Comparing estimates of fishing effort and lake choice derived from aerial creel surveys and smartphone application data in Ontario, Canada

A Thesis
SUBMITTED TO THE FACULTY OF
UNIVERSITY OF MINNESOTA
BY

Timothy James Martin

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

Advised by Paul A. Venturelli

September 2017

Acknowledgements

I thank the Great Lakes Fishery Commission for funding this study. The Minnesota Department of Natural Resources and the Water Resources Science graduate degree program also provided me with additional funding. The app companies Fishidy, C-Map, and iFish graciously provided me with the data used in this study. Len Hunt of the Ontario Ministry of Natural Resources and Forestry (OMNRF) provided me creel survey and lake attribute data I used. He was also integral in helping guide the direction of the study. Cindy Chu of the OMNRF also provided me with data and help in obtaining some of the geospatial data I used. I am grateful to my fellow Venturelli lab mates for helping me with coding and other project needs. I am also grateful to my committee members, Ray Newman and Nick Phelps for the guidance they provided me. I owe much to my friends and family, especially my parents, who helped and supported me throughout my graduate school experience despite living on the other side of the world. Finally, I thank Paul Venturelli for his direction and patience with me as I proceeded through the grad school journey.

Abstract

Anglers make decisions that have consequences for the fish stocks, ecosystems, and socio-economics with which they interact. Smartphone angling applications (apps), are a potentially less expensive and more comprehensive data source than conventional methods, but their utility has not been evaluated. In this study, I compared results from app and aerial creel survey data from Ontario, Canada. A standard major axis regression found low agreement between effort estimates (n=111, R^2=0.20, p=8.2458e-07) and app-based effort was poorly explained by lake characteristics in a random forest analysis (7.66% vs. 29.52% for creels). Explained variation improved when I included more lakes, but province-wide effort prediction did not agree with those based on creel data. I attribute these inconsistent results to low app data volumes and inherent differences between collection and analyses. Until more app data are generated, I recommend using app data to supplement conventional surveys and gain novel insights into angler behavior.

Table of Contents

List of Tables

List of Figures

Introduction

Understanding recreational angler behavior is an important aspect of fisheries management and research. The choices that anglers make can impact the natural systems with which they interact (Hunt 2005). For example, angler effort has been link to resource over-exploitation and the spread of aquatic invasive species and diseases (De Kerckhove et al. 2015, Drake & Mandrak 2010, Goodwin et al. 2004). Angler effort has also been show to affect regional economies. The presence of angling and angling-related activities can attract significant amounts of recreational tourism to an area (Lew & Seung 2010, Kauppila & Karjalainen 2012).

Angler effort can be predicted from models that link angler site choice to waterbody features. Hunt (2005) reviewed 47 published studies and found that anglers chose fishing locations based on costs (e.g., travel distance, lodging expenses, ease of access), fishing quality (e.g., fish density), environmental quality (e.g., water quality, surrounding scenery), facility development (e.g., access sites), encounters with other anglers, and fishing regulations. Recent examples include Kaufmann et al. (2008), who highlighted the importance of surface area, access, and travel times in 589 Lake trout lakes in Northeastern Ontario and Mkwara and Marsh (2011) who found that water clarity, lake size, fish weight, facilities and forest cover were important in 11 lakes in New Zealand.

Most site choice models are specific to a space and time because obtaining the broad-scale data required to conduct wide-ranging studies is both expensive and labor intensive (National Research Council 2006). The surveys upon which many of these models rely can also be biased by intentional or accidental misreporting by anglers (Sullivan 2003, Mallison & Cichra 2004). Digital technologies such as online fishing fora, internet volume search tools, geo-tagged digital photographs and smartphone apps are potential alternatives to conventional surveys. These technologies allow anglers to record their own activity, which can reduce the costs associated with data collection (especially over wide spatial scales), and reduce recall bias (Venturelli et al. 2017).

Several studies have demonstrated the ability of these technologies to provide fisheries data. For example, Martin et al. (2014) found that the number of posts on an online fishing forum for reservoirs in Nebraska closely mirrored creel based estimates of fishing effort. Shiffman et al. (2017) also used online fishing fora to gauge how anglers perceive and respond to shark-fishing regulations. Carter et al. (2014) used an internet search volume tool to improve catch predictions for the Gulf of Mexico Red snapper fishery, and Keeler et al. (2015) analyzed geo-tagged photographs from a photo sharing website and found that lake size, clarity, near-lake population, and the presence of a boat ramp predicted the recreational use of 1,000 lakes in Minnesota and Iowa.

Smartphone angling apps are particularly promising because anglers can record a diversity of data in ‘real time.’ There are already dozens of consumer-level apps available to anglers. Although few of these apps were designed to collect scientific data, they collect information at geographic and temporal scales that are beyond the ability of traditional means. Stunze et al. (2014) analyzed app data from a deep-sea charter boat fishery in the Gulf of Mexico and found that they compared favorably to traditional creel surveys conducted at the same time. Papenfuss et al. (2015) found that app data accurately predicted relative angler effort among ten fishery management zones in Alberta, Canada, and also showed that app data provided insight into angler behavior (e.g., seasonal trends, network connectivity). Finally, Jiorle et al. (2016) used data from an app to estimate catch rates that were statistically similar to an access-point creel survey in Florida.

In this study, I determine if app effort data are related to creel effort data, if these two sources of angler effort are influenced by similar lake attributes, and if they predict similar effort trends province-wide in Ontario, Canada. App effort data are relatively inexpensive and have few limits in terms of time and space; therefore, they have the potential to increase the efficiency, timeliness, and spatial coverage of fisheries research and management. It is important to demonstrate the utility of app data by comparing them to more established and accepted sources to determine how this new data source can fit within the spectrum of available fisheries data tools.

Methods

Aerial Creel Survey Data

I obtained aerial creel survey data from the Ontario Ministry of Natural Resources and Forestry (OMNRF), which were collected as a part of the Broad Scale Monitoring (BsM) program for both open water and ice fishing seasons between the years 2010 and 2016. This program focused on inland lakes and did not include rivers or the Great Lakes. Additionally, it did not include some of the larger, popular lakes (e.g., Lake Simcoe, Lake of the Woods, Lake St. Claire) in the province because these lakes are monitored separately. Aerial creel surveys involved mid-day flights over the lake of interest, counting the number of fishing vessels, shore anglers, ice huts, and open ice anglers present. These data ranged between one and 15 flights per day type during each season. To ensure that I used representative samples for each lake, I set a minimum sample size of five flights per lake per season during the summer and four flights per lake per season during the winter. At least one flight had to occur on a weekend and at least two flights had to occur on a weekday. I calculated angler effort hours as

$$\mathbf{(((MAC_V*p_V)+MAC_S)*T*D)/K,} \qquad (1)$$

where MAC_V is the mean number of vessels in the summer or huts in the winter counted on the waterbody at mid-day, p_V refers to the party size (the mean number of anglers in summer, assumed to be two in winter), MAC_S is the number of shoreline or open ice anglers, T is the number of hours during the day (14 for summer and 10 for winter), K is a season- and day-specific corrective factor (1.1, 1.3, 1.2, and 1.5 for summer weekends, summer weekdays, winter weekends, and winter weekdays, respectively), and D is the number of days in summer (35 weekend and holiday days and 74 weekday days from the third Saturday in May - the start of walleye season- to Labor Day) or winter (16 weekend days and 43 weekday days from January 15 to March 15). I used this formula to calculate effort hours for each day type, season, and lake, and then summed for each lake to generate an annual estimate of effort hours (Kaufman et al. 2008). I averaged annual effort over years if a lake was surveyed more than one year.

App Data

I combined angler activity and catch data from three consumer angler app companies into a single dataset. The iFish Ontario and Fishidy angler apps were both “typical” in that they allowed users to log their catches and other items of interest such as favorite fishing spots or waterbody hazards. The C-Map Genesis app was a mapping companion that anglers use when operating their fish-finding boat sonar.

Each app company provided user-sourced input data that identified the angler (through an anonymous user id), the date, and the lake for every data point collected. I ensured that each data point was associated with a lake centroid, sourced either from the app company itself, or, if those were not available, from centroid coordinates calculated from a lake polygon geospatial dataset that combined data from the Ontario Hydrological Network (OHN-Waterbody 2011), Minnesota Department of Natural Resources Hydrography (DNR Hydrological Dataset 2012), and National Hydrography Dataset (USGS TNM Hydrography (NHD) 2017). I removed data points if I could not identify which lake they were associated with, the GPS/user-input coordinates were not near a lake and there was no associated lake id, or a date was abnormal (e.g., before the app was created or on a day in the future). The iFish Ontario dataset had a default user id for guest users (representing <6% of all trips in the combined dataset) which I considered to be a single user to ensure that I did not over-count trips. Because I was interested in how angler effort was distributed among lakes, I simplified the data by converting them to angler “trips”. I defined a “trip” (the unit of app angler effort) as one angler visiting one lake on one day. I omitted trips from rivers, the Great Lakes, and the large, popular lakes because these features were not included in the aerial creel survey data. I limited the data to the years 2014-2016 because the data volumes levelled off during these years and omitted any data that fell outside of the aerial creel survey seasons.

Lake Attribute Data

I obtained lake attributes (or likely correlates thereof) based upon the attributes that have been found to be important in previous studies (e.g., Hunt 2005, Kaufmann et. al 2008, Mkwara & Marsh 2011, Keeler et al. 2015) and could be obtained on a broad-scale, whether from previous sampling efforts or geospatial analyses (Table 1). I obtained depth data (maximum, mean, and Secchi) from the OMNRF's BsM program (Sandstrom 2013) and Aquatic Health Index (Dodge 1987). I favored BsM data when Index data were also available because the former were more recent.

I calculated a fishing regulation metric by scaling the standard, species-specific 2017 bag limits for fish that could be expected in lakes (i.e., omitted most salmonids) in each Fisheries Management Zone (FMZ) in the province from 0 to 1, with 0 indicating conservative catch limits and 1 indicating more liberal catch limits (OMNRF 2016), using the formula:

$$\mathbf{(x\text{-}\min(x)/\max(x)\text{-}\min(x)),} \qquad (2)$$

where x is the value to be scaled and min(x) and max(x) are the minimum and maximum values in the data. If a bag limit was unlimited for a given fish type, I set it to the largest bag limit for that fish type across the other FMZs. I then averaged scores across all fish types within an FMZ and applied these averages to all lakes within the corresponding FMZ.

I used the combined lake polygon geospatial data set to calculate spatial attributes in ArcMap 10.3.1. (ESRI 2015). I calculated lake area (m^2) using the built-in geometry calculator. The Shoreline Development Index (SDI) is a metric that increases from 1 as shoreline irregularity increases from a perfect circle. I calculated SDI as:

$$\mathbf{P/(2*sqrt(\pi*A)}, \qquad (3)$$

where P is the lake perimeter (m) and A is the lake area (m^2) (Cole & Weihe 2016). For the highway distance metric, I determined the distance (m) from the edge of each lake to the nearest highway, using highway lines extracted from a roads dataset (Road Network 2011). I derived fishing access site counts for each lake from an access site point geospatial data (Stuart 2015).

To gauge the “scenic” value of a lake, I calculated the percent of surrounding natural land cover using data from the North American Land Change Monitoring System (NALCMS), a 250-meter resolution continent-wide land cover raster dataset (2005 North American Land Cover). I calculated a 500-meter buffer around each lake using the Canada Albers Equal Area projected coordinate system and used the corresponding values to classify the underlying land cover types as natural or unnatural (Table A1).

I used a modified version of a proximity metric developed by Hunt and Lester (2009) to determine the effect of human populations on effort. This metric utilizes the number of households in a community and the distance from these communities to a destination to calculate the influence that the distance and size of these communities have on the destination. I included origins from outside of Ontario to minimize boundary effects, but placed penalties on all locations based on the proportion of anglers in Ontario that originate from the province itself, or a surrounding jurisdiction (OMNRF 2014). Origins within Ontario, the United States, and the surrounding Canadian provinces were penalized by multiplying the original values by 0.849, 0.133, and 0.018 respectively. For origins, I used Canadian forward sortation administrative unit centroids (517 within Ontario and 469 within Manitoba and Quebec) (Forward Sortation Areas 2011), and US zip code centroids that fell within 200 kilometers of the Canada-US border (722 within Minnesota, Michigan, Ohio, Pennsylvania and New York) (Minnesota Population Center 2016). To minimize the processing time required to calculate this metric I created a province-wide inverse distance weighted (IDW) raster with a cell size of 6102.2 to calculate the metric for 9,971 destination points evenly spaced along a geospatial grid. I then assigned metric values for individual lakes based upon the corresponding raster values at the lake centroid.

To estimate how fish abundance affected angler effort, I obtained catch per unit effort (CPUE) in kg per net of fish data for walleye and all fish for 91 of the BsM lakes with effort data. These CPUE data were based on catches made by overnight sets of standard North American or Ontario small mesh gill nets during the summer (Sandstrom 2013).

Analysis

Comparisons using lakes with both app and creel data

I used standard major axis regression via the smatr package in R (Warton et al. 2012, R Core Team 2017) to directly compare effort estimates among lakes for which both app and creel data were available. I transformed these effort estimates by taking the natural log of both sources. I also created two random forest models (Breiman 2001) using the randomForest package in R (Liaw &Wiener 2002) -one for each effort data source- to compare how they related to the lake attributes. The random forest algorithm uses machine learning to generate a regression tree for each *n* random subset of data and then uses these trees to produce an average tree (i.e., forest) that also estimates the relative importance of variables to response data. This type of analysis does not require data to be continuous or normally distributed, and therefore, allowed me to evaluate all lake attributes simultaneously. I used the default settings of the randomForest package, but increased the number of trees grown from 500 to 1,000.

Comparisons using lakes with either app or creel data

I generated a second pair of random forest models that used all lakes for which all app or creel data were available. This approach increased the number of lakes in the analysis (accordingly, I increased the number of trees grown to 5,000), but reduced the extent to which the models were based on the same lakes. My logic in adopting this approach was that a researcher or manager was more likely to base analyses on one data type than the subset of lakes for which both data types were available. I compared the results of these two random forest models, as well as the effort that these models predicted when applied to all lakes for which complete attributes were available. To ensure that I did not make predictions beyond the extents of my explanatory variables, I only predicted effort for lakes with attributes that fell inside the spatial extent of the app data and within 10% above and below the attribute ranges of each respective data source (Table 2). I used a linear regression to diagnose the relationship between the predicted and actual data values for both sources and used a t-statistic to check whether the slope of the regression indicated that there was a one to one relationship between the predicted and actual values. I used formula (2) to rescale the predicted app trips and creel hours data from 0-1, so that I could directly compare the two sources.

I used the scaled results to run a kernel density analysis on both data sources in ArcGIS with an output cell size of 5152.25, density units in km^2, output values to densities, and used the planar method to determine where in the province each source predicted the highest density of angler effort. I conducted a raster calculation to determine the relative percent difference from the app to the creel data using the formula:

$$\mathbf{(App-Creel)/[(App+Creel)/2]*100.} \quad (4)$$

According to this formula, 0% indicates perfect agreement, and positive and negative percentages indicate that the app data are over- and under-estimating effort relative to creel data, respectively. Because kernel density analyses take into account both the degree of effort at lakes and the proximity of lakes to each other, I ran an IDW spatial interpolation, which did not lake proximity into consideration. I used all the default settings with a cell size of 5152.25 and limited the resulting rasters to display the top decile of the scaled prediction data, thereby indicating which lakes had the highest predicted effort.

Results

Comparisons using lakes with both app and data

The dataset that I used to calculate the standard major regression comprised 694 app trips and 3,988,684 creel effort hours for 111 lakes (Figure 1). The standard major axis regression of the natural logs of the two data sources was positive and significant (p=8.2458e-07) but only explained 20% of the variation (Figure 2). The dataset that I used to directly compare the variable imporance for app and creel lake effort was limited to the 91 lakes with CPUE data (Figure 1), comprised 670 app trips and 3,931,661 creel effort hours. The random forest models explained 7.66% and 29.52% of variation in the trips (app data) and effort hours (creel data), respectively. Lake area was the most important variable for both data sources. The remaining variables differed in their degree of importance, though access sites, SDI, and max depth were among the top-five most important variables for both sources (Figure 3). Partial dependence plots show similar trends between the two data sources for many of the variables (Figure 4). For example, effort increased with lake area, SDI, and access sites, but was somewhat U-shaped for maximum depths.

Comparisons using lakes with either app or creel data

The combined app dataset comprised 1,739 trips to 368 lakes and the combined aerial creel survey dataset comprised 6,501,355 effort hours at 559 lakes in both the summer and winter seasons between the years 2010 and 2016 (Figure 5). The random forest models derived from all available data explained 28.62% and 45.93% of variation in trip numbers and effort hours respectively. Consistent with the direct comparison, lake area was the most important attribute for both models. Only maximum depth and the regulation metric were also common among the top-five most important attributes for both sources (Figure 6). The partial dependence plots showed less consistent variable trends between sources than they did in the previous analysis (Figure 7). Effort increased with lake area, but decreased with maximum depth for the app-based model and was U-shaped for the creel-based model. The slope of the relationship between predicted and observed effort for both data sources was positive (Figures 8 & 9), but significantly different from 1 for both the app (t: -32.15502, df :366, p: 1.158754e-108) and creel (t: -32.7252, df: 557, p: 7.92152e-132), suggesting that a correction factor may be necessary to accurately predict effort for either source (Table 3A).

Models derived from app and creel data predicted different patterns of effort across Ontario. Relative to the creel-based model, the app-based model predicted higher activity in rural areas and lower activity closer to urban centers and in the southern, more populous part of the province (Figure 10). The top deciles of the IDW spatial interpolation rasters also indicated that the creel-based model predicted high effort for most of southern Ontario (Figure 11). The app predicted more isolated areas of high effort in this same region, but predicted more activity in the western part of the province than the creel.

