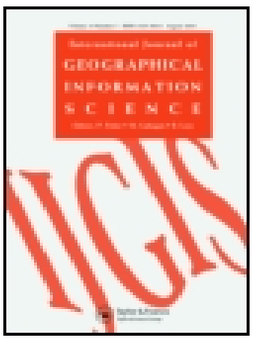

International Journal of Geographical Information Science

ISSN: 1365-8816 (Print) 1362-3087 (Online) Journal homepage: http://www.tandfonline.com/loi/tgis20

The 'cottage effect' in citizen science? Spatial bias in aquatic monitoring programs

Edward E. Millar, E. C. Hazell & S. J. Melles

To cite this article: Edward E. Millar, E. C. Hazell & S. J. Melles (2018): The 'cottage effect' in citizen science? Spatial bias in aquatic monitoring programs, International Journal of Geographical Information Science, DOI: 10.1080/13658816.2018.1423686

To link to this article: https://doi.org/10.1080/13658816.2018.1423686

Published online: 09 Jan 2018.

INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE, 2018
https://doi.org/10.1080/13658816.2018.1423686

Taylor & Francis
Taylor & Francis Group

ARTICLE

The 'cottage effect' in citizen science? Spatial bias in aquatic monitoring programs

Edward E. Millar[a], E. C. Hazell[a] and S. J. Melles[b]

[a]Environmental Applied Science and Management, Ryerson University, Toronto, Canada; [b]Chemistry and Biology, Ryerson University, Toronto, Canada

ABSTRACT

Citizen science aquatic monitoring programs often rely on opportunistic, incidental contributions, which can lead to spatial bias, the uneven geographical distribution of sample sites. It is less known how this spatial bias compares to professional monitoring activities, or how geospatial biases (e.g. terrain slope, population density, road density) influence aquatic citizen science and professional lake monitoring programs. This paper compares sample sites in Ontario's volunteer Lake Partner Program, against those identified by a stratified random sampling method currently used by the Province of Ontario, Ministry of Natural Resources and Forestry. Exploration of spatial bias within each sampling method was conducted using Kernel Density Estimation, a nonparametric approach to interpolating the spatial trend of a given variable. Results indicate that two distinct patterns of sampling clusters exist between the two datasets, suggesting a 'cottage effect' in which volunteers are more likely to sample accessible locations associated with recreation and summer home ownership. Although professional monitoring programs are not exempt from spatial bias, our research suggests that citizen science lake monitoring programs in Ontario are more influenced by natural and demographic biases related to the location, accessibility, size and general attractiveness of lakes.

ARTICLE HISTORY
Received 31 August 2017
Accepted 31 December 2017

KEYWORDS
Citizen science data; community-based monitoring; spatial bias; lake monitoring; methods comparison

1. Introduction

Over the past two decades, public participation in data collection has transformed the fields of spatial ecology, ecological mapping and geographic environmental science. The popularity of citizen science, in which volunteers actively participate in scientific research, is driven by the availability of easily operated networked technologies and user-friendly software, as well as recognition by professionals that volunteers can provide an efficient source of labor (Silvertown 2009). Citizen science is considered a form of crowd-science (Franzoni and Sauermann 2014), or crowd-sourced geographic information (See *et al.* 2016), following Brabham's discussion of crowd-sourcing as 'an online, distributed problem solving and production model whereby an organization leverages the collective intelligence of an online community for a specific purpose' (Brabham 2012, p. 394). While web-based tools allow scientists to mine this public collective intelligence, the growth in scientific crowd-sourcing also involves direct collaboration between professionals and

CONTACT Edward E. Millar Edward.millar@ryerson.ca

amateurs in the field (Dickinson *et al*. 2012). This form of citizen science is sometimes referred to as community-based monitoring (CBM) and involves volunteers travelling to physical locations to monitor the status of an area or ecosystem, thus providing environmental scientists and managers with the data required to govern natural resources and evaluate existing management strategies (Conrad and Hilchey 2011).

Though CBM does not always involve networked mobile devices or GIS tools, the practice nevertheless corresponds with Goodchild's (2007) vision of participatory geography that sees citizens as sensors of environmental change. One emerging opportunity for professional ecologists is to consolidate information gathered by volunteer monitors, scaling up local monitoring by centralizing findings into larger databases to build broader baseline data (Gouveia and Fonseca 2008). This practice remodels 'classic citizen science' (in which sparsely distributed volunteers conduct monitoring efforts as a hobby), into 'geographical citizen science', which demands greater spatial precision and accuracy, and in which the precise geographic location of the data point is essential (Haklay 2013).

1.1. *Data quality in citizen science*

Effective protocols, well-designed technological platforms and training activities can improve the accuracy of volunteer-generated data for all types of citizen science (Wiggins *et al*. 2011, Bonter and Cooper 2012, Jacobs 2016); however, the unstructured and ad hoc nature of data gathered using these methods means that crowd-sourcing is particularly susceptible to the challenges of sampling bias as a source of error. Crowd-sourced ecological data have been referred to as opportunistic data, since it is typically gathered in the absence of standardized sampling design or established field protocols (Van Strien *et al*. 2013). According to Isaac *et al*. (2014) and Geldmann *et al*. (2016), sampling bias in opportunistic data can come in four primary forms:

(1) temporal bias, referring to irregular recording effort over time;
(2) geographical bias, or spatial bias, referring to irregular coverage across the space of a given area;
(3) observation bias, referring to irregular or uneven recording per site visit;
(4) detection bias, referring to differences in volunteer abilities to detect species, leading to selective or incomplete reports.

Bias in sampling effort and intensity has the potential to generate statistical noise, obscuring true indicators of change, or generating the appearance of patterns in the data which do not, in fact, exist (Isaac *et al*. 2014). Sampling bias in species distribution data can lead to miscalculations, skewed results or false conclusions (Geldmann *et al*. 2016). Our research focuses on the second type of sampling bias: geographic or spatial bias and is from here on referred to as spatial bias.

1.2. *Spatial bias*

Availability of online citizen science programs, platforms and data repositories has increased the volume of opportunistic biodiversity data (Van Strien *et al*. 2013).

However, in the absence of a thoroughly designed sampling collection method that controls for sampling intensity, results may be skewed by survey effort and method of data collection (Fernández and Nakamura 2015). Even when data verification protocols are used to enhance the validity of citizen-generated data, geographic inaccuracy can result from positional errors in GPS technology (Rocchini *et al*. 2011).

Most sites sampled by volunteers in large-scale citizen science projects do not explicitly control for spatial bias and are unevenly distributed through space (Van Strien *et al*. 2013). Spatial sampling bias is also known to be influenced by proximity to roads and larger cities (Kadmon *et al*. 2004, Fernández and Nakamura 2015, Geldmann *et al*. 2016). A study of citizen science projects surveying 13 different taxonomic groups revealed that citizen sampling is heavily biased toward accessibility and that sample sites are influenced not only by roads and population density but also by landscape features like steepness and elevation (Mair and Ruete 2016). In a study of four major Danish citizen science projects, each with different objectives, protocols, study areas and sampling designs, volunteers oversampled agricultural areas and undersampled forests and grasslands, suggesting that citizen science databases may be more susceptible to spatial bias toward human-modified land (Geldmann *et al*. 2016). Volunteers are also known to focus their effort on protected areas such as national parks or designated wildlife zones, which are both publicly accessible and often have attractive landscapes known for biodiversity value (Boakes *et al*. 2010).

The problem of spatial bias is not unique to citizen science, and ecological data collected by professionals are often correlated with the accessibility of sample sites. In studies of biodiversity and species distributions, this can emerge from a focus on hot spots in which areas known to contain greater biodiversity are sampled more frequently by biologists (Hurlbert and Jetz 2007). These hot spots are also more likely to be close to the recorders' places of residence (Dennis and Thomas 2000) and often occur in hyperdiverse regions that are of specific interest to scientists and natural historians at the time of collection (Boakes *et al*. 2010). Factors known to influence sampling bias in studies conducted by professionals include minimum, maximum, mean elevation, elevation range, land-use variables (urban and industrial areas, irrigated croplands, non-irrigated croplands, pasturelands), distance to the home of the researcher and even 'topographically varied areas with comparatively pleasing summer temperatures, and more varied landscapes' (Romo *et al*. 2006, p. 883).

