are in **Z'** (the second stage ℓ is still a Poisson PDF given that summer Secchi is truncated at 0). The parsimonious **Z** in this case includes variables with nonzero coefficients in either solution to (12) *and* average summer Secchi depth. Therefore, the double LASSO-informed **Z** includes average summer Secchi depth, variables that strongly predict **Y**, and variables that strongly predict average summer Secchi depth. The inclusion of this last set of variables means the estimate of (11) with the double LASSO-informed **Z** or $\mathbf{Z_{DL}}$ is less likely to be affected by omitted variable bias.

We also use a random forest algorithm to form an alternative parsimonious **Z**. Unlike LASSO, random forest (RF) analysis does select a subset of the strongest predictors of **Y**. Instead, among other output, RF analysis ranks all variables from **Z'** in order of prediction importance. In RF analysis, the importance of a variable is measured by the average increase in out-of-sample predicted mean square error across the forest of decision trees when a variable's values are converted to random noise. In this algorithm a subset of data observations are used to build the trees (the "train" set) and the remaining observations are used to evaluate tree precision (the "test" set).

We use the R package VSURF to select the subset of the most important variables to include in RF-informed **Z**s. VSURF iteratively builds a RF or series of RFs over **Y** = f(**Z**) where at each iteration, **Z** has been winnowed down to a smaller set of variables based on a variable importance score (Genuer et al. 2015). In VSURF's first RF iteration, the variables from **Z'** that

meet a variable importance threshold are retained (see Genuer et al. 2015 for threshold details). Let this smaller set of variables be given by $\mathbf{Z}^{\#}$. In the next step, called the interpretation step, the VSURF routine builds a series of RFs over subsets of $\mathbf{Z}^{\#}$, the first subset being made up of the most important variable in $\mathbf{Z}^{\#}$**,** the second subset being made up of the two most important variables in $\mathbf{Z}^{\#}$, etc., and then chooses the subset of variables that creates the trees with the smallest out-of-sample error. Let this winnowed set of variables be given by $\mathbf{Z}^{\#\#}$. Finally, VSURF constructs an ascending sequence of RF models, starting with the most important variable in $\mathbf{Z}^{\#\#}$, and then adding one variable at a time, until the decrease in out-of-sample error created by a variable's addition no longer passes a threshold. Therefore, this so-called prediction set, is made up of the first *m* variables from $\mathbf{Z}^{\#\#}$ where $\mathbf{Z}^{\#\#}$'s variable order is determined by variable importance. Let this last winnowed set of explanatory variables form the VSURF-informed **Z** or $\mathbf{Z}_{\mathbf{VSURF}}$.

4.d. The statistical model we use to estimate Y = f(Z)

No matter the makeup of **Z**, the dependent variable in every case is a count variable. Therefore, we estimate the probability that $Y_j = k$ given the exogenous variable set $\mathbf{Z}_j$ with the Poisson PDF,

$$\Pr(Y_j = k|\mathbf{Z}_j) = \begin{cases} \frac{e^{-\lambda_j}\lambda_j^k}{k!} & k = 0,1,2,\ldots \\ 0 & \text{otherwise} \end{cases} \quad (13)$$

where we set λ_j equal to $\exp(\mathbf{Z}_j\boldsymbol{\beta})$. The coefficient vector $\boldsymbol{\beta}$ that maximizes the log likelihood function of (13) given **Z** is the value of $\boldsymbol{\beta}$ that makes the observed **Y** the most probable. We found statistical evidence that the count data is over-dispersed after estimating various

iterations of **Y** = f(**Z**) with (13) (estimated E(**Y**) is not equal to estimated Var(**Y**)). Estimated Poisson models with over-dispersed data can generate biased results. However, we use robust standard errors with our Poisson estimator. This step renders Poisson estimators robust to over-dispersion (i.e., estimates of (11) using the Poisson estimator are similar to estimates of (11) using the negative binomial estimator; see Wooldridge 2001).

5. Results

*5.a. The explanatory variables we include in parsimonious **Z***

We generated two sets of $\mathbf{Z}_{SL}$, two sets of $\mathbf{Z}_{DL}$, and two sets of $\mathbf{Z}_{VSURF}$**.** The first **Z**'s in each set were generated using *half* of the lakes with a complete set of observed data, including average summer Secchi and maximum lake depth (there are 6,553 lakes with a complete set of data; therefore 3,726 lakes were used to generate the first set of ML-informed **Z**'s). See Tables 4-5 for the variables in $\mathbf{Z}_{SL,Limited}$, $\mathbf{Z}_{DL,Limited}$, and $\mathbf{Z}_{VSURF,Limited}$ where 'Limited' refers to the dataset with a complete set of observed data.

We generated the second **Z**'s in each set using *half* of the dataset that included imputed average summer Secchi and maximum lake depth when these data were missing (the IM-augmented dataset has 39,582 complete cases; therefore 19,791 lakes were used to generate the second set of ML-informed **Z**'s). Recall there are 20 estimates of average summer Secchi and maximum lake depth at each lake in the augmented dataset. If average summer Secchi or maximum lake depth was observed at a lake then each of its 20 estimates of average summer Secchi or maximum lake depth are equal to the observed values. For all other lakes, the average summer Secchi or maximum lake depth values varies. Therefore, when using the MI-augmented

dataset, we solve the single LASSO, double LASSO, and VSURF problems twenty times; each time using a unique set of imputed summer Secchi and maximum lake depth values. Let $\mathbf{Z}_{SL,Augmented}$, $\mathbf{Z}_{DL,Augmented}$ and $\mathbf{Z}_{VSURF,Augmented}$ be comprised of variables that were selected at least 15 out of the 20 times by the respective ML methods when run over the 20 versions of the augmented dataset. See Tables 4-5 for the variables in each of these ML-informed **Z**'s.

5.b. Estimates of $\mathbf{Y} = f(\mathbf{Z})$

In Table 6 we present the Poisson estimates of $\mathbf{Y} = f(\mathbf{Z}_{Lit,Limited})$, $\mathbf{Y} = f(\mathbf{Z}_{DL,Limited})$, $\mathbf{Y} = f(\mathbf{Z}_{VSURF,Limited})$, and $\mathbf{Y} = f(\mathbf{Z}_{DL+VSURF,Limited})$ (the **Z** formed by the union of $\mathbf{Z}_{DL,Limited}$ and $\mathbf{Z}_{VSURF,Limited}$) over the half of the limited dataset (the dataset of lakes with observed depth measures) not used to generate the ML-informed **Z**'s. We estimate the visitation model over the remaining half of dataset to avoid the incorrect standard error estimation generated by using the same data to select variables and estimate a model (Leamer 1983, Egan et al. 2009).[13]

In Table 7 we present the Poisson estimates of $\mathbf{Y} = f(\mathbf{Z}_{Lit,Augmented})$, $\mathbf{Y} = f(\mathbf{Z}_{DL,Augmented})$, $\mathbf{Y} = f(\mathbf{Z}_{VSURF,Augmented})$, and $\mathbf{Y} = f(\mathbf{Z}_{DL+VSURF,Augmented})$ generated with the half of the augmented dataset not used to construct the ML-informed **Z'**s. The coefficients reported in Table 7 when **Z** is equal to $\mathbf{Z}_{Lit}$, $\mathbf{Z}_{Lit,Augmented}$, or $\mathbf{Z}_{DL+VSURF,Augmented}$ represent the mean of 20 model estimates (recall there are 20 unique sets of summer Secchi or maximum lake depth values in the augmented dataset). The MI routine in Stata generates robust standard errors that account for this averaging

[13] We could estimate $\mathbf{Y} = f(\mathbf{Z}_{Lit})$ over all 6,553 observations in the limited dataset given we did not use the data to select $\mathbf{Z}_{Lit}$. However, using the same set of data to estimate all models makes comparisons easy.

process. The coefficients associated with $\mathbf{Z}_{\mathbf{VSURF,Augmented}}$ come from a single estimation given that this version of **Z** does not include average summer Secchi or maximum lake depth data.

5.c. Interpreting results across lakes with measured summer Secchi and maximum depths

Not surprisingly, the ML informed-**Z** are much more parsimonious than the literature informed **Z**. Over the dataset limited to lakes with observed summer Secchi and maximum lake depths (the "limited dataset"), the single LASSO and VSURF-informed **Z**'s, $\mathbf{Z}_{\mathbf{SL,Limited}}$ and $\mathbf{Z}_{\mathbf{VSURF,Limited}}$, include 2 and 5 variables, respectfully, whereas $\mathbf{Z}_{\mathbf{Lit}}$ contains 42 variables. Neither ML algorithm selects average summer Secchi depth selected, indicating that it is not a particularly strong predictor of summer PUD counts. Instead lake area and amenity feature counts make up the bulk of the single LASSO and VSURF-informed **Z**'s.

Of the 16 covariates selected by the double LASSO over the limited dataset, 13 were selected in the stage with average summer Secchi depth (the focal independent variable) as the explanatory variable. Many of the variables selected in this stage describe the land cover and hydrological features around the lakes. Historic mean annual rainfall in the lake's subwatershed was also selected as a second stage covariate. These selections are not surprising given 1) land cover and use and rate of water flow in areas surrounding a lake largely determines nutrient flow into the lake (e.g., Vanni et al. 2011) and 2) under most circumstances, the rate of nutrient loading impacts water clarity (Mazumder and Lean 1994).[14]

[14] We can also say that average summer Secchi depth is not a "randomized treatment" across the lakes measured for summer Secchi depth, as the double LASSO found 13 covariates that predict average summer Secchi depth.

Our preferred **Z**s are those that best predict PUD counts at a lake according to 10-fold cross validation. Over the limited dataset, the double LASSO variable selection procedure generates the **Z** that returns the lowest mean root mean square error (RMSE) (see Table 8; we use the mean RMSE after removing the two largest RMSE outliers). Using this **Z** in our summer PUD count model, we find that an additional meter of Secchi depth, based on 1995 and 2014 measurements, was associated with 11.2% increase in summer PUD count during the 2005 to 2014-time period at lakes with measured summer Secchi and maximum depths. Assuming a linear relationship with a non-zero y-axis intercept between summer visits to these lakes and summer PUD counts, this is a lower bound estimate of the impact of higher Secchi depth on summer visits to these lakes during the 2005 to 2014 period.

Of the 3,276 lakes in the "testing" subset of the limited dataset, 130 lakes had one beach feature and the rest had 0 beach features. Therefore, we can treat the beach feature count variable in the double LASSO-informed **Z** as a dummy variable: a lake with a beach feature had 1.14 times the summer PUD counts between 2005 and 2014, all else equal, relative to a lake without a beach feature. Again, this is a lower bound on the average difference in summer visits to summer Secchi and maximum-depth measured lakes with and without beaches. No other lake amenity feature count was included in $\mathbf{Z}_{\mathbf{DL,Limited}}$.

Our summer PUD count model estimate with $\mathbf{Z}_{\mathbf{DL,Limited}}$ also indicates that Secchi and maximum-depth measured lakes surrounded by denser networks of forest, wetlands and streams had less summer PUDs than lakes with less dense networks of these feature, all else equal. We suspect that this means that people find lakes surrounded by development (e.g., roads, parking lots, and buildings) and grasslands, the land covers that persist in the absence of

forests, wetlands and streams, easier to access, all else equal. In addition, the estimated model indicates that lakes measured for Chlorophyll at least once between 1995 and 2014 had 1.73 times the summer PUD counts than lakes not measured for Chlorophyll levels. We suspect this indicates that water quality at popular lakes was more scrutinized than at less popular lakes.

Further, the double LASSO, when run over the limited dataset, selected several spatial lag covariates. Lakes that were surrounded by cleaner lakes (i.e., lakes with higher *lag Secchi*) and that were closer to lakes measured for summer Secchi (i.e., lakes with lower average distance to the nearest 5 lakes with a summer Secchi measurement) had less summer PUD counts than other lakes, all else equal. In other words, when lake *j*'s nearby competition had relatively clearer water or was at least monitored for its clarity, then lake *j* got less visits, all else equal.

However, this does not mean all forms of nearby competition meant lower summer PUD counts, on average, for lakes in the limited dataset. We also found that the closer lake *j* was to lakes with higher summer PUD (*lag Y*) and toilet and boat launch feature counts, the more summer PUDs it had, all else equal. These relationships suggest that "star" lakes generated positive spillover effects that ameliorated the negative effect nearby competition had on lake *j*'s summer PUD count: visitors to popular lakes, potentially because of their facilities, tended to visit nearby lakes at a higher rate than was otherwise expected.

5.d. Interpreting results across all lakes

The ML techniques run over the dataset that includes imputed depth measures (the "augmented dataset") also generates much more parsimonious $\mathbf{Z}$ then $\mathbf{Z}_{\mathbf{Lit}}$. The single LASSO,

double LASSO, and VSURF-informed **Z'**s have 13, 28, and 5 variables, respectfully, compared to the 42 variables in $\mathbf{Z}_{\mathbf{Lit}}$. Lake size, amenity feature counts, landscape conditions around the lake, socioeconomic conditions in a lake's subwatershed, and the spatial lags of amenity feature counts make up the bulk of the augmented dataset's ML-informed **Z**s.

Other than the double LASSO-informed **Z**, which includes average summer Secchi depth by definition, the ML techniques run over the augmented dataset do not select the water clarity measure. The finding that average summer Secchi based on 1995 to 2013 measures was not a strong predictor of 2005 to 2014 summer PUD counts is common across both the limited and augmented datasets. Unique to the $\mathbf{Z}_{\mathbf{SL,Augmented}}$ and $\mathbf{Z}_{\mathbf{DL,Augmented}}$ versus their limited dataset analogs are the presence of some socioeconomic variables. Another difference relative to the previous analysis: $\mathbf{Z}_{\mathbf{SL,Augmented}}$ and $\mathbf{Z}_{\mathbf{DL,Augmented}}$ are longer (contain more variables) than their analogs created from the limited dataset.

Our preferred **Z** over the augmented dataset is the one that generates the lowest average RMSE over 20 sets of 10-fold cross validations or 200 RMSEs (recall we estimate model (11) over each version of the augmented dataset **Z** twenty times; once for each unique iteration of the average summer Secchi and maximum lake depth vectors).[15] In this case the VSURF generates the preferred **Z** (as measured by mean RMSE less outliers; Table 9). However, because $\mathbf{Z}_{\mathbf{VSURF,Augmented}}$ does not include average summer Secchi depth we use the next best **Z** according to average RMSE criteria, the double LASSO + VSURF-informed **Z**. We do not sacrifice

[15] Except for VSURF-informed **Z** from the augmented dataset. This **Z** does not include the imputed average summer Secchi or maximum lake depth.

much predictive power by using $\mathbf{Z}_{\mathbf{DL+VSURF,Augmented}}$ rather than $\mathbf{Z}_{\mathbf{VSURF,Augmented}}$ as the difference in average RMSE between the two is very small.

We found that for every additional meter of average summer Secchi depth based on 1995 to 2013 measures, summer PUD count between 2005 and 2014 across all lakes in the dataset increased 7.3% (and actual visits likely even more). Please note that the marginal impact of water clarity on summer PUD counts was smaller when considering all lakes versus lakes with measured depths (at least across our preferred models). Further, across the entire dataset, deeper lakes and lakes with more amenities were also visited more than shallower and less amenity-dense lakes, all else equal. The impacts of beach and hotel amenities summer PUD counts in the augmented dataset are particularly impressive. If we treat the beach and hotel variables as dummy variables, a lake with a beach or a hotel had 1.54 times or 6.57 times the number of PUDs than lake without these amenities. These are larger amenity affects than we saw across the lakes in the limited dataset.

Across the augmented dataset, lakes with more nearby competition (as measured by average distance to *j*'s nearest 5 lakes) had less PUDs, all else equal. (Recall that across the lakes in the limited dataset the competitive pressure came from nearby lakes that were cleaner and measured for clarity, not all lakes in general.) However, once we accounted for the relative popularity of the neighboring lakes and their amenity features, the competitive pressure of nearby lakes on lake *j*'s PUD count was less pronounced. In other words, high levels of spatial clustering meant less PUDs for all lakes in the cluster, on average. However, the deleterious effects of clustering on *j*'s PUD count was ameliorated if one or more nearby competitors is a "star". (We saw similar spillover effects across lakes in the limited dataset.) Further, across the

augmented dataset, lakes with more developed covers (e.g., pavement, buildings) tended to have more summer PUDs than lakes with less developed space (e.g., forest, wetlands, and agriculture). We found a similar impact of developed land extent on PUD counts across the lakes in the limited dataset as well.

$\mathbf{Z}_{\mathbf{DL+VSURF,Augmented}}$ includes three variables that describe socioeconomic conditions in a lake's subwatershed. If we assume most PUDs and, therefore, visits to a lake, are made by local residents, these variables can provide some clues on the popularity of lake recreating across socioeconomic classes (or at least across the subset of the population that post pictures on social media). Of the three socioeconomic variables $\mathbf{Z}_{\mathbf{DL+VSURF,Augmented}}$, only subwatershed-level median household income is statistically significant. However, the magnitude of the income effect is small: for every $1,000 increase in a subwatershed's median household income, visits to a lake in the subwatershed increased by at least 1.1%. Despite the presence of socioeconomic variables in $\mathbf{Z}_{\mathbf{DL+VSURF,Augmented}}$, summer visits, or at least summer PUD counts, were not strongly associated with any socioeconomic group.