Discussion

I found that analyses using current app data did not generate the same results as creel data and that it would not be appropriate to use current app data in place of aerial creel data when estimating or predicting effort. These results might stem from the weak correlation between app trips and creel effort hours. Regardless of the strength of this correlation, I observed similar effort-attribute relationships between pairs of random forest models. These results suggest that we can use app data to gain insight into the way that effort changes with most individual lake attributes. However, app- and creel-based random forest models produced different predictions of effort across all lake attributes. The fact that app data over-predicted effort in rural areas and under-predicted effort in urban areas was surprising given that smartphone usage is highest in more populated areas (“Mobile Fact Sheet” 2017). This may be due to high app usage by out-of-town anglers in the northwestern part of the province.

I hypothesize that the relatively low performance of the app data was largely a result of low sample sizes. Most notably, 43% of the lakes for which we had app data were only represented by a single trip (Figure 1A). This issue was exacerbated by the fact that I limited the seasonal extent of the app data, only used data from 2014-2016, and was unable to include the most popular lakes in the analysis. The net result of these limits was ≈61 and 56% reductions in the number of trips and lakes that were available for analysis. When I ran the random forest without these restrictions, the percent of variation explained increased and four of the top-five variables were the same (Figures 2A & 3A). Relative percent difference and IDW rasters showed more agreement between the app and creel data than the analyses with restricted app data (Figures 4A & 5A). Additionally, leaving out the most popular lakes likely decreased performance. For example, Papenfuss et al. 2015 found there was a significant relationship between summer app and creel data for 36 of the most popular lakes in Alberta, Canada (R^2: 0.74 vs. R^2:0.2 in this study). As the volume of app data continues to grow, the percentage of variation explained is likely to improve (especially as the number of lakes with single app visits decreases) and may converge on results from aerial creels.

My assessment of app-based results assumed that aerial creel survey results were accurate, which is unlikely to be the case. For example, aerial creel surveys cannot be conducted during inclement weather (Lockwood & Rakoczy 2005); therefore, these surveys will not record if angling effort is lower at these times (Malvestuto et al. 1979). Creel surveys will also generate inaccurate estimates of effort if samples are not representative or there are flaws in the algorithm that converts samples to annual estimates.

Until more app data are available for Ontario, it may be best to conduct analyses at different scales, use the data to supplement existing data, or to generate novel insights. An option for re-scaling includes limiting the analysis to southern Ontario (where sample sizes are larger). Another option is to analyze the data at a smaller scale (e.g., aggregated by FMZ). For example, Jiorle et al. (2016) addressed the issue of low sample sizes by comparing app and creel data at the county level. App data also collect data year-round, and so could be used to reduce aerial survey costs by helping designers to plan when flights should occur (Béliveau et al. 2015) or to improve algorithms for extrapolating survey results to an entire season or year. Other tools such as webcams, motion sensors, and traffic counters have been successfully employed for these purposes (Hartill et al. 2016, Steffo et al. 2008), but app data can be obtained from more locations and are likely to be cheaper (certainly if they are obtained from an existing company). App data can also reveal diel trends of angler effort that can be used to more accurately predict angler effort throughout the day. Diogo and Pereira (2016) found that 17% of the fishing activity in a mixed marine recreational fishery occurred at night.

Angler apps collect data that cannot be obtained through aerial creel surveys. Most of these apps allow users to input the fish species captured, making them similar to angler diaries, in that both are economically feasible and rely on anglers themselves to self-report (Cooke et al. 2000, Bray & Schramm 2001, Jiorle et al. 2016). Because aerial creel surveys are not capable of recording catch data, lakes with larger app data activity could provide insight into catch effort for specific species. App data may also be able to help predict the spread of aquatic invasive species. Davis et al. (2017) determined that fishing activity was the most important predictor of aquatic invasive species. App data could be used to identify where hotspots of angler activity are (and therefore propagule pressure) and how these vary over time.

I note several ways in which app companies can increase the value of their data for fisheries professionals. At a minimum, apps should record the type of data point (e.g., catch, waterbody feature, etc.), unique user id, date and time, location, waterbody id, and waterbody centroid coordinates. These inputs should be complete, through manual or automatic means, before a point can be logged. All data points should also be associated with a waterbody that has centroid coordinates. These criteria will eliminate the need for analysts to interpret which lake a data point is associated with and will diminish the need to eliminate points due to incomplete data. Additionally, most of the apps do not collect user demographic data; information that is useful for understanding the extent to which app users represent the overall angler population, and for filtering or subsampling. For example, Jiorle et al. (2016) found that app data over-represented avid anglers. Ontario anglers are middle aged and primarily originate from the more populous Southern Ontario region (OMNRF 2014), while smartphone users tend to be younger and from either urban or suburban areas (“Mobile Fact Sheet” 2017). It would be useful to know how these two populations intersect to understand the demographic biases inherent in the app data.

Angler apps can be a timely, broad-scale, and inexpensive source of data for monitoring and understanding angler activity. However, my results suggested that the app data that are currently available for Ontario generate results that are largely inconsistent with the results of aerial creel surveys. Although it is unreasonable to expect all anglers to always use an app, or creel-based estimates of effort to be perfectly accurate, it is reasonable to assume that consistent results can be obtained for some degree of app use and survey accuracy.

I encourage further research to realize this ideal, and efforts to identify ways that app data can benefit fisheries management and research (e.g., in support of conventional methods and for novel insight such as angler movement amongst lakes). It is also important to remember that the collection and analysis of app data are in the early stages of development and that the amount of data and our understanding of how to use it are still growing. Once there is a greater volume of app data, this comparison between app and creel data should be revisited to see if there is a stronger relationship between the two sources.

Table 1. The lake attributes used in the random forest models for app and creel data in Ontario, Canada. These attributes were calculated using statistical and geospatial tools or obtained from lake survey data.

Name (units)	Definition	Notes	Sources
Natural Log Area ($\ln(m^2)$)	Natural Log of Lake surface area	Canada Albers Equal Area Conic projected coordinate system.	NA
Shoreline Development Index (SDI)	Measurement of the irregularity of a lakes shoreline	Perimeter/(2*sqrt(π*Area) 1 – circular lake >1 – increasingly irregular shoreline	Florida LAKEWATCH 2001
Surrounding Land Cover (%)	Percentage of LC that can be considered natural	LC dataset buffered 500 meters around each lake. Canada Albers Equal Area Conic projected coordinate system	2005 North American Land Cover
Highway Distance (m)	Distance of the edge of the lake to the nearest highway.	North America Equidistant Conic projected coordinate system.	Road Network 2011
Fishing Access sites	Number of fishing access sites		Stuart 2015
Proximity Metric	Lake accessibility to human populations	Calculates the influence that total community sizes and distances have on a destination	Hunt & Lester 2009
Secchi Depth (m)	Water clarity depth		Sandstrom 2013, Dodge 1987
Maximum Depth (m)	Measurement of the deepest part of the lake		Sandstrom 2013, Dodge 1987
Mean Depth (m)	Measurement of the average depth of the lake		Sandstrom 2013, Dodge 1987
Regulation Metric (m)	Calculation of how restrictive regulations are within a fisheries management zone	The average of scaled bag limit values for FMZs where 0 is the lowest bag limit and 1 is the highest.	OMNRF 2016
Walleye and All Fish CPUE	Measurement of fish density	Data for 91 Lakes with both app and creel data	Sandstrom 2013

Table 2. The ranges, means, and standard deviations of the attributes used in the random forest models and predictions of app and creel data in Ontario, Canada.

Direct Comparison Lakes (n=91)			
	Ranges	**Mean**	**Standard Deviation**
Natural Log Area (ln(m^2))	13.29658 - 19.31746	16.73944	1.151835
SDI	1.598631 - 20.12087	5.839508	3.39003
Surrounding Land Cover (%)	0.4 - 1	0.9666443	0.08835472
Highway Distance (m)	0 – 25,760.86	3812.53	5532.614
Access Sites	0 - 19	1.505495	2.746204
Proximity Metric	40.92354 – 17,984.1	5,433.514	4,280.666
Maximum Depth (m)	4.3 - 125.95	31.9612	24.74278
Mean Depth (m)	1.3 - 38.4	8.645055	6.878182
Secchi Depth (m)	0 - 8.65	3.495055	1.691679
Regulation Metric	0.2916667 – 0.8125	0.5724977	0.1663449
All Fish CPUE	0.94 - 21.64	4.807473	4.016194
Walleye CPUE	0.02 - 12.65	1.385286	2.158784

App Data Lakes (n=368)			
	Ranges	**Mean**	**Standard Deviation**
Natural Log Area (ln(m^2))	10.18432 - 20.86637	15.36584	1.96803
SDI	1.097631 - 28.45331	4.582843	3.510374
Surrounding Land Cover (%)	0 - 1	0.9544572	0.1450783
Highway Distance (m)	0 – 170,798.3	5,987.635	12,779.55
Access Sites	0 - 19	0.7038043	1.669338
Proximity Metric	40.92354 – 24,953.11	5,670.386	4,396.615
Maximum Depth (m)	0 – 213.5	26.54051	22.1187
Mean Depth (m)	0 – 38.4	7.785598	5.900042
Secchi Depth (m)	0 - 9.2	3.495411	1.800584
Regulation Metric	0.125 - 0.8125	0.559265	0.1447697

Table 2. (Continued).

Creel Data Lakes (n=559)			
	Ranges	**Mean**	**Standard Deviation**
Natural Log Area (ln(m^2))	12.25604 - 20.7453	15.77345	1.469262
SDI	1.212671 - 20.36005	4.815797	3.105693
Surrounding Land Cover (%)	0.4 - 1	0.9910589	0.04873859
Highway Distance (m)	0 - 81236.23	10,193.32	12,502.19
Access Sites	0 - 19	0.5241503	1.415116
Proximity Metric	31.2838 – 17,984.1	4,221.684	3,648.72
Maximum Depth (m)	1.4 - 186.1	32.9224	24.3652
Mean Depth (m)	0 - 40.1	9.71127	7.267453
Secchi Depth (m)	0 – 12.5	3.888602	1.926745
Regulation Metric	0.2916667 - 0.8125	0.506349	0.1737366

Prediction Data Lakes (n=9,451)			
	Ranges	**Mean**	**Standard Deviation**
Natural Log Area (ln(m^2))	6.535383 - 20.86637	13.23593	1.784734
SDI	1.007921 - 28.45331	2.687534	1.746887
Surrounding Cover (%)	0 - 1	0.9868511	0.08588746
Highway Distance (m)	0 – 170,798.3	11,772.25	14,467.15
Access Sites	0 - 19	0.1285578	0.5042468
Proximity Metric	29.4676 – 24,953.11	4,765.816	3,860.098
Maximum Depth (m)	0 – 213.5	17.00168	14.27345
Mean Depth (m)	0 – 47.5	5.557422	4.466175
Secchi Depth (m)	0 – 22.65	3.597981	1.971863
Regulation Metric	0.125 - 0.8125	0.5497632	0.1539463

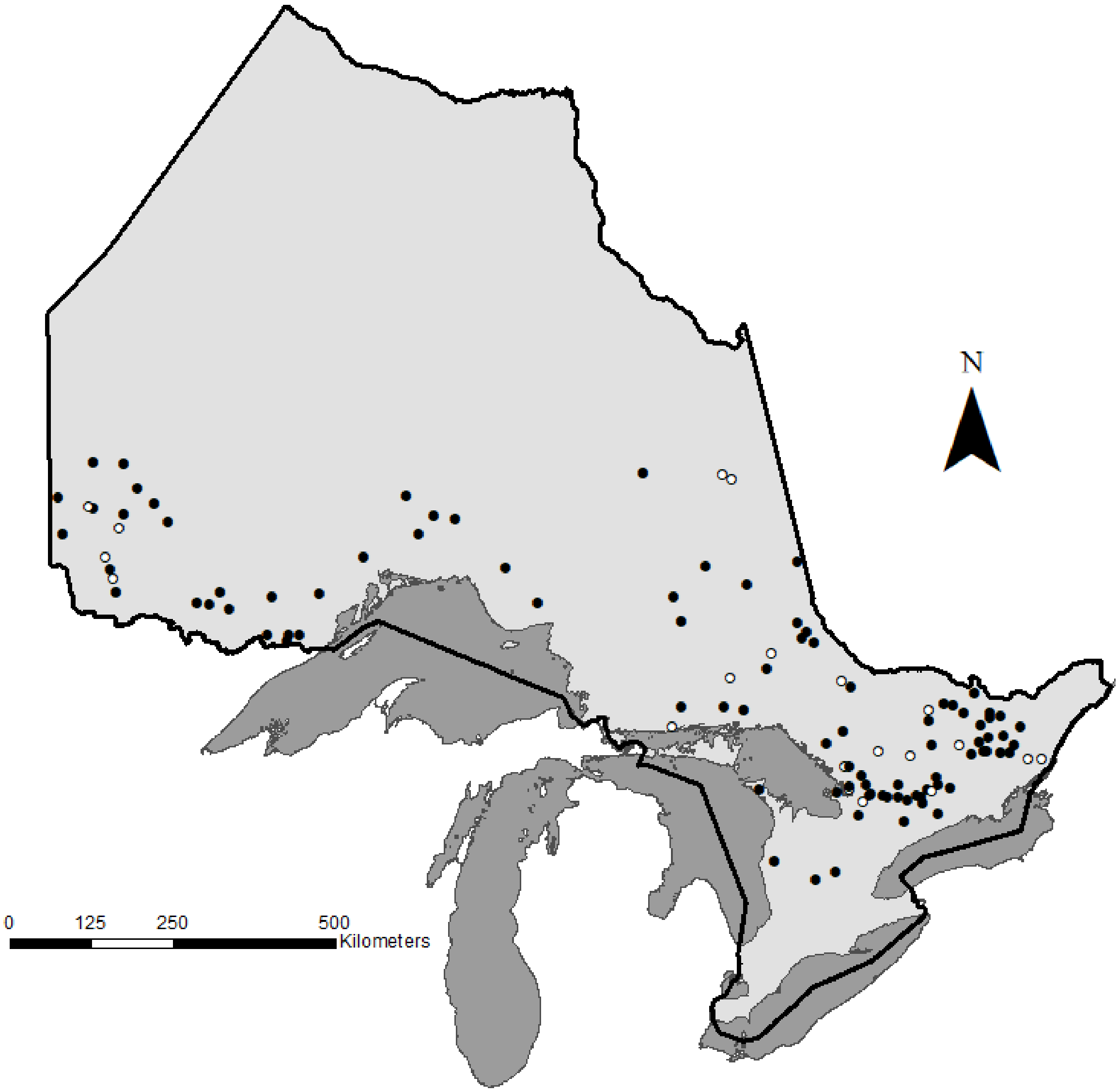

Figure 1. The locations of lakes with both app and creel data in Ontario, Canada. All points represent lakes (n=111) that were used in the standard major axis regression analysis, and black points represent lakes (n=91) that were used in the random forest models.

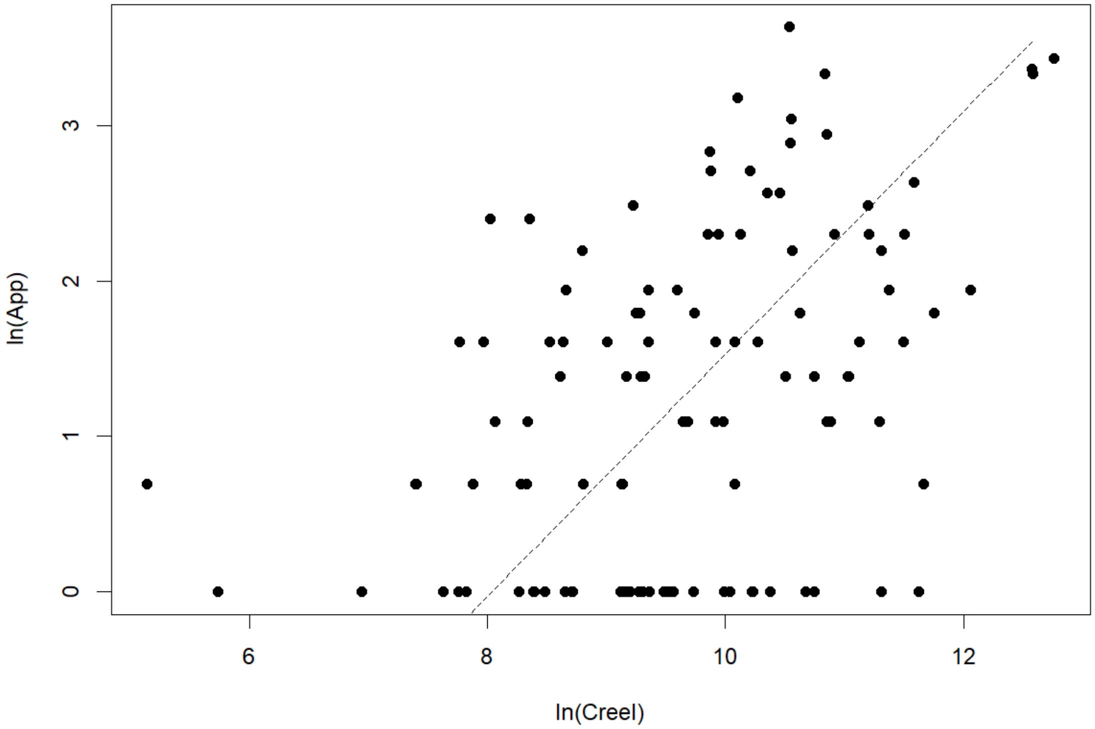

Figure 2. Standard major axis regression for the natural log of app trips and creel survey effort hours for Ontario lakes with data from both app and creel data (n=111, R^2=0.20, p=8.2458e-7).