Spatial bias in volunteer sampling can lead to misguided or inappropriate observations. In a citizen science study of large-scale bird distributions, spatial bias in sampling location and effort led to results that erroneously suggested forest species prefer non-forested areas (Higa *et al*. 2015). Spatial bias can also impact conservation policy as the spatiotemporal variability of citizen science can lead to under-sampling of remote areas, which could in turn impact conservation management (Tulloch *et al*. 2013). If spatial bias is not taken into consideration, it is possible that conservation measures or environmental management strategies may miss key areas of concern, simply because these remote regions are unstudied or understudied.

Spatial bias can result in an overall picture of species distribution that is incomplete, unbalanced and potentially misleading, as the datasets from which we build our knowledge about the natural world can be 'skewed by cultural, socioeconomic, and policy constraints' (Romo *et al*. 2006, p. 873). Although spatial bias is typically studied in areas of ecology that

focus on species counts and species distribution, the impact of accessibility and attractiveness of sample sites are less understood in CBM and water resource management initiatives. To date, studies assessing data quality in aquatic citizen science and CBM focus on comparing the accuracy of physical and chemical samples collected by volunteers to those collected by professionals (Fore *et al*. 2001, Loperfido *et al*. 2010, Hoyer *et al*. 2012, Storey *et al*. 2016). Though these studies suggest that under the right conditions, volunteers can collect useful aquatic data, they do not determine how the accessibility and attractiveness of sample locations can influence *where* information is collected. Furthermore, existing research discussing spatial bias in volunteer aquatic monitoring programs does not specifically assess how spatial bias in aquatic citizen science compares to traditional methods for site selection (Deutsch *et al*. 2009, Deutsch and Ruiz-Cordova 2015, Jollymore *et al*. 2017). For instance, Deutsch et al. (2009) find that volunteer water monitoring groups may cluster around specific regions based on population dynamics and socioeconomic factors but leave open the question of the degree to which this spatial bias might differ in a program administered entirely by professional scientists.

This paper aims to explore the presence of spatial bias in two distinct sampling programs in Ontario, Canada. We compare the sampling distribution of a citizen science project (the Lake Partner Program [LPP]), which applies crowd-sourcing methods to help identify sample sites, with the sampling distribution of a large-scale government-run and professionally designed lake monitoring program, which uses a stratified random sampling method to identify sample sites (Ontario's Broad-scale Monitoring program, BsM). Both programs monitor physical and chemical characteristics of inland lakes and aim to monitor and assess the relative health of Ontario's freshwater ecosystems, though the BsM is much more comprehensive. Using spatial analysis techniques, distinct clusters (or hot spots) of each sampling lake dataset were identified. To further demonstrate explicit bias in select sampling locations, predictive models were used to explore if any statistically significant correlations exist between the number and location of lakes sampled and indicators of lake accessibility and attractiveness.

We hypothesized that both datasets will demonstrate significant spatial bias given accessibility constraints, but that (1) spatial bias will be more pronounced in the citizen science dataset, and (2) that the spatial bias in the professionally managed program will be correlated predominantly with factors related to accessibility, whereas the spatial bias in the LPP citizen science dataset will be more strongly correlated with factors related to lake attractiveness. This article is the first to directly compare the location of sampling sites predominately identified by volunteers in an aquatic citizen science program against a professionally led program, which monitors the same characteristics and which surveys the same geographic area.

2. Methods

2.1. *Study area and data sources*

The Canadian Province of Ontario has over 250,000 inland lakes, which together may contain up to one-fifth of the planet's total freshwater resources (OMNRF 2017). Lakes are vital to the culture, history and economy of the province, providing opportunities for recreation and transportation, as well as water for households, agriculture and industry.

Ontario lakes are also fundamental to the province's ecological health and biodiversity, providing habitat for thousands of native aquatic plant species, over 150 native fish species, and hundreds of species of birds.

However, the sheer size of the province and the number of lakes located within it make it extremely challenging for the government to monitor consistently and comprehensively. Ontario has an area covering 1076,395 sq. km, consisting of 917,741 sq. km of land and 158,654 sq. km of freshwater (Statistics Canada 2005). Nearly 95% of inland Ontario lakes have a surface area of under 100 ha, and approximately 50% have a surface area of under 10 ha (DESC 2015). The geographical features of the province make it expensive and difficult to monitor inland lakes, and Natural Resources Canada estimates that only 1–2% of lakes are sampled for water quality (DESC 2015). Furthermore, the human population in the province is subject to extreme spatial bias, with 98% of the population concentrated around the southern regions, within the Great Lakes Basin and Ottawa River Basin watersheds. The north of the province is remote, spatially expansive and largely inaccessible by road, which makes it difficult and expensive to sample. The spatial distribution of human population does not correspond to the spatial distribution of lakes. Roughly 18% of northwestern Nelson Basin and roughly 8% of northern Hudson basin are covered by lakes, whereas in the Great Lakes/Ottawa river basins, inland lakes (excluding the Great Lakes) cover roughly 6% of the land area (Dove-Thompson *et al*. 2011).

2.2. *BsM program*

The Ontario Ministry of Natural Resources and Forestry (OMNRF) approach to fisheries management has traditionally focused on individual lakes and was historically reactive rather than proactive (Lester *et al*. 2003). Environment managers would conduct lake monitoring and develop water-related policies primarily in response to indications from anglers, cottagers or other stakeholders that fish levels were declining (Lester *et al*. 2003). Though this reactive management approach incorporated input from public stakeholders to identify key areas in need of attention, it was less capable of producing a general assessment of the health of fish stocks throughout the province. In 2004, the OMNRF established the Ecological Framework for Fisheries Management, which sought to help the Ministry increase public participation in fisheries management and simplify existing regulations (OMNRF 2005, 2009). Part of this initiative involved long-term monitoring of inland lakes at the landscape scale, shifting from a reactive management approach where lakes were monitored in response to feedback from stakeholders, and toward an effort to provide comprehensive and long-term monitoring of inland lakes. The resulting BsM program assesses lakes for water quality, water chemistry, fish levels and productive capacity during 5-year cycles (Sandstrom *et al*. 2013), the first cycle of which was completed in 2012. The BsM monitors Ontario's 20 Fisheries Management Zones (FMZs). FMZs are administrative boundaries that were established by the province to facilitate fisheries management at the landscape level, which each has unique regulations, strategies and evaluation processes (OMNRF 2012). These boundaries were determined based on a combination of human factors (such as angling activity) and ecological factors (watersheds, climate, road networks) (OMNRF 2012). The BsM applies a stratified random sampling method to select sample sites, identifying lakes with surface areas between 50 and 250,000 ha, and randomly selecting a representative number of

lakes from each of the twenty FMZs. The OMNR calculates that more than 18,000 lakes in the province meet these size criteria, 92% of which are situated within northeast and northwest Ontario, and 8% of which are located within southern Ontario (OMNRF 2016).

2.3. *LPP*

The LPP is a citizen-science initiative that has monitored water quality levels in Ontario lakes since 1996. The LPP began as a project initiated by the Federation of Ontario Cottagers' Associations (FOCA), the Lake of the Woods District Property Owners' Association, and, in 2002, these groups established a partnership with Ontario Ministry of the Environment at the Dorset Environmental Science Centre (DESC). Volunteers collect total phosphorus and Secchi depth samples from lakes across Ontario, which are then analyzed by the DESC's chemistry laboratory. The program attracts roughly 600 volunteers annually, who collect information about water quality and water clarity at approximately 800 sampling locations within nearly 550 lakes (DESC 2015). Volunteers measure concentrations of total phosphorous (TP) as a measure of nutrient status and potential for algal growth. Volunteers also measure turbidity, or water clarity, which can be an indicator of dissolved organic carbon and can complement information gained from TP samples (DESC 2015). In 2008, the LPP also began measuring concentrations of calcium as a means of detecting trends and discovering correlations with TP concentrations.

The LPP provides volunteers with the necessary materials to collect their samples in the form of a 'Lake Stewards' sampling kit. The LPP also provides a PDF of sampling instructions for volunteers to follow, which describes the protocols for collecting phosphorus and Secchi depth samples. The FOCA provides a video with further instructions on how to collect samples, and DESC provides volunteers with 100 ml jars, tubes, a filter and a funnel. Volunteers collect samples according to the protocol and then mail the samples and materials back to the DESC laboratory. Water chemistry protocols administered by the LPP are generally considered to be robust. The LPP has compared the accuracy of samples generated by volunteers against the accuracy of samples generated by Ministry of Environment scientists, Conservation Authority scientists or municipal representatives and found no statistically significant difference, indicating that volunteers can collect meaningful field samples (DESC 2013).