5.e. Summary of main results

To summarize our main results, the lakes measured for average summer Secchi and maximum lake depth are not a random sample of all lakes in the dataset. Therefore, our ML constructed covariate vectors and lake visit model results differ across the two sets of lakes. According to our analysis, at lakes measured for average summer Secchi and maximum lake depth, an additional meter of average summer Secchi depth during the period 1995 to 2013 was associated with 4% to 11% more summer PUDs from 2005 to 2014 (this is the range across

$Z_{Lit,Limited}$, $\mathbf{Z}_{DL,Limited}$, and $\mathbf{Z}_{DL+VSURF,Limited}$). Conversely, the marginal impact of a meter in average summer Secchi depth across all lakes in the 17-state region ranged from 7.0% to 8.5% (this is the range across $Z_{Lit,\ Augmented}$, $\mathbf{Z}_{DL,\ Augmented}$, and $\mathbf{Z}_{DL+VSURF,Augmented}$). As we mentioned above, these are lower bound estimates on average summer Secchi's impact on visits assuming a linear relationship between summer PUDs and summer visits. Further, In the smaller set of measured lakes, an additional meter of maximum lake depth – another indicator of lake water quality – led to a 1.4 to 1.9% increase in the summer PUD count (this is the range across $Z_{Lit,Limited}$, $\mathbf{Z}_{DL,Limited}$, and $\mathbf{Z}_{DL+VSURF,Limited}$). Across the full set of lakes, an additional meter in maximum depth meant a 3% in 2005 to 2014 summer PUDs, all else equal (this is the range across $Z_{Lit,\ Augmented}$, $\mathbf{Z}_{DL,\ Augmented}$, and $\mathbf{Z}_{DL+VSURF,Augmented}$).

While the impact of water quality variables on summer PUD counts is generally the same across both sets of lakes, the impact of lake amenities on PUD counts differs substantially between the two sets of lakes. First, the ML-informed **Z**'s created with the augmented dataset more often included amenity variables as important predictors of summer PUD counts. Second, the coefficients on the lake amenity variables in the lake visit model tended to be larger when estimated over the full set of lakes. We believe these modeling dissimilarities can be explained by the differences in amenity supply across the two sets of lakes. Lakes measured for maximum lake depth and Secchi depth between 1995 to 2013 had greater amenity density and less variability in amenity density than lakes from the full set.[16] Therefore, given that we have shown that amenities tend to draw visitors to lakes, the relative scarcity of amenities across all

[16] 14.5% of lakes in the measured set of lakes had one or more amenity counts whereas the rate was 3.9% across the full set of lakes.

lakes in the 17-state region means lake amenities will explain PUD counts much more strongly across the larger set of lakes than across the smaller set of lakes.

The two sets of lakes also have different competitive pressures. While both sets indicate that nearby lakes with high amenity and PUD counts create a positive spillover effect, only the larger dataset indicated that nearby competition in of itself reduced PUD counts at a lake. Namely, an additional 1 km in the average distance to *j*'s nearest 5 lakes increased PUD counts at *j* by 9.7% to 12.9% across all lakes but had little to no effect across the smaller set of lakes with measured Secchi and maximum depth. Further, cleaner lakes unequivocally created a positive PUD count spillover effect in the augmented dataset. This was not the case across the smaller set of lakes.

Otherwise, results from both sets of data were relatively consistent. For example, across both sets of lakes we find that lakes surrounded by greater amounts of developed land covers – roads, parking lots, buildings, etc. – had higher PUD counts, all else equal. This suggests that easy access to lakes is an important driver of visits. Further, not surprisingly, larger lakes had more PUDs, all else equal, across both sets of data.

6. Robustness checks

6.a. Instrumenting for average Summer Secchi depth

In our estimates of **Y** = f(**Z**) when **Z** includes summer Secchi information we assumed that summer water quality impacted summer lake visitation, or more appropriately, summer PUD counts, but lake visitation rates did not influence water quality. However, more popular lakes may face more pollution pressures. Therefore, the causal links between lake visitation and

lake water quality may have run both ways. To test whether endogeneity issues could be affecting the estimate of **Y** = f(**Z**) when **Z** includes average summer Secchi depth we can use the instrumental variable (IV) approach. In this case we assume that Secchi depth at a lake is the endogenous variable explained by the other variables in **Z** *and* an excluded instrument or two.

A valid instrument helps explain summer PUD count *via* average summer Secchi depth but not directly. [17] As previous research has found and as the second stage of our double LASSO analyses confirmed, a lake's average summer Secchi depth is greatly influenced by its surrounding land use and water flow regime. Therefore, we experiment with instrumenting average summer Secchi depth with stream density in the lake's subwatershed and the amount of agriculture land use in its 500-meter buffer. Greater stream density around a lake means polluted runoff can more easily be conveyed into a lake. Further, lakes surrounded by agricultural land – a source of phosphorus and nitrogen – tend to have lower lake quality as well, all else equal (cite). We surmise that while stream density in a lake's subwatershed and the presence of agricultural land in a lake's 500-meter buffer directly impacts the lake's water quality, and therefore, indirectly impacts its summer PUD count, it does *not directly* affect visits.

Therefore, we re-estimate the visit model with stream density and percentage of agricultural land in the 500-meter buffer as instruments. However, because the limited and augmented dataset $\mathbf{Z}_{DL}$'s and $\mathbf{Z}_{DL+VSURF}$'s include stream density we cannot use the IV method to re-estimate our preferred visit models. However, given that the estimated impact of average summer Secchi depth on summer PUD counts is fairly consistent across all versions of **Z**, IV

[17] A valid instrumental variables approach requires only that the instruments (i) be sufficiently correlated with the endogenous variable of interest and (ii) not be correlated with any unobserved determinants of the outcome of interest.

analyses limited to $\mathbf{Z_{Lit}}$'s should still be able to tell us if 1) reverse causality may be an issue in the summer PUD count – water quality relationship and 2) the direction of the bias if reverse causality is an issue.

The IV Poisson estimated coefficient on average summer Secchi depth from $\mathbf{Z_{Lit,Limited}}$ is 2.47, or, for every one meter increase in average Summer Secchi depth, summer PUD count increases, on average, by 1178%. While this coefficient is obviously larger than its non-IV analog, we found that the mean of the IV Poisson estimated coefficients on average summer Secchi depth from $\mathbf{Z_{Lit,Augmented}}$ is less than its non-IV analog (recall that there are 20 coefficient estimates for each variable in $\mathbf{Z_{Lit,Augmented}}$). Namely, in the non-IV version of the Poisson model estimated with $\mathbf{Z_{Lit,Augmented}}$, a one meter increase in average summer Secchi depth was associated with a 2005 to 2014 summer PUD count increase, on average, of 8.5%. In the IV estimate this percentage increase is only 2.5% (See SI Table 1 for all IV results).

Therefore, if the causal links between lake visitation, or more appropriately summer PUD counts, and lake water quality run both ways then we have found some evidence that the default estimate of the positive relationship between PUD counts and water quality across the lakes in the limited dataset is biased downward but is biased upward when we account for all lakes in the 17-state region.

6.b. Using average summer Secchi depth from the 2005 to 2013 period

We used summer Secchi depth measurements from 1995 to 2013 in our default approach to increase the number of lakes with observed Secchi and maximum depth measurements (6551 versus 5412 lakes when summer Secchi measures from 2005 to 2013 are

used instead). However, if a substantial number of lakes had average summer Secchi depth measures during the 2005 to 2013 period that were noticeably different than their 1995 to 2013 averages then our default lake visitation estimates may be spurious: observed photo-posting behavior from 2005 to 2014 could be a function of observed and imputed Secchi measures not accurately defined in our default datasets. Therefore, we redo our aforementioned analysis using average summer Secchi based on 2005 to 2013 measures instead of 1995 to 2013 measures (however, the spatial lag of average summer Secchi depth is still based on 1995 to 2013 summer-time measures). In this analysis the number of lakes with observed average summer Secchi measures fall and the number of lakes with imputed average summer Secchi measures rise.

We find fairly similar relationships between summer PUD counts and the covariate vector **Z** when using 2005 to 2013 average Secchi measures instead of 1995 to 2013 measures (for modeling results using 2005 to 2014 average Secchi measures see SI Tables 2-5). First, we compare the explanatory variables selected by the ML approaches given the two different measures of average summer Secchi depth. In all four comparisons of unique ML – dataset combinations (DL – limited dataset, DL – augmented dataset, VSURF – limited dataset, VSURF – augmented dataset) we find that the number of commonly selected variables is always equal to or greater than the number of uniquely selected variables. For example, seventeen of the same variables are selected by the double LASSO no matter how Secchi is represented in the augmented dataset. An additional 11 unique variables are selected when average summer Secchi is measured with 1995 to 2013 data and 2 unique variables are selected when average summer Secchi is measured with 2005 to 2013 data. Just as with the default data, the single

LASSO and VSURF methods never select average summer Secchi as an important predictor (see Table 10 for a complete comparison of variable selection across datasets and ML methods).

Regardless of which dataset was used – the limited or augmented – the double LASSO selected more covariates when average summer Secchi was based on 1995 to 2013 measures versus 2005 to 2013 data. This suggests, not surprisingly, that covariates measured circa 2010 (e.g., land cover, feature counts, etc.) could more precisely predict contemporaneous average summer Secchi (averages based on 2005-2013 measurements) than averages based on both historical and contemporaneous measurements (1995 to 2013 measurements).

The estimated models' ability to predict 2005 to 2014 summer PUD counts using 2005 to 2013 summer Secchi measures does not dominate the predictive ability of models estimated with 1995 to 2013 summer Secchi measures. In this case predictive ability is measured with 10-fold cross validations. For example, $\mathbf{Z}_{\mathbf{DL,Limited}}$ with average summer Secchi depth based on 1995 to 2013 observations better predicts 2005 to 2014 summer PUD counts than $\mathbf{Z}_{\mathbf{DL,Limited}}$ with average summer Secchi depths based on 2005 to 2013 observations. However, $\mathbf{Z}_{\mathbf{DL+VSURF,Augmented}}$ with average summer Secchi depths based on 1995 to 2013 measures and $\mathbf{Z}_{\mathbf{DL+VSURF,Augmented}}$ with average summer Secchi depths based on 2005 to 2013 measures are similarly predictive of 2005 to 2013 summer PUD counts (compare average RMSEs in Tables 8-9 and SI Tables 6-7).

Finally, there are differences in the estimated coefficients on average summer Secchi depths based on 2005 to 2013 measures versus 1995 to 2013 measures (Table 11). At lakes in the limited dataset, an additional meter of average summer Secchi depth based on 2005 to 2013 measures was associated with 14.9% to 25.9% more PUDs from 2005 to 2014 (this is the range across $\mathbf{Z}_{\mathbf{Lit,Limited}}$, $\mathbf{Z}_{\mathbf{DL,Limited}}$, and $\mathbf{Z}_{\mathbf{DL+VSURF,Limited}}$). This range was 4% to 11% when we used

Secchi depth measures based on 1995 to 2013 measures. Further, the impact of a one-meter increase in average Secchi depth as measured or imputed from 2005 to 2014 across all lakes in the 17-state region ranged from 10.3% to 12.7% (this is the range across $\mathbf{Z_{Lit,Augmented}}$, $\mathbf{Z_{DL,\ Augmented}}$, and $\mathbf{Z_{DL+VSURF,\ Augmented}}$). This range was 7% to 8.5% when we used 1995 to 2013 Secchi depth measures. (The impact of an additional meter of maximum lake depth on summer PUD count is the same across both approaches). Therefore, by choosing a default approach that relied on summer Secchi measurements from a greater time frame, thereby generating a larger number of observed Secchi measures and relying less on imputed Secchi numbers, we likely have underestimated the impact of clean water on summer PUD counts, and therefore, summer visits.

6.c. Limiting the dataset to lakes with water-based recreation amenity features

It may be reasonable to assume that water quality matters more for lake recreators that use the water directly. For example, people that swim, boat, or fish on lakes may care more about water quality than lake visitors that only walk or run around lakes or BBQ near lakes (although Ziv et al. (2016) does not find this to be the case in England). To determine if the PUD count – water quality relationship is demonstrably different at lakes with recreational infrastructure that let people directly access lakes we re-conduct our analysis only considering lakes with at least one beach, boat launch, or marina feature (see SI Tables 8-13 for full results of this analysis). While we cannot separate the PUD counts of those that recreated on or in the lake from those that didn't at these select lakes, presumably a fair number of PUDs at these lakes were generated by direct users of the lakes.

Surprisingly, PUD counts tended to fall as water quality improved at these lakes, all else equal. At lakes in the limited dataset that had at least one water recreation feature, an additional meter of average summer Secchi depth based on 1995 to 2013 measures was associated with 15.5% to 33.6% less 2005 to 2014 summer PUDs (this is the range across $\mathbf{Z}_{\text{Lit,Limited}}$, $\mathbf{Z}_{\text{DL,Limited}}$, and $\mathbf{Z}_{\text{DL+VSURF,Limited}}$). Further, the marginal 2005 to 2014 summer PUD impact of a meter in average Secchi depth as measured or imputed from 1995 to 2014 across all lakes in the 17-state region ranged from -1.17% to 7.8% (this is the range across $\mathbf{Z}_{\text{Lit,Augmented}}$, $\mathbf{Z}_{\text{DL, Augmented}}$, and $\mathbf{Z}_{\text{DL+VSURF, Augmented}}$). However, none of these marginal impacts are statistically significant at p = 0.05. Therefore, our hypothesis that water quality was especially coveted by users of lakes that let them get on or in the lakes is not supported by the data. In fact, at lakes in the limited dataset, summer PUD counts are much higher among the lakes that have worse water quality, all else equal.

The particularly perverse finding for lakes in the limited dataset could be explained by previously mentioned reverse causality issues: swimmers, boaters, and fishermen bring pollution with them and therefore, the more popular lakes for swimming, boating, and fishing, as proxied by overall PUD count, had less clarity. To test for this possibility we instrument for average summer Secchi depth in the $\mathbf{Z}_{\text{Lit,Limited}}$ covariate vector when model (11) is estimated over lakes with water-recreation infrastructure. We again used subwatershed stream density and percentage of a lake's 500-meter buffer area in agriculture cover as instruments. We found that the negative coefficient on average summer Secchi depth from the non-IV Poisson estimated model with $\mathbf{Z}_{\text{Lit,Limited}}$ became positive (albeit, statistically insignificant at p = 0.05; SI Table 14) when we instrumented for average Secchi depth. This result makes us question the

result that worse water clarity was associated with higher summer PUD counts at the water-recreation equipped lakes in the limited dataset.

7. Discussion and conclusion

In this study we have identified some of the lake features that attract Americans to lakes. Past research on this question has often been limited by data availability. We do not have this problem. Instead we have data on a lake visit proxy and more than 100 attributes for 51,107 lakes (4 ha or greater) across 17 states. However, this data richness comes with its own curse: how do we generate accurate, succinct, and easily digestible conclusions on the relationships between lake visits and lake attributes?

A further complication was introduced by the lack of water quality measures at a majority of the 51,107 lakes in our dataset. Based on prior evidence that lake visitation rates are affected by water quality, not including data on water quality at each lake would introduce omitted variable bias into any econometric estimate we made of lake visit and lake attribute relationships. Further, the relationship between lake visitation rates and lake water quality is of great policy interest. For these reasons, summer Secchi and lake depth measures needed to be part of our visit model.

Normally the solution to our problem would be to drop the lakes without summer Secchi or maximum depth measures. However, we found statistical evidence that the subset of lakes with measured Secchi and lake depths was not a random draw from all 51,107 lakes. This unrepresentativeness meant that we could not use the estimated visit model across the lakes with measured Secchi and lake depths to explain visit and attribute relationships across the

population of lakes. Therefore, to be able to say something about relationships between lake visitation rates and lake attributes, including water quality, across the entire population of lakes we imputed Secchi and maximum lake depths at lakes missing this data. In other words, we worked with two sets of data: a subset of lakes that were measured for Secchi and maximum lake depth (the limited dataset) and all lakes where Secchi and maximum lake depth was imputed 20 times when these data were missing (the augmented dataset)

We used ML techniques to determine the parsimonious set of lake attribute covariates that best predict 2005 to 2014 summer PUD counts at each lake in each dataset where PUD counts proxies for lake visits. By using these variable selection algorithms, we 1) kept our covariate vectors to a manageable and interpretable size; 2) identified some strongly predictive explanatory variables not identified by the lake and park visitation literature; and 3) generated visit models that minimized summer PUD count prediction error. Once we generated the covariate vectors we then used a Poisson count model (with robust standard errors to account for over dispersion in the data) to estimate the impact that a small change in each covariate has on summer 2005 to 2014 PUD counts. Assuming a linear relationship between actual summer visits and summer PUD counts between 2005 and 2014, the measured marginal effects give the lower bound on change in summer visits from 2005 to 2014.

Based on our analysis we conclude the following about lake visit and lake attribute relationships. First, according to ML algorithms, average summer Secchi depth is not one of the top predictors of summer PUD counts. No matter the dataset, the single LASSO and VSURF algorithms did not select average summer Secchi as a covariate in our visit model (the VSURF algorithm run over the augmented dataset of lakes with water-recreation features being the

one exception). Average summer Secchi depth only shows up in double LASSO-informed covariates vectors because we made it the focus of the second step in the double LASSO process.