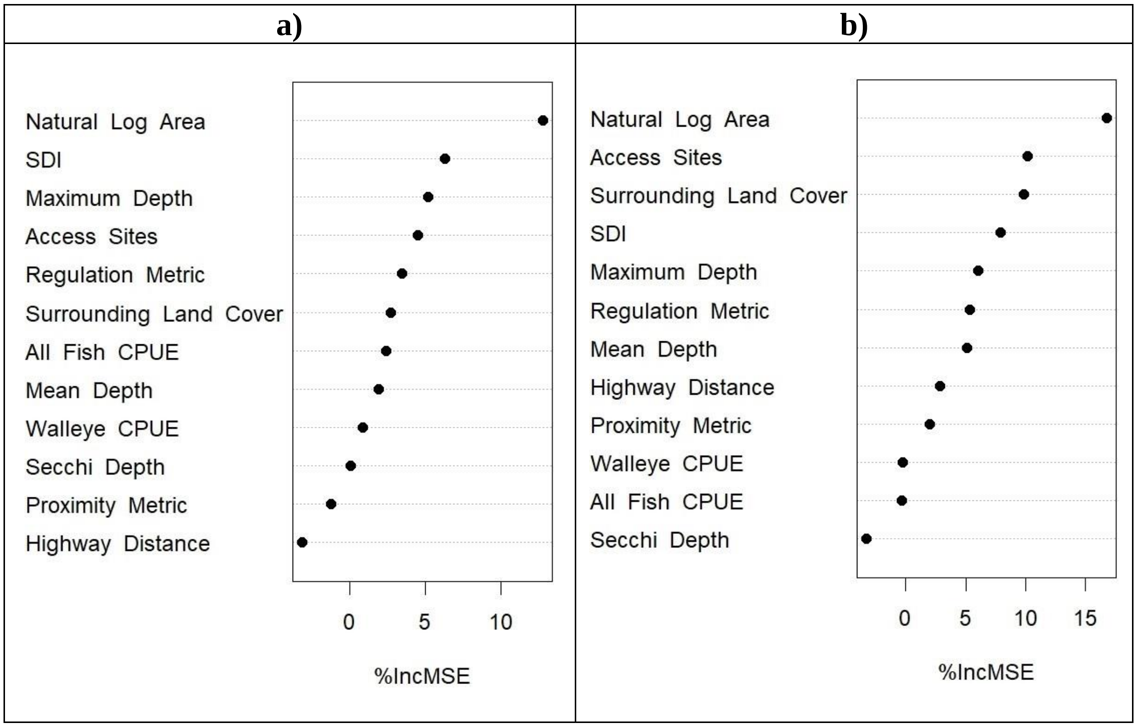

Figure 3. Random forest variable importance plot for a) app trip data and b) creel survey effort hour data with the twelve variables analyzed for the 91 lakes with both app and creel survey data in Ontario, Canada. %IncMSE is the percent increase in the mean square error. This metric refers to the mean decrease in model accuracy if a variable is removed.

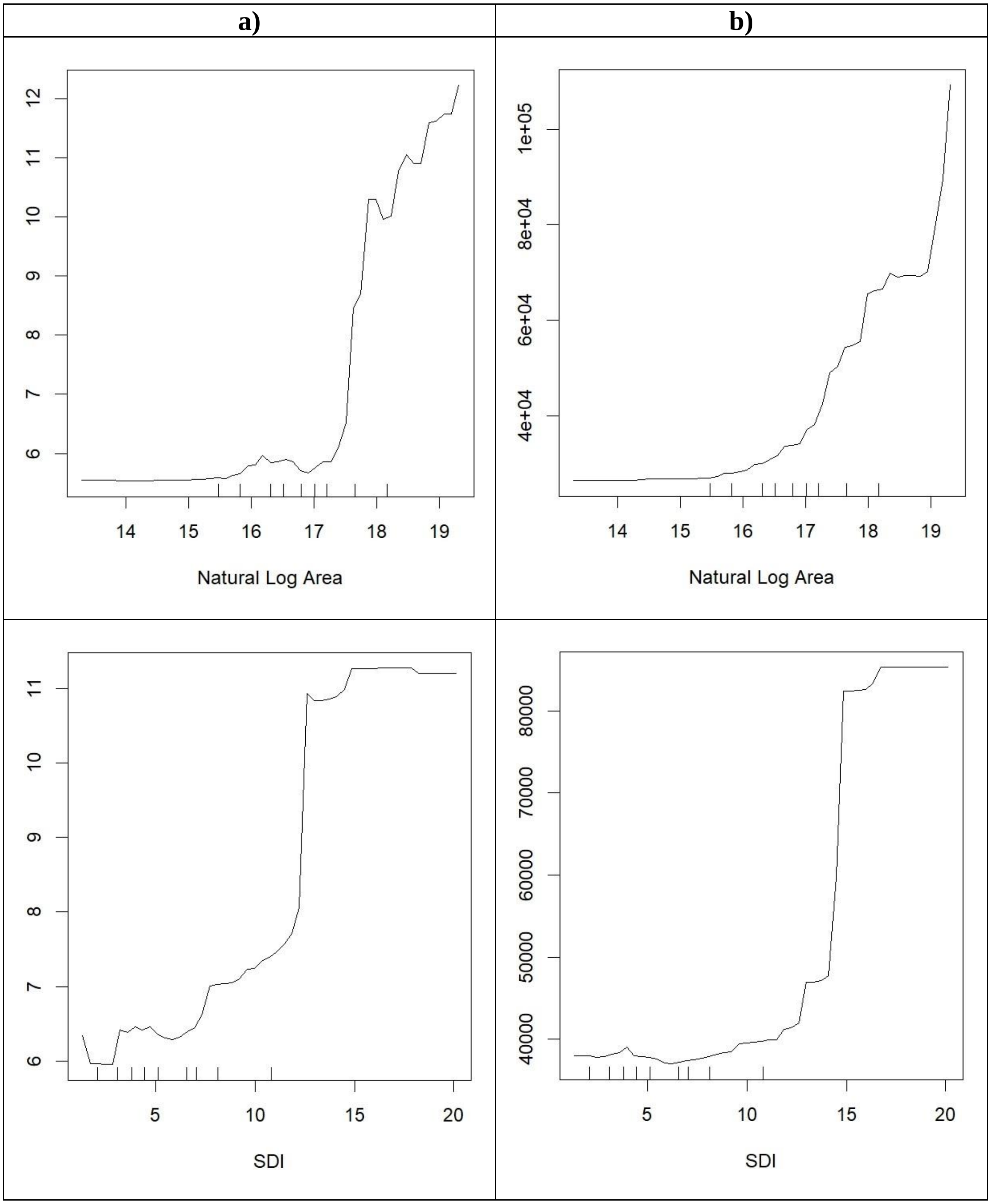

Figure 4. Random forest partial dependence plots for a) app (units of trip numbers) and b) creel data (units of aerial creel survey hours). These are based upon the 91 lakes in Ontario with both app and creel survey data and CPUE data detailing how changes in the variable affect the amount of angler effort.

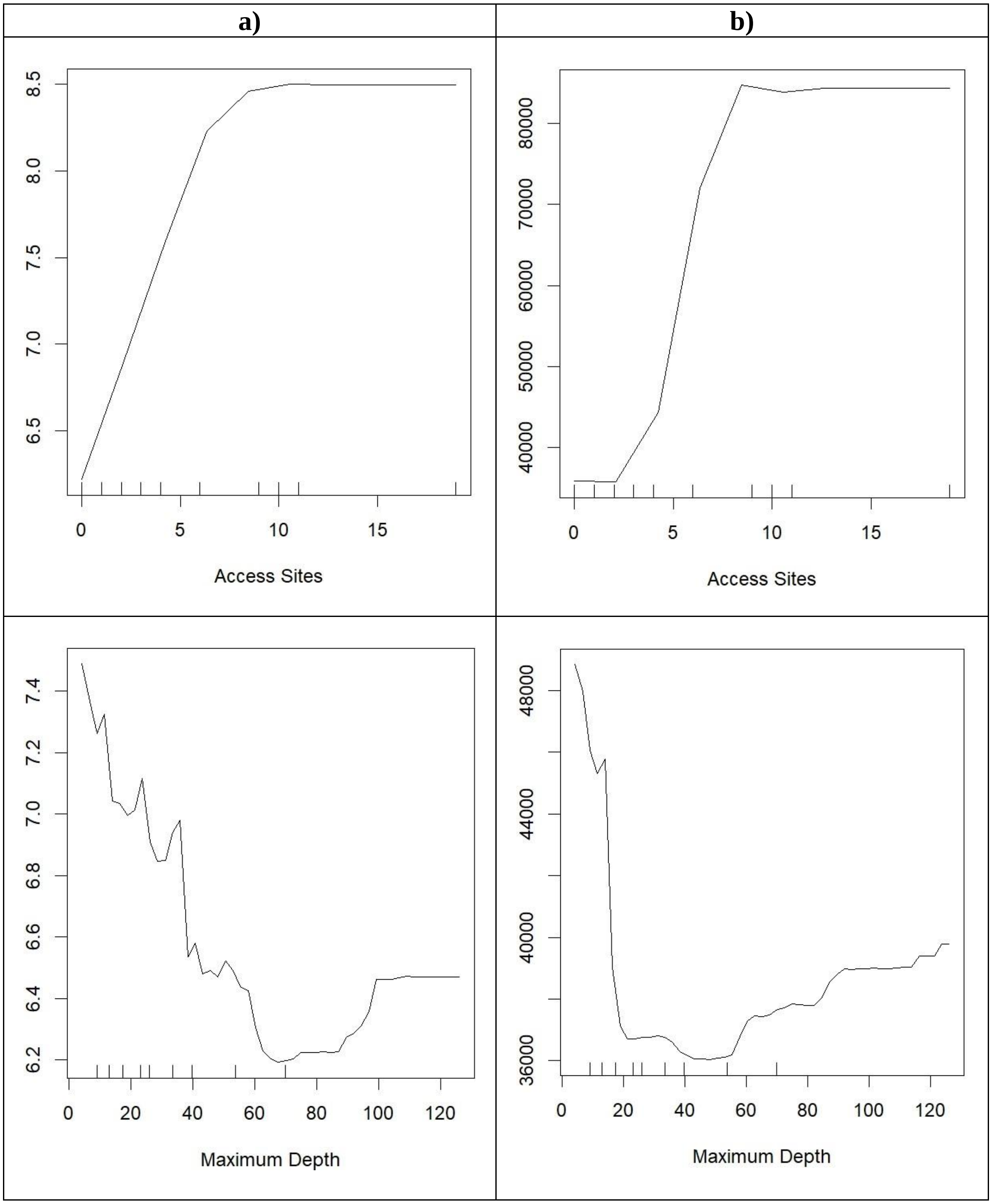
a)
b)
8.5
8.0
7.5
7.0
6.5
0
5
10
15
Access Sites
80000
70000
60000
50000
40000
0
5
10
15
Access Sites
7.4
7.2
7.0
6.8
6.6
6.4
6.2
0
20
40
60
80
100
120
Maximum Depth
48000
44000
40000
36000
0
20
40
60
80
100
120
Maximum Depth

Figure 4. (Continued).

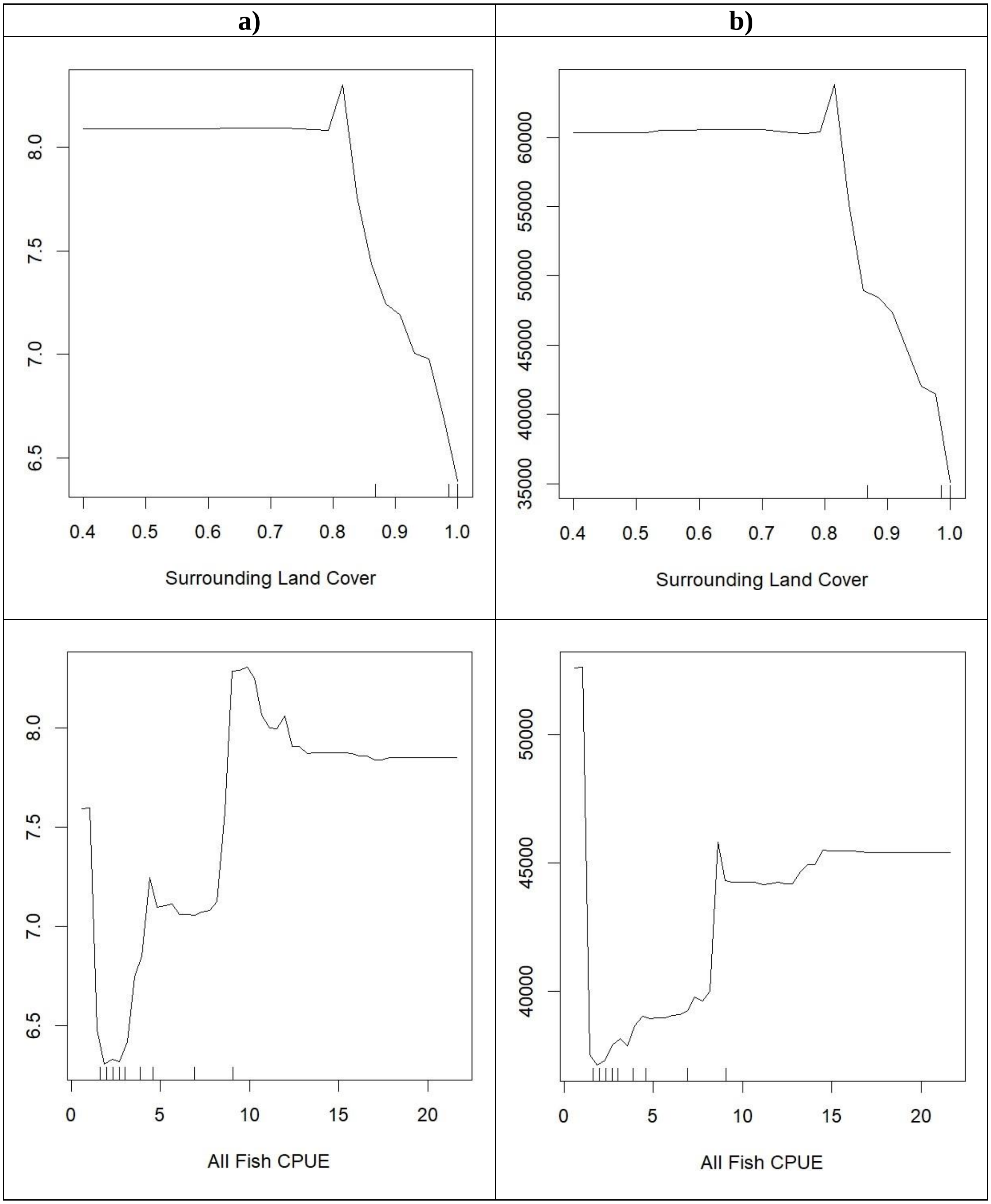

Figure 4. (Continued).

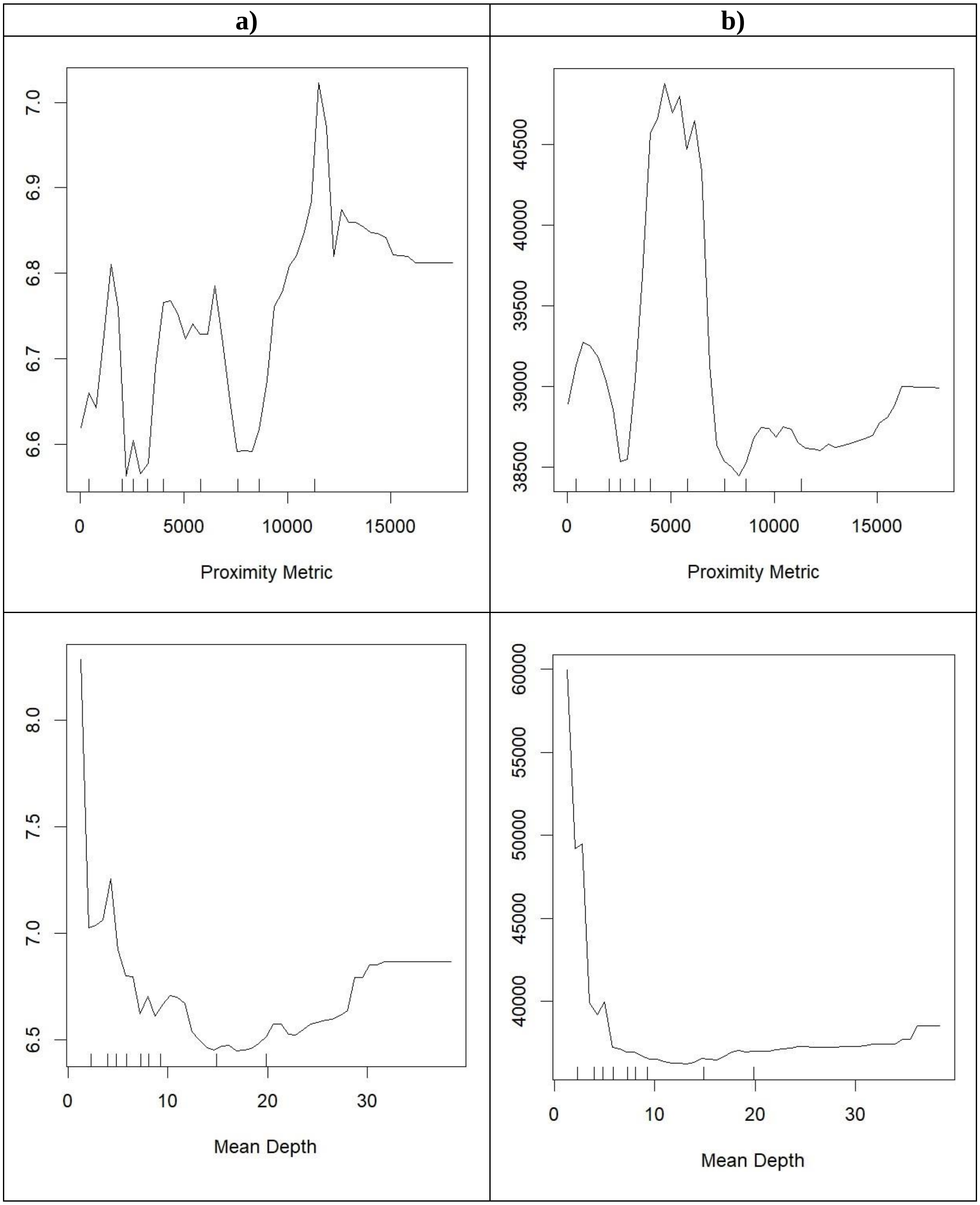

Figure 4. (Continued).