2.4. *Identifying spatial bias*

To test for the presence of spatial bias within the LPP or BsM sampling programs, we divided the Province of Ontario into square (800 m × 800 m) zones of equal size, 6400 ha squares. To minimize the number of zones with no data, the study area was limited to the Great Lakes Basin. This resulted in 4920 zones that contained 398 BsM and 2818 LPP sample sites. Kernel Density Estimation (KDE) was used to explore the degree of clustering present in each dataset, by providing a graphic depiction of the density of point samples within a given geographic area. The Mann–Whitney *U* test was used to test whether the median proportion of lakes sampled per (6400 ha) zone by the BsM and LPP datasets was equal. Principle component analysis (PCA) was used to create and summarize 'cottage effect' indices to reduce collinearity among descriptor variables (e.g. population density and road density) and to extract unique indices that capture the

accessibility and attractiveness components of landscape features in a given zone. Poisson generalized linear model (GLM) was used to model how the count of lakes per zone was related to dominant cottage effect predictor variables. The following Poisson GLM (Zuur *et al*. 2009) was applied to each sampling dataset using R (version 3.3.2), where the response variable Y_i, the number of sampled lakes per zone i, has a Poisson distribution with mean intensity (μ):

$$\mathrm{Ln}\,(\mu_i) = \eta\left(X_{i1}, \ldots, X_{iq}\right)$$

The systematic component is given by $\eta(X_{i1}, \ldots, X_{iq}) = \alpha + \beta_1 \times X_{i1} + \ldots + \beta_q \times X_{iq}$, for any set of explanatory variables, in this case the cottage effect predictor variables (Zuur *et al*. 2009).

2.5. *KDE*

KDE is a spatial approach to measuring both the intensity and extent of a specific feature (i.e. sample sites). Given a set of samples, $S = \{x_i\}$, where $i = 1 \ldots n$, an estimate of density $p(x)$ can be calculated as

$$p(x) = 1/N \sum_{k=1}^{n} K\sigma(x - xi)$$

where $K\sigma$ is a kernel function with a bandwidth scale σ (Elgammal *et al*. 2002). KDE uses the presence of specific feature locations to interpolate a continuous density surface by fitting a series of kernels over each feature (King *et al*. 2015). The bandwidth of the kernel is predetermined by the user to depict the surrounding neighborhood of each feature, whereby the density of the feature is smoothed out across the radius of the kernel (Thornton *et al*. 2011, Chen 2015) and nearby observations are given greater weight (Charpentier and Gallic 2016). Due to the exploratory nature of this method, the optimal bandwidth was calculated using a variant that accounts for spatial outliers, specifically Silverman's rule of thumb or Gaussian approximation that minimizes the mean integrated squared error (or L_2 risk function, Silverman 1986). This is an accepted approach to predicting spatial trends across a variety of disciplines including relative health outcomes (Maroko *et al*. 2009, King *et al*. 2015), environmental monitoring (Lin *et al*. 2011) and targeted sampling efforts (Yenilmez *et al*. 2015). KDE analyses were conducted using ArcGIS 10.4 spatial analyst extension.

2.6. *Mann–Whitney U test*

The Mann–Whitney *U* test can be used to compare the central tendencies of two independent samples; we used this test to compare the distribution of sampled lakes per 6400 ha square zone in the LPP and BsM datasets. This nonparametric approach tests whether a random variable from one cumulative sample distribution is likely to be greater than a random variable from a second cumulative sample distribution (Mann and Whitney 1947) and the Mann–Whitney *U* test is thought to be more robust than the parametric equivalent *T* test against uneven sample distributions and outliers in general. The total number of sample sites per zone was summarized for each dataset and the

proportion (% of total sites) was used for the Mann–Whitney *U* test to determine if the proportions of lakes sampled per zone were equal for BsM and LPP methods.

2.7. *Cottage desirability index*

To explore potential spatial bias in citizen science sampling techniques, we determined geographic variables that may exert an influence on the likelihood that any given lake will be sampled by either the citizen-based method (LPP) or the stratified random sampling method (BsM). Four high-level indicators were identified that are thought to contribute to accessibility and attractiveness of sample sites and include infrastructure, landscape/terrain, remoteness and protected/natural cover areas. These indicators were further broken down into 13 variables (see Table 1). These variables were selected as a measure of the likelihood that an area would be a desirable location for a recreational summer home, or cottage. Summer cottages located within wilderness areas have a symbolic importance within Ontario culture, and regular trips to these secondary recreational cabins and homes have been a major feature of Ontario urban life since the Post-war era (Stevens 2013).

To quantify spatial bias, a composite index of cottage desirability was created using these geographic variables and PCA. PCA is a widely accepted multivariate statistical technique used to reduce dimensionality of large datasets. PCA reduces the number of variables, while maximizing explained variance among the original variables (Jolliffe 2002, Mandelik *et al*. 2010, Bro and Smilde 2014). Ultimately, the method combines many variables into a small number of principal components that account for varying amounts of variance within the dataset (Ou *et al*. 2012).

2.8. *Variable selection*

Road length and road density were both included as measures of cottage accessibility. Roadside bias is a recognized problem in ecology, with areas closer to major roads facing a greater likelihood of being sampled, a phenomenon that is referred to as the 'highway effect' (Fernández and Nakamura 2015). This can lead to a distortion in data and can reduce the accuracy of predictive models given that road networks are not representative of the overall geographic and ecological conditions in a region (Kadmon

Table 1. Descriptive statistics of explanatory variables included in the principal component analysis.

Indicator	Description	Mean	Std. Dev.	Range
Elevation	Ave. elevation (m)	302	112	39–550
	Min. elevation (m)	259	103	4–497
	Max. elevation (m)	371	134	52–688
Slope	Ave. slope (°)[a]	2	2	0–10
	Max. slope (°)[a]	19	9	0–70
Road length	High volume roads (m)[a]	21,465	28,406	0–286,115
	Local urban roads (m)[a]	35,306	45,414	0–517,269
	Resource roads (m)[a]	21,155	23,239	0–162,263
Lakes	Lake size (sq. m)[a]	5113,589	8699,052	0–64,160,100
	Number of lakes[a]	57	55	0–526
	Length of shoreline (m)[a]	72,341	61,054	0–602,896
Pop. density	People per zone[a]	2575	16,276	0–394,664
Protected area	Total area (sq. m)[a]	14,207,788	20,057,830	0–64,160,100

[a]Variable was log transformed for statistical analysis.

et al. 2004). The Ontario road network features were obtained from Land Inventory Ontario (LIO), an open data source that contains geographic data for the province. The road network is updated on a weekly basis and includes attribute information such as road class, which was used to further categorize the road network into three classes (see Table 1), summarized by total length per zone.

Lake size and available shoreline were included to provide further measures of accessibility and attractiveness, as recreational lake visitors tend to favor larger, deeper lakes which can be easily accessed through entry points along the shoreline. Lake features were derived from LIO, specifically the Ontario Hydro Network (OHN) waterbody shapefile that provides geographic features of surface water polygons. Only waterbodies categorized as permanent lakes (i.e. containing water for more than 9 months of the year) were extracted and summarized per zone. Shoreline features were also obtained from LIO, specifically the OHN shoreline shapefile that represents a dividing line where water meets land. Only lake shoreline features were included and further summarized by total length for each zone.

Total protected area (m^2) within each zone was included as a measure of attractiveness as well as tourist accessibility. Infrastructure provided in protected areas, such as walking trails, camping sites and boat put-ins, provide additional opportunities for citizen scientists to marry volunteer efforts with vacation time. A protected lands shapefile was created by merging shapefiles that delineate provincial/federal parks, environmentally sensitive areas, conservation areas and natural heritage systems.

Population density accounts for potential number of citizen volunteers as well as the relative remoteness of sample sites, which relates to site attractiveness. Population density was derived from Statistics Canada 2016 Census. Using dissemination areas (small geographic regions of approximately 400–700 persons), the total population per zone was estimated.