Second, despite not being the among the most important predictors, lake PUD counts increase in water quality. Considering all lakes in our dataset we estimated that a one meter increase in average summer Secchi as measured between 1995 and 2013 increased 2005 to 2014 summer PUD counts by 7.0% to 8.5%, all else equal.

Third, lake amenities and access appear to have the most powerful impact on summer PUD counts. For example, considering all lakes in our dataset, lakes with beaches get at least 1.5 times the visits that lakes without beaches get, all else equal. The rate is 1.25 times for lakes with boat launches and at least 5 times for lakes with hotels. (Although the impact of amenities in summer PUD counts is weaker when we use 2005 to 2013 measures of Secchi depth instead of the 1995 to 2013 measures.) Further, the extent of a lake's buffer that is in developed cover has a strong effect on summer PUD counts. Considering all lakes in our dataset, for every 10-percentage point increase in the buffer area that is in developed cover as of 2011, summer 2005 to 2014 PUD counts increase 12 to 23%. Given that we believe that density of developed cover around a lake is a proxy for the extent of lake access we contend that lake access plays a very important part in lake visitation.

Fourth, future lake recreation policy focused on lake *j* must account for the recreation substitutability and complementarity impacts of lakes near *j*. In of itself, competition decreases visits to lake *j*: considering all lakes in our dataset, a lake nearer other lakes have lower summer PUD counts than more spatially isolated lakes, all else equal. However, once we account for the

popularity of nearby lakes and their amenity density this substitution effect can be reversed. Popular lakes with amenities seem to generate more visits than expected for their neighboring lakes no matter the neighbor's attribute density.

Considering all lakes in our dataset, we found little evidence that lakes in richer, whiter, and more educated subwatersheds had higher summer PUD counts, all else equal. Therefore, assuming most PUDs of a lake are taken by people that live near a lake, there was little to no socioeconomic divide in 2005 to 2014 summer PUD counts, and by extension, 2005 to 2014 summer lake visits.

While policy makes and recreation managers will find our analysis useful, they should view are results with some skepticism. First, a large percentage of lakes have zero 2005 to 2014 summer PUDs. We suspect that the histogram of actual visits to these lakes over this time period is much less concentrated at 0. We do not know how much the clustering of our visits proxy at 0 biases our results relative to an estimate of **V** = f(**Z**) where visit counts (**V**) has replaced summer PUD counts (**Y**).

Second our population-level estimates rely on extensive data imputation. We provided evidence that imputation of average summer Secchi and maximum lake depth are not particularly accurate. It is unclear how much imputation inaccuracy affects our results.

Third, policy makers and concerned citizens are interested in how lake attributes impact lake visits not lake PUD counts. Prior research that has had access to both site visit and PUD count data have shown that the relationship between visits (y-axis) and PUD counts (x-axis) can generally be represented by a linear curve with a positive y-axis intercept and slope. Assuming this linear relationship generally holds in our case we have shown that estimated explanatory

variable marginal impacts on summer PUD counts represent the lower bound on summer visit marginal impacts. However, is our assumption about a generally linear relationship between summer visits and summer PUD counts valid? And if it is, what is the gap between the lower bound estimates we provide and the central tendency estimate?

Data and computer code

A zip file with data and computer code (R and Stata) for this paper can be found at the link https://www.dropbox.com/s/817jakh4x6ie4jc/DataandCodeLakePaper.zip?dl=0.

Acknowledgements

The authors would like to thank Aaron Gilbreath of Bowdoin College for help creating the dataset used in this paper.

References

Center for International Earth Science Information Network (CIESIN). 2016. 2015 Gridded Population of the World, Version 4 (GPWv4): Population Density. Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC). https://doi.org/10.7927/H4NP22DQ.

Collins, S. M., S. K. Oliver, J. F. Lapierre, E. H. Stanley, J. R. Jones, T. Wagner, and P. A. Soranno. 2017. Lake nutrient stoichiometry is less predictable than nutrient concentrations at regional and sub-continental scales. Ecological Applications. 27(5): 1529-1540. https://doi.org/10.1002/eap.1545.

Deryugina, Tatyana, Garth Heutel, Nolan H. Miller, David Molitor, and Julian Reif. 2019. The Mortality and Medical Costs of Air Pollution: Evidence from Changes in Wind Direction. American Economic Review. 109 (12): 4178-4219. https://doi.org/10.1257/aer.20180279.

Di Minin, Enrico, Tenkanen Henrikki, Toivonen Tuuli. 2015. Prospects and challenges for social media data in conservation science. Frontiers in Environmental Science. 3: 63. https://doi.org/10.3389/fenvs.2015.00063.

Donahue, Marie L., Bonnie L. Keeler, Spencer A. Wood, David M. Fisher, Zoé A. Hamstead, and Timon McPhearson. 2018. Using social media to understand drivers of urban park visitation in the Twin Cities, MN. Landscape and Urban Planning. 175: 1-10. https://doi.org/10.1016/j.landurbplan.2018.02.006.

Genuer, Robin, Jean-Michel Poggi, and Christine Tuleau-Malot. 2015. VSURF: An R Package for Variable Selection Using Random Forests. The R Journal, R Foundation for Statistical Computing 7(2): 19-33. hal-01251924v1.

Harari, G. M., N. D. Lane, R. Wang, B. S. Crosier, A. T. Campbell, and S. D. Gosling. 2016. Using Smartphones to Collect Behavioral Data in Psychological Science: Opportunities, Practical Considerations, and Challenges. Perspect Psychol Sci. 11(6): 838–854. https://doi.org/10.1177/1745691616650285.

Harrell Jr., F. E., K. L. Lee, and D. B. Mark. 1996. Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors. Statistics in Medicine. 15: 361-387. https://doi.org/10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4.

Hofacker, C., E. Malthouse, and F. Sultan. 2016. Big Data and consumer behavior: imminent opportunities. Journal of Consumer Marketing. 33(2): 89-97. https://doi.org/10.1108/JCM-04-2015-1399.

Hunt, Len M., Adam Dyck. 2011. The Effects of Road Quality and Other Factors on Water-Based Recreation Demand in Northern Ontario, Canada. Forest Science. 57(4): Pages 281–291, https://doi.org/10.1093/forestscience/57.4.281

Ilieva, R.T. and T. McPhearson. 2018. Social-media data for urban sustainability. Nat Sustain 1: 553–565. https://doi.org/10.1038/s41893-018-0153-6.

Jakobsen, J. C., C. Gluud, J. Wetterslev, et al. 2017. When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts. BMC Med Res Methodol 17, 162. https://doi.org/10.1186/s12874-017-0442-1.

James, G., D. Witten, T. Hastie, and R. Tibshirani. 2013. An introduction to statistical learning. New York: Springer.

von Haefen, R. H. and D. J. Phaneuf. 2005. Kuhn-Tucker Demand System Approaches to Non-Market Valuation. In: Scarpa R., Alberini A., Eds. Applications of Simulation Methods in Environmental and Resource Economics. The Economics of Non-Market Goods and Resources, Vol 6. Springer, Dordrecht.

Hausmann, A., T. Toivonen, R. Slotow, H. Tenkanen, A. Moilanen, V. Heikinheimo, and E. Di Minin. 2018. Social Media Data Can Be Used to Understand Tourists' Preferences for Nature-Based Experiences in Protected Areas. Conservation Letters. 11: e12343. https://doi.org/10.1111/conl.12343.

Karl, Herman A., Lawrence E. Susskind, and Katherine H. Wallace. 2007. A dialogue, not a diatribe: effective integration of science and policy through joint fact finding. Environment: Science and Policy for Sustainable Development 49.1: 20-34. https://doi.org/10.3200/ENVT.49.1.20-34.

Leamer, Edward E. 1983. Let's Take the Con Out of Econometrics. The American Economic Review, 73(1): 31–43. https://www.jstor.org/stable/1803924.

Levin, Noam, Alex Mark Lechner, and Greg Brown. 2017. An evaluation of crowdsourced information for assessing the visitation and perceived importance of protected areas. Applied Geography. 79: 115-126. https://doi.org/10.1016/j.apgeog.2016.12.009.

Matz, Sandra C and Oded Netzer. 2017. Using Big Data as a window into consumers' psychology. Current Opinion in Behavioral Sciences. 18: 7-12. https://doi.org/10.1016/j.cobeha.2017.05.009.

Mazumder, Asit, David R. S. Lean. 1994. Consumer-dependent responses of lake ecosystems to nutrient loading. Journal of Plankton Research. 16(11): 1567–1580. https://doi.org/10.1093/plankt/16.11.1567

Milne, Dave and David Watling. 2019. Big data and understanding change in the context of planning transport systems. Journal of Transport Geography 76: 235-244. https://doi.org/10.1016/j.jtrangeo.2017.11.004.

OpenStreetMap contributors. 2015. Planet dump [Data file from 10/28/16]. Retrieved from https://planet.openstreetmap.org.

Parsons, G. R., E. C. Helm, and T. Bondelid. 2003. Measuring the Economic Benefits of Water Quality Improvements to Recreational Users in Six Northeastern States: An Application of the Random Utility Maximization Model. University of Delaware Manuscript. http://works.bepress.com/george_parsons/25/.

Phaneuf, D. J. 2002. A random utility model for total maximum daily loads: Estimating the benefits of watershed-based ambient water quality improvements. Water Resour. Res. 38(11): 1254. https://doi.org/10.1029/2001WR000959, 2002.

Pullin, Andrew S., and Teri M. Knight. 2005. Assessing conservation management's evidence base: a survey of management-plan compilers in the United Kingdom and Australia. Conservation Biology 19(6): 1989-1996. https://doi.org/10.1111/j.1523-1739.2005.00287.x.

Qin, B., J. Zhou, J. J. Elser, W. S. Gardner, J. Deng, and J. D. Brookes. 2020. Water Depth Underpins the Relative Role and Fates of Nitrogen and Phosphorus in Lakes. Environmental Science & Technology. 54(6): 3191–3198. https://doi.org/10.1021/acs.est.9b05858.

Richards, Daniel R. and Daniel A. Friess. 2015. A rapid indicator of cultural ecosystem service usage at a fine spatial scale: Content analysis of social media photographs. Ecological Indicators. 53: 187-195. https://doi.org/10.1016/j.ecolind.2015.01.034.

Roberge, J. M. 2014. Using data from online social networks in conservation science: which species engage people the most on Twitter? Biodiversity Conservation 23: 715. https://doi.org/10.1007/s10531-014-0629-2.

Ruckelshaus, Mary, Emily McKenzie, Heather Tallis, Anne Guerry, Gretchen Daily, Peter Kareiva, Stephen Polasky, Taylor Ricketts, Nirmal Bhagabati, Spencer A. Wood, and Joanna Bernhardt. 2015. Notes from the field: Lessons learned from using ecosystem service approaches to inform real-world decisions. Ecological Economics. 115: 11-21. https://doi.org/10.1016/j.ecolecon.2013.07.009.

Sessions, Carrie, Spencer A. Wood, Sergey Rabotyagov, and David M. Fisher. 2016. Measuring recreational visitation at U.S. National Parks with crowd-sourced photographs. Journal of Environmental Management. 183, Part 3: 703-711. https://doi.org/10.1016/j.jenvman.2016.09.018.

Siart, Christoph Markus Forbriger and Olaf Bubenzer. 2017. Digital Geoarchaeology: New Techniques for Interdisciplinary Human-Environmental Research. Springer, 269 pages

Smirnov, Oleg A. and Kevin J. Egan. 2012. Spatial random utility model with an application to recreation demand. Economic Modelling. 29(1): 72-78. https://doi.org/10.1016/j.econmod.2010.09.026.

Sonter, L. J., K. B. Watson, S. A. Wood, and T. H. Ricketts. 2016. Spatial and Temporal Dynamics and Value of Nature-Based Recreation, Estimated via Social Media. PLOS ONE 11(9): e0162372. https://doi.org/10.1371/journal.pone.0162372

Soranno, P.A., L.C. Bacon, M. Beauchene, K.E. Bednar, E.G. Bissell, C.K. Boudreau, M.G. Boyer, M.T. Bremigan, S.R. Carpenter, J.W. Carr, K.S. Cheruvelil, S.T. Christel, M. Claucherty, S.M. Collins, J.D. Conroy, J.A. Downing, J. Dukett, C.E. Fergus, C.T. Filstrup, C. Funk, M.J. Gonzalez, L.T. Green, C. Gries, J.D. Halfman, S.K. Hamilton, P.C. Hanson, E.N. Henry, E.M. Herron, C. Hockings, J.R. Jackson, K. Jacobson-Hedin, L.L. Janus, W.W. Jones, J.R. Jones, C.M. Keson, K.B.S. King, S.A. Kishbaugh, J.-F. Lapierre, B. Lathrop, J.A. Latimore, Y. Lee, N.R. Lottig, J.A. Lynch, L.J. Matthews, W.H. McDowell, K.E.B. Moore, B.P. Neff, S.J. Nelson, S.K. Oliver, M.L. Pace, D.C. Pierson, A.C. Poisson, A.I. Pollard, D.M. Post, P.O. Reyes, D.O. Rosenberry, K.M. Roy, L.G. Rudstam, O. Sarnelle, N.J. Schuldt, C.E. Scott, N.K. Skaff, N.J. Smith, N.R. Spinelli, J.J. Stachelek, E.H. Stanley, J.L. Stoddard, S.B. Stopyak, C.A. Stow, J.M. Tallant, P.-N. Tan, A.P. Thorpe, M.J. Vanni, T. Wagner, G. Watkins, K.C. Weathers, K.E. Webster, J.D. White, M.K. Wilmes, S. Yuan. 2017. LAGOS-NE: A multi-scaled geospatial and temporal database of lake ecological context and water quality for thousands of U.S. lakes. Gigascience 6(12). https://doi.org/10.1093/gigascience/gix101.

Soranno, P. and K. Cheruvelil. 2017. LAGOS-NE-GEO v1.05: A module for LAGOS-NE, a multi-scaled geospatial and temporal database of lake ecological context and water quality for thousands of U.S. Lakes: 1925-2013. Environmental Data Initiative. http://dx.doi.org/10.6073/pasta/b88943d10c6c5c480d5230c8890b74a8. Dataset accessed 9/26/2017.

Soranno P. A., N. R. Lottig, A. D. Delany, and K. S. Cheruvelil. 2019. LAGOS-NE-LIMNO v1.087.3: A module for LAGOS-NE, a multi-scaled geospatial and temporal database of lake ecological context and water quality for thousands of U.S. Lakes: 1925-2013. Environmental Data Initiative. https://doi.org/10.6073/pasta/08c6f9311929f4874b01bcc64eb3b2d7. Dataset accessed 12/09/2019.

Tenkanen, H., E. Di Minin, V. Heikinheimo, et al. 2017. Instagram, Flickr, or Twitter: Assessing the usability of social media data for visitor monitoring in protected areas. Sci Rep 7, 17615. https://doi.org/10.1038/s41598-017-18007-4.

Urminsky, Oleg, Christian Hansen, and Victor Chernozhukov. 2016. Using Double-Lasso Regression for Principled Variable Selection. http://dx.doi.org/10.2139/ssrn.2733374.

US Census 2017. P053. MEDIAN HOUSEHOLD INCOME IN 1999 (DOLLARS).

US Environmental Protection Agency (US EPA). 2013. EnviroAtlas - Dasymetric Population in the Conterminous United States Web Service. U.S. EPA Office of Research & Development (ORD) - National Exposure Research Laboratory (NERL). Research Triangle Park, NC.

van Zanten, Boris T., Derek B. Van Berkel, Ross K. Meentemeyer, Jordan W. Smith, Koen F. Tieskens, and Peter H. Verburg. 2016. Mapping landscape values using social media Proceedings of the National Academy of Sciences. 113(46) 12974-12979. http://dx.doi.org/10.1073/pnas.1614158113.

Vanni, M. J., W. H. Renwick, A. M. Bowling, M. J. Horgan, and A.D. Christian. 2011. Nutrient stoichiometry of linked catchment-lake systems along a gradient of land use. Freshwater Biology. 56: 791-811. http://dx.doi.org/10.1111/j.1365-2427.2010.02436.x.

Wooldridge, J. M. 2010. Econometric analysis of cross section and panel data. MIT press.

Wood, S. A., A. D. Guerry, J. M. Silver, and M. Lacayo. 2013. Using social media to quantify nature-based tourism and recreation. Scientific Reports 3: 2976. https://doi.org/10.1038/srep02976.

Yi, D., and J. A. Herriges. 2017. Convergent Validity and the Time Consistency of Preferences: Evidence from the Iowa Lakes Recreation Demand Project. Land Economics. 93(2): 269-291. https://doi.org/10.3368/le.93.2.269.