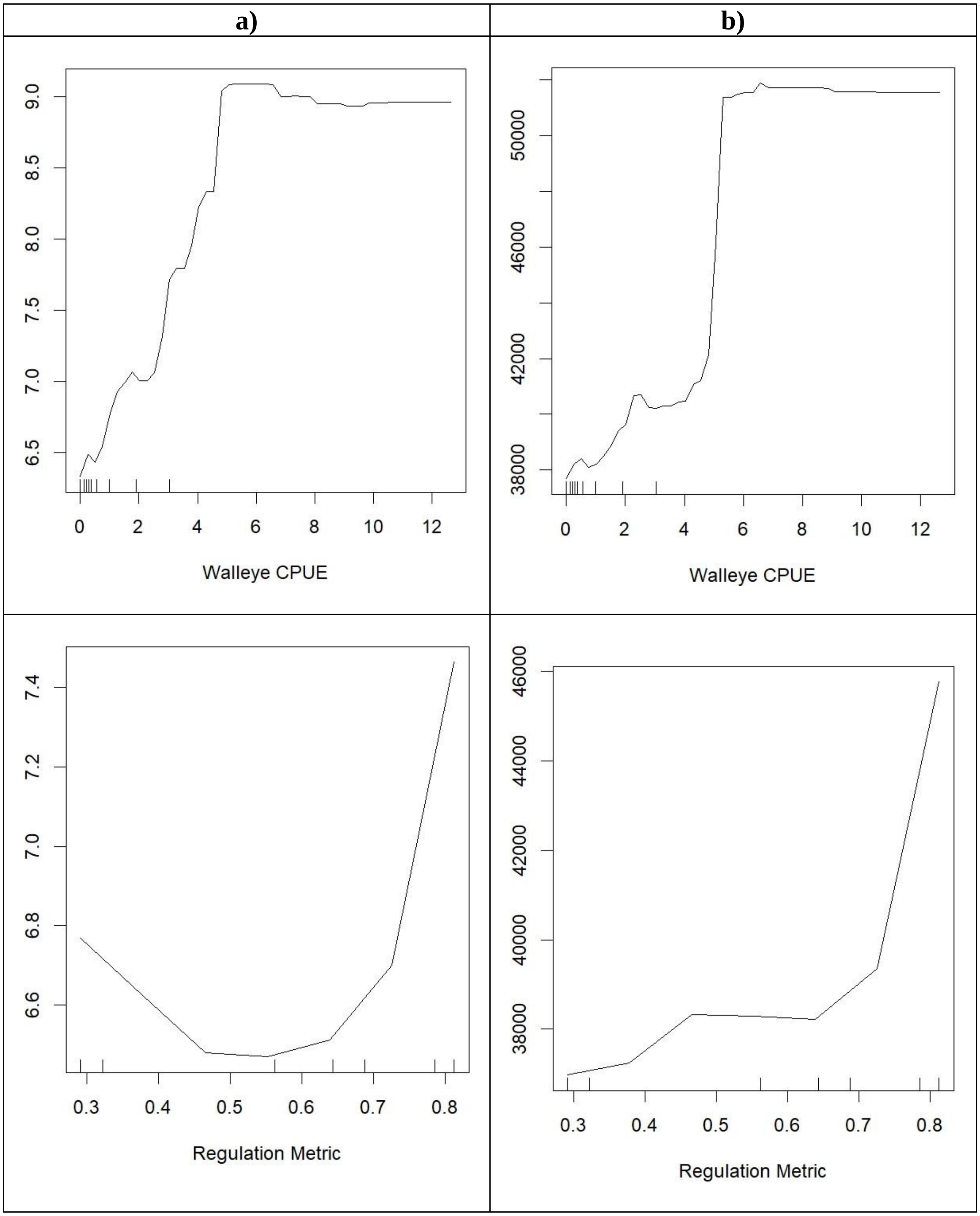

Figure 4. (Continued).

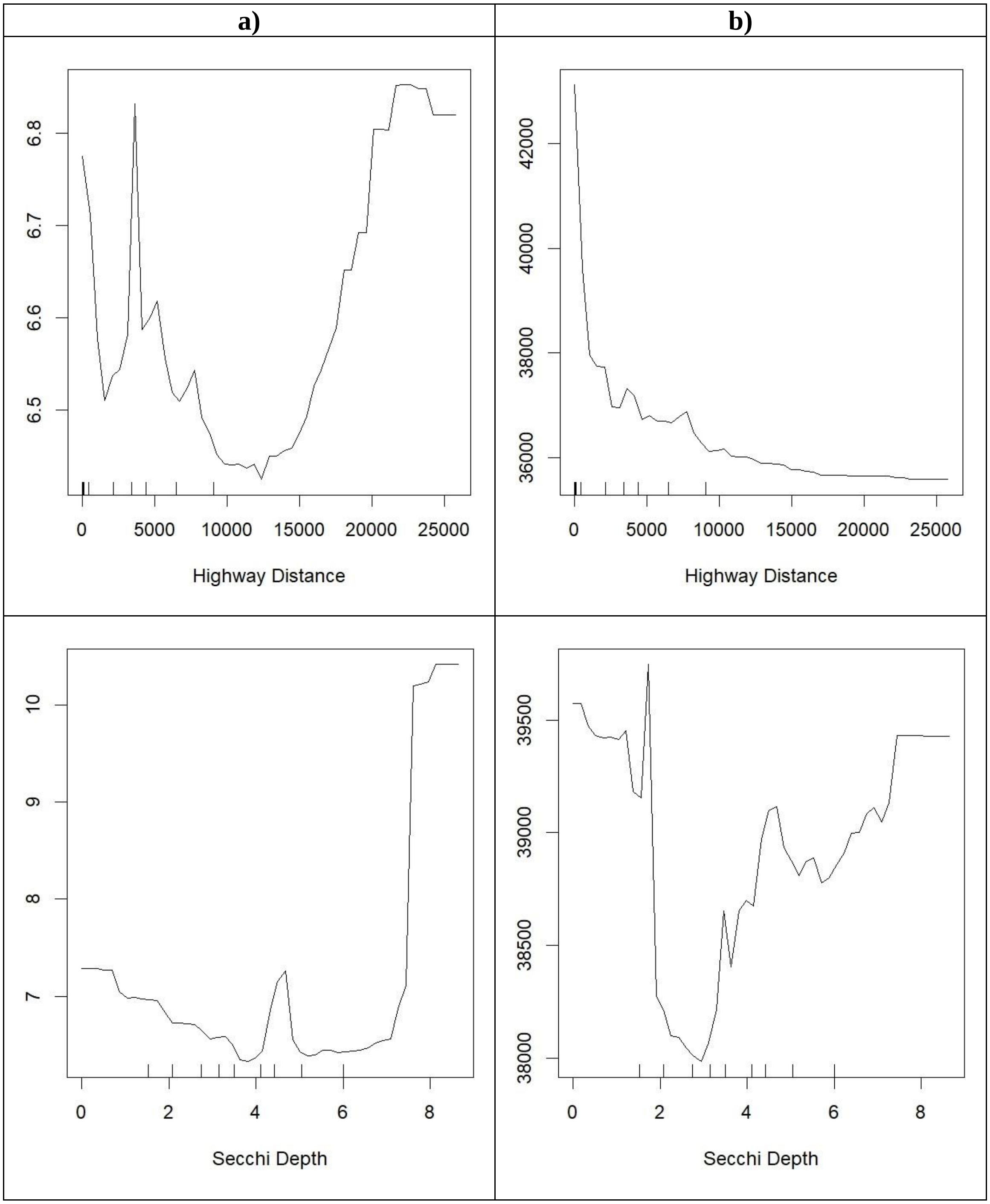

Figure 4. (Continued).

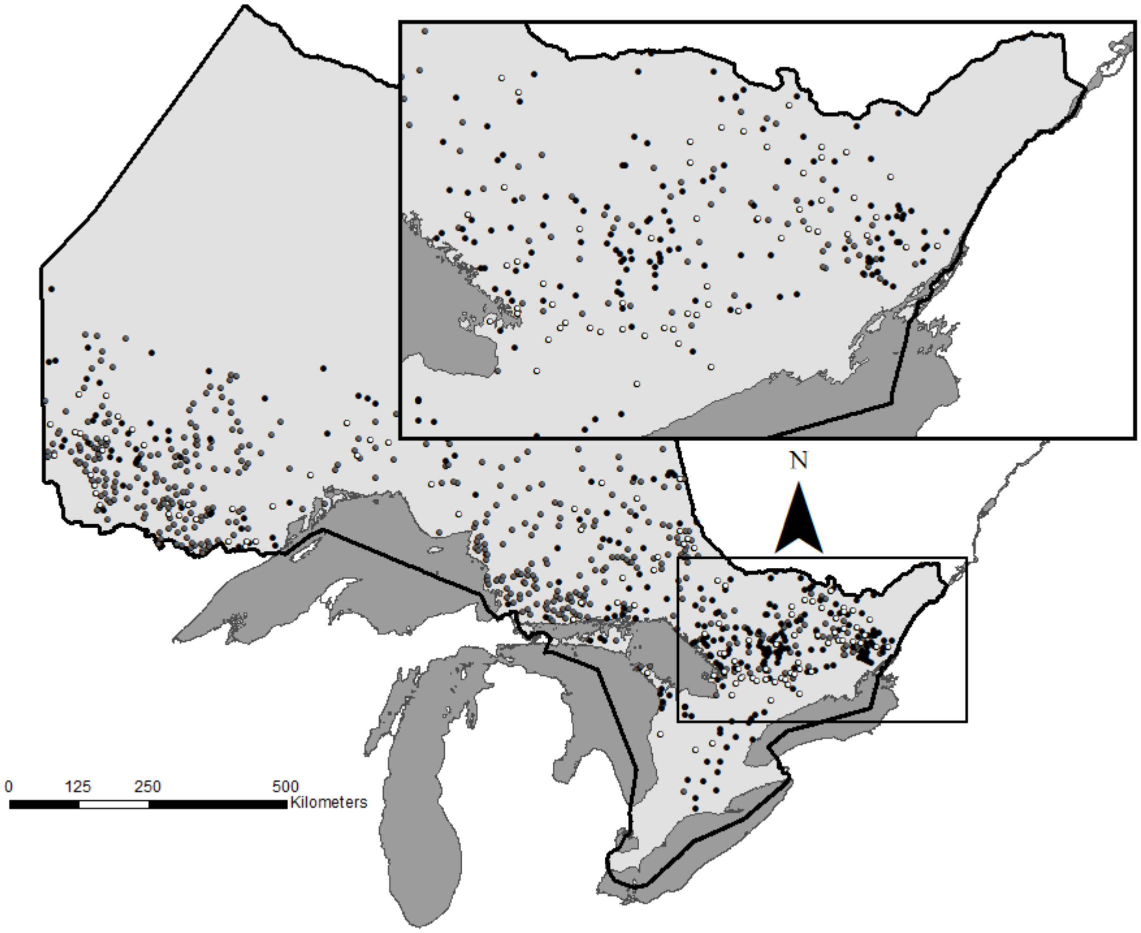

Figure 5. The locations of lakes in Ontario, Canada used in the random forest analyses on lakes with creel and/or app data. Black points indicate lakes with only creel data (n=448), gray points indicate lakes with only app data (n=257), and white points indicate lakes with both creel and app data (n=111). In total, there were 559 lakes with creel data and 368 lakes with app data.

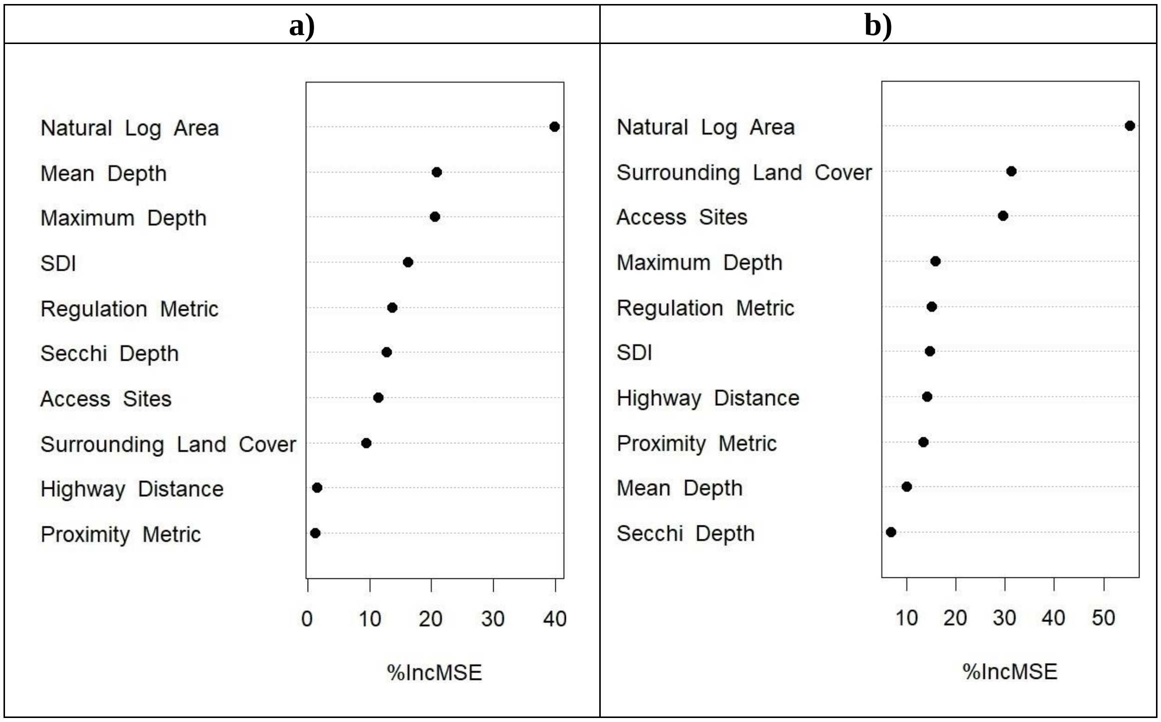

Figure 6. Variable importance plots of the random forest models built with ten variables and all available data from both app and creel data including 368 lakes with a) app data and 559 lakes with b) creel data in Ontario, Canada. %IncMSE is the percent increase in the mean square error. This metric refers to the mean decrease in model accuracy if a variable is removed.

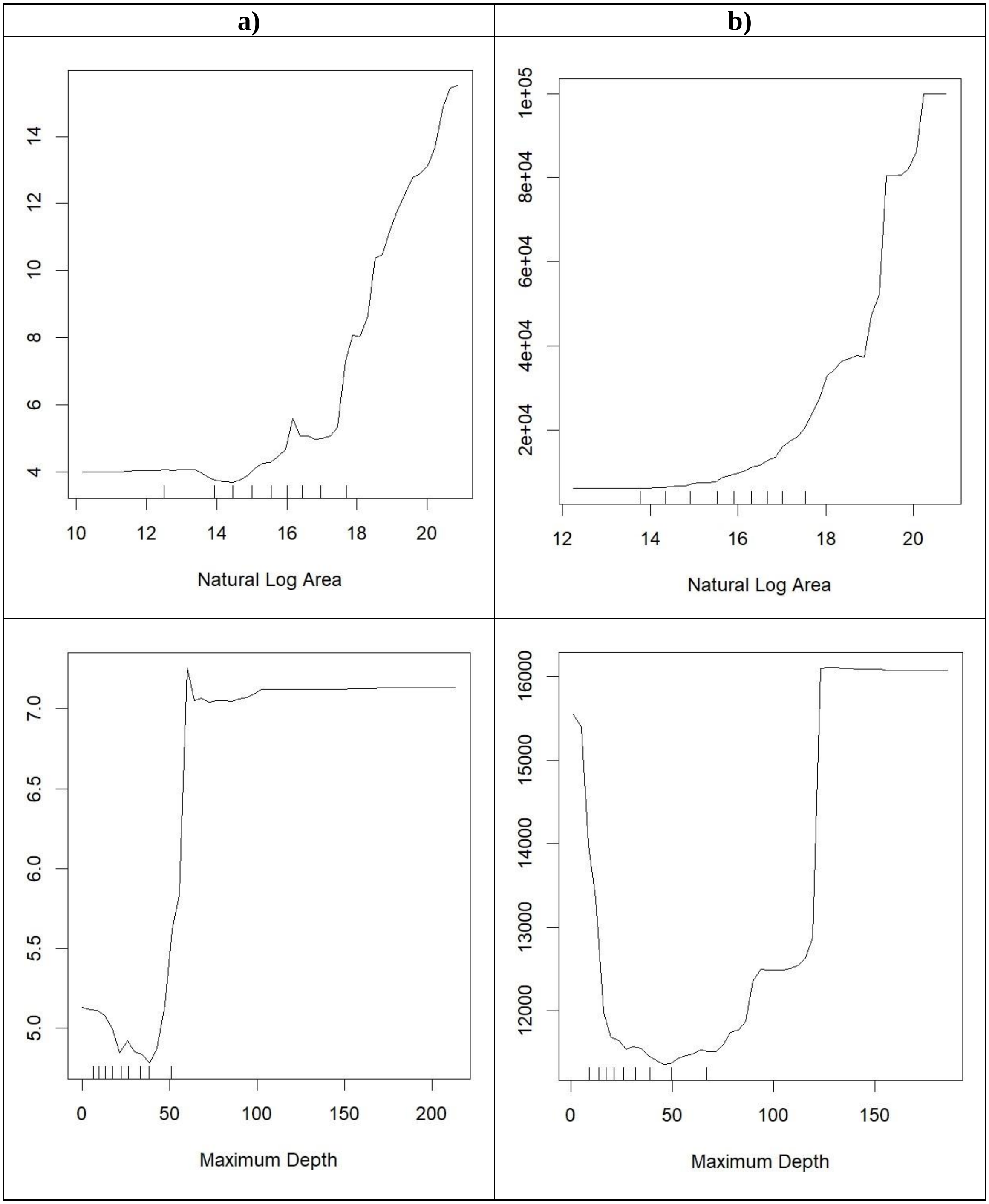

Figure 7. Random forest partial dependence plots for ten variables for 368 lakes with a) app data and 559 lakes with b) creel data in Ontario, Canada detailing how changes in the variable affect the amount of angler effort.