3. Results

3.1. *Spatial clustering*

Results from the density probability mapping exercise (KDE) for both the BsM and the LPP sample lakes are depicted in Figure 1. Hot-spot mapping enables the reader to quickly and easily visualize significant clusters across the study area and compare the geographic distributions of two sampling techniques. The resulting surface maps suggest that the extent of clustering for the BsM sampling technique is greater or more spread out, while the intensity of clustering is more concentrated for the LPP dataset.

While the distribution of surface lakes across the Canadian Shield is fairly even, both sampling distributions demonstrate a nonrandom pattern of site locations concentrated mainly within the Great Lakes Basin, which also adheres to relative accessibility (i.e. roads) within the Province of Ontario. However, the LPP dataset indicates an explicit clustering pattern of sites near popular areas for outdoor tourism and seasonal homes, a region which is known in Ontario as 'cottage country'. This region is a popular recreational area based on its accessibility, as it is situated within a comfortable driving distance of major metropolitan areas and population centers. The region is also attractive for its natural features as it is characterized by coniferous boreal forests, wilderness

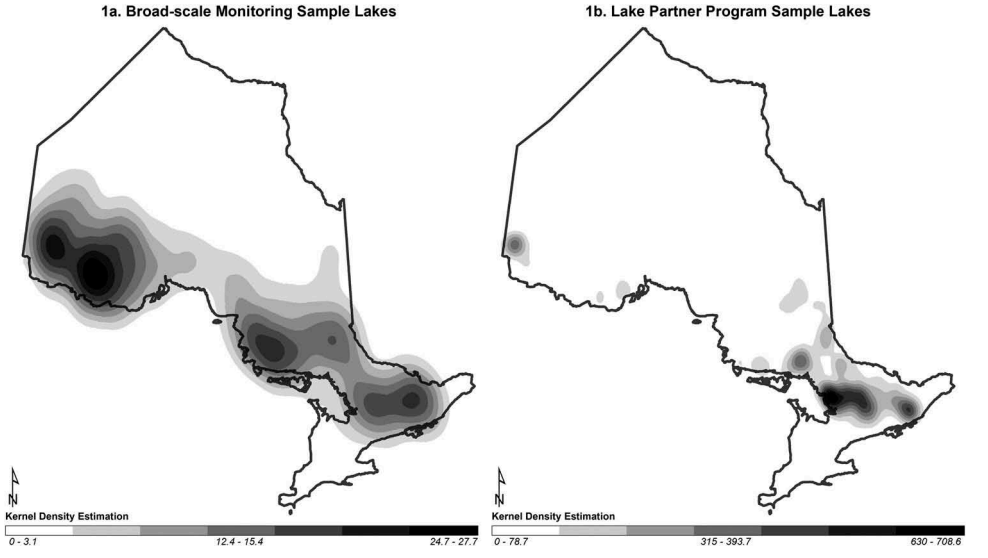

Figure 1. Kernel density estimation models of two sampling techniques in Ontario ((a) Broad-scale Monitoring Sample Lakes and (b) the Lake Partner Program Sample Lakes), whereby the darker zones represent increased probability of lake sampling.

areas and a picturesque Canadian Shield landscape that is an important feature of the province's culture and identity.

These findings supported our hypothesis that sample lakes monitored by volunteers from the LPP citizen science program exhibited more intense clustering than lake locations selected by the BsM's stratified random sampling method. The presence of significant nonrandom clustering within the BsM dataset can be understood in the context of the BsM's protocol, which selects a representative number of lakes from each of the 20 FMZs. These zones are not equal in size. Four larger zones cover the Northern region of the province, whereas 13 smaller zones cover the middle section of the province, i.e. the area that exhibits clustering in the BsM dataset (OMNRF 2012). These results were also reinforced by the Mann–Whitney *U* test that suggests the median rank proportion of LPP lakes sampled per zone was significantly ($p < 0.001$) larger than the proportion of BsM lakes sampled per zone. This allowed us to reject the null hypothesis that the spatial sampling intensity was equivalent in the two programs.

3.2. *Modeling the cottage effect*

Using an open source statistical package, version 3.3.2 of the *R* Project for Statistical Computing, a PCA was run on 13 variables that define accessibility and attractiveness for Ontario lakes. A correlation matrix was chosen for the PCA to ensure that all variables included had equal weight and that variables with greater variance or range (see Table 1) did not dominate the first principal component.

There are two important outputs from the PCA, including component scores and loadings. Component scores (see Figure 2) represent the newly transformed variable where each value corresponds to a particular case or record within the dataset (Jolliffe

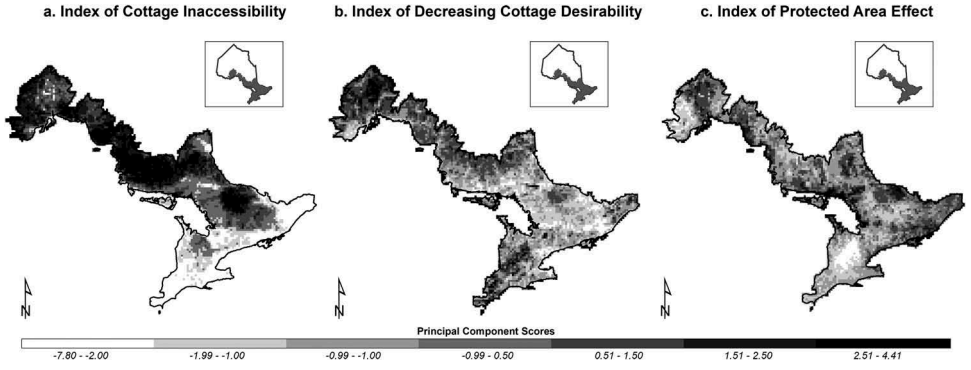

Figure 2. Composite indices describing the 'cottage effect' for Ontario lakes. Each map depicts the relative composite index score (i.e. PC1, PC2, PC3), within each zone and darker zones represent a higher component scores. Specifically, (a) depicts PC1 factor scores, whereby the darker zones indicate areas of high elevation and low population density. (b) depicts PC2 factor scores, whereby the darker zones indicate zones with fewer lakes and less shoreline. Finally, (c) depicts PC3 factors scores, whereby the darker zones indicate zones with more environmentally protected areas present. (a) Index of cottage inaccessibility, (b) Index of decreasing cottage desirability, (c) Index of protected area effect.

2002, Li *et al*. 2012), while the principal loadings represent the relative weight that a variable has on the principal scores. Specifically, the loadings can help to describe the principal components within the context of the original variables. Table 2 depicts the loadings from the first three principal components that account for approximately 68.7% of total variance in the data cumulatively.

The first principal component (PC1) accounts for 43.9% of the variance in the original descriptor variables and is predominantly an index of (lake or cottage) inaccessibility. The regression models indicate that for both datasets, the probability of sampling increases with our indicators of inaccessibility. This can be interpreted as a function of the physical geography of the province and the distribution of lakes. In Ontario, population density is concentrated within the southern region, which sits at a lower elevation and has a lower density of inland lakes. As PC1 scores increase, elevation and

Table 2. Principal component loadings for PC1, PC2 and PC3.

Indicator	Description	PC1	PC2	PC3
Elevation	Ave. elevation (m)	0.357	0.116	−0.401
	Min. elevation (m)	0.323	0.182	−0.447
	Max. elevation (m)	0.377	–	−0.274
Slope	Ave. slope (°)[a]	0.332	−0.227	–
	Max. slope (°)[a]	0.303	−0.196	0.231
Road length	High volume roads (m)[a]	−0.264	−0.343	−0.287
	Local urban roads (m)[a]	−0.245	−0.374	−0.328
	Resource roads (m)[a]	0.154	−0.119	–
Lakes	Lake size (sq. m)[a]	0.256	−0.276	–
	Number of lakes[a]	0.205	−0.489	–
	Length of shoreline (m)[a]	0.245	−0.37	0.23
Pop. density	People per zone[a]	−0.305	−0.357	−0.189
Protected area	Total area (sq. m)[a]	0.1	–	0.464

[a]Variable was log transformed for statistical analysis.
Loading coefficients correspond with the original explanatory variables included, whereby higher coefficients explain more variation in the transformed variables or principal component.

slope increase, whereas population and road density decrease (see Figure 2(a)). We interpreted this as a measure of cottage inaccessibility, where sampled lakes with high PC1 scores are farther from roads and population centers, and these areas tended to have steeper slopes at higher elevations.