Figures

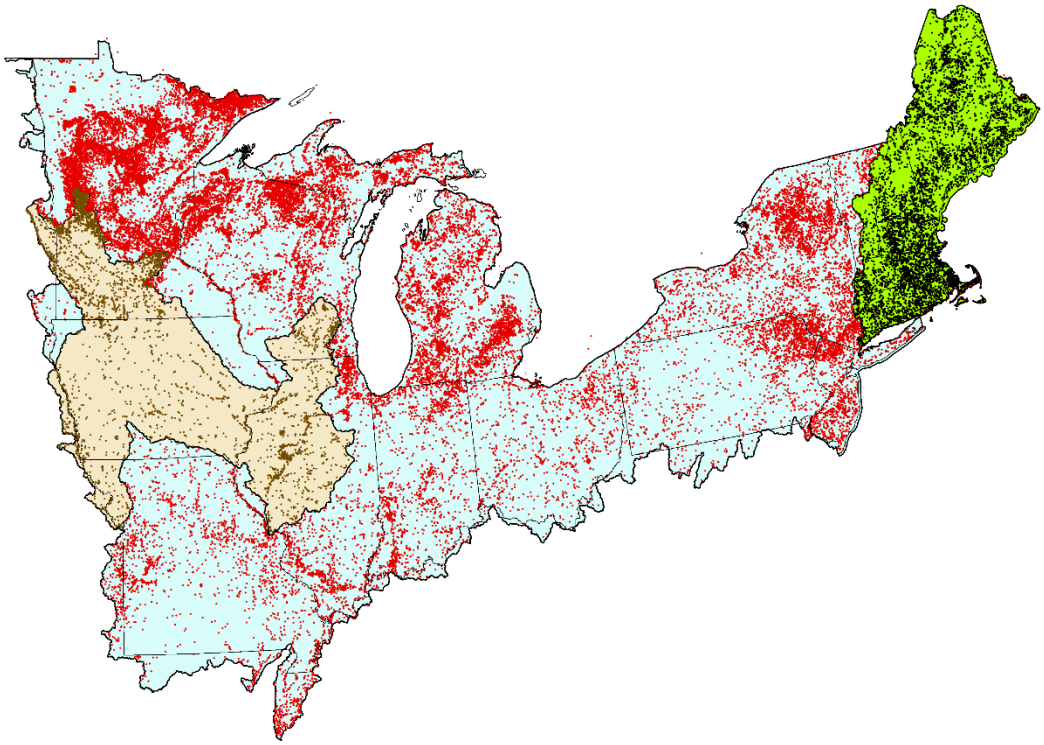


Figure 1. The 51,107 lakes in our dataset. Each dot represents a lake. The data comes from LAGOS (Soranno et al. 2017). Our dataset includes the entire population of lakes 4 ha or greater in the LAGOS region. The LAGOS region can be broken into three (sub) regions (defined using 4-digit hydrologic units (HUC4)). In the 'Low Agriculture' region (green background, black dots; N = 6,673) agricultural land cover constitutes less than 10% of land area in each HUC4. In the 'High Agriculture' region (tan background, brown dots; N = 4,584) agricultural land cover constitutes more than 75% of the land area in each HUC4. In the 'Moderate Agriculture' region (blue background, red dots; N = 39,850) agricultural land cover constitutes between 10% and 75% of land area in each HUC4. These regions were defined in Collins et al. (2017). The 17 states entirely covered by the LAGOS region include Maine, New Hampshire, Vermont, New York, Ohio, Michigan, Wisconsin, Minnesota, Iowa, Illinois, Missouri, Indiana, Pennsylvania, New Jersey, Connecticut, Massachusetts, and Rhode Island.

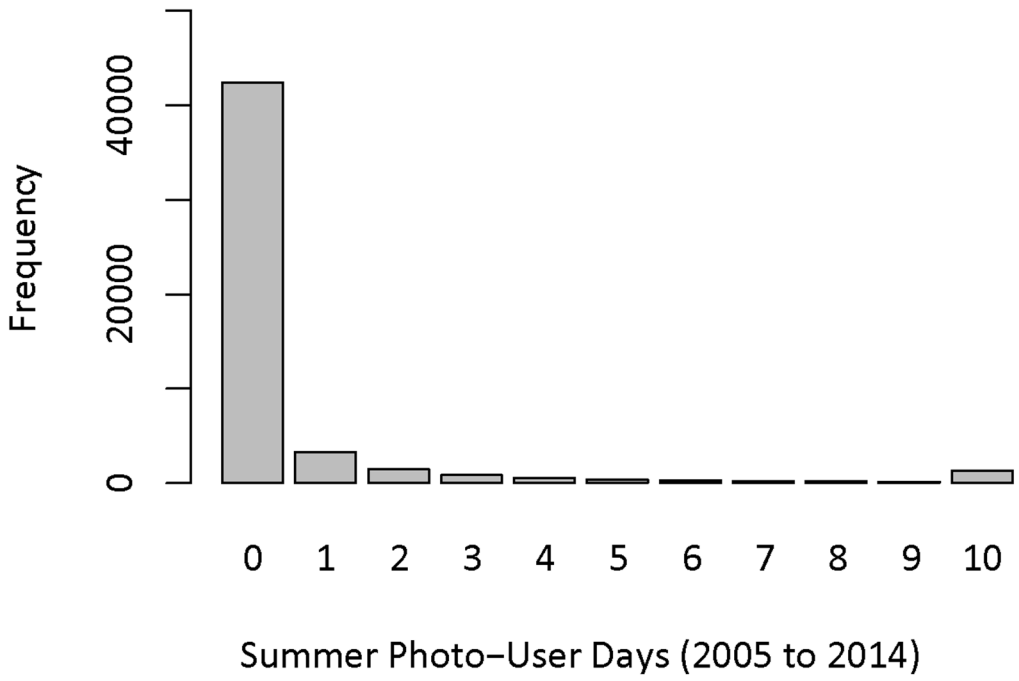


Figure 2. Histogram of 2005 to 2014 summer photo-user days (PUDs) across the 51,107 lakes in our database. Mean summer PUD count: 2.12; Median summer PUD count: 0.00; Standard deviation of summer PUD count: 104.98; Minimum of summer PUD count: 0; Maximum of summer PUD count: 21,705. **R code:** Figure2.R

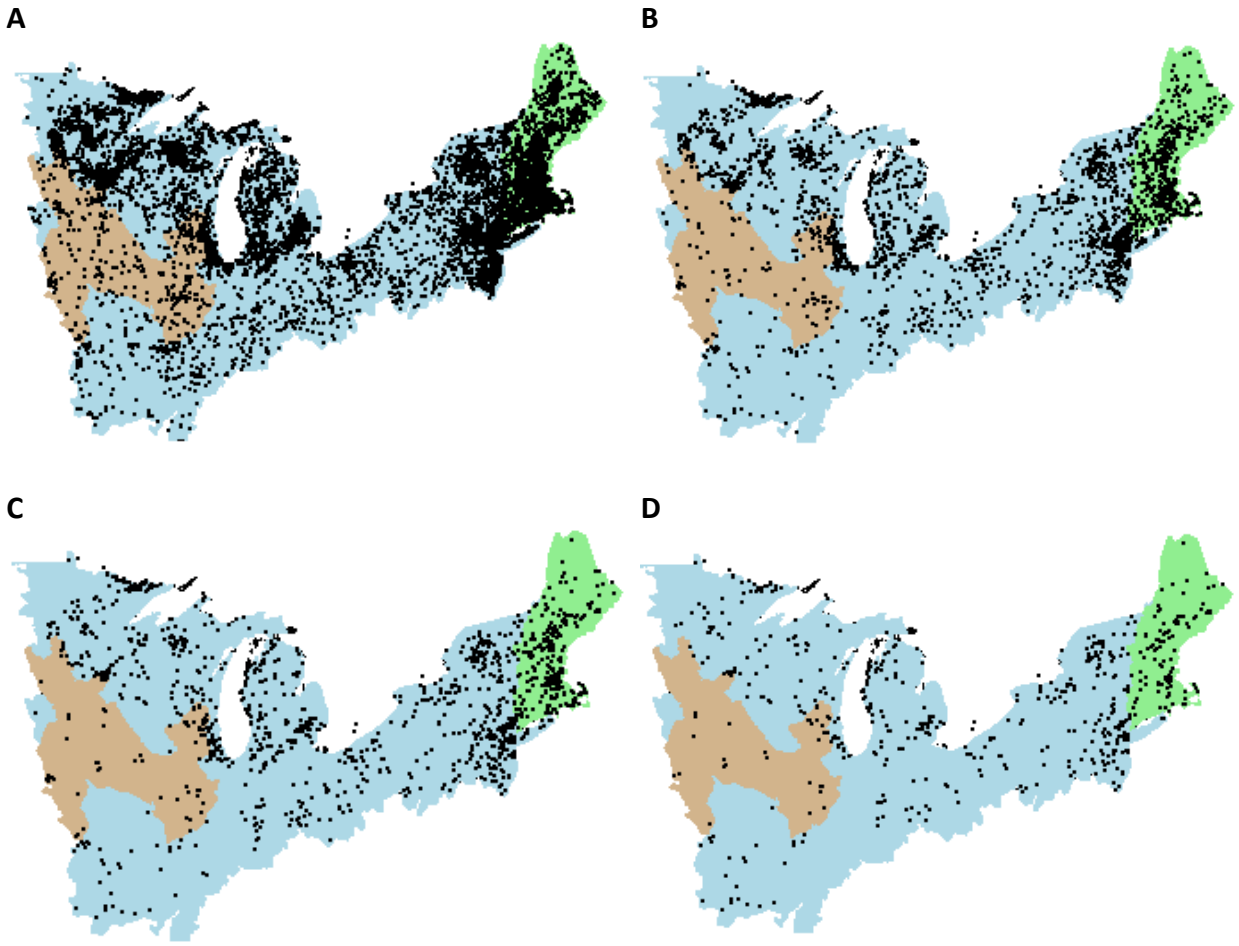
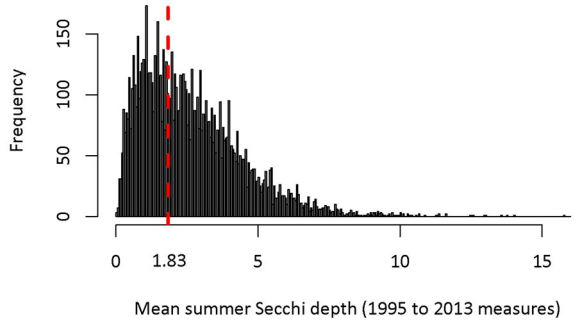
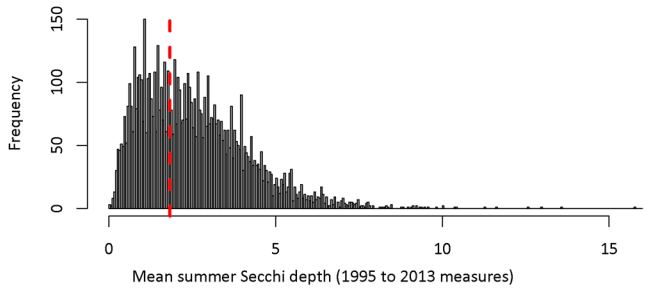


Figure 3. (A) Lakes with one or more summer PUDs between 2005 and 2014. Entire LAGOS region: N = 8,728 (17.1% of all lakes in region). Low agriculture region: N = 1,915 (28.7% of all lakes in region). High agriculture region: N = 504 (11.0% of all lakes in region). Moderate agriculture region: N = 6,309 (15.8% of all lakes in region). **(B) Lakes with five or more summer PUDs between 2005 and 2014.** Entire LAGOS region: N = 2,557 (5.0% of all lakes in region). Low agriculture region: N = 558 (8.4% of all lakes in region). High agriculture region: N = 145 (3.2% of all lakes in region). Moderate agriculture region: N = 1,854 (4.7% of all lakes in region). **(C)** Lakes with ten or more summer PUDs between 2005 and 2014. Entire LAGOS region: N = 1,302 2.5% of all lakes in region). Low agriculture region: N = 278 (4.2% of all lakes in region). High agriculture region: N = 69 (1.5% of all lakes in region). Moderate agriculture region: N = 955 (2.4% of all lakes in region). **(D)** Lakes with twenty or more summer PUDs between 2005 and 2014. Entire LAGOS region: N = 580 (1.1% of all lakes in region). Low agriculture region: N = 119 (1.8% of all lakes in region). High agriculture region: N = 31 (0.7% of all lakes in region). Moderate agriculture region: N = 430 (1.1% of all lakes in region). **R code:** Figure3.R.

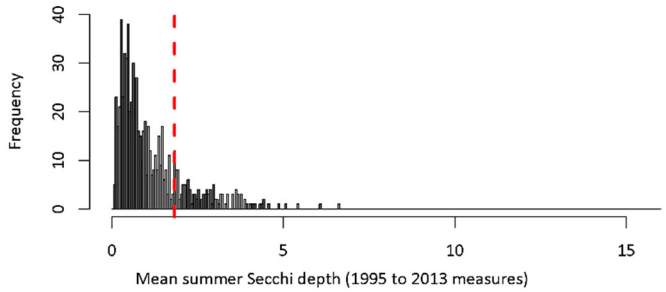
A



Moderate agriculture region



High agriculture region



Low agriculture region

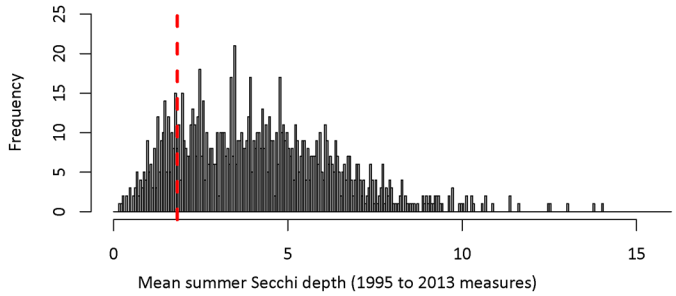


Figure 4 (A). Histogram of mean summer Secchi depths (m) observed between 1995 to 2013 at the 9,005 lakes in our dataset with at least one summer Secchi measurement. A lake is considered eutrophic (on average) if the mean summer Secchi depth is 1.83 meters or less (the red dashed line; personal communication with Patricia A. Soranno from 6/12/2017). According to this threshold, 3,503 of the 9,005 Secchi-measured lakes (40.0%) were eutrophic (on average) between 1995 and 2013. Mean: 2.67 m; Median: 2.35 m; Standard deviation: 1.81 m. **(B) Histograms of mean summer Secchi depths (m) observed between 1995 to 2013 in the regions.** Between 1995 and 2013 the number of Secchi-measured lakes that were eutrophic (on average) were: 2,745 (38.5%) in the moderate agriculture region; 576 (81.2%) in the high agriculture region; and 182 (15.7%) in the low agriculture region. **R code:** Figure4.R.

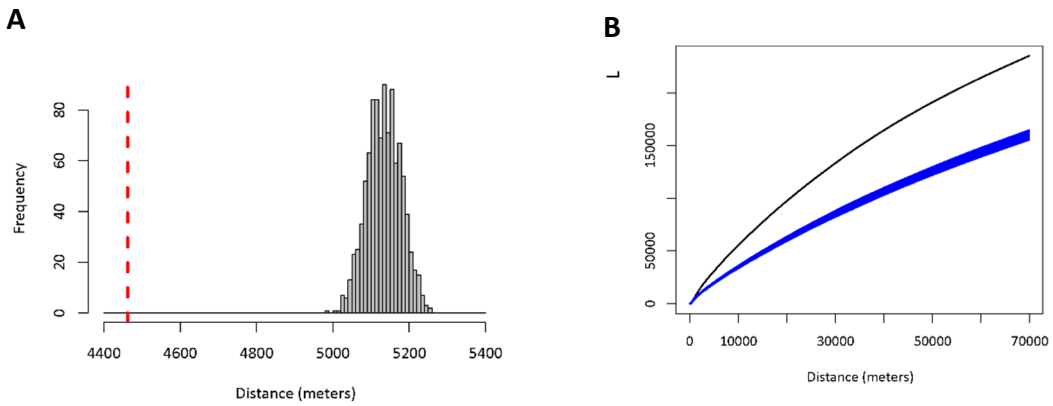


Figure 5. (A). The distribution of mean nearest lake distances across 1,000 random samples of 9,005 lakes and the mean nearest lake distances across the 9,005 lakes with measured Secchi depth. The mean nearest lake distances across 1,000 random samples of 9,005 lakes (without replacement) is 5,133.9 $\left(\bar{\bar{d}}_{min} = 5{,}133.9\right)$ and 45.0 meters $\left(sd\left(\bar{d}_{min}\right) = 45.0\right)$, respectively. The distribution of mean nearest lake distances across 1,000 random samples of 9,005 lakes is given by the histogram. The mean nearest lake distance across the 9,005 lakes with Secchi depth summer measurement data is 4,462.7 meters $\left(\bar{d}_{min} = 4{,}462.7\right)$, as indicated by the dashed red line. **(B). The normalized *K* function for the 9,005 lakes with measured Secchi depth (the thin black line) and the range of normalized *K* functions across the 1,000 random samples of 9,005 lakes (the blue area).** *L* indicates the (normalized) average lake density found around 9,005 lakes. The density is measured in a series of circles of radius *d* meters drawn around each lake in the sample. The thin black line gives the average density in the series of circles drawn around the 9,005 lakes with average summer Secchi measurements. The blue area contains all *L* functions for the 1,000 random samples of 9,005 lakes. Note that the black curve lies above the blue area. This would indicate that spatial pattern of Secchi lakes is more clustered in space than a spatial pattern of lakes randomly drawn from the entire dataset (Siart et al. 2017). The normalized *K*(*d*) function is known as the *L*(*d*) function, $L(d) = \sqrt{K(d)/\pi} - d$. **R code:** Figure5and6.R.