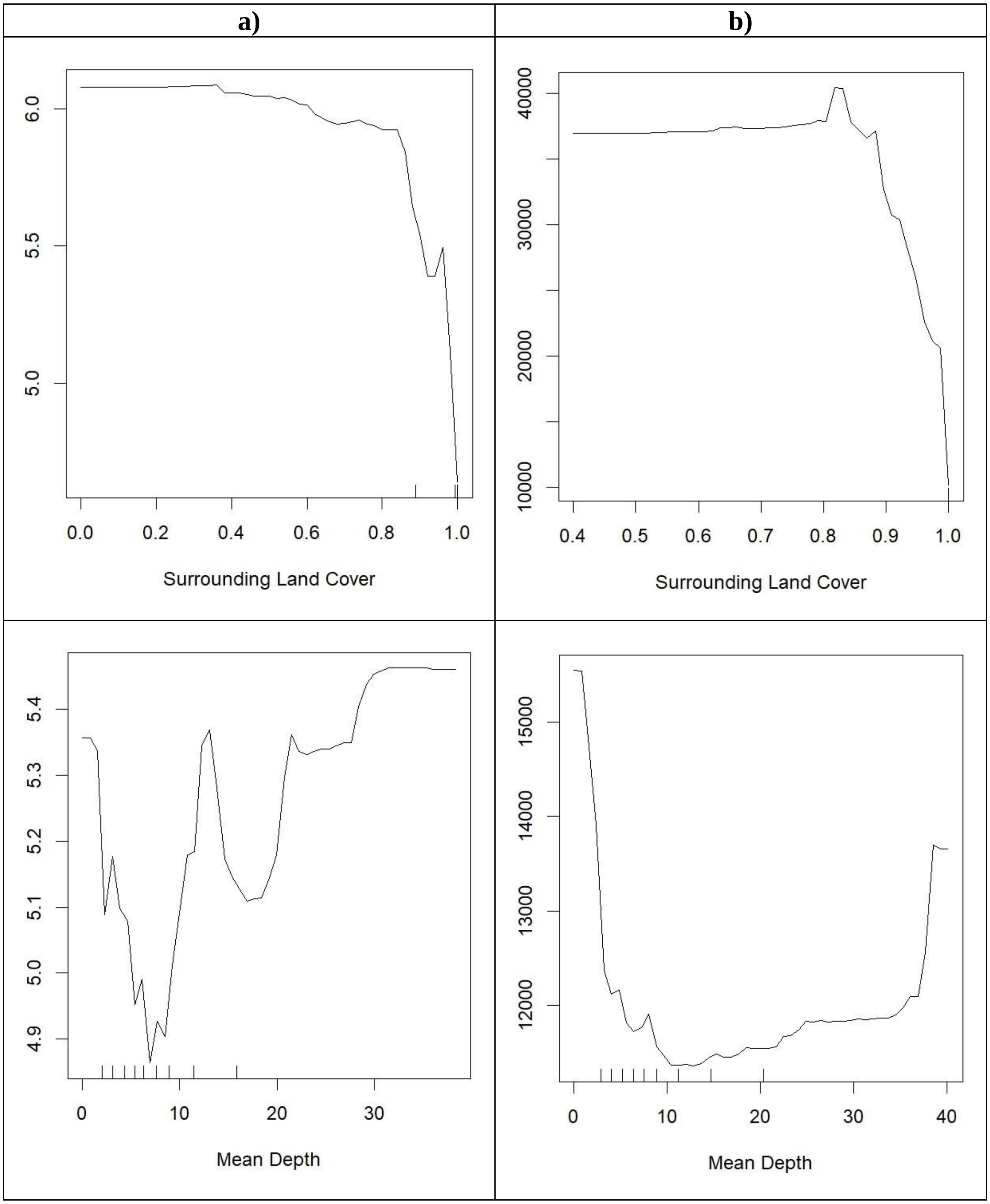

Figure 7. (Continued)

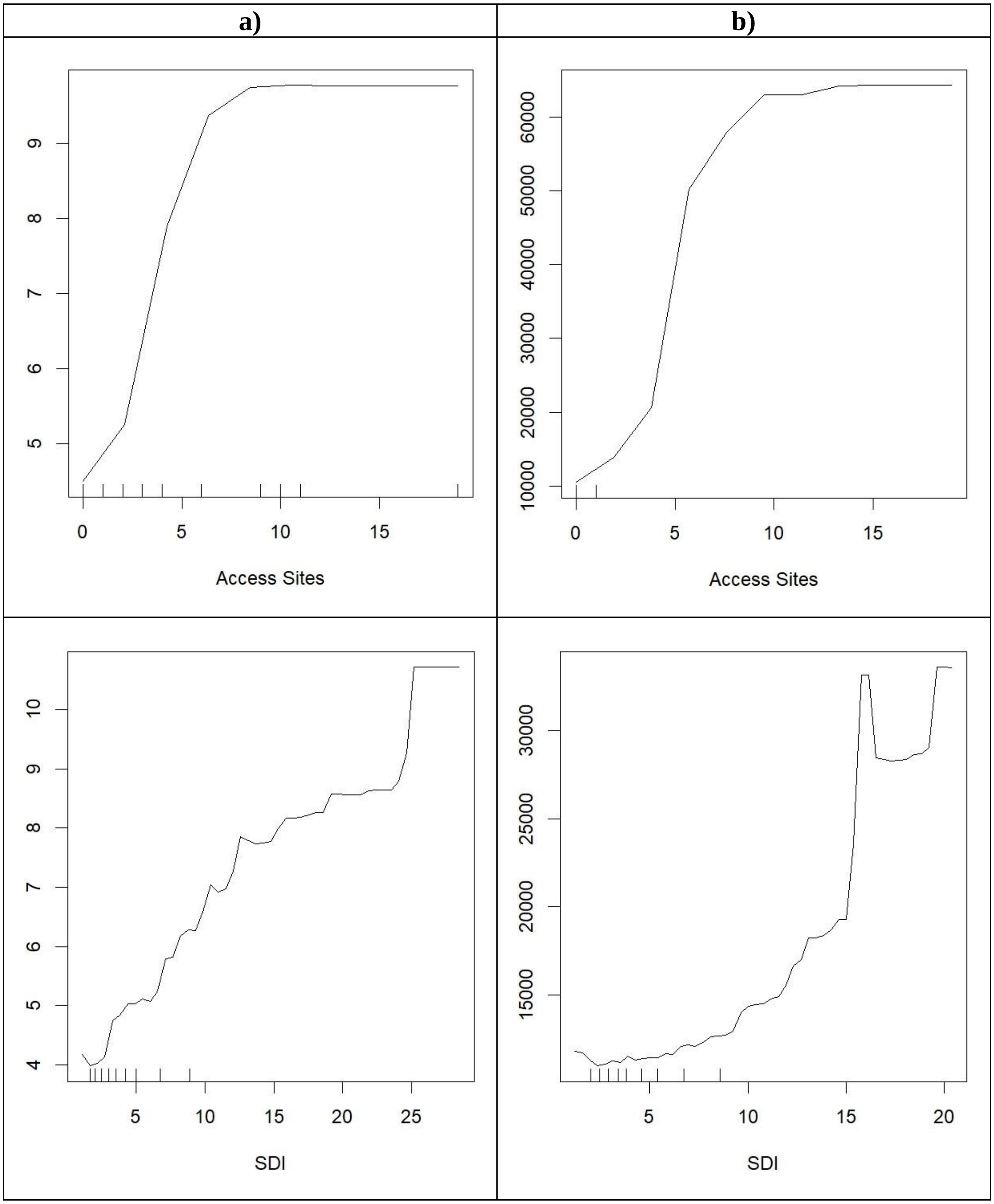

Figure 7. (Continued)

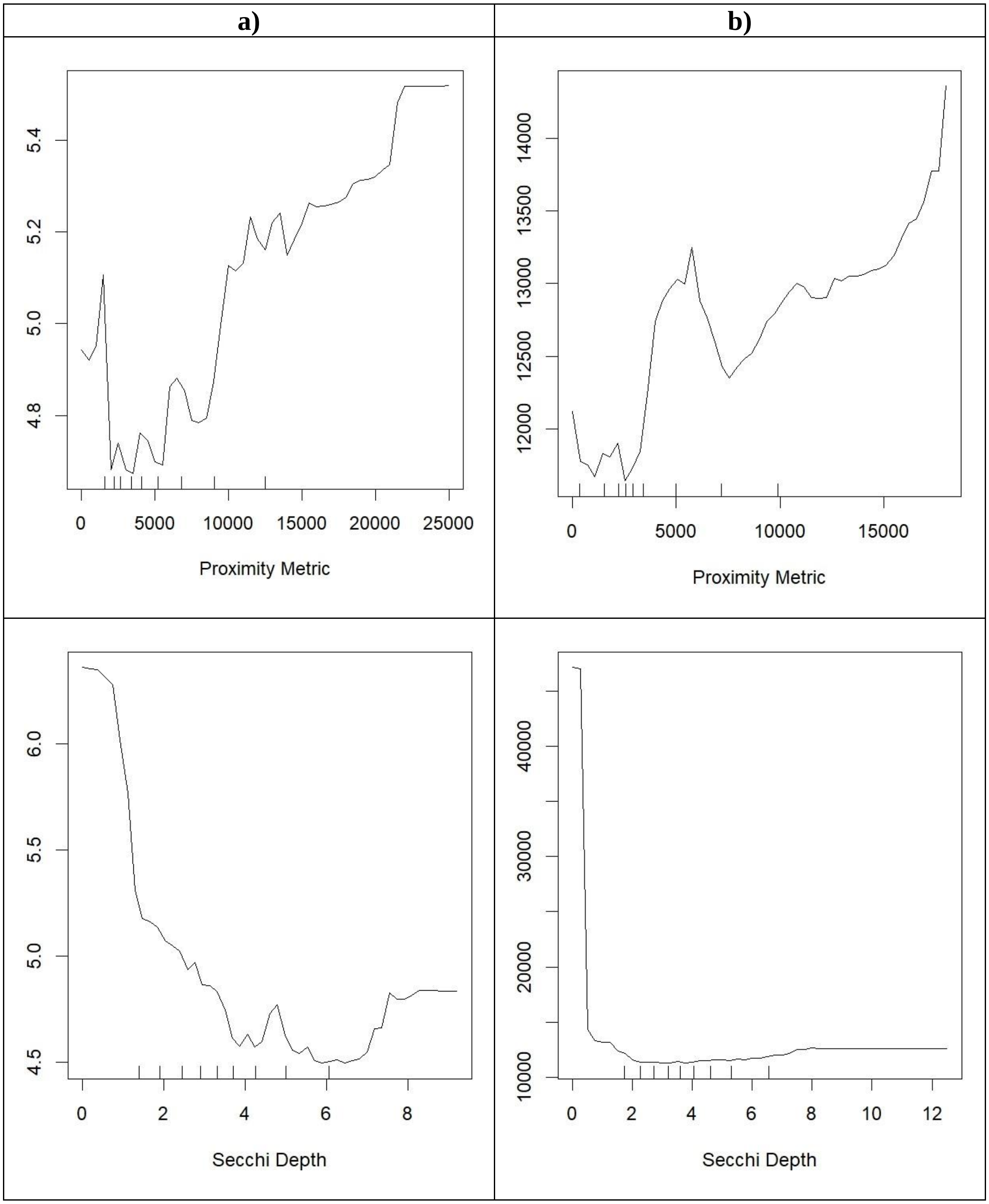

Figure 7. (Continued)

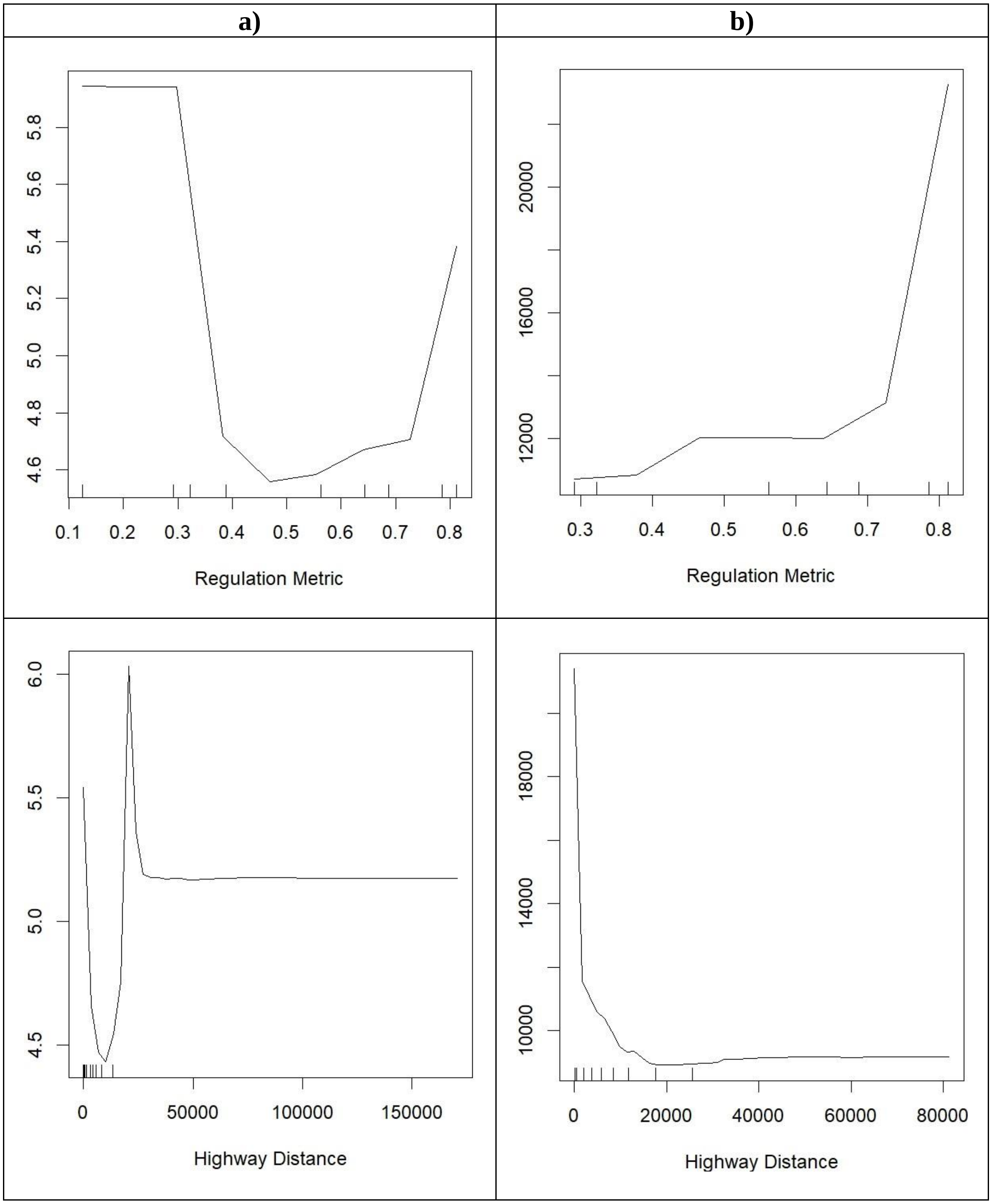

Figure 7. (Continued)