The second component (PC2) accounts for approximately 15.1% of the total variation in the original descriptor variables and is predominantly an index of lake and shoreline density. As PC2 increases, the number of lakes in a zone decreases as does the relative amount of shoreline, making this an index of decreasing cottage desirability (see Figure 2(b)). The third and final component included in this analysis (PC3) is primarily an index of the protected area effect and accounts for approximately 9.7% of total variation in the descriptor variables. Figure 2(c) depicts a map of this index. Presumably, cottages are considered more attractive if situated near protected areas (i.e. provincial parks and environmentally significant areas), even if they are more difficult to access (i.e. at a higher elevation).

Tables 3 (BsM) and 4 (LPP) depict the regression models for the two response variables including the BsM and LPP sampling techniques, respectively. PC1 had twice as much influence on model estimates of the number of BSM sampled lakes than the number of LPP sampled lakes in a given zone (coefficients 0.363 and 0.155, respectively) suggesting that the BSM sampling design is not as influenced by lake accessibility as LPP sampled lakes. The coefficient is positive, meaning that as lakes become more inaccessible, the likelihood of sampling a lake via BsM is greater than the likelihood of sampling lakes via LPP.

PC2 had a smaller influence on model estimates of the number of BsM sampled lakes than on number of LPP sampled lakes in a given zone or square (coefficients −0.417 and −1.183, respectively). This suggests more of a cottage effect in the LPP sampling design.

Table 3. Poisson regression coefficients for a model of the cottage effect on number of lakes sampled per 6400 ha zone in the broad-scale monitoring program, Ontario.

Coefficients	Estimate	Std. error	*z* Value
(Intercept)	−2.945***	0.083	−35.488
PC1	0.363***	0.033	10.928
PC2	−0.417***	0.060	−6.939
PC3	0.186**	0.066	2.800
PC1:PC2	0.0129	0.030	0.428
PC1:PC3	−0.110***	0.030	−3.616
PC2:PC3	−0.051	0.047	−1.105
PC1:PC2:PC3	0.013	0.026	0.482

**p Value < 0.01; ***p value < 0.001.
Coefficients represent the relative effect or predictive capability each composite variable has on the response variable.

Table 4. Poisson regression coefficients for a model of the cottage effect on number of lakes sampled per 6400 ha zone in the Lake Partner Program, Ontario.

Coefficients	Estimate	Std. error	*z* Value
(Intercept)	−1.515***	0.038	−40.014
PC1	0.155***	0.016	9.708
PC2	−1.183***	0.023	−51.996
PC3	0.047	0.033	1.418
PC1:PC2	−0.059***	0.012	−4.853
PC1:PC3	−0.050**	0.016	−3.200
PC2:PC3	0.110***	0.020	5.451
PC1:PC2:PC3	−0.004	0.012	−0.355

**p Value < 0.01; ***p value < 0.001.
Coefficients represent the relative effect or predictive capability each composite variable has on the response variable.

Specifically, as the number of lakes and the length of the shoreline within a zone decreases, the number of sampled lakes also decreases, but the decrease was far sharper in the LPP sampling design.

Finally, the index of protected area (or PC3) had a greater effect on the BsM sample sites compared to the LPP lakes (coefficients 0.186 and 0.047, respectively). Moreover, the effect estimate of PC3 was not statistically significant for the LPP sample lakes, implying that the protected area index was a better predictor of BsM sampling regimes. However, total protected area was thought to represent a lake attractiveness score, whereby lakes situated in or near protected areas are thought to be more attractive than those outside these areas. One explanation for this discrepancy could be the measure by which protected area was included in this study. Some 6400 ha zones in our study area were entirely covered by environmentally sensitive or protected area, and most cottages would not be located directly within such areas due to development restrictions. Therefore, we believe that a proximity variable (such as distance to protected area) may have provided more predictive power in terms of the LPP sample lakes.

4. Discussion

Our results confirm existing research indicating that ecological data collection is spatially biased (Dennis and Thomas 2000, Kadmon *et al*. 2004, Romo *et al*. 2006, Phillips *et al*. 2009, Rocchini *et al*. 2011, Fernández and Nakamura 2015). The results contribute to findings from previous studies that indicate data collected using citizen science methods may be even more biased by geographic, demographic and sociocultural factors such as accessibility and attractiveness of sampling sites (Tulloch *et al*. 2013, Ruete 2015, Maldonado *et al*. 2015, Mair and Ruete 2016). Existing studies investigating spatial bias in citizen science have concentrated primarily on species counts, suggesting that the location and recording intensity by volunteers has influenced available data about biodiversity and species distribution. Our results extend this discussion to suggest that volunteer water monitoring programs are also subject to an uneven distribution of sampling effort.

4.1. *Impact of spatial bias on lake monitoring*

To date, studies on data quality in aquatic citizen science have placed more focus on volunteer skill and ability (Fore *et al*. 2001, Sharpe and Conrad 2006, Loperfido *et al*. 2010, Cox *et al*. 2012, Hoyer *et al*. 2012), with less attention to spatial bias, although notable exceptions exist (Deutsch *et al*. 2009, Deutsch and Ruiz-Cordova 2015, Jollymore *et al*. 2017). In a comparison of citizen-collected and researcher-collected water samples from active streams and rivers, Jollymore et al. (2017) found that samples collected by members of the public had higher concentrations of NO_3, which could have resulted from volunteers deciding to sample environmentally degraded sites, such as urban storm-water lagoons. A study of the Alabama Water Watch, a long-running CBM initiative, found that volunteer monitoring efforts were correlated with factors such as education, income and population density (Deutsch and Ruiz-Cordova 2015). Volunteers tended to be wealthy and well educated, and their monitoring efforts were concentrated around regions that were closer to

population centers, and which had higher levels of wealth and education (Deutsch and Ruiz-Cordova 2015).

Correspondingly, participation tended to wane in rural areas with fewer resources, unless these areas contained specific lakes whose size or location meant that they were unique and of significant concern (Deutsch and Ruiz-Cordova 2015). While our study did not directly account for the impacts of wealth and education on sampling sites, we nevertheless observed that the highest concentration of sampling sites for the LPP was clustered around the lower Canadian shield, within the Muskoka region. This is an area known within Ontario as 'cottage country'. As a popular summer recreation destination for wealthy families in the more populous region of southern Ontario, the concentration of sampling sites around this area suggests that volunteer monitoring programs may be correlated with wealth.

4.2. *Motivations for monitoring*

The difference in sampling design between the government-run stratified random sampling design and the methods used by the LPP, which relies on data submitted by volunteers, also corresponds with previous studies suggesting that volunteers and scientists may approach collaborative research projects with different objectives and motivations (Ellis and Waterton 2004, Lawrence and Turnhout 2010, Cornwell and Campbell 2012, Jalbert and Kinchy 2016, Kinchy 2017). Whereas scientists are likely to prioritize robustness and reliability of data and natural resource, managers may be motivated by obtaining a neutral and comprehensive overview of the state of a given resource so that they may make informed policy decisions; citizen volunteers often approach projects with their own motivations, interests and agendas (Lawrence 2006). Professional scientists and citizen scientists may share common values and goals, but volunteers participating in citizen science may be more driven by the desire to see a specific result come to fruition than they are to participate in basic research objectives for the sake of science (Weng 2015). This may be particularly true in aquatic citizen science projects, which have the potential to lead to policy change or conservation outcomes.

Indeed, volunteers may experience frustration if they perceive that they are 'monitoring for the sake of monitoring', without the possibility of a more concrete outcome from their efforts (Sharpe and Conrad 2006, p. 403). This may be particularly true in aquatic citizen science, since water is a resource with deeply political dimensions related to community health and environmental justice (Sharpe and Conrad 2006, Jalbert and Kinchy 2016, Kinchy 2017). These motivations may, consciously or unconsciously, shape the apparently innocuous decision of where a volunteer takes their water sample. To an extent, this subjectivity and sense of personal meaning are embedded within the design of the LPP project, which reaches out to cottage owners, who may have a vested interest in monitoring the quality of their cottage lake. Our results demonstrate how this approach can have a cumulative effect of concentrating sampling locations toward areas that are of interest to volunteers. Volunteers may be motivated to participate in sampling based on their personal connection to a lake, as well as a desire to obtain information about the water quality for drinking or recreation. While geographic bias could be reduced by placing-specific limitations on where volunteers can collect their lake water samples, such constraints may result in decreased numbers of participants (Jollymore *et al.* 2017).