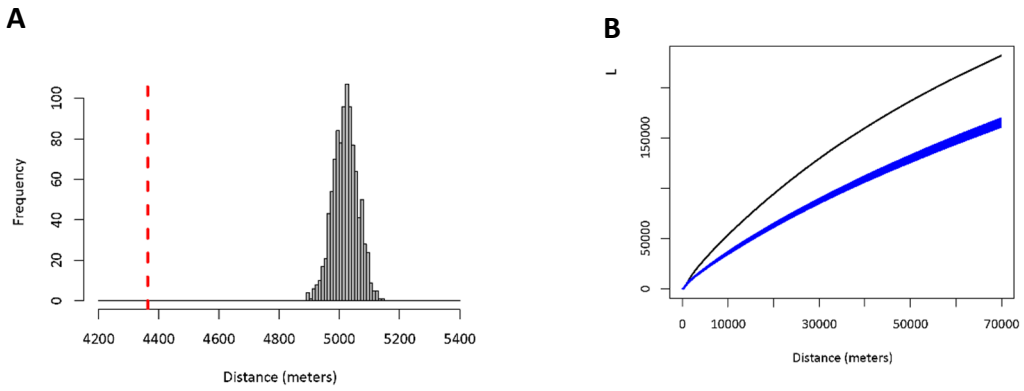


Figure 6. (A). The distribution of mean nearest lake distances across 1,000 random samples of 9,371 lakes and the mean nearest lake distances across the 9,371 lakes with measured maximum depth. The mean and standard deviation of the mean nearest lake distances across 1,000 random samples of 9,371 lakes (without replacement) is 5,019.7 $\left(\bar{\bar{d}}_{min} = 5{,}019.7\right)$ and 41.1 meters $\left(sd\left(\bar{d}_{min}\right) = 41.1\right)$, respectively. The distribution of mean nearest lake distances across 1,000 random samples of 9,371 lakes is given by the histogram. The mean nearest lake distance across the 9,371 lakes with maximum depth measurement data is 4,363.7 meters $\left(\bar{d}_{min} = 4{,}363.7\right)$, as indicated by the dashed red line. **(B). The normalized *K* function for the 9,371 lakes with measured maximum depth (the thin black line) and the range of *L* functions across the 1,000 random samples of 9,371 (the blue area).** *L* indicates the (normalized) average lake density found around 9,371 lakes. The density is measured in a series of circles of radius *d* meters drawn around each lake in the sample. The thin black line gives the average density in the series of circles drawn around the 9,371 lakes with average summer Secchi measurements. The blue area contains all *L* functions for the 1,000 random samples of 9,371 lakes. Note that the black curve lies above the blue area. This would indicate that spatial pattern of lakes measured for maximum depth is more clustered in space than a spatial pattern of lakes randomly drawn from the entire dataset (Siart et al. 2017). The normalized *K*(*d*) function is known as the *L*(*d*) function, $L(d) = \sqrt{K(d)/\pi} - d$. **R code:** Figure5and6.R.

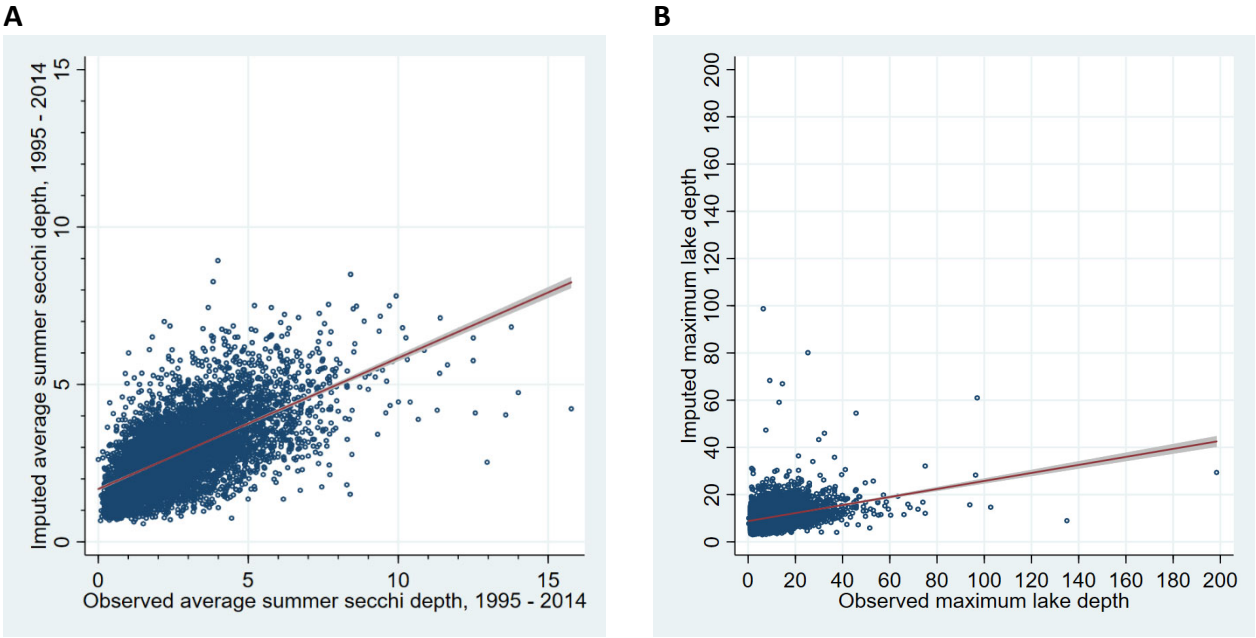


Figure 7. (A) Observed average summer Secchi depth versus mean imputed average summer Secchi depth and (B) observed maximum lake depth versus mean imputed maximum lake depth across the 5,570 lakes with observed average summer Secchi and maximum lake depths. See the text for details on how a lake's mean imputed average summer Secchi and maximum lake depths were calculated. Observed versus imputed data summary statistics are given in the table below. **Stata code:** Figure7.do

	Average summer Secchi depth		Maximum lake depth	
	Mean	**Std. dev.**	**Mean**	**Std. dev.**
Observed	2.869	1.765	10.600	9.072
Imputed	2.871	1.185	10.591	4.498

Notes: Average summer Secchi depth is based on 1995 to 2013 summer measures. Average summer Secchi RMSE: 1.387. Maximum depth RMSE: 8.634.

Tables

Table 1. Lake count and average summer Secchi depth in LAGOS region.

	Entire study area	Moderate ag. region	High ag. region	Low ag. region
All lake count	51,107	39,850	4,584	6,673
	Lakes with at least one summer Secchi depth measurement			
Lake count	8,733	6,869	710	1,154
Avg. summer Secchi depth (m)	2.668	2.570	1.127	4.196
Number (fraction) eutrophic	3,418 (0.39)	2,656 (0.39)	576 (0.81)	186 (0.16)

Notes: The data only includes lakes that are 4 ha or larger. Average summer Secchi depth and eutrophic counts are based on 1995 to 2013 summer measures of Secchi depth. The New England region is comprised of the HUC4 units with IDs of 1, 2, 3, 4, 6, 7, 8, 9, 10, and 11. This is a region of the country where agricultural land cover is less than 10% of all land cover. The Corn Belt region is comprised of the HUC4 units with IDs of 34, 50, 53, 56, 57, 61, and 63. This is a region of the country where agricultural land cover is greater than 75% of all land cover. All other HUC4 units in the dataset comprise the remaining region, which we call the 'Other' region. These regions were defined in Collins et al. (2017).

Table 2. Predicting which lakes have been measured for summer Secchi and maximum depth.
R code: RFSecchiYesNo.R and RFMaxDepthYesNo.R

	Predictions based on Random Forests		**Predictions based on random selection**	
			Weighted coin that flips to 1 with 0.1852 probability	*Weighted coin that flips to 1 with 0.1937 probability*
Dependent variable	Avg. summer Secchi depth measured = 1; 0 otherwise	Max. depth measured = 1; 0 otherwise	Avg. summer Secchi depth measured = 1; 0 otherwise	Max. depth measured = 1; 0 otherwise
Test data N	36,256	36,256	NA	NA
Train data N	9,065	9,065	9,065	9,065
	Prediction confusion matrices		We flipped a weighed coin 9,065 times and then compared to observed patterns of missing data. The stats below are averages over the 500 iterations.	
	Actual	**Actual**		
Prediction	**0** **1**	**0** **1**		
0	7298 192	7169 340		
1	94 1481	169 1387		
Accuracy	0.9685	0.9438	0.6985	0.6893
Sensitivity	0.9873	0.9770	0.8148	0.8059
Specificity	0.8852	0.8031	0.1850	0.1936

Notes: After dropping lakes with missing observations for variables in **Z** we are left with 45,321 of the 51,107 lakes for this analysis. Sensitivity = fraction of missing summer Secchi or maximum lake depth observations predicted correctly. Specificity = fraction of observed summer Secchi or maximum lake depth observation predicted correctly. When using the random selection method for predicting missing summer Secchi depth we used a weighted coin that flips to 1 with 0.1852 probability given that 18.52% of the 45,321 lakes have measured summer Secchi depth. When using the random selection method for predicting missing summer Secchi depth we used a weighted coin that equals 1 on 17.62% of coin flips given that 17.62% of the 51,107 lakes have measured summer Secchi depth.

Table 3. Variables in literature review-informed Z (Z_{Lit}). The column indicates the type of variable we assume each covariate is.

q_j variables	Q_{-j} variables	p_j variables	p_{-j} variables	Other variables
Lake area (ha)	Spat. lag of avg. summer Secchi depth	Dist. to nearest Core Business Statistical Area (CBSA) (km)	Avg. dist. to closest CBSA of the nearest 5 lakes (km)	% of pop. in lake's subwatershed with a BA or higher, 2011-2015
Max. lake depth (m)	*Spat. lag of count of each amenity feature*	Pop. density in lake's subwatershed (people sq km^{-1})	Count of lakes ≥ 4 ha in lake's subwatershed	% of pop. in lake's subwatershed that is poor, 2011-2015
Avg. summer Secchi depth (m)	Avg. dist. to the nearest 5 lakes with Secchi measurements (km)	% of 500 m buffer around lake in dev. cover in 2011	% lake area within lake's subwatershed (lakes ≥ 4 ha)	% of pop. in lake's subwatershed that is white, 2011-2015
Secchi depth sampled more than once*	Avg. size of the nearest 5 lakes (ha)	Avg. distance to nearest 5 lakes (km)		Median age in lake's subwatershed, 2011-2015
30-yr avg. temp. in lake's subwatershed (Celsius)				Median HH inc. in lake's subwatershed, 2011-2015
% of 500 m buffer around lake in forest in 2011				*Region dummies*
% of 500 m buffer around lake in wetlands in 2011				Latitude and longitude of lake
Count of each amenity feature				Spat. lag of summer PUD count

Notes: * indicates only available when using limited dataset.

Table 4. Variables in single (SL) and double LASSO (DL)-informed Z's. *R code:* LASSO.R with the Limited dataset and LASSOImputed1.R – LASSOImputed20.R with the Augmented datasaet. See ImputeSecchiMaxLakeDepth.do for Stata code to create the Augmented dataset.

	$Z_{SL,Limited}$	$Z_{DL,Limited}$	$Z_{SL,Augmented}$	$Z_{DL,Augmented}$
	Lake area	Lake area	Lake area	Lake area
	Beach feature count	Max. lake depth	Max. lake depth	Max. lake depth
		Avg. summer Secchi depth	Beach feature count	Avg. summer Secchi depth
		30-yr precip. avg. in lake's subwatershed	Boat launch feature count	30-yr precip. avg. in lake's subwatershed
		% of 500 m buffer around lake in forest in 2011	Hotel feature count	% of 500 m buffer around lake in forest in 2011
		% of 500 m buffer around lake in wetlands in 2011	Marina feature count	% of 500 m buffer around lake in wetlands in 2011
		Total stream density in lake's subwatershed	Toilets feature count	% of 500 m buffer around lake in agriculture in 2011
		Avg. area of forested wetland in lake's subwatershed	Pop. density in lake's subwatershed	Total stream density in lake's subwatershed
		Beach feature count	% of 500 m buffer around lake in developed cover in 2011	Avg. area of forested wetland in lake's subwatershed
		Spat. lag of avg. summer Secchi depth	% of pop. in lake's subwatershed with a BA degree or higher, 2011-2015	Beach feature count
		Avg. dist. to the nearest 5 lakes with Secchi measurements	Lake has been sampled for NO_2NO_3	Boat launch feature count
		Avg. dist. to nearest 5 lakes	Lake has been sampled for total phosphorous	Hotel feature count
		Spat. lag of boat launch feature count	Spat. lag of summer PUD count	Marina feature count
		Spat. lag of toilet feature count		Picnic feature count
		Lake has been sampled for Chlorophyll		Spat. lag of average summer Secchi depth
		High ag. region dummy		Spat. lag of beach feature count

	$Z_{SL,Limited}$	$Z_{DL,Limited}$	$Z_{SL,Augmented}$	$Z_{DL,Augmented}$
				Spat. lag of boat launch feature count
				Spat. lag of hotel feature count
				Spat. lag of toilet feature count
				Avg. dist. to the nearest 5 lakes with Secchi measurements
				Avg. dist. to nearest 5 lakes
				Pop. density in lake's subwatershed
				Avg. distance to closest CBSA across the nearest 5 lakes
				% lake area within lake's subwatershed (lakes ≥ 4 ha)
				Median age in lake's subwatershed, 2011-2015
				% of pop. in lake's subwatershed that is white, 2011-2015
				Lake has been sampled for NO_2NO_3
				High ag. region dummy
λ_{min}	4.53		0.93 (0.00)	
$\lambda_{min,DL,FS}$		4.97		3.11 (0.00)
$\lambda_{min,DL,SS}$		0.07		0.06 (0.14)
CVM	142.22		143.36 (0.65)	
$CVM_{DL,FS}$		142.22		143.36 (0.65)
$CVM_{DL,SS}$		3.27		2.63 (0.04)

Notes: CVM = Cross-validation mean. Subscript SL indicates single LASSO. Subscript DL,FS indicates double LASSO, first stage. Subscript DL,SS indicates double LASSO, second stage.

Table 5. Variables in VSURF-informed Zs. *R code:* VSURF.R with the Limited dataset and VSURFImputed1.R – VSURFImputed20.R with the Augmented datasaet.

$\mathbf{Z}_{\mathbf{VSURF,Limited}}$	$\mathbf{Z}_{\mathbf{VSURF,Augmented}}$
Lake area	Beach feature count
Spat. lag of summer PUD count	Lake area
Marina feature count	Spat. lag of summer PUD count
% of 500 m buffer in dev. cover in 2011	% of 500 m buffer in dev. cover in 2011
Toilet feature count	30-yr precip. avg in lake's subwatershed[$]

Notes: Selected variables in $\mathbf{Z}_{\mathbf{VSURF,Limited}}$ column are listed in order of predictive importance. Selected variables in $\mathbf{Z}_{\mathbf{VSURF,Augmented}}$ column are listed in mean order of predictive importance over the 20 sets of selected variables.
[$]Only 7 of the 20 VSURF iterations included this variable. However, it was needed for the Poisson count model to converge to a solution over $\mathbf{Z}_{\mathbf{VSURF,Augmented}}$.

Table 6. Estimated Poisson count model with $Z_{Lit,Limited}$, $Z_{DL,Limited}$, $Z_{VSURF,Limited}$, and $Z_{DL+VSURF,Limited}$. Robust standard errors are in parentheses. *Stata code:* ObservedSecchi.do.