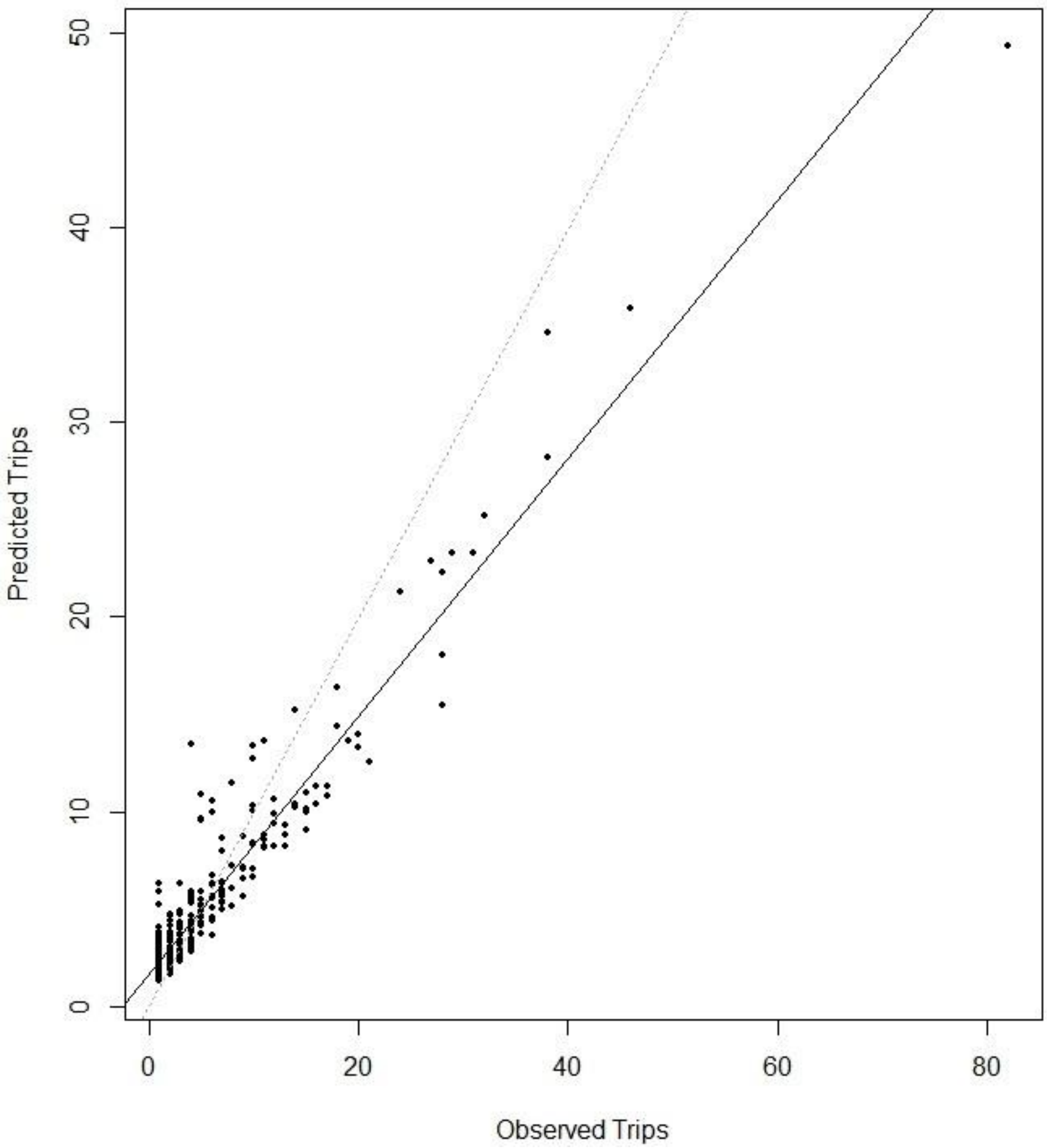

Figure 8. A linear regression plot comparing the relationship between predicted app trips and observed app trips (n=368, R^2=0.92, p= 2.2e-16) in Ontario, Canada. The solid line is the linear regression trend line (y=1.6987+0.65607x); whereas, the grey, dashed line indicates the location of the 1:1 slope with a zero origin.

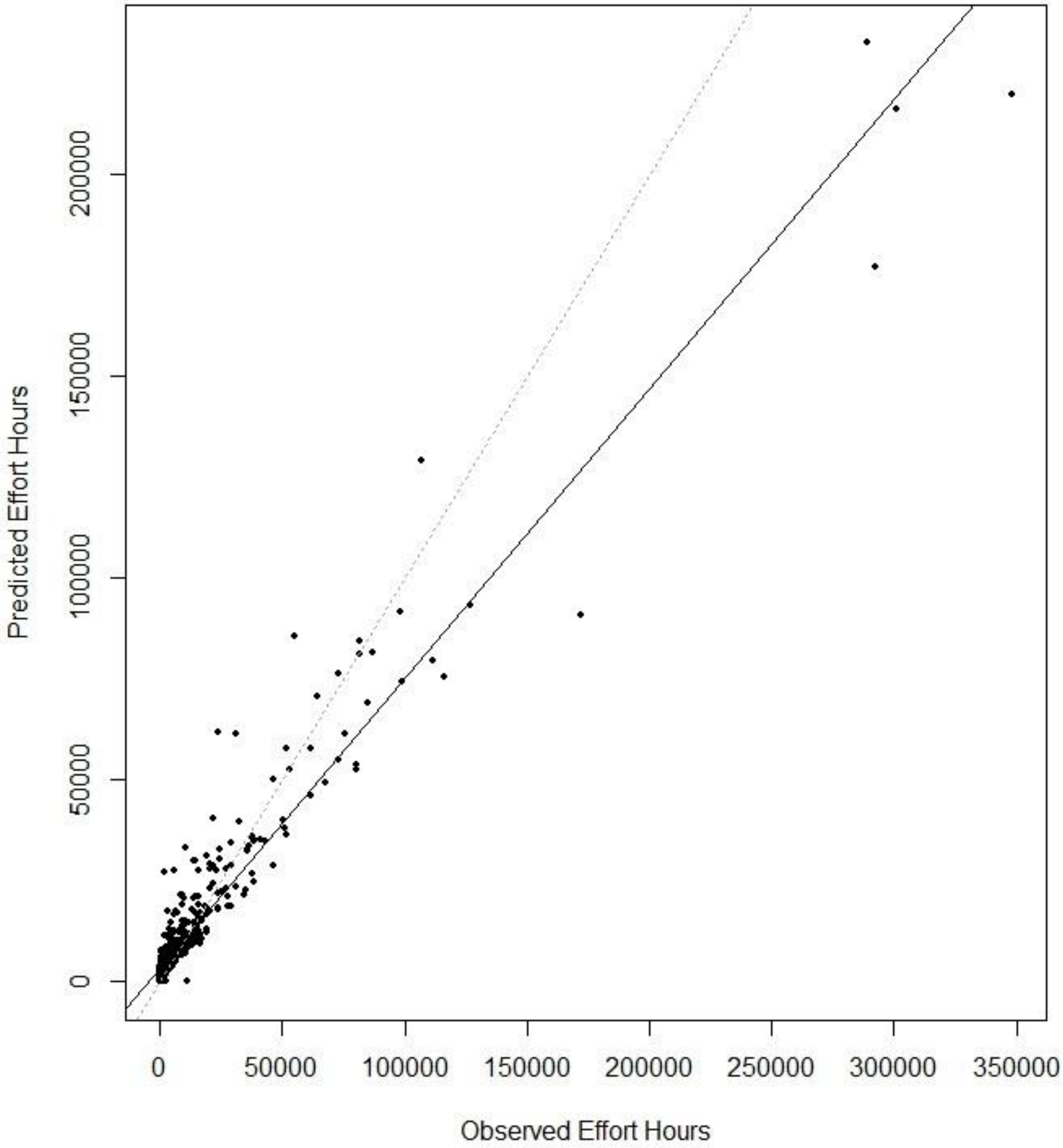

Figure 9. A linear regression plot comparing the relationship between predicted creel effort hours and observed effort hours (n=559, R^2=0.93, p= 2.2e-16) in Ontario, Canada. The solid line is the linear regression trend line (y=3.214e3+7.169e-1x); whereas, the grey, dashed line indicates the location of the 1:1 slope with a zero origin.

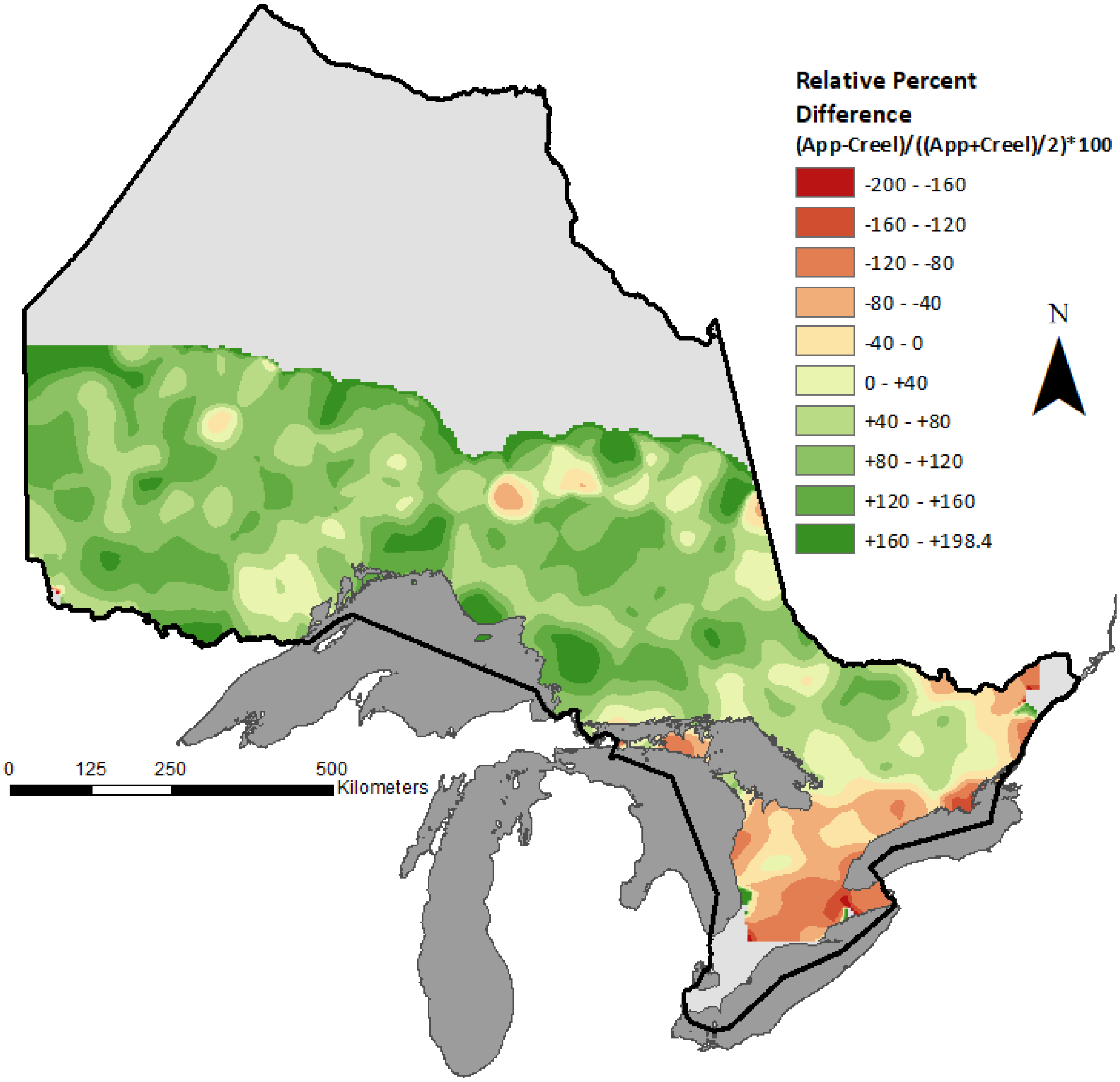

Figure 10. The relative percent difference analysis between kernel density analyses conducted on the scaled app and creel data in Ontario, Canada. The app data has a mean percent difference of 71.25% compared to the creel data.

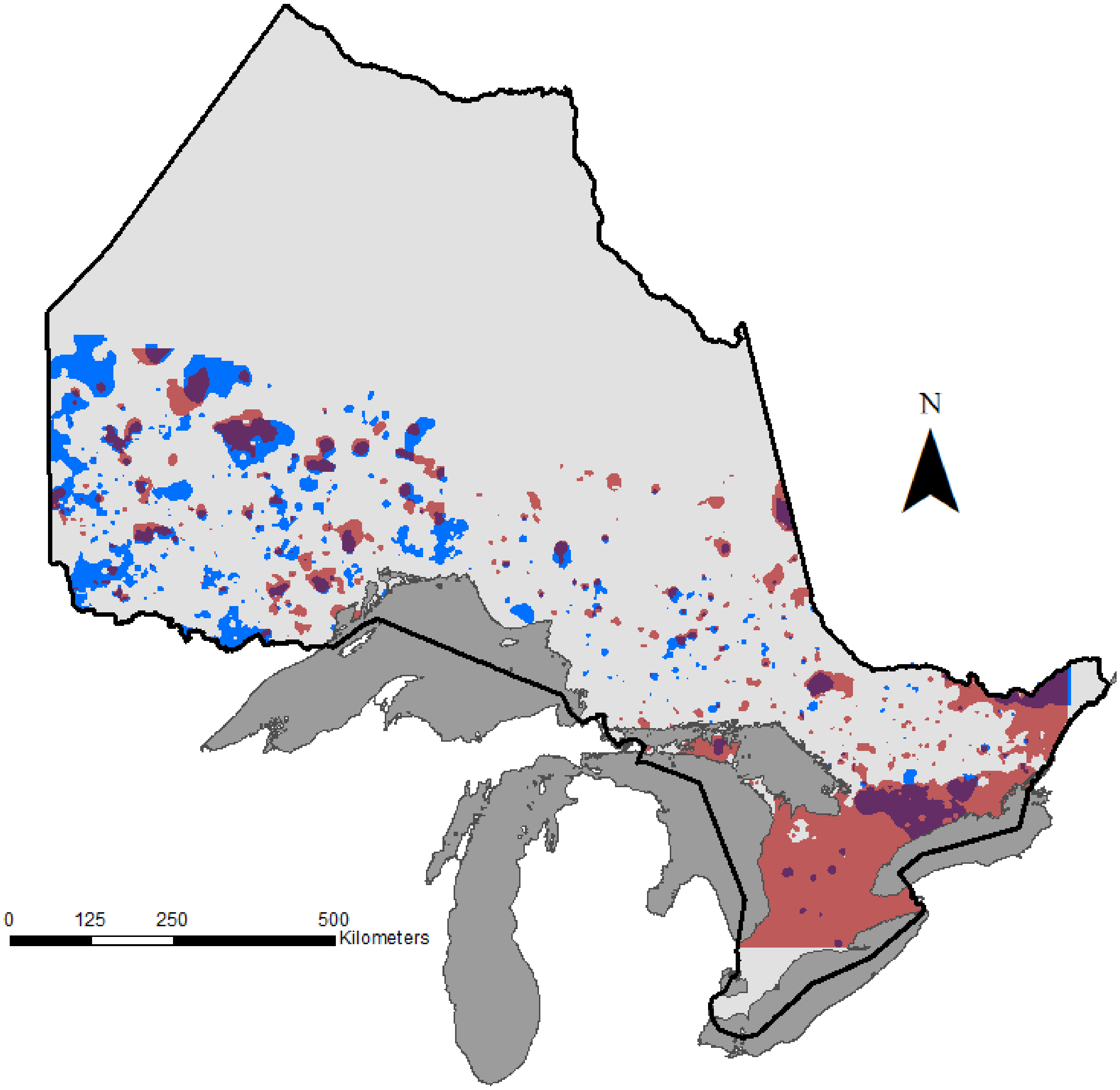

Figure 11. The IDW spatial interpolations of the scaled app (blue) and creel (red) data indicating the location of values within the top deciles of the data (app > 0.063, creel > 0.035). Purple indicates areas where the two sources overlap. This indicates where the random forest models predicted the highest levels of activity in Ontario, Canada.

References

2005 North American Land Cover at 250 m spatial resolution. Produced by Natural Resources Canada/Canadian Center for Remote Sensing (NRCan/CCRS), United States Geological Survey (USGS); Insituto Nacional de Estadística y Geografía (INEGI), Comisión Nacional para el Conocimiento y Uso de la Biodiversidad (CONABIO) and Comisión Nacional Forestal (CONAFOR).

Béliveau, Audrey, Richard A. Lockhart, Carl J. Schwarz, and Steven K. Arndt. "Adjusting for Undercoverage of Access-points in Creel Surveys with Fewer Overflights." *Biometrics* 71.4 (2015): 1050-059.

Bray, Gregory S., and Harold L. Schramm. "Evaluation of a Statewide Volunteer Angler Diary Program for Use as a Fishery Assessment Tool." *North American Journal of Fisheries Management* 21.3 (2001): 606-15.

Breiman, L. "Random Forests." *Machine Learning* 45 (2001): 5-32.

Carter, David W., Scott Crosson, and Christopher Liese. "Nowcasting Intraseasonal Recreational Fishing Harvest with Internet Search Volume." *PLoS ONE* 10.9 (2015).

Cole, Gerald A., and Paul E. Weihe. *Textbook of Limnology*. Long Grove, IL: Waveland, (2016).

Cooke, S. J., W. I. Dunlop, D. Macclennan, and G. Power. "Applications and Characteristics of Angler Diary Programmes in Ontario, Canada." *Fisheries Management and Ecology* 7.6 (2000): 473-87.

De Kerckhove, Derrick Tupper, Charles Kenneth Minns, and Cindy Chu. "Estimating Fish Exploitation and Aquatic Habitat Loss across Diffuse Inland Recreational Fisheries." *PloS ONE* 10.4 (2015).

Diogo, H., and Jg Pereira. "Fishing in the Dark: The Importance of Integrating a Nocturnal Component into Recreational Fishing Surveys." *Marine Ecology Progress Series* 542 (2016): 187-93.

DNR Hydrography Dataset. Saint Paul: Minnesota Department of Natural Resources, (2012). GDB.

Dodge, Douglas P. Manual of Instructions: Aquatic Habitat Inventory Surveys. Toronto: Ontario Ministry of Natural Resources, (1987).

Drake, D. Andrew R., and Nicholas E. Mandrak. "Least-cost Transportation Networks Predict Spatial Interaction of Invasion Vectors." *Ecological Applications* 20.8 (2010): 2286-299.

ESRI. ArcGIS Desktop: Release 10.3.1. Redlands, CA: Environmental Systems Research Institute. (2015).

Forward Sortation Areas, 2011 Census. Statistics Canada Catalogue no. 92-179-X. (2011).