4.3. *Fitness of purpose*

These findings also demonstrate how scale can impact the fitness of purpose for volunteer-generated data. Proponents of crowd-sourced science often emphasize the capability for volunteer-derived data to help scientists work at larger temporal and spatial scales that would not be possible if professionals were responsible for both data collection and analysis (Goodchild 2007, Devictor *et al*. 2010, Franzoni and Sauermann 2014). Devictor et al. (2010, p. 354) pointed out that volunteers have enabled spatial ecologists, environmental geographers and conservation biologists to move 'beyond scarcity' and investigate large-scale patterns and processes.

The issue of scale is sometimes considered to enhance the reliability of citizen science, based on the assumption that any individual errors generated by less experienced volunteers will be eclipsed by the total volume of observations. Kolok et al. (2011, p. 629) referred to this as an inherent 'self-correctiveness' of citizen science, in which false positives, false negatives and erroneous results are often amended by subsequent recording from additional volunteers. However, the promise of self-correctiveness in crowd-sourcing does not necessarily apply to datasets that are spatially biased. If sampling locations are influenced by factors related to accessibility and attractiveness, then greater numbers of volunteers may amplify spatial bias rather than correcting it. While the results from the phosphorous, calcium and turbidity tests may provide a reliable indication of the status of any given sampled lake, the results may not offer a representative overview of the overall physical and chemical properties of Ontario lakes at broad spatial scales if sampling is concentrated and clustered based on accessibility and attractiveness of sampling sites.

Within Ontario, the testing of physical and chemical characteristics of lakes at the provincial scale has largely been conducted by government employees, in relation to reactive fisheries management (Lester *et al*. 2003). The BsM program was developed partly out of a recognition that government monitoring programs had traditionally been subject to spatial bias, which had in turn impacted natural resource management. In their paper proposing a shift toward a broad-scale adaptive management approach to regulating recreational fisheries, Lester *et al*. (2003) acknowledged that effective management requires accounting for the behavioral dynamics of anglers in addition to the population dynamics of fish. The historical management approach concentrated management action around heavily used and economically significant lakes, which generated large blind spots when it came to obtaining an objective assessment of the overall state of the quality of Ontario lakes.

In a region that is as large and lake-dense as Ontario, obtaining a comprehensive picture of the state of water quality is impractical to achieve province-wide using either professionals or citizen volunteers. In general, aquatic and hydrological monitoring is expensive and requires repeated measurements over long time periods, specialized knowledge and advanced technologies beyond the scope of a normal budget (Buytaert *et al*. 2014). Over the past few decades, government funding for environmental science has decreased, and community-based aquatic monitoring groups have formed in response to a perceived retreat of the state in its role as an environmental monitor and regulator (Savan *et al*. 2004, Sharpe and Conrad 2006).

In the early stages of community-based water quality monitoring, some natural resource administrators were reticent about using citizen-generated data out of concern

that members of the public lacked the expertise required to competently gather information relevant to scientists and policy-makers (Sharpe and Conrad 2006). There is some indication that this perspective is shifting as more evidence emerges that citizen science can produce comparable or nearly equivalent data about physical and chemical characteristics of lake water (Fore *et al*. 2001, Loperfido *et al*. 2010, Hoyer *et al*. 2012). However, if citizen science data are to be used to inform policy, it is important to recognize that spatial bias in sampling effort may influence resulting analyses. We fully acknowledge that citizen science-based programs such as the LPP often have a different suite of goals and objectives than government-led programs, and we further acknowledge that the data collected are valuable and reliable. Our results pertain to the use of these data for broad scale analysis, which we infer would require estimates of spatial uncertainty, and an in-depth understanding of the implications of that uncertainty on the resulting inferences made.

Furthermore, statistical methods exist which can account for spatial bias in sampling effort, including subsampling (Phillips *et al*. 2009, Isaac *et al*. 2014) and sample weighting (Stolar and Nielsen 2015). Modeling approaches such as generalized linear and additive mixed models, mixed-effects models and hierarchical models can be used when systematic bias can be found in the sampling design (Bird *et al*. 2014). Ruete (2015) and Mair and Ruete (2016) developed algorithms to generate ignorance scores and 'maps of ignorance' that can help quantify presence of spatial bias in recorder distributions when biodiversity datasets rely on citizen science methods. It should also be noted that the presence of spatial bias at large scales does not reduce the importance of reaching out to citizens to assist in lake monitoring. The LPP has conducted tests that indicate citizen scientists are capable of sampling with comparable accuracy to professionals (Dorset Environmental Science Centre (DESC) 2013). For individuals who participate in the program to help monitor lakes that they use, cottage at or live near, the LPP data and protocols are suitable.

We suggest avenues for future research that explore how the presence of spatial bias might impact inferences made about water quality at broad scales. We expect to see that spatial bias may lead to misinterpretation or over and under-emphasis of pollution hot spots and changes through time (i.e. with fluctuating high phosphorous readings), particularly in situations where there are insufficient processed-based data or knowledge about how environmental factors can influence water chemistry and water clarity. For example, landscape and geologic factors such as lake position within lake chains (Kratz *et al*. 1997) and soil type can influence water chemistry (Dillon and Kircher 1975, Soranno *et al*. 1996), but these data are not readily available for lake-sheds at broad spatial scales.

5. Conclusion

Volunteers are becoming increasingly recognized as an important and cost-effective source of data for ecologists and natural resource managers. This may be especially true in aquatic monitoring, which often requires a widespread geographic distribution of sample sites and expansive spatial coverage that is cost-prohibitive when undertaken by professionals. While government agencies across North America have fostered partnerships with community-based aquatic monitoring groups and have established protocols and training mechanisms to ensure the quality of citizen-generated data, these programs nevertheless rely to some degree on opportunist and incidental contributions by

volunteers. This can lead to a clustering of sample sites toward areas that are accessible to the public, with more road infrastructure, and which are within a comfortable driving distance of population centers. Additionally, crowd-sourced aquatic monitoring can be influenced by the degree to which a given sample site is attractive as a leisure destination, and volunteers may be more drawn toward larger, picturesque lakes. While professional monitoring programs are also subject to spatial bias, results indicate that there was more spatial bias in the citizen science dataset than in the government-run lake monitoring program. Furthermore, our study indicates that while both datasets were biased toward the variables associated with accessibility (elevation, slope, road density), the citizen science dataset was also more heavily biased toward variables related to the subjective attractiveness of lakes, including lake density and shoreline access. Overall our findings contribute to the existing literature on citizen science, which suggests that greater attention must be paid to the sociological, demographic and cultural influences that shape both the characteristics of volunteer-based research and the data that these methods generate.

Acknowledgments

The authors would like to thank the three anonymous reviewers whose insightful comments and feedback helped improve this article.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This research was supported in part by a Natural Sciences and Engineering Research Council of Canada Discovery grant to SJM: [Grant Number 03834-2015], and to EM through participation in the Social Science and Humanities Research Council of Canada partnership grant on 'How the Geospatial Web 2.0 is Reshaping Government – Citizen Interactions' [Grant Number 895-2015-1023].

References

Bird, T.J., *et al*., 2014. Statistical solutions for error and bias in global citizen science datasets. *Biological Conservation*, 173, 144–154. doi:10.1016/j.biocon.2013.07.037

Boakes, E.H., *et al*., 2010. Distorted views of biodiversity: spatial and temporal bias in species occurrence data. *PLoS Biology*, 8, 6. doi:10.1371/journal.pbio.1000385

Bonter, D.N. and Cooper, C.B., 2012. Data validation in citizen science : a case study from Project FeederWatch. *Frontiers in Ecology and the Environment*, 10 (6), 305–307. doi:10.1890/110273

Brabham, D.C., 2012. The myth of amateur crowds: critical discourse analysis of crowdsourcing coverage. *Information, Commuication & Society*, 15 (3), 394–410. doi:10.1080/1369118X.2011.641991

Bro, R. and Smilde, A.K., 2014. Principal component analysis. *Analytical Methods*, 6 (9), 2812–2831. doi:10.1039/C3AY41907J

Buytaert, W., *et al*., 2014. Citizen science in hydrology and water resources: opportunities for knowledge generation, ecosystem service management, and sustainable development. *Frontiers in Earth Science*, 2 (October), 1–21. doi:10.3389/feart.2014.00026

Charpentier, A. and Gallic, E., 2016. Kernel density estimation based on Ripley's correction. *Geoinformatica*, 20 (1), 95–116. doi:10.1007/s10707-015-0232-z

Chen, S., 2015. Optimal bandwidth selection for kernel density estimation. *Journal of Probability and Statistics*, 19 (4), 1883–1905.