	$Z_{Lit,Limited}$	$Z_{DL,Limited}$	$Z_{VSURF,Limited}$	$Z_{DL+VSURF,Limited}$
Lake area	4.45e-05***	9.35e-05***	9.83e-05***	7.16e-05***
	(1.22e-05)	(1.11e-05)	(1.42e-05)	(1.00e-05)
Avg. summer Secchi depth	0.0413	0.106**		0.0760*
	(0.0331)	(0.0446)		(0.0454)
Max. depth	0.0192***	0.0141***		0.0162***
	(0.00436)	(0.00301)		(0.00329)
Secchi depth sampled more than once	0.480***			
	(0.158)			
30-yr avg. temp. in lake's subwatershed	-0.0918			
	(0.129)			
% of 500 m buffer in forest in 2011	0.00808**	-0.0203***		0.0112***
	(0.00322)	(0.00383)		(0.00297)
% of 500 m buffer in wetlands in 2011	0.00460	-0.0374***		0.00102
	(0.00454)	(0.00646)		(0.00565)
Boat feature count	0.350***			
	(0.0575)			
Beach feature count	0.0374	0.128***		0.103***
	(0.0289)	(0.0194)		(0.0217)
Hotel feature count	0.309			
	(0.856)			
Shelter feature count	0.0584			
	(0.642)			
Toilets feature count	0.0597**		0.101***	0.0533**
	(0.0248)		(0.0222)	(0.0253)
Picnic feature count	-1.182**			
	(0.468)			
BBQ feature count	0.822**			
	(0.350)			
Marina feature count	0.125		0.146***	0.0358
	(0.143)		(0.0156)	(0.0225)
Spat. lag of avg. summer Secchi depth	0.813**	-0.294		-0.131
	(0.392)	(0.259)		(0.236)
Spat. lag of boat feature count	-2.355	56.34***		30.36***
	(6.090)	(5.462)		(5.311)
Spat. lag of beach feature count	2.821			
	(7.846)			
Spat. lag of hotel feature count	150.9			
	(114.2)			
Spat. lag of shelter feature count	6.246			
	(65.43)			
Spat. lag of toilets feature count	-1.591	18.04***		12.58***
	(4.272)	(3.173)		(3.778)
Spat. lag of picnic feature count	55.92**			
	(27.56)			
Spat. lag of BBQ feature count	-134.8			
	(88.93)			
Spat. lag of marina feature count	9.331			
	(30.98)			

	$Z_{Lit,Limited}$	$Z_{DL,Limited}$	$Z_{VSURF,Limited}$	$Z_{DL+VSURF,Limited}$
Avg. dist. to nearest 5 lakes w/ Secchi measure	0.0219***	0.0349***		0.0371***
	(0.00851)	(0.00931)		(0.00931)
Avg. size of nearest 5 lakes	0.000338			
	(0.000314)			
Dist. to nearest CBSA	9.55e-06			
	(2.37e-05)			
Pop. den. in subwatershed	0.000342***			
	(9.19e-05)			
% of 500 m buffer in dev. cover in 2011	0.0254***		0.0292***	0.0361***
	(0.00321)		(0.00323)	(0.00355)
Avg. dist. to nearest 5 lakes	0.0210	-0.0215		0.00864
	(0.0265)	(0.0268)		(0.0286)
Avg. dist. to closest CBSA of the nearest 5 lakes	-7.78e-06			
	(2.41e-05)			
Lake count in subwatershed	-0.0145			
	(0.00949)			
Lake area in subwatershed	0.0270***			
	(0.00745)			
% of pop. in lake's subwatershed with a BA	0.0433***			
	(0.00676)			
% of pop. in lake's subwatershed that is poor	0.00934			
	(0.0187)			
% of pop. in lake's subwatershed that is white,	0.00626			
	(0.00774)			
Median age in lake's subwatershed, 2011-2015	-0.0156			
	(0.0157)			
Median HH inc. in lake's subwatershed, 2011-2015	1.61e-06			
	(9.57e-06)			
Lag of summer PUD count	0.274**		0.410***	0.283***
	(0.119)		(0.0604)	(0.0593)
Latitude	-0.151			
	(0.143)			
Longitude	-0.0313			
	(0.0282)			
Low ag. region	-0.727**			
	(0.339)			
High ag. Region	-0.743***	0.106		0.344
	(0.249)	(0.310)		(0.265)
30-yr avg. precip. in subwatershed (mm)		0.00327***		0.00166***
		(0.000548)		(0.000553)
Chlorophyll a measured at lake (µg/l)		0.549***		0.567***
		(0.181)		(0.173)
Total stream density in subwatershed (m/m sq.)		-0.193***		-0.148***
		(0.0280)		(0.0280)
Avg. forested wetland area in subwatershed (ha)		-0.00304		-0.00201
		(0.00453)		(0.00450)
Constant	-0.583	-2.066***	-0.116	-3.501***
	(7.023)	(0.617)	(0.151)	(0.632)
Observations	3,276	3,276	3,276	3,276

Table 7. Estimated Poisson count model with $Z_{Lit,Augmented}$, $Z_{DL,\ Augmented}$, $Z_{VSURF,\ Augmented}$, and $Z_{DL+VSURF,\ Augmented}$. Robust standard errors are in parentheses *Stata code:* ImputedSecchi.do.

	$Z_{Lit,Aug.}$	$Z_{DL,Aug.}$	$Z_{VSURF,Aug.}$	$Z_{DL+VSURF,Aug.}$
Lake area	0.00005***	0.00006***	0.00012***	0.00005***
	(0.00001)	(0.00001)	(0.00002)	(0.00001)
Avg. summer Secchi depth	0.08127**	0.07**		0.07**
	(0.03373)	(0.034)		(0.034)
Max. depth	0.0301***	0.029***		0.029***
	(0.00407)	(0.004)		(0.004)
30-yr avg. temp. in lake's subwatershed	-0.064			
	(0.064)			
% of 500 m buffer in forest in 2011	0.006**	-0.016***		-0.0067*
	(0.002)	(0.002)		(0.0034)
% of 500 m buffer in wetlands in 2011	0.011***	-0.012***		-0.0019
	(0.003)	(0.004)		(0.004)
Boat feature count	0.306***	0.200***		0.216***
	(0.049)	(0.052)		(0.054)
Beach feature count	0.517***	0.429***	0.341***	0.431***
	(0.056)	(0.065)	(0.022)	(0.069)
Hotel feature count	2.148***	2.007***		1.868***
	(0.506)	(0.491)		(0.463)
Shelter feature count	0.261*			
	(0.158)			
Toilets feature count	0.085***			
	(0.024)			
Picnic feature count	0.198***	-0.37**		-0.35**
	(0.06625)	(0.162)		(0.169)
BBQ feature count	-0.667***			
	(0.097)			
Marina feature count	0.398***	0.19 (0.169)		0.159
	(0.121)			(0.176)
Spat. lag of avg. summer Secchi depth	0.274	0.255**		0.26**
	(0.23)	(0.126)		(0.125)
Spat. lag of boat feature count	-1.73468	10.301**		4.452
	(4.719)	(4.927)		(4.872)
Spat. lag of beach feature count	4.94	5.586**		-7.333
	(4.55433)	(2.202)		(5.979)
Spat. lag of hotel feature count	-328.08	-146.747		-341.665
	(202.38)	(132.196)		(217.513)
Spat. lag of shelter feature count	-4.339			
	(9.73)			
Spat. lag of toilets feature count	1.743	4.72**		5.933***
	(2.33)	(2.273)		(2.295)
Spat. lag of picnic feature count	27.25**			
	(12.14)			
Spat. lag of BBQ feature count	30.95			
	(28.01)			
Spat. lag of marina feature count	-110.04***			
	(34.53)			

	$Z_{Lit,Aug.}$	$Z_{DL,Aug.}$	$Z_{VSURF,Aug.}$	$Z_{DL+VSURF,Aug.}$
Avg. dist. to nearest 5 lakes w/ Secchi measure	0.011*	0.005 (0.006)		0.008
	(0.00592)			(0.006)
Avg. size of nearest 5 lakes	-0.00006			
	(0.00008)			
Dist. to nearest CBSA	0.00003			
	(0.00003)			
Pop. den. in subwatershed	0.00011	0.00041***		0.00025**
	(0.00013)	(0.00009)		(0.00012)
% of 500 m buffer in dev. cover in 2011	0.023***		0.022***	0.012***
	(0.003)		(0.002)	(0.003)
Avg. dist. to nearest 5 lakes	0.093***	0.122***		0.114***
	(0.021)	(0.024)		(0.025)
Avg. dist. to closest CBSA of the nearest 5 lakes	-0.00003	-0.000001		-0.0000004
	(0.00003)	(0.000002)		(0.0000019)
Lake count in subwatershed	-0.011***			
	(0.004)			
Lake area in subwatershed	0.028***	0.023***		0.023***
	(0.005)	(0.006)		(0.006)
% of pop. in lake's subwatershed with a BA	0.038***			
	(0.004)			
% of pop. in lake's subwatershed that is poor	-0.015*			
	(0.009)			
% of pop. in lake's subwatershed that is white	-0.00175	-0.005		-0.004
	(0.00479)	(0.005)		(0.005)
Median age in lake's subwatershed	-0.0076	0.006 (0.008)		0.005
	(0.0081)			(0.008)
Median HH inc. in lake's subwatershed	-0.00001***	0.000013***		0.000011***
	(0.000004)	(0.000002)		(0.000002)
Latitude	-0.126*			
	(0.069)			
Longitude	0.0297***			
	(0.0116)			
Low ag. Region	-0.632***			
	(0.162)			
High ag. region	0.131	0.11		0.239
	(0.222)	(0.211)		(0.216)
Spat. lag of summer PUD count	0.454***		0.275***	0.322***
	(0.131)		(0.024)	(0.097)
30-yr min temp. in lake's subwatershed (Celsius)			0.031	
			(0.019)	
30-yr avg. precip. in subwatershed (mm)		0.00128***		0.0011***
		(0.00035)		(0.0003)
% of 500 m buffer in ag. in 2011		-0.0295***		-0.02***
		(0.00307)		(0.004)
Total stream density in subwatershed (m/m sq.)		-0.00495		-0.0097
		(0.01882)		(0.01871)
Avg. forested wetland area in subwatershed (ha)		-0.015**		-0.016**
		(0.007)		(0.008)
NO_2-NO_3 measured at lake (µg/l as N)		1.195***		1.16***
		(0.082)		(0.082)

	$Z_{Lit,Aug.}$	$Z_{DL,Aug.}$	$Z_{VSURF,Aug.}$	$Z_{DL+VSURF,Aug.}$
Constant	6.093*	-3.208***	-1.008***	-3.693***
	(3.41)	(0.527)	(0.064)	(0.627)
Observations	19,790	19,790	19,790	19,790

Table 8. Ten-cross fold validation of Poisson count model with $Z_{Lit,Limited}$, $Z_{DL,Limited}$, $Z_{VSURF,Limited}$, and $Z_{DL+VSURF,Limited}$. Each cell gives the RMSE of for the given fold. *Stata code:* ObservedSecchi.do.

Fold	$Z_{Lit,Limited}$	$Z_{DL,Limited}$	$Z_{VSURF,Limited}$	$Z_{DL+VSURF,Limited}$
1	34	21	10	355
2	48,220	46	84	63
3	26	12	593,000,000	15
4	14	47	171	40
5	34	14	11	1,224
6	407,000,000,000	18	23	12
7	9	38	15	39
8	19	15	24	21
9	34	313	38	16
10	473	1,626	32	16,900,000
Mean RMSE	**$40.7x10^9$**	**215**	**59,300,041**	**1,690,178**
SD of RMSE	*$122.1x10^9$*	*478*	*177,899,986*	*5,069,941*
Mean RMSE less largest RMSE	**5,429.4**	**58.3**	**45.3**	**198.3**
Mean RMSE less two largest RMSEs	**80.6**	**26.4**	**29.5**	**70.0**

Table 9. Ten-cross fold validation of Poisson count model with $Z_{Lit,Augmented}$, $Z_{DL,Augmented}$, $Z_{VSURF,Augmented}$, and $Z_{DL+VSURF,Augmented}$. For each model we conduct 20 cross fold validation analyses, one for each model estimated over a unique set of imputed average summer Secchi and maximum lake depths. The exception is the model estimated with $Z_{VSURF,Augmented}$ as this covariate vector does not contain any imputed data. *Stata code:* ImputedSecchi.do.

Fold	Literature	D. LASSO	VSURF	DLASSO+VSURF
Mean RMSE	**66.0**	**$2.21x10^{11}$**	**27,380**	**$7.71x10^{10}$**
SD of RMSE	263.4	$1.59x10^{12}$	82,120	$3.56x10^{11}$
Mean RMSE less the ten largest RMSE	**20.7**	**40.8**	**6.7**	**18.6**
Mean RMSE less the twenty largest RMSEs	**10.6**	**7.7**	**6.3**	**7.7**

Table 10. Comparison of variables selected by the Machine Learning algorithms when average summer Secchi is measured with 1995 to 2013 observations versus 2005 to 2013 observations.

	Double LASSO			VSURF		
		# of unique variables with…			# of unique variables with…	
	# of common variables	1995-2013 measures	2005-2013 measures	# of common variables	1995-2013 measures	2005-2013 measures
Limited dataset	8	8	4	5	0	5
Augmented dataset	17	11	2	3	2	3

Table 11. Comparison of estimated coefficients using preferred models and 1995-2013 versus 2005-2013 measures of summer Secchi depth. Robust standard errors are in parentheses. *Stata code:* ObservedSecchi.do, ImputedSecchi.do, ObservedSecchi0513.do, ImputedSecchi0513.do.

	$Z_{DL,Limited}$		$Z_{DL+VSURF,Augmented}$	
	w/ 95-13 measures	**w/ 05-13 measures**	**w/ 95-13 measures**	**w/ 05-13 measures**
Lake area	$9.4x10^{-5}$***	$7.6x10^{-5}$***	$5.0x10^{-5}$***	$5.0x10^{-5}$***
	($-1.1x10^{-5}$)	($-1.2x10^{-5}$)	($1 x10^{-5}$)	($1 x10^{-5}$)
Avg. summer Secchi depth	0.106**	0.230***	0.070**	0.109***
	(-0.045)	(-0.047)	(0.034)	(0.038)
Max. depth	0.014***	0.008***	0.029***	0.021***
	(-0.003)	(-0.003)	(0.004)	(0.004)
30-yr avg. precip. in subwatershed (mm)			0.001***	0.0014***
			(0.0003)	(0.0004)
% of 500 m buffer in forest in 2011	-0.020***	-0.011***	-0.007*	0.004
	(-0.004)	(-0.004)	(0.003)	(0.006)
% of 500 m buffer in wetlands in 2011			-0.0019	0.002
			(0.004)	(0.005)
% of 500 m buffer in ag. in 2011			-0.020***	-0.017***
			(0.004)	(0.005)
Total stream den. in subwatershed (m/m sq)			-0.010	-0.089***
			(0.019)	(0.03)
Beach feature count	0.128***	0.065***		
	(-0.019)	(-0.025)		
Spat. lag of avg. summer Secchi depth	-0.294	0.190	0.260**	-0.064 (0.163)
	(-0.259)	(-0.206)	(0.125)	
Spat. lag of boat launch feature count			4.452	13.625***
			(4.872)	(4.766)
Spat. lag of beach feature count			-7.333	-11.826*
			(5.979)	(6.099)
Spat. lag of toilets feature count			5.933***	3.301
			(2.295)	(2.087)
Avg. dist. to nearest 5 lakes w/ Secchi measure	0.035***	0.039***	0.008	-0.003
	(-0.009)	(-0.009)	(0.006)	(0.006)
Pop. den. in subwatershed			0.0003**	-0.00003
			(0.0001)	(0.00015)
% of 500 m buffer in dev. cover in 2011			0.012***	0.017***
			(0.003)	(0.005)
Avg. dist. to nearest 5 lakes	-0.022	-0.015	0.114***	0.182***
	(-0.027)	(-0.028)	(0.025)	(0.017)
Avg. dist. to closest CBSA from the nearest 5 lakes			$-4x10^{-7}$	$-4x10^{-6}$**
			($1.9 x10^{-6}$)	($2x10^{-6}$)

High Ag. Region	0.239	0.997*
	(0.216)	(0.513)
Spat. lag of summer PUD count	0.322***	0.379***
	(0.097)	(0.085)

Appendix

SI Table 1. Estimated IV Poisson count model with $Z_{Lit,Limited}$ and $Z_{Lit,Augmented}$. Robust standard errors are in parentheses**.** *Stata Code:* ObservedSecchiIV.do and ImputedSecchiIV.do.

	$Z_{Lit,Limited}$	$Z_{Lit,Augmented}$
Avg. summer Secchi depth	2.467***	0.025
	(0.905)	
Lake area	0.00083**	0.007
	(0.0004)	
Max. depth	-0.091**	0.04
	(0.036)	
Secchi depth sampled more than once	0.234	
	(0.367)	
30-yr avg. temp. in lake's subwatershed	-0.519***	-0.259
	(0.202)	
% of 500 m buffer in forest in 2011	0.014**	0.003
	(0.006)	
% of 500 m buffer in wetlands in 2011	0.031***	0.00044
	(0.009)	
Boat feature count	0.636***	0.738
	(0.13)	
Beach feature count	0.072	1.002
	(0.141)	
Hotel feature count	0.816	2.122
	(0.805)	
Shelter feature count	0.680	0.95
	(0.454)	
Toilets feature count	0.036	0.242
	(0.134)	
Picnic feature count	-0.594	0.058
	(0.414)	
BBQ feature count	-0.074	-1.49
	(0.362)	
Marina feature count	-0.515*	0.223
	(0.274)	
Spat. lag of avg. summer Secchi depth	-4.664***	0.086
	(1.666)	
Spat. lag of boat feature count	-10.356	4.843
	(11.913)	
Spat. lag of beach feature count	39.834**	13.645
	(19.664)	

	$Z_{Lit,Limited}$	$Z_{Lit,Augmented}$
Spat. lag of hotel feature count	-210.912	-96.831
	(296.566)	
Spat. lag of shelter feature count	34.552	-18.053
	(120.298)	
Spat. lag of toilets feature count	18.001***	1.324
	(5.104)	
Spat. lag of picnic feature count	-22.047	20.24
	(76.069)	
Spat. lag of BBQ feature count	166.29*	29.066
	(88.992)	
Spat. lag of marina feature count	-115.274*	-136.273
	(69.829)	
Avg. dist. to nearest 5 lakes w/ Secchi measure	0.022	-0.006
	(0.018)	
Avg. size of nearest 5 lakes	-0.00081	0.000003
	(0.00072)	
Dist. to nearest CBSA	-0.0001**	0.00001
	(0.00005)	
Pop. den. in subwatershed	0.001**	0.00041
	(0.001)	
% of 500 m buffer in dev. cover in 2011	0.034***	0.029
	(0.006)	
Avg. dist. to nearest 5 lakes	0.191***	0.114
	(0.073)	
Avg. dist. to closest CBSA across the nearest 5 lakes	0.0001**	-0.00001
	(0.00005)	
Lake count in subwatershed	-0.036**	0.005
	(0.017)	
Lake area in subwatershed	0.081**	0.006
	(0.04)	
% of pop. in lake's subwatershed with a BA, 2011-15	0.019	0.043
	(0.013)	
% of pop. in lake's subwatershed that is poor, 2011-15	0.017	0.024
	(0.022)	
% of pop. in lake's subwatershed that is white, 2011-15	0.025*	-0.004
	(0.014)	
Median age in lake's subwatershed, 2011-2015	-0.026	0.008
	(0.018)	
Median HH inc. in lake's subwatershed, 2011-2015	0.000001	-0.000002
	(0.00001)	
Spatial lag of summer PUD count	0.058	0.788
	(0.371)	

	$\mathbf{Z}_{\mathbf{Lit,Limited}}$	$\mathbf{Z}_{\mathbf{Lit,Augmented}}$
Latitude	-0.674***	-0.391
	(0.225)	
Longitude	0.083	0.013
	(0.051)	
Low ag. Region	0.074	-0.421
	(0.686)	
High ag. Region	0.046	-0.738
	(0.35)	
Constant	44.641***	15.054
	(15.362)	
Observations	3,276	19,790

Notes: Average summer Secchi instrumented with subwatershed-level stream density and percentage of 500-m buffer in agriculture land cover. $\mathbf{Z}_{\mathbf{Lit,Augmented}}$ results are mean coefficient values over 18 model estimates. Test of overidentifying restriction in IV with $\mathbf{Z}_{\mathbf{Lit,Limited}}$: Hansen's J chi2(1) = 0.00003 (p = 0.995). Mean of tests of overidentifying restriction in IV with $\mathbf{Z}_{\mathbf{Lit,Augmented}}$: mean Hansen's J chi2(1) = 38.02 (p = 0.000).