Goodwin, Andrew E., James E. Peterson, Theodore R. Meyers, and David J. Money. "Transmission of Exotic Fish Viruses." *Fisheries* 29.5 (2004): 19-23.

Hartill, Bruce W., George W. Payne, Nicola Rush, and Richard Bian. "Bridging the Temporal Gap: Continuous and Cost-effective Monitoring of Dynamic Recreational Fisheries by Web Cameras and Creel Surveys." *Fisheries Research* 183 (2016): 488-97.

Hunt, Len M. "Recreational Fishing Site Choice Models: Insights and Future Opportunities."*Human Dimensions of Wildlife* 10.3, (2005): 153-172.

Hunt, Len M. and Nigel Lester. "The Effect of Forestry Roads on Access to Remote Fishing Lakes in Northern Ontario, Canada." *North American Journal of Fisheries Management* 29.3 (2009): 586-97.

Jiorle, Ryan P., Robert N.M. Ahrens, and Michael S. Allen. "Assessing the Utility of a Smartphone App for Recreational Fishery Catch Data. *Fisheries*, 41:12, (2016): 758-766.

Kaufman, Scott D., Ed Snucins, John M. Gunn, and Wayne Selinger. "Impacts of Road Access on Lake Trout (Salvelinus Namaycush) Populations: Regional Scale Effects of Overexploitation and the Introduction of Smallmouth Bass (Micropterus Dolomieu)." *Canadian Journal of Fisheries and Aquatic Sciences* 66.2 (2009): 212-23.

Kauppila, Pekka, and Timo P. Karjalainen. "A Process Model to Assess the Regional Economic Impacts of Fishing Tourism: A Case Study in Northern Finland." *Fisheries Research* 127-128 (2012): 88-97.

Keeler, B. L., Wood, S. A., Polasky, S., Kling, C., Filstrup, C. T. and Downing, J. A. Recreational demand for clean water: evidence from geotagged photographs by visitors to lakes. *Frontiers in Ecology and the Environment*, (2015): 13: 76–81.

Lew, Daniel K., and Chang K. Seung. "The Economic Impact of Saltwater Sportfishing Harvest Restrictions in Alaska: An Empirical Analysis of Nonresident Anglers." *North American Journal of Fisheries Management* 30.2 (2010): 538-51.

Liaw, A. and M. Wiener. Classification and Regression by randomForest. *R News*, (2002): 2(3), 18-22.

Lockwood, Roger N., and Gerald P. Rakoczy. "Comparison of Interval and Aerial Count Methods for Estimating Fisher Boating Effort." *North American Journal of Fisheries Management* 25.4 (2005): 1331-340.

Mallison, Craig T., and Charles E. Cichra. "Accuracy of Angler-Reported Harvest in Roving Creel Surveys." *North American Journal of Fisheries Management* 24.3 (2004): 880-89.

Malvestuto, Stephen P., William D. Davies, and William L. Shelton. "Predicting the Precision of Creel Survey Estimates of Fishing Effort by Use of Climatic Variables." *Transactions of the American Fisheries Society* 108.1 (1979): 43-45.

Martin, Dustin R., Christopher J. Chizinski, Kent M. Eskridge, and Kevin L. Pope. "Using Posts to an Online Social Network to Assess Fishing Effort." *Fisheries Research* 157 (2014): 24-27.

Minnesota Population Center. National Historical Geographic Information System: Version 11.0. Zip Codes 2010 Census. Minneapolis: University of Minnesota. (2016). http://doi.org/10.18128/D050.V11.0.

Mkwara, Lena, and Dan Marsh. "Valuing Trout Angling Benefits of Water Quality Improvements While Accounting for Unobserved Lake Characteristics: An Application to the Rotorua Lakes." Proc. of 2011 New Zealand Agricultural and Resource Economics Conference, Tahuna Conference Centre, Nelson, New Zealand. (2011).

“Mobile Fact Sheet” Pew Research Center, Washington, D.C. (2017) http://www.pewinternet.org/fact-sheet/mobile/, August 2017.

National Research Council. Review of Recreational Fisheries Survey Methods. Washington, DC: *The National Academies Press,* (2006).

OHN - Waterbody. Peterborough: Ontario Ministry of Natural Resources and Forestry, (2011). GDB.

Ontario Ministry of Natural Resources and Forestry. 2010 Survey of Recreational Fishing in Canada: Selected Results for Ontario Fisheries. Biodiversity Branch. Ontario

Ministry of Natural Resources and Forestry. Peterborough, Ontario. 40 p. + appendices, (2014).

Ontario Ministry of Natural Resources and Forestry. 2017 Fishing Ontario: Recreational Fishing Regulations Summary. Ministry of Natural Resources and Forestry, (2016).

Papenfuss, Jason T., Nicholas Phelps, David Fulton, and Paul A. Venturelli. "Smartphones Reveal Angler Behavior: A Case Study of a Popular Mobile Fishing Application in Alberta, Canada." *Fisheries* 40.7 (2015): 318-27.

R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/. (2017).

Road Network, 2011 Census. Statistics Canada Catalogue no. 92-151-X. (2012).

Sandstrom, S, M. Rawson and N. Lester. Manual of Instructions for Broad-scale Fish Community Monitoring; using North American (NA1) and Ontario Small Mesh (ON2) Gillnets. Ontario Ministry of Natural Resources. Peterborough, Ontario. Version 2013.2 35 p. + appendices. (2013).

Shiffman, David S., Catherine Macdonald, Harry Y. Ganz, and Neil Hammerschlag. "Fishing Practices and Representations of Shark Conservation Issues among Users of a Land-based Shark Angling Online Forum." *Fisheries Research* 196 (2017): 13-26.

Stuart, Chloe. Fishing Access Points. Ontario, Ontario Ministry of Natural Resources and Forestry: (2015).

Stunz, G.W., M.J. Johnson, D. Yoskowitz, M. Robillard, and J. Wetz. “iSnapper: Design, Testing, and Analysis of an iPhone-based Application as an Electronic Logbook in the for-hire Gulf of Mexico Red Snapper Fishery”. National Oceanic and Atmospheric Administration Final Report. NA10NMF4540111, (2014): 64.

Sullivan, Michael G. "Exaggeration of Walleye Catches by Alberta Anglers." North American Journal of Fisheries Management 23.2 (2003): 573-80.

USGS TNM Hydrography (NHD). Reston: U. S. Geological Survey - National Geospatial Program, 27 Feb. 2017. GDB.

Venturelli, Paul A., Kieran Hyder, and Christian Skov. "Angler Apps as a Source of Recreational Fisheries Data: Opportunities, Challenges and Proposed Standards." *Fish and Fisheries* 18.3, (2017): 578-95.

Warton, David I., Remko A. Duursma, Daniel S Falster. and Sara Taskinen. smatr 3 - an R package for estimation and inference about allometric lines Methods in Ecology and Evolution 3(2), (2012): 257-259.

Appendix A

Table A1. The reclassification of the NALCMS land cover dataset. I used this dataset to determine the percentage of natural land cover surrounding the lakes in Ontario, Canada.

Land Cover Type	Classification
Temperate or sub-polar needleleaf forest	Natural
Sub-polar taiga needleleaf forest	Natural
Tropical or sub-tropical broadleaf evergreen forest	Natural
Tropical or sub-tropical broadleaf deciduous forest	Natural
Temperate or sub-polar broadleaf deciduous forest	Natural
Mixed forest	Natural
Tropical or sub-tropical shrubland	Natural
Temperate or sub-polar shrubland	Natural
Tropical or sub-tropical grassland	Natural
Temperate or sub-polar grassland	Natural
Sub-polar or polar shrubland-lichen-moss	Natural
Sub-polar or polar grassland-lichen-moss	Natural
Sub-polar or polar barren-lichen-moss	Natural
Wetland	Natural
Cropland	Unnatural
Barren land	Natural
Urban and built-up	Unnatural
Water	Removed
Snow and ice	Natural

Table 2A. The percent mean squared error values derived from the random forest models run on 91 lakes with both a) seasonally-limited app data and b) creel data, 368 lakes with c) seasonally-limited app data, 559 lakes with d) creel data, and 670 lakes with e) all app data, all in Ontario, Canada. The top five highest variables for each model are in bold.

	a)	**b)**	**c)**	**d)**	**e)**
Access Sites	**4.50**	**10.16**	11.35	**29.51**	**29.48**
All Fish CPUE	2.36	-0.33	NA	NA	NA
Area	**12.77**	**16.79**	**39.86**	**55.24**	**61.15**
Highway Distance	-3.19	2.86	1.40	14.29	2.56
Land Cover	2.70	**9.88**	9.37	**31.31**	**24.31**
Maximum Depth	**5.18**	**6.06**	**20.48**	**15.88**	**24.39**
Mean Depth	1.90	5.12	**20.89**	10.00	**23.16**
Proximity Metric	-1.26	1.97	1.13	13.41	4.96
Regulation Metric	**3.45**	5.34	**13.54**	**15.15**	12.35
Secchi Depth	0.05	-3.32	12.77	6.88	13.73
SDI	**6.27**	**7.90**	**16.14**	14.75	23.11
Walleye CPUE	0.82	-0.27	NA	NA	NA

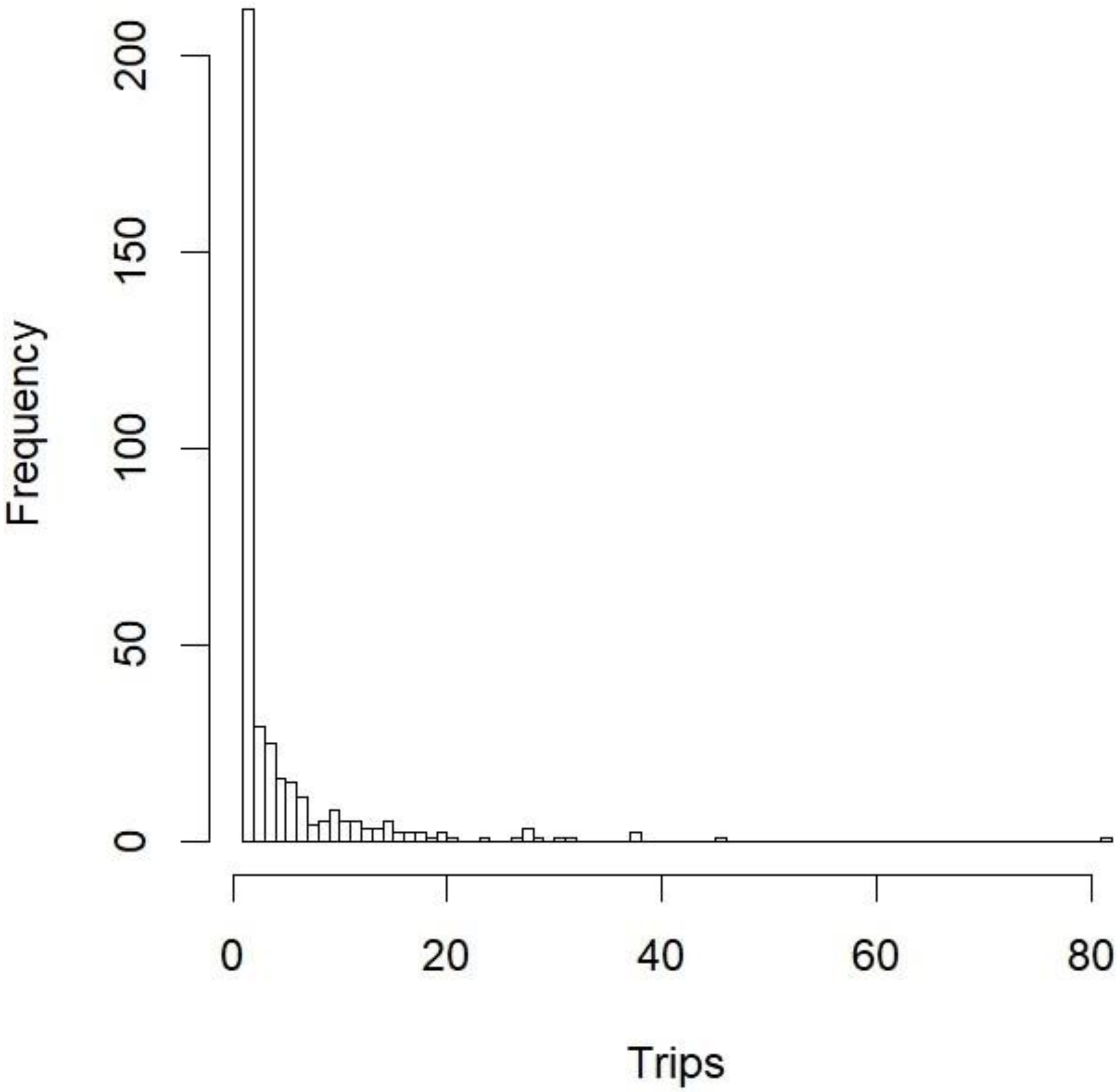

Figure 1A. A histogram of the frequency of seasonally- and yearly-adjusted app trip values at lakes in Ontario, Canada. The vast majority of lakes only recorded a single trip.

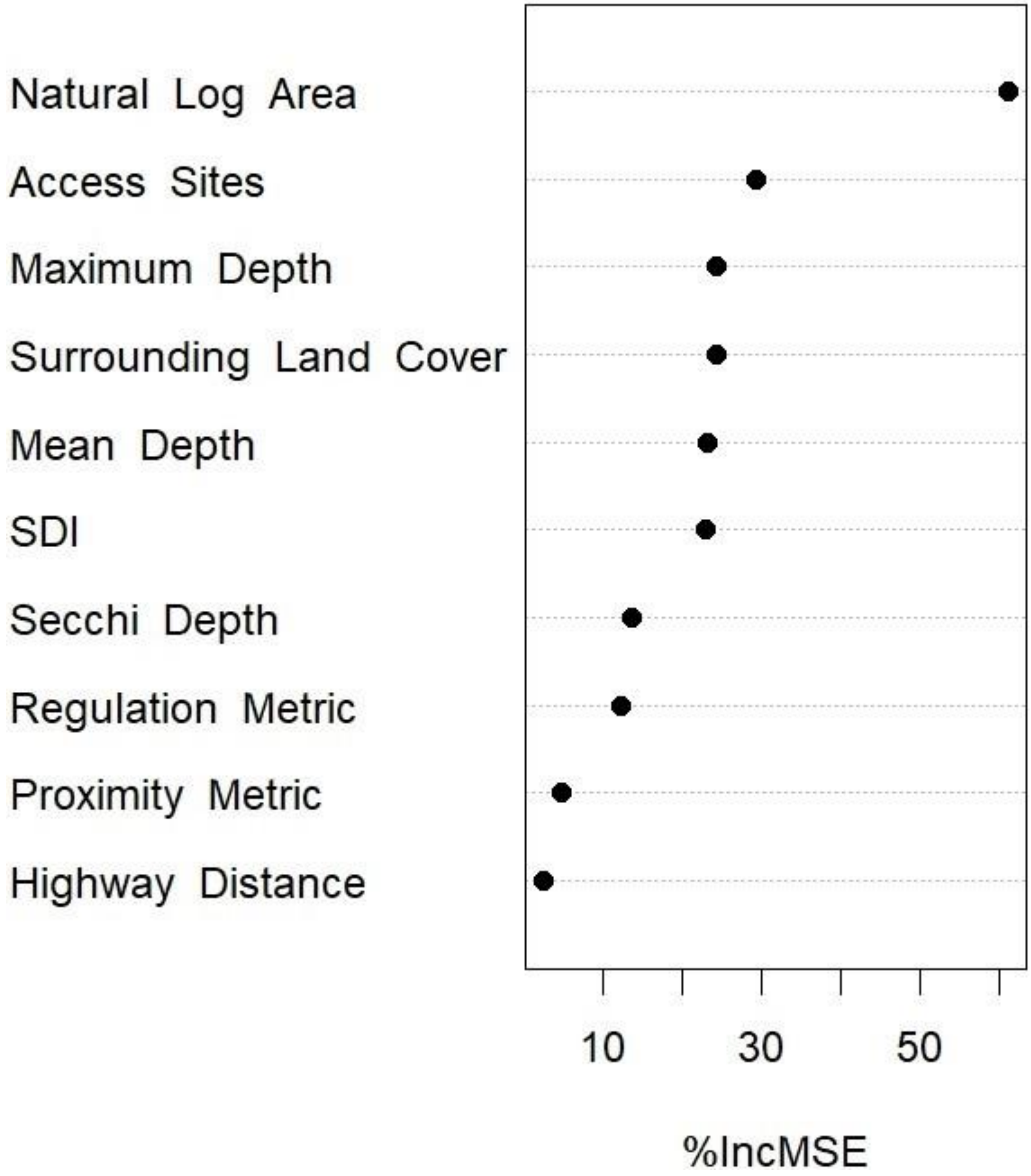

Figure 2A. Variable importance plots of a random forest model built with ten variables and all available app data, unrestricted by season or year, for 670 lakes in Ontario, Canada. %IncMSE is the percent increase in the mean square error. This metric refers to the mean decrease in model accuracy if a variable is removed. These app data explained 36.94% of variation.