Conrad, C.C. and Hilchey, K.G., 2011. A review of citizen science and community-based environmental monitoring: issues and opportunities. *Environmental Monitoring and Assessment*, 176 (1–4), 273–291. doi:10.1007/s10661-010-1582-5

Cornwell, M.L. and Campbell, L.M., 2012. Co-producing conservation and knowledge: citizen-based sea turtle monitoring in North Carolina, USA. *Social Studies of Science*, 42 (1), 101–120. doi:10.1177/0306312711430440

Cox, T.E., *et al*., 2012. Expert variability provides perspective on the strengths and weaknesses of citizen-driven intertidal monitoring program. *Ecological Applications*, 22 (4), 1201–1212. doi:10.1890/11-1614.1

Dennis, R.L.H. and Thomas, C.D., 2000. Bias in butterfly distribution maps: the influence of hot spots and recorder's home range. *Journal of Insect Conservation*, 4 (2), 73–77. doi:10.1023/A:1009690919835

Deutsch, W.G., Lhotka, L., and Ruiz-Cordova, S., 2009. Group dynamics and resource availability of a long-term volunteer water-monitoring program. *Society and Natural Resources*, 22 (7), 637–649. doi:10.1080/08941920802078216

Deutsch, W.G. and Ruiz-Cordova, S., 2015. Trends, challenges, and responses of a 20-year, volunteer water monitoring. *Ecology and Society*, 20 (3), Art. 14. doi:10.5751/ES-07578-200314

Devictor, V., Whittaker, R.J., and Beltrame, C., 2010. Beyond scarcity: citizen science programmes as useful tools for conservation biogeography. *Diversity and Distributions*, 16 (3), 354–362. doi:10.1111/j.1472-4642.2009.00615.x

Dickinson, J.L., *et al*., 2012. The current state of citizen science as a tool for ecological research and public engagement. *Frontiers in Ecology and the Environment*, 10 (6), 291–297. doi:10.1890/110236

Dillon, P.J. and Kircher, W.B., 1975. The effect of geology and land use on the export of phosphorus from watersheds. *Water Research*, 9 (2), 135–148. doi:10.1016/0043-1354(75)90002-0

Dorset Environmental Science Centre (DESC), 2013. Guide to interpreting total phosphorus and Secchi depth data from the Lake Partner Program. Available from:http://www.desc.ca/sites/default/files/Guide%20to%20Interpreting%20TP%20and%20Secchi%20Data.pdf [Accessed 24 Aug 2017].

Dorset Environmental Science Centre (DESC), 2015. Lake Partner Program report card 2015. Available from: http://www.desc.ca/sites/default/files/LakePartnerReportCardRevised5%20FINAL-RS_LowResolution.pdf [Accessed 24 Aug 2017].

Dove-Thompson, D., *et al*., 2011. Climate Change Research Report: a summary of the effects of climate change on Ontario's aquatic ecosystems. Ontario Ministry of Natural Resources (OMNR). Available from: http://files.ontario.ca/environment-and-energy/aquatics-climate/stdprod_088243.pdf [Accessed 24 Aug 2017].

Elgammal, A., *et al*., 2002. Background and foreground modeling using nonparametric kernel density estimation for visual surveillance. *Proceedings of the IEEE*, 90 (7), 1151–1163. doi:10.1109/JPROC.2002.801448

Ellis, R. and Waterton, C., 2004. Environmental citizenship in the making: the participation of volunteer naturalists in UK biological recording and biodiversity policy. *Science and Public Policy*, 31 (2), 95–105. doi:10.3152/147154304781780055

Fernández, D. and Nakamura, M., 2015. Estimation of spatial sampling effort based on presence-only data and accessibility. *Ecological Modelling*, 299, 147–155. doi:10.1016/j.ecolmodel.2014.12.017

Fore, L.S., Paulsen, K.I.T., and Laughlin, K.O., 2001. Assessing the performance of volunteers in monitoring. *Freshwater Biology*, 46 (1), 109–123. doi:10.1111/j.1365-2427.2001.00640.x

Franzoni, C. and Sauermann, H., 2014. Crowd science: the organization of scientific research in open collaborative projects. *Research Policy*, 43 (1), 1–20. doi:10.1016/j.respol.2013.07.005

Geldmann, J., *et al*., 2016. What determines spatial bias in citizen science? Exploring four recording schemes with different proficiency requirements. *Diversity and Distributions*, 22 (11), 1139–1149. doi:10.1111/ddi.12477

Goodchild, M.F., 2007. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69 (4), 211–221. doi:10.1007/s10708-007-9111-y

Gouveia, C. and Fonseca, A., 2008. New approaches to environmental monitoring: the use of ICT to explore volunteered geographic information. *GeoJournal*, 72 (3–4), 185–197. doi:10.1007/s10708-008-9183-3

Haklay, M., 2013. Citizen Science and volunteered geographic information – overview and typology of participation. *In*: D.Z. Sui, S. Elwood, and M.F. Goodchild, eds. *Crowdsourcing geographic knowledge: volunteered Geographic Information (VGI) in Theory and Practice*. Berlln: Springer, 105–122.

Higa, M., *et al*., 2015. Mapping large-scale bird distributions using occupancy models and citizen data with spatially biased sampling effort. *Diversity and Distributions*, 21 (1), 46–54. doi:10.1111/ddi.12255

Hoyer, M.V., *et al*., 2012. A comparison between professionally (Florida Department of Environmental Protection) and volunteer (Florida LAKEWATCH) collected trophic state chemistry data in Florida. *Lake and Reservoir Management*, 28 (4), 277–281. doi:10.1080/07438141.2012.736016

Hurlbert, A.H. and Jetz, W., 2007. Species richness, hotspots, and the scale dependence of range maps in ecology and conservation. *Proceedings of the National Academy of Sciences*, 104 (33), 13384–13389. doi:10.1073/pnas.0704469104

Isaac, N.J., *et al*., 2014. Statistics for citizen science: extracting signals of change from noisy ecological data. *Methods in Ecology and Evolution*, 5 (10), 1052–1060. doi:10.1111/2041-210X.12254

Jacobs, C., 2016. Data quality in crowdsourcing for biodiversity research: issues and examples. *In*: C. Capineri, *et al*., eds. *European handbook of crowdsourced geographic information*. London: Ubiquity Press, 75–86.

Jalbert, K. and Kinchy, A.J., 2016. Sense and influence: environmental monitoring tools and the power of citizen science. *Journal of Environmental Policy & Planning*, 18 (3), 379–397. doi:10.1080/1523908X.2015.1100985

Jolliffe, I.T., 2002. *Principal component analysis*. 2nd ed. New York, NY: Springer.