SI Table 2. Variables in single (SL) and double LASSO (DL)-informed Z's with average summer Secchi based on summer 2005 to 2013 measures. *R code:* LASSO0513.Rwith Limited dataset and LASSOImputed10513.R – LASSOImputed200513.R with the Augmented dataset.

	$Z_{SL,Limited}$	$Z_{DL,Limited}$	$Z_{SL,Augmented}$	$Z_{DL,Augmented}$
	Lake area	Lake area	<<None>>	Lake area
	Beach feature count	Average summer Secchi depth		Average summer Secchi depth
	Boat launch feature count	Maximum lake depth		Maximum lake depth
	BBQ feature count	% of 500 m buffer in forest in 2011		30-yr avg. precip. in lake's subwatershed
	Pop. den. in lake's subwatershed	Beach feature count		% of 500 m buffer in wetlands in 2011
	Spat. lag of summer PUD count	Boat launch feature count		% of 500 m buffer in ag. in 2011
		BBQ feature count		% of 500 m buffer in forest in 2011
		Spat. lag of average summer Secchi depth		Total stream density in lake's subwatershed
		Avg. dist. to nearest 5 lakes		Avg. scrub-shrub wetland area in lake's subwatershed
		Avg. dist. to the nearest 5 lakes w/ Secchi measurements		BBQ feature count
		Pop. den. in lake's subwatershed		Spat. lag of average summer Secchi depth
		Spat. lag of summer PUD count		Spat. lag of beach feature count
				Spat. lag of boat feature count
				Spat. lag of toilet feature count
				Avg. dist. to the nearest 5 lakes w/ Secchi measurements
				Pop. den. in lake's subwatershed
				Avg. dist. to nearest 5 lakes
				Avg. dist. to closest CBSA across the nearest 5 lakes
				High ag. region
λ_{min}	4.30		4.19 (0.00)	
$\lambda_{min,DL,FS}$		4.30		2.89 (0.00)
$\lambda_{min,DL,SS}$		0.21		2.67 (0.05)
CVM	374.14		145.64 (0.00)	
$CVM_{DL,FS}$		374.14		145.64 (0.00)
$CVM_{DL,SS}$		3.39		2.67 (0.05)

Notes: CVM = Cross-validation mean. Subscript SL indicates single LASSO. Subscript DL,FS indicates double LASSO, first stage. Subscript DL,SS indicates double LASSO, second stage.

SI Table 3. Variables in VSURF-informed Z's with average summer Secchi based on summer 2005 to 2013 measures. *R code:* VSURF0513.R with the Limited dataset and VSURFImputed10513.R - VSURFImputed200513.R with the Augmented dataset.

$\mathbf{Z_{VSURF,Limited}}$	$\mathbf{Z_{VSURF,Augmented}}$
Spat. lag of summer PUD count	Lake area
Lake area	% of 500 m buffer in dev. cover in 2011
Spat. lag of marina feature count	Spat. lag of shelter feature count
% of pop. in lake's subwatershed with a BA, 2011-15	Spat. lag of summer PUD count
Pop. den. in lake's subwatershed	Pop. den. in lake's subwatershed
Longitude	% of population in lake's subwatershed with a BA, 2011-15
% of 500 m buffer in dev. cover in 2011	
Marina feature count	
Toilet feature count	
Maximum depth	

Notes: Selected variables in $\mathbf{Z_{VSURF,Limited}}$ column are listed in order of predictive importance. Selected variables in $\mathbf{Z_{VSURF,Augmented}}$ column are listed in mean order of predictive importance over the 20 sets of selected variables.

SI Table 4. Estimated Poisson count model with $Z_{Lit,Limited}$, $Z_{DL,Limited}$, $Z_{VSURF,Limited}$, and $Z_{DL+VSURF,Limited}$ and using average summer Secchi based on measurements from 2005 to 2013.
Robust standard errors are in parentheses. *Stata code:* ObservedSecchi0513.do.

	$Z_{Lit,Limited}$	$Z_{DL,Limited}$	$Z_{VSURF,Limited}$	$Z_{DL+VSURF,Limited}$
Lake area	2.81e-05**	7.60e-05***	8.48e-05***	7.63e-05***
	(1.27e-05)	(1.22e-05)	(1.05e-05)	(1.13e-05)
Avg. summer Secchi depth	0.139**	0.230***		0.199***
	(0.0569)	(0.0474)		(0.047)
Max. depth	0.0143***	0.008***	0.011***	0.006**
	(0.003)	(0.003)	(0.004)	(0.003)
Secchi depth sampled more than once	0.209			
	(0.170)			
30-yr avg. temp. in lake's subwatershed	-0.041			
	(0.114)			
% of 500 m buffer around lake in forest in 2011	0.003	-0.011***		6.70e-05
	(0.004)	(0.004)		(0.00333)
% of 500 m buffer around lake in wetlands in 2011	0.0016			
	(0.0051)			
Boat feature count	0.182***	0.343***		0.239***
	(0.060)	(0.042)		(0.0517)
Beach feature count	0.068***	0.065***		0.065***
	(0.024)	(0.025)		(0.024)
Hotel feature count	1.524***			
	(0.318)			
Shelter feature count	-0.0642			
	(0.316)			
Toilets feature count	0.156***		0.161***	0.160***
	(0.0358)		(0.0349)	(0.0312)
Picnic feature count	-0.107			
	(0.238)			
BBQ feature count	2.105***	1.628***		2.014***
	(0.256)	(0.168)		(0.153)
Marina feature count	0.372*		1.006***	0.500***
	(0.219)		(0.100)	(0.133)
Spat. lag of avg. summer Secchi depth	0.229	0.190		0.807***
	(0.297)	(0.206)		(0.205)
Spat. lag of boat feature count	-0.405			
	(6.175)			
Spat. lag of beach feature count	-3.708			
	(8.798)			
Spat. lag of hotel feature count	-141.0			
	(133.9)			
Spat. lag of shelter feature count	3.655			
	(30.32)			
Spat. lag of toilets feature count	-0.954			
	(3.226)			
Spat. lag of picnic feature count	44.99**			
	(21.00)			
Spat. lag of BBQ feature count	-196.5*			
	(108.8)			
Spat. lag of marina feature count	-47.11		56.41	9.352

	$Z_{Lit,Limited}$	$Z_{DL,Limited}$	$Z_{VSURF,Limited}$	$Z_{DL+VSURF,Limited}$
	(37.95)		(45.71)	(43.40)
Avg. dist. to nearest 5 lakes w/ Secchi measure	0.044***	0.039***		0.0418***
	(0.008)	(0.009)		(0.00797)
Avg. size of nearest 5 lakes	0.0006			
	(0.0004)			
Dist. to nearest CBSA	6.3e-05***			
	(2.1e-05)			
Pop. den. in subwatershed	0.0002*	0.0007***	0.0002	0.0002**
	(0.0001)	(7.77e-05)	(0.0001)	(9.12e-05)
% of 500 m buffer around lake in dev. cover in 2011	0.0245***		0.0171***	0.0246***
	(0.00384)		(0.00309)	(0.00326)
Avg. dist. to nearest 5 lakes	-0.0236	-0.0154		0.00925
	(0.0237)	(0.0281)		(0.0254)
Avg. dist. to closest CBSA across the nearest 5 lakes	-6.1e-5***			
	(2.0e-05)			
Lake count in subwatershed	-0.007			
	(0.006)			
Lake area in subwatershed	0.038***			
	(0.007)			
% of pop. in lake's subwatershed with a BA, 2011-15	0.045***		0.0266***	0.0295***
	(0.006)		(0.00404)	(0.00401)
% of pop. in lake's subwatershed that is poor, 2011-15	-0.006			
	(0.011)			
% of pop. in lake's subwatershed that is white, 2011-15	0.0007			
	(0.0057)			
Median age in lake's subwatershed, 2011-15	-0.021**			
	(0.011)			
Median HH inc. in lake's subwatershed, 2011-15	-2.2e-5***			
	(5.8e-6)			
Lag of summer PUD count	0.499***	0.282***	0.171	0.266**
	(0.148)	(0.0333)	(0.126)	(0.117)
Latitude	-0.0892			
	(0.104)			
Longitude	0.0159		0.0144	-0.0390***
	(0.0201)		(0.00936)	(0.0102)
Low ag. region	-0.536**			
	(0.268)			
High ag. region	-0.109			
	(0.230)			
Constant	4.545	-0.602	0.576	-7.472***
	(5.693)	(0.534)	(0.899)	(1.410)
Observations	2,706	2,706	2,706	2,706

SI Table 5. Estimated Poisson count model with $Z_{Lit,Augmented}$, $Z_{DL,Augmented}$, $Z_{VSURF,Augmented}$, and $Z_{DL+VSURF,Augmented}$ and using average summer Secchi based on measurements from 2005 to 2013. Robust standard errors are in parentheses. Robust standard errors are in parentheses
Stata code: ImputedSecchi0513.do.

	$Z_{Lit,Aug.}$	$Z_{DL,Aug.}$	$Z_{VSURF,Aug.}$	$Z_{DL+VSURF,Aug.}$
Lake area	-0.00001	0.00006***	0.00011***	0.00005***
	(0.00003)	(0.00001)	(0.00001)	(0.00001)
Avg. summer Secchi depth	0.12***	0.098**		0.109***
	(0.039)	(0.039)		(0.038)
Max. depth	0.021***	0.022***		0.021***
	(0.008)	(0.004)		(0.004)
30-yr avg. temp. in lake's subwatershed	0.00485			
	(0.08088)			
% of 500 m buffer in forest in 2011	0.00958***	-0.014***		0.004
	(0.00287)	(0.002)		(0.006)
% of 500 m buffer in wetlands in 2011	0.00852**	-0.019***		0.002
	(0.00356)	(0.003)		(0.005)
Boat feature count	0.29***			
	(0.071)			
Beach feature count	0.024			
	(0.127)			
Hotel feature count	0.701			
	(2.053)			
Shelter feature count	0.137			
	(0.240)			
Toilets feature count	0.195***			
	(0.031)			
Picnic feature count	-0.405			
	(1.051)			
BBQ feature count	-1.287	0.804		0.87
	(1.177)	(0.527)		(0.544)
Marina feature count	0.401**			
	(0.193)			
Spat. lag of avg. summer Secchi depth	0.050	-0.141		-0.064
	(0.238)	(0.155)		(0.163)
Spat. lag of boat feature count	3.267	25.194***		13.625***
	(4.360)	(5.345)		(4.766)
Spat. lag of beach feature count	-0.028	4.62**		-11.826*
	(6.412)	(2.065)		(6.099)
Spat. lag of hotel feature count	123.237*			
	(70.167)			
Spat. lag of shelter feature count	10.50		6.661	-10.997
	(9.464)		(7.716)	(16.768)
Spat. lag of toilets feature count	-4.155*	9.5***		3.301
	(2.242)	(2.765)		(2.087)

	$Z_{Lit,Aug.}$	$Z_{DL,Aug.}$	$Z_{VSURF,Aug.}$	$Z_{DL+VSURF,Aug.}$
Spat. lag of picnic feature count	40.802**			
	(16.338)			
Spat. lag of BBQ feature count	-5.507			
	(34.825)			
Spat. lag of marina feature count	-108.6***			
	(31.23)			
Avg. dist. to nearest 5 lakes w/ Secchi measure	0.00205	-0.02***		-0.003
	(0.00614)	(0.007)		(0.006)
Avg. size of nearest 5 lakes	-0.00077			
	(0.00074)			
Dist. to nearest CBSA	0.00003			
	(0.00003)			
Pop. den. in subwatershed	0.00004	0.00029***	0.00012	-0.00003
	(0.00019)	(0.00011)	(0.00015)	(0.00015)
% of 500 m buffer in dev. cover in 2011	0.024***		0.015***	0.017***
	(0.003)		(0.002)	(0.005)
Avg. dist. to nearest 5 lakes	0.139***	0.192***		0.182***
	(0.018)	(0.018)		(0.017)
Avg. dist. to closest CBSA of the nearest 5 lakes	$-3x10^{-5}$	$-1x10^{-5}$***		$-4x10^{-6}$**
	($3x10^{-5}$)	($1x10^{-6}$)		($2x10^{-6}$)
Lake count in subwatershed	-0.015***			
	(0.006)			
Lake area in subwatershed	0.04***			
	(0.006)			
% of pop. in lake's subwatershed with a BA, 2011-15	0.04***		0.037***	0.033***
	(0.005)		(0.006)	(0.006)
% of pop. in lake's subwatershed that is poor, 2011-15	-0.029***			
	(0.01)			
% of pop. in lake's subwatershed that is white, 2011-15	-0.0031			
	(0.00504)			
Median age in lake's subwatershed, 2011-2015	-0.00277			
	(0.00911)			
Median HH inc. in lake's subwatershed, 2011-2015	$-2\ x10^{-5}$***			
	(0)			
Latitude	-0.044			
	(0.09)			
Longitude	0.02*			
	(0.011)			
Low ag. Region	-0.424***			
	(0.159)			
High ag. Region	-0.42	1.062*		0.997*
	(0.339)	(0.609)		(0.513)
Spat. lag of summer PUD count	0.446***		0.286***	0.379***
	(0.126)		(0.021)	(0.085)
30-yr avg. precip. in subwatershed (mm)		0.002***		0.00138***
		(0.0005)		(0.00038)

	$Z_{Lit,Aug.}$	$Z_{DL,Aug.}$	$Z_{VSURF,Aug.}$	$Z_{DL+VSURF,Aug.}$
% of 500 m buffer in ag. in 2011		-0.038***		-0.017***
		(0.003)		(0.005)
Total stream density in subwatershed (m/m sq.)		-0.078**		-0.089***
		(0.032)		(0.03)
Avg. scrub-shrub wetland area in lake's subwatershed		-0.001		0.004
		(0.009)		(0.005)
Constant	1.687	-2.065***	-1.891***	-3.706***
	(4.005)	(0.557)	(0.176)	(0.819)
Observations	19,705	19,705	19,705	19,705

SI Table 6. Ten-cross fold validation of Poisson count model with $Z_{Lit,Limited}$, $Z_{DL,Limited}$, $Z_{VSURF,Limited}$, and $Z_{DL+VSURF,Limited}$ and using average summer Secchi based on measurements from 2005 to 2013. Each cell gives the RMSE of for the given fold. *Stata code:* ObservedSecchi0513.do.

Fold	$Z_{Lit,Limited}$	$Z_{DL,Limited}$	$Z_{VSURF,Limited}$	$Z_{DL+VSURF,Limited}$
1	28	25	2,835	28
2	11	28,300,000	2,467	7
3	916	13	11	10
4	13	12	23	19
5	18	23,986	49	1,771
6	26	808	24	12
7	136,905	17	9	15
8	17	11	21	11
9	434	29	14	256,329
10	11	16	12	56
Mean RMSE	**13,838**	**2,832,492**	**546**	**25,826**
SD of RMSE	41,023	8,489,172	1,056	76,836
Mean RMSE less largest RMSE	**163.9**	**2,768.5**	**292.0**	**214.2**
Mean RMSE less largest two RMSEs	**69.9**	**116.3**	**20.1**	**19.7**

SI Table 7. Ten-cross fold validation of Poisson count model with $Z_{Lit,Augmented}$, $Z_{DL,Augmented}$, $Z_{VSURF,Augmented}$, and $Z_{DL+VSURF,Augmented}$ and using average summer Secchi based on measurements from 2005 to 2013. Each cell gives the RMSE of for the given fold. For each model we conduct 20 cross fold validation analyses, one for each model estimated over a unique set of imputed average summer Secchi and maximum lake depth. The exception is the model estimated with $\mathbf{Z_{VSURF,Augmented}}$ given this covariate vector does not contain any imputed data. *Stata code:* ImputedSecchi0513.do.