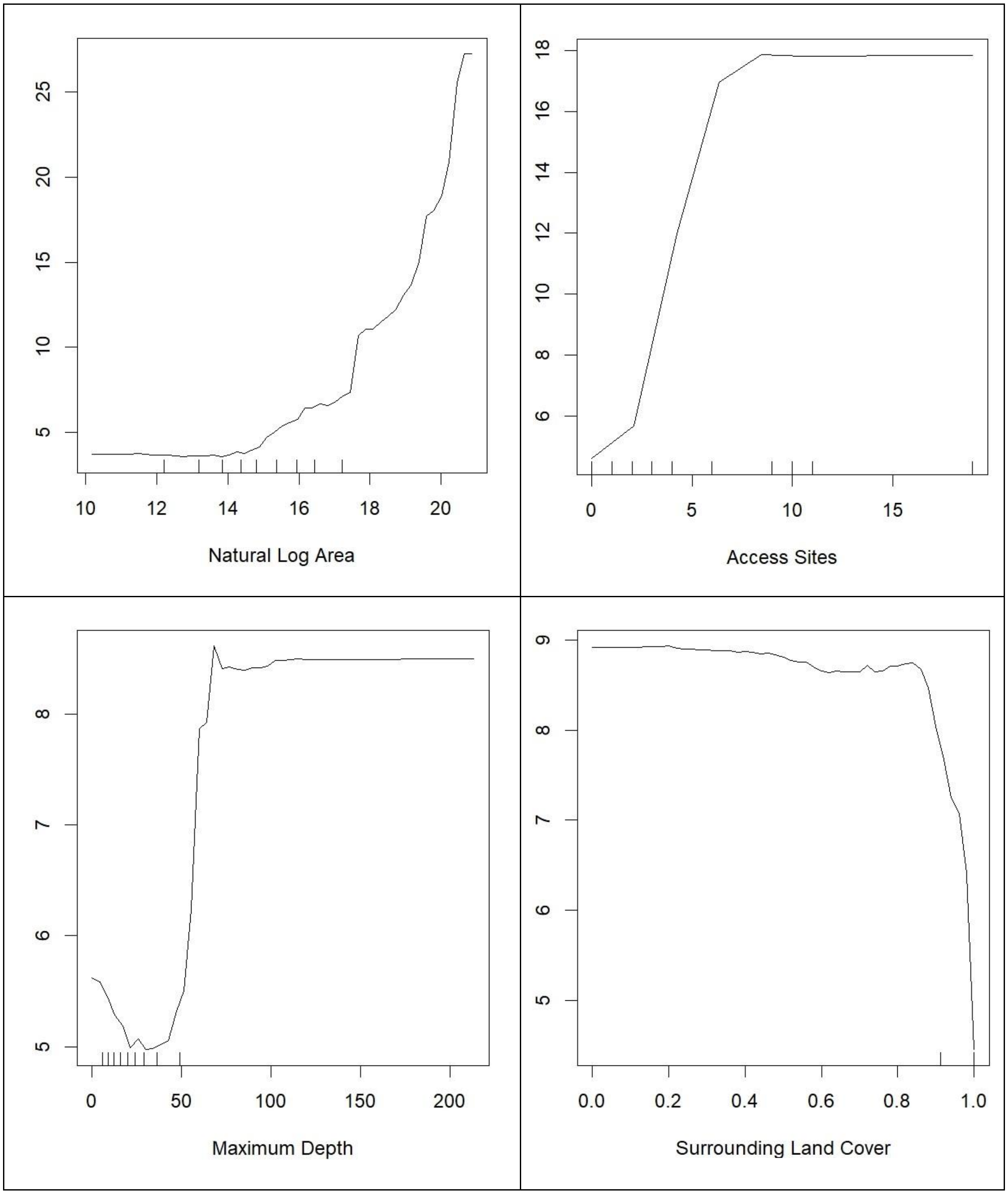

Figure 3A. Random forest partial dependence plots for ten variables for 670 lakes with app data in Ontario, Canada detailing how changes in the variable affect the amount of angler effort.

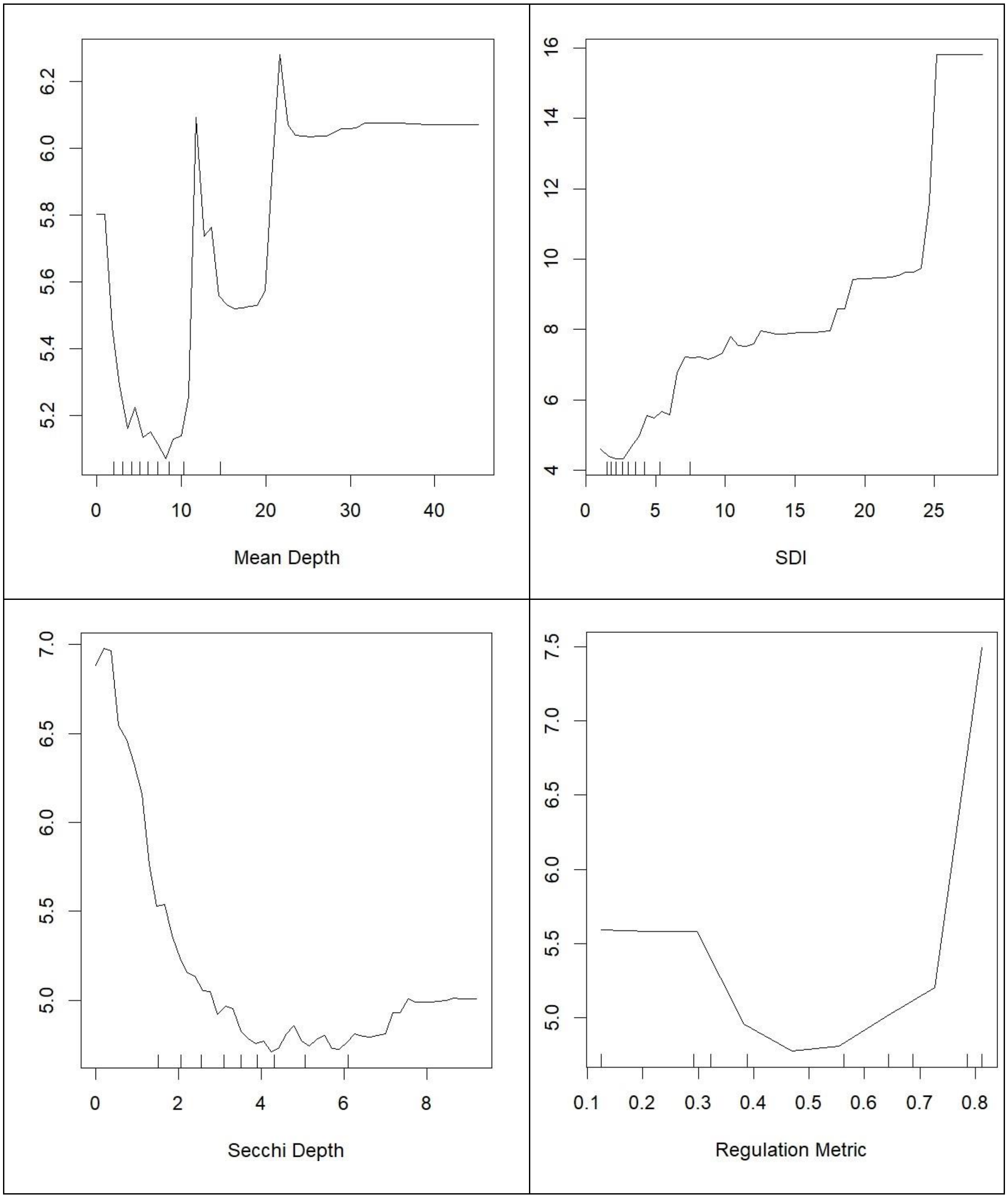

Figure 3A. (Continued)

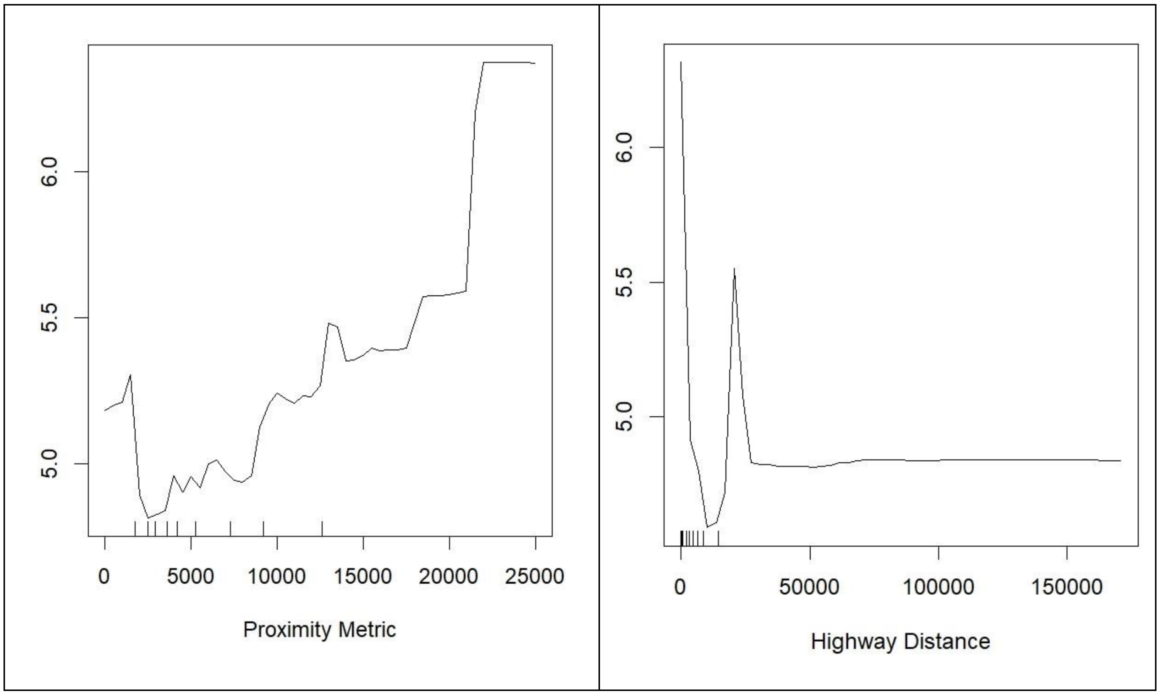

Figure 3A. (Continued)

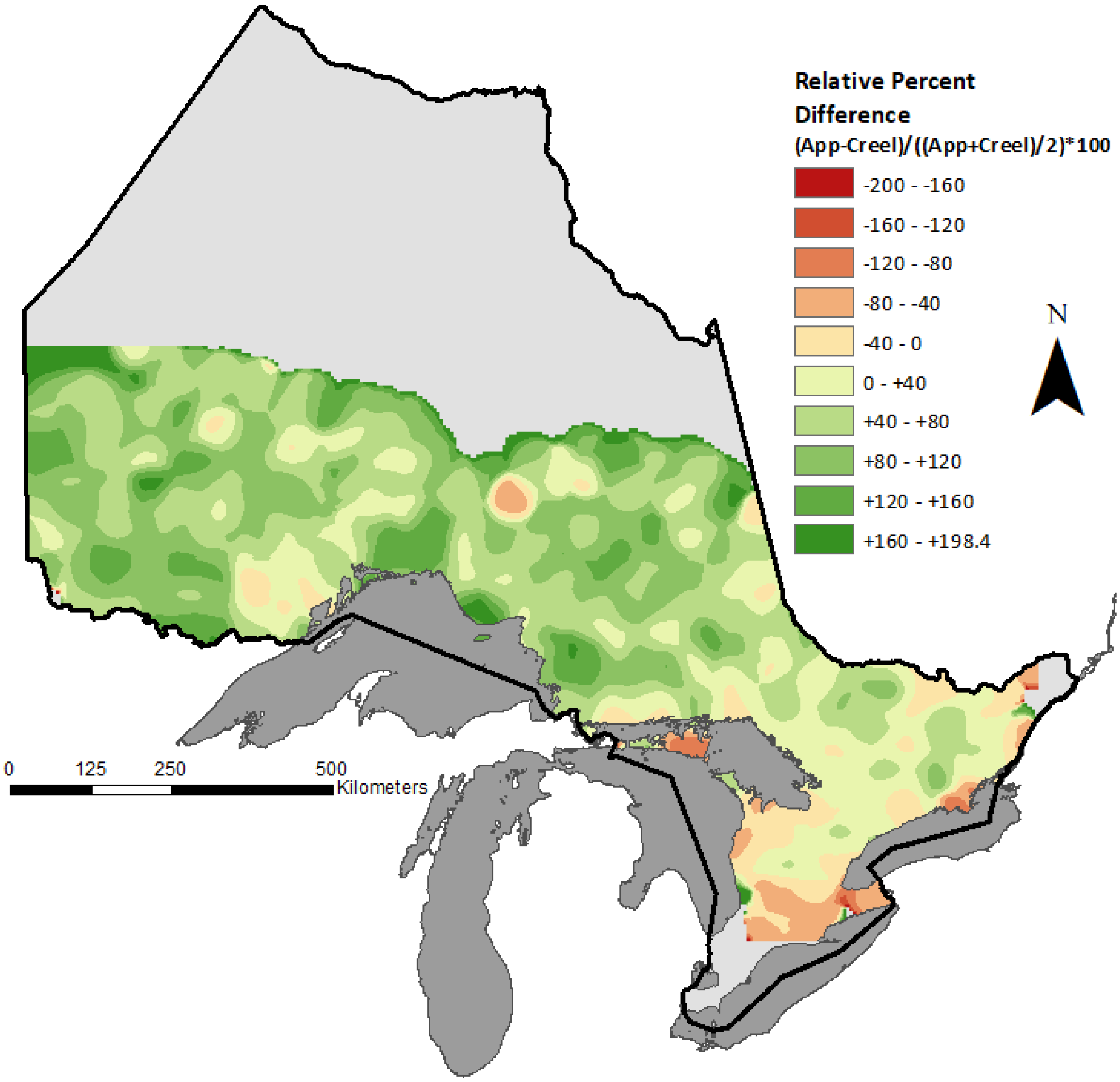

Figure 4A. The relative percent difference between kernel density rasters calculated from predictions made by random forests built using all available app data and creel data in Ontario, Canada. The average relative percent difference across the province was 58.89%.

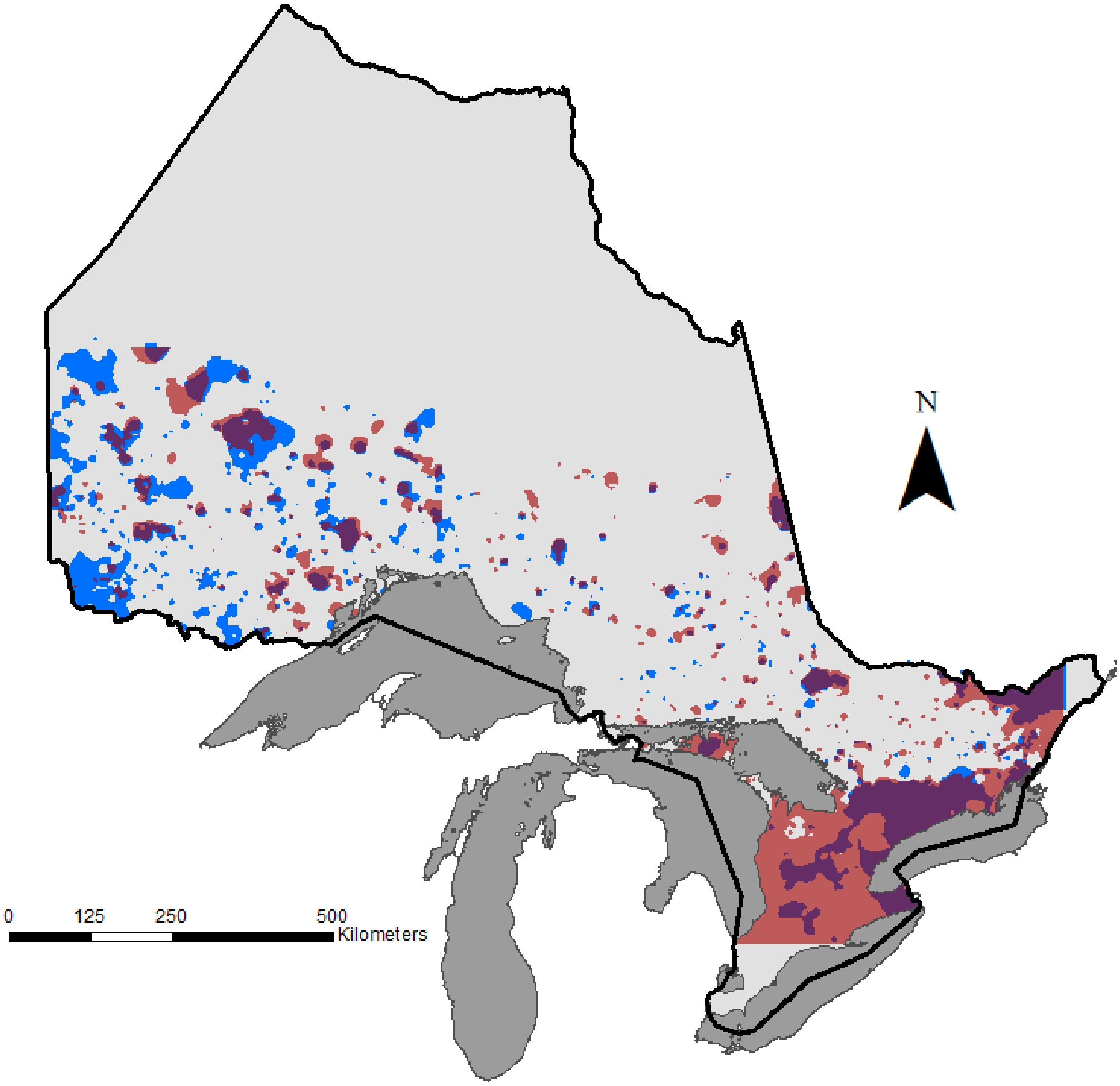

Figure 5A. The IDW spatial interpolations of the scaled, unrestricted app data (blue) and the creel data (red) indicating the location of values within the top deciles of the data (app > 0.051, creel > 0.035). Purple indicates areas where the two sources overlap. This indicates where the random forest models predicted the highest levels of activity in Ontario, Canada.