Fold	$Z_{Lit,Aug.}$	$Z_{DL,Aug.}$	$Z_{VSURF,Aug.}$	$Z_{DL+VSURF,Aug.}$
Mean RMSE	**75,503,402**	**258**	**13.17**	**661**
SD of RMSE	*1,065,058,326*	*1,138*	*13.57*	*3,103*
Mean RMSE less largest RMSE	**204**	**12.65**	**9.16**	**15.99**
Mean RMSE less largest two RMSEs	**73.54**	**8.25**	**6.95**	**8.26**

SI Table 8. Variables in single (SL) and double LASSO (DL)-informed Z's over lakes with water recreation features. *R code:* LASSOWaterRec.R with the Limited dataset and LASSOWaterRecImputed1.R – LASSOWaterRecImputed20.R with the Augmented dataset.

	$Z_{SL,Limited}$	$Z_{DL,Limited}$	$Z_{SL,Augmented}$	$Z_{DL,Augmented}$
	Beach feature count	Maximum lake depth	Lake area	Lake area
		Average summer Secchi depth	Marina feature count	Maximum lake depth
		Beach feature count		Average summer Secchi depth
		Spatial lag of average summer Secchi depth		% of 500 m buffer around lake in forest in 2011
				% of 500 m buffer around lake in Agriculture in 2011
				Average forested wetland area in lake's subwatershed
				Marina feature count
				BBQ feature count
				Spatial lag of average summer Secchi depth
				Spatial lag of boat launch feature count
				Average distance to the nearest 5 lakes with Secchi measurements
				Average distance to the nearest 5 lakes
				Lake latitude
λ_{min}	46.58		9.363 (0.000)	
$\lambda_{min,DL,FS}$		46.58		9.36 (0.000)
$\lambda_{min,DL,SS}$		0.602		0.13 (0.041)
CVM	6348.73		583.65 (0.093)	
$CVM_{DL,FS}$		6348.73		583.65 (0.093)
$CVM_{DL,SS}$		3.84		3.41 (0.087)

Notes: CVM = Cross-validation mean. Subscript SL indicates single LASSO. Subscript DL,FS indicates double LASSO, first stage. Subscript DL,SS indicates double LASSO, second stage.

SI Table 9. Variables in VSURF-informed Z's over lakes with water recreation features. *R code:* VSURFWaterRec.R with the Limited dataset and VSURFWaterRecImputed1.R – VSURFWaterRecImputed20.R with the Augmented dataset.

$Z_{VSURF,Limited}$	$Z_{VSURF,Augmented}$
Beach feature count	Lake area
Lake area	Beach feature count
Toilet feature count	Toilets feature count
Marina feature count	% of lake's subwatershed area covered by lakes
	% of 500 m buffer around lake in developed cover in 2011
	Marina feature count
	Average summer Secchi depth
	Spatial lag of shelter feature count
	% of pop. in lake's subwatershed with a bachelor's degree or higher, 11-15

Notes: Selected variables in "Dataset without imputed data" column are listed in order of predictive importance. Selected variables in "Dataset with imputed data" column are listed in mean order of predictive importance over the 20 sets of selected variables.

SI Table 10. Estimated Poisson count model with $Z_{Lit,Limited}$, $Z_{DL,Limited}$, $Z_{VSURF,Limited}$, and $Z_{DL+VSURF,Limited}$ across lakes with water recreation features. Robust standard errors are in parentheses. *Stata code:* ObservedSecchiH2OAct.do.

	$Z_{Lit,Limited}$	$Z_{DL,Limited}$	$Z_{VSURF,Limited}$	$Z_{DL+VSURF,Limited}$
Lake area	0.000135*		0.000118***	0.000105***
	(8.14e-05)		(2.56e-05)	(2.37e-05)
Avg. summer Secchi depth	-0.169**	-0.409***		-0.340***
	(0.0832)	(0.0999)		(0.111)
Max. depth	0.0303***	0.0370***		0.0335***
	(0.00723)	(0.00746)		(0.00720)
Secchi depth sampled more than once	0.173			
	(0.244)			
30-yr avg. temp. in lake's subwatershed	0.0490			
	(0.170)			
% of 500 m buffer in forest in 2011	0.0137*			
	(0.00715)			
% of 500 m buffer in wetlands in 2011	0.00359			
	(0.0118)			
Boat feature count	0.109			
	(0.0678)			
Beach feature count	0.0509**	0.0424*	0.0502**	0.0416
	(0.0233)	(0.0250)	(0.0248)	(0.0254)
Hotel feature count	0.219			
	(0.713)			
Shelter feature count	0.370***			
	(0.140)			
Toilets feature count	0.342***		0.147***	0.148***
	(0.118)		(0.0442)	(0.0447)
Picnic feature count	0.0569			
	(0.324)			
BBQ feature count	-2.483***			
	(0.890)			
Marina feature count	0.150		0.647***	0.543***
	(0.174)		(0.123)	(0.136)
Spat. lag of avg. summer Secchi depth	1.353**	0.403		0.371
	(0.540)	(0.314)		(0.280)
Spatial lag of boat feature count	-9.853			
	(9.056)			
Spatial lag of beach feature count	14.50			
	(12.42)			
Spatial lag of hotel feature count	82.05			
	(105.0)			
Spatial lag of shelter feature count	121.8**			
	(47.92)			
Spatial lag of toilets feature count	-21.73			
	(14.55)			
Spatial lag of picnic feature count	59.23			
	(52.20)			
Spatial lag of BBQ feature count	-6.800			
	(60.68)			
Spatial lag of marina feature count	-49.85			
	(51.28)			
Avg. dist. to nearest 5 lakes w/ Secchi measure	0.0250**			
	(0.0127)			
Avg. size of nearest 5 lakes	0.000375			
	(0.000338)			

	$Z_{Lit,Limited}$	$Z_{DL,Limited}$	$Z_{VSURF,Limited}$	$Z_{DL+VSURF,Limited}$
Dist. to nearest CBSA	1.71e-05			
	(3.34e-05)			
Pop. den. in subwatershed	-0.000601*			
	(0.000322)			
% of 500 m buffer in dev. cover in 2011	0.0334***			
	(0.00677)			
Avg. dist. to nearest 5 lakes	0.0264			
	(0.0307)			
Avg. dist. to closest CBSA across the nearest 5 lakes	-1.48e-05			
	(3.47e-05)			
Lake count in subwatershed	-0.0191			
	(0.0129)			
Lake area in subwatershed	0.0181			
	(0.0170)			
% of pop. in lake's subwatershed with a BA or higher, 2011-15	0.0452***			
	(0.00806)			
% of pop. in lake's subwatershed that is poor, 2011-15	0.0320			
	(0.0217)			
% of pop. in lake's subwatershed that is white, 2011-15	-0.0125			
	(0.00874)			
Median age in lake's subwatershed, 2011-2015	0.0167			
	(0.0240)			
Median HH inc. in lake's subwatershed, 2011-2015	5.35e-06			
	(8.84e-06)			
Spat. lag of summer PUD count	0.0532			
	(0.136)			
Latitude	0.0322			
	(0.195)			
Longitude	-0.0431			
	(0.0370)			
Low ag. region	-0.521			
	(0.569)			
High ag. region	-0.491*			
	(0.267)			
Constant	-11.04	2.197***	2.440***	1.886***
	(8.943)	(0.814)	(0.158)	(0.711)
Observations	435	435	435	435

SI Table 11. Estimated Poisson count model with $Z_{Lit,Augmented}$, $Z_{DL,Augmented}$, $Z_{VSURF,Augmented}$, and $Z_{DL+VSURF,Augmented}$ across lakes with water recreation features. Robust standard errors are in parentheses. *Stata code:* ImputedSecchiH2OAct.do.

	$Z_{Lit,Aug.}$	$Z_{DL,Aug.}$	$Z_{VSURF,Aug.}$	$Z_{DL+VSURF,Aug.}$
Lake area	0.00001	0.00007***	0.00002*	0.00001
	(0.00002)	(0.00001)	(0.00001)	(0.00001)
Avg. summer Secchi depth	0.00113	0.075	0.068	-0.017
	(0.07026)	(0.072)	(0.047)	(0.074)
Max. depth	0.01216***	0.011***		0.009***
	(0.0038)	(0.003)		(0.002)
30-yr avg. temp. in lake's subwatershed	0.040			
	(0.115)			
% of 500 m buffer in forest in 2011	0.00908*	-0.019***		0.016***
	(0.00523)	(0.005)		(0.005)
% of 500 m buffer in wetlands in 2011	-0.00705			
	(0.00702)			
% of 500 m buffer in ag. in 2011		-0.032***		0.005
		(0.007)		(0.007)
Average forested wetland area in lake's subwatershed		-0.069**		-0.003
		(0.028)		(0.004)
Boat feature count	0.156***			
	(0.056)			
Beach feature count	0.113**		0.086***	0.088***
	(0.045)		(0.021)	(0.022)
Hotel feature count	-0.776			
	(0.628)			
Shelter feature count	0.144			
	(0.175)			
Toilets feature count	0.475***		0.53***	0.543***
	(0.064)		(0.075)	(0.076)
Picnic feature count	-0.07208			
	(0.12897)			
BBQ feature count	0.011	-0.003		0.448
	(0.324)	(0.322)		(0.354)
Marina feature count	0.24*	0.697***	0.552***	0.393***
	(0.126)	(0.221)	(0.106)	(0.11)
Spat. lag of avg. summer Secchi depth	0.539	-0.360		0.358
	(0.423)	(0.302)		(0.246)
Spat. lag of boat feature count	-25.652***	12.764		
	(7.288)	(9.045)		-3.05 (5.721)
Spat. lag of beach feature count	-3.912			
	(5.795)			
Spat. lag of hotel feature count	514.918**			
	(212.517)			
Spat. lag of shelter feature count	-22.408*		2.992	-2.927
	(11.758)		(11.369)	(10.648)
Spat. lag of toilets feature count	-4.386			
	(6.084)			
Spat. lag of picnic feature count	57.881**			
	(23.604)			
Spat. lag of BBQ feature count	15.244			
	(39.643)			
Spat. lag of marina feature count	-5.932			
	(29.195)			
Avg. dist. to nearest 5 lakes w/ Secchi measure	0.029***	0.016*		0.033***
	(0.008)	(0.009)		(0.008)

	$Z_{Lit,Aug.}$	$Z_{DL,Aug.}$	$Z_{VSURF,Aug.}$	$Z_{DL+VSURF,Aug.}$
Avg. size of nearest 5 lakes	-0.00028			
	(0.00046)			
Dist. to nearest CBSA	0.00005*			
	(0.00002)			
Pop. den. in subwatershed	-0.00009			
	(0.00024)			
% of 500 m buffer in dev. cover in 2011	0.018***		0.009***	0.025***
	(0.005)		(0.003)	(0.005)
Avg. dist. to nearest 5 lakes	-0.00036	-0.019		0.014
	(0.02353)	(0.047)		(0.023)
Avg. dist. to closest CBSA across the nearest 5 lakes	-0.00004*			
	(0.00003)			
Lake count in subwatershed	-0.00182			
	(0.00817)			
Lake area in subwatershed	0.028***		0.024***	0.03***
	(0.008)		(0.007)	(0.007)
% of pop. in lake's subwatershed with a BA or higher, 2011-15	0.043***		0.025***	0.028***
	(0.008)		(0.006)	(0.005)
% of pop. in lake's subwatershed that is poor, 2011-15	-0.029*			
	(0.015)			
% of pop. in lake's subwatershed that is white, 2011-15	-0.011			
	(0.007)			
Median age in lake's subwatershed, 2011-2015	0.002			
	(0.018)			
Median HH inc. in lake's subwatershed, 2011-2015	-0.00002***			
	(0.00001)			
	0.35***			
Spat. lag of summer PUD count	(0.093)			
Latitude	0.016	0.053		
	(0.124)	(0.050)		0.032 (0.04)
Longitude	-0.016			
	(0.021)			
Low ag. Region	-0.396			
	(0.279)			
High ag. Region	-0.195			
	(0.237)			
Constant	-2.387	1.342	0.815***	-3.064
	(6.258)	(2.352)	(0.238)	(2.005)
Observations	696	696	696	696

SI Table 12. Ten-cross fold validation analysis with $Z_{Lit,Limited}$, $Z_{DL,Limited}$, $Z_{VSURF,Limited}$, and $Z_{DL+VSURF,Limited}$ across lakes with water recreation features. Each cell gives the RMSE of for the given fold. *Stata code:* ObservedSecchiH2OAct.do.

Fold	$Z_{Lit,Limited}$	$Z_{DL,Limited}$	$Z_{VSURF,Limited}$	$Z_{DL+VSURF,Limited}$
1	18	75	47	14
2	65	13	27	14
3	51	97	77	41
4	46	10,300,000	24	98
5	189	41	105	98
6	11,808	30	54	37
7	31	18	19	8,388,530
8	105	40	15	120
9	252	43	2,780,000,000	39
10	27	18	23	16
Mean RMSE	1,259	1,030,038	278,000,039	838,901
SD of RMSE	3,517	3,089,987	833,999,987	2,516,543
Mean RMSE less largest RMSE	87.2	41.8	43.6	53.2
Mean RMSE less largest two RMSEs	66.5	34.9	36.0	44.7

SI Table 13. Ten-cross fold validation analysis with $Z_{Lit,Augmented}$, $Z_{DL,Augmented}$, $Z_{VSURF,Augmented}$, and $Z_{DL+VSURF,Augmented}$ across lakes with water recreation features. For each model we conduct 20 cross fold validation analyses, one for each model estimated over a unique set of imputed average summer Secchi and maximum lake depth. *Stata code:* ImputedSecchiH2OAct.do.

Fold	$Z_{Lit,Aug.}$	$Z_{DL,Aug.}$	$Z_{VSURF,Aug.}$	$Z_{DL+VSURF,Aug.}$
Mean RMSE	685	43,974	408	198
SD of RMSE	2,448	606,851	2,079	316
Mean RMSE less largest RMSE	189	145	77.3	113
Mean RMSE less largest two RMSEs	105	66.2	48.0	80.0

SI Table 14. Estimated IV Poisson count model with $Z_{Lit,Limited}$ across lakes with water recreation features. Robust standard errors are in parentheses. *Stata code*: ObservedSecchiH2OActIV.do

	$Z_{Lit,Limited}$
Avg. summer Secchi depth	0.177
	(2.386)
Lake area	0.00005
	(0.00016)
Max. depth	0.426
	(0.311)
Secchi depth sampled more than once	0.005
	(0.136)
30-yr avg. temp. in lake's subwatershed	0.023
	(0.093)
% of 500 m buffer in forest in 2011	0.009
	(0.018)
% of 500 m buffer in wetlands in 2011	0.002
	(0.038)
Boat feature count	0.21
	(0.324)
Beach feature count	0.238**
	(0.114)
Hotel feature count	0.567
	(2.616)
Shelter feature count	0.217
	(0.321)
Toilets feature count	0.115
	(0.211)
Picnic feature count	0.091
	(0.276)
BBQ feature count	-0.331
	(0.299)
Marina feature count	0.167
	(0.439)
Spat. lag of avg. summer Secchi depth	-0.121
	(4.865)
Spat. lag of boat feature count	-25.389**
	(12.271)
Spat. lag of beach feature count	-10.877
	(32.475)
Spat. lag of hotel feature count	548.929
	(644.874)
Spat. lag of shelter feature count	-34.572
	(27.373)
Spat. lag of toilets feature count	-3.055
	(9.289)
Spat. lag of picnic feature count	70.65
	(62.217)
Spat. lag of BBQ feature count	-55.465
	(142.254)

	$Z_{Lit,Limited}$
Spat. lag of marina feature count	41.975
	(282.767)
Avg. dist. to nearest 5 lakes w/ Secchi measure	0.025
	(0.048)
Avg. size of nearest 5 lakes	0.00021
	(0.00178)
Dist. to nearest CBSA	0.00001
	(0.00002)
Pop. den. in subwatershed	0.00016
	(0.00042)
% of 500 m buffer in dev. cover in 2011	0.019***
	(0.006)
Avg. dist. to nearest 5 lakes	-0.008
	(0.15)
Avg. dist. to closest CBSA across the nearest 5 lakes	-0.000003
	(0.000026)
Lake count in subwatershed	-0.004
	(0.023)
Lake area in subwatershed	0.045
	(0.032)
% of pop. in lake's subwatershed with a BA, 2011-15	0.028
	(0.025)
% of pop. in lake's subwatershed that is poor, 2011-15	0.013
	(0.037)
% of pop. in lake's subwatershed that is white, 2011-15	0.001
	(0.008)
Median age in lake's subwatershed, 2011-2015	-0.006
	(0.016)
Median HH inc. in lake's subwatershed, 2011-2015	0.000002
	(0.000031)
Spatial lag of summer PUD count	0.481
	(0.504)
Latitude	-0.044
	(0.137)
Longitude	0.002
	(0.079)
Low ag. Region	-0.484
	(1.699)
High ag. Region	-0.172
	(0.216)
Constant	0.395
	(5.841)
Observations	870

SI Text 1. List of all potential covariates

Variable name	Variable definition

<<To be added>>