Jollymore, A., *et al*., 2017. Citizen science for water quality monitoring: data implications of citizen perspectives. *Journal of Environmental Management*, 200, 456–467. doi:10.1016/j.jenvman.2017.05.083

Kadmon, R., Farber, O., and Danin, A., 2004. Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. *Ecological Applications*, 14 (2), 401–413. doi:10.1890/02-5364

Kinchy, A., 2017. Citizen science and democracy: participatory water monitoring in the Marcellus shale fracking boom. *Science as Culture*, 26 (1), 88–110. doi:10.1080/09505431.2016.1223113

King, T.L., *et al*., 2015. The use of kernel density estimation to examine associations between neighborhood destination intensity and walking and physical activity. *PLoS ONE*, 10 (9), e0137402. doi:10.1371/journal.pone.0137402

Kolok, A.S., *et al*., 2011. Empowering citizen scientists: the strength of many in monitoring biologically active environmental contaminants. *Bioscience*, 61 (8), 626–630. doi:10.1525/bio.2011.61.8.9

Kratz, T., *et al*., 1997. The influence of landscape position on lakes in northern Wisconsin. *Freshwater Biology*, 37 (1), 209–217. doi:10.1046/j.1365-2427.1997.00149.x

Lawrence, A., 2006. 'No personal motive ?' volunteers, biodiversity, and the false dichotomies of participation. *Ethics, Place, and Environment*, 9 (3), 279–298. doi:10.1080/13668790600893319

Lawrence, A. and Turnhout, E., 2010. Personal meaning in the public sphere: the standardisation and rationalisation of biodiversity data in the UK and the Netherlands. *Journal of Rural Studies*, 26 (4), 353–360. doi:10.1016/j.jrurstud.2010.02.001

Lester, N.P., *et al*., 2003. A broad-scale approach to management of Ontario's recreational fisheries. *North American Journal of Fisheries Management*, 23 (4), 1312–1328. doi:10.1577/M01-230AM

Li, T., *et al*., 2012. A PCA-based method for construction of composite sustainability indicators. *The International Journal of Life Cycle Assessment*, 17 (5), 593–603. doi:10.1007/s11367-012-0394-y

Lin, Y., *et al*., 2011. Hotspot analysis of spatial environmental pollutants using kernel density estimation and geostatistical techniques. *International Journal of Environmental Research and Public Health*, 8 (1), 75–88. doi:10.3390/ijerph8010075

Loperfido, J.V., *et al*., 2010. Uses and biases of volunteer water quality data. *Environmental Science and Technology*, 44 (19), 7193–7199. doi:10.1021/es100164c

Mair, L. and Ruete, A., 2016. Explaining spatial variation in the recording effort of citizen science data across multiple taxa. *PLoS ONE*, 11 (1), 1–13. doi:10.1371/journal.pone.0147796

Maldonado, C., *et al*., 2015. Estimating species diversity and distribution in the era of big data: to what extent can we trust public databases? *Global Ecology and Biogeography*, 24 (8), 973–984. doi:10.1111/geb.12326

Mandelik, Y., Roll, U., and Fleischer, A., 2010. Cost-efficiency of biodiversity indicators for Mediterranean ecosystems and the effects of socio-economic factors. *Journal of Applied Ecology*, 47 (6), 1179–1188. doi:10.1111/jpe.2010.47.issue-6

Mann, H.B. and Whitney, D.R., 1947. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18 (1), 50–60. doi:10.1214/aoms/1177730491

Maroko, A.R., *et al*., 2009. The complexities of measuring access to parks and physical activity sites in New York City: a quantitative and qualitative approach. *International Journal of Health Geographics*, 8, 1. doi:10.1186/1476-072X-8-34

Ontario Ministry of Natural Resources & Forestry (OMNRF), 2005. A new ecological framework for recreational fisheries management in Ontario. Available from: http://www.ontla.on.ca/library/repository/mon/12000/256625.pdf [Accessed 24 Aug 2017].

Ontario Ministry of Natural Resources & Forestry (OMNRF), 2009. Ecological framework for fisheries management: monitoring the health of Ontario's inland lakes fact sheet. Available from: https://dr6j45jk9xcmk.cloudfront.net/documents/2964/273246.pdf [Accessed 24 Aug 2017].

Ontario Ministry of Natural Resources & Forestry (OMNRF), 2012. The broad-scale monitoring program: monitoring the status of inland lakes in Ontario. Available from: http://sobr.ca/_biosite/wp-content/uploads/BsM_ProgPublic_v10_FINAL_2012-01-22.pdf [Accessed 7 Nov 2017].

Ontario Ministry of Natural Resources & Forestry (OMNRF), 2016. Broad-scale monitoring program. Available from: https://www.ontario.ca/page/broad-scale-monitoring-program [Accessed 24 Aug 2017].

Ontario Ministry of Natural Resources & Forestry (OMNRF), 2017. About Ontario. Available from: https://www.ontario.ca/page/about-ontario [Accessed 24 Aug 2017].

Ou, C., *et al*., 2012. Coupling geostatistical approaches with PCA and fuzzy optimal model (FOM) for the integrated assessment of sampling locations of water quality monitoring networks (WQMNs). *Journal of Environmental Monitoring*, 14 (12), 3118–3128. doi:10.1039/c2em30372h

Phillips, S.J., *et al*., 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological Applications*, 19 (1), 181–197. doi:10.1890/07-2153.1

Rocchini, D., *et al*., 2011. Accounting for uncertainty when mapping species distributions: the need for maps of ignorance. *Progress in Physical Geography*, 35 (2), 211–226. doi:10.1177/0309133311399491

Romo, H., García-Barros, E., and Lobo, J.M., 2006. Identifying recorder-induced geographic bias in an Iberian butterfly database. *Ecography*, 29 (6), 873–885. doi:10.1111/eco.2006.29.issue-6

Ruete, A., 2015. Displaying bias in sampling effort of data accessed from biodiversity databases using ignorance maps. *Biodiversity Data Journal*, 3, e5361. doi:10.3897/BDJ.3.e5361

Sandstrom, S., Rawson, M., and Lester., N., 2013. 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. Available from: https://dr6j45jk9xcmk.cloudfront.net/documents/2578/stdprod-103359.pdf [Accessed 24 Aug 2017].

Savan, B., Gore, C., and Morgan, A.J., 2004. Shifts in environmental governance in Canada: how are citizen environment groups to respond? *Environment and Planning C: Government and Policy*, 22 (4), 605–619. doi:10.1068/c12r

See, L., *et al*., 2016. Crowdsourcing, citizen science or volunteered geographic information? The current state of crowdsourced geographic information. *ISPRS International Journal of Geo-Information*, 5 (5), 55. doi:10.3390/ijgi5050055

Sharpe, A. and Conrad, C., 2006. Community based ecological monitoring in Nova Scotia: challenges and opportunities. *Environmental Monitoring and Assessment*, 113 (1–3), 395–409. doi:10.1007/s10661-005-9091-7

Silverman, B.W., 1986. *Density estimation for statistics and data analysis*. London: Chapman and Hall.

Silvertown, J., 2009. A new dawn for citizen science. *Trends in Ecology and Evolution*, 24 (9), 467–471. doi:10.1016/j.tree.2009.03.017

Soranno, P.A., *et al*., 1996. Phosphorus loads to surface waters: a simple model to account for spatial pattern of land use. *Ecological Applications*, 6 (3), 865–878. doi:10.2307/2269490

Statistics Canada, 2005. Land and freshwater area, by province and territory. Available from: http://www.statcan.gc.ca/tables-tableaux/sum-som/l01/cst01/phys01-eng.htm [Accessed 24 Aug 2017].

Stevens, P.A., 2013. 'Roughing it in comfort': family cottaging and consumer culture in postwar Ontario. *Canadian Historical Review*, 94 (2), 234–262. doi:10.3138/chr.600

Stolar, J. and Nielsen, S.E., 2015. Accounting for spatially biased sampling effort in presence-only species distribution modelling. *Diversity and Distributions*, 21 (5), 595–608. doi:10.1111/ddi.12279

Storey, R.G., *et al*., 2016. Volunteer stream monitoring: do the data quality and monitoring experience support increased community involvement in freshwater decision making? *Ecology and Society*, 21, 4. doi:10.5751/ES-08934-210432

Thornton, L., Pearce, J.R., and Kavanagh, A., 2011. Using geographic information systems (GIS) to assess the role of the built environment in influencing obesity: a glossary. *International Journal of Behavioral Nutrition and Physical Activity*, 8 (1), 71–79. doi:10.1186/1479-5868-8-71

Tulloch, A.I.T., *et al*., 2013. Realising the full potential of citizen science monitoring programs. *Biological Conservation*, 165, 128–138. doi:10.1016/j.biocon.2013.05.025

Van Strien, A.J., Van Swaay, C.A.M., and Termaat, T., 2013. Opportunistic citizen science data of animal species produce reliable estimates of distribution trends if analysed with occupancy models. *Journal of Applied Ecology*, 50 (6), 1450–1458. doi:10.1111/1365-2664.12158

Weng, Y.C., 2015. Contrasting visions of science in ecological restoration: expert-lay dynamics between professional practitioners and volunteers. *Geoforum*, 65, 134–145. doi:10.1016/j.geoforum.2015.07.023

Wiggins, A., *et al*., 2011. Mechanisms for data quality and validation in citizen science. In "*Computing for Citizen Science" workshop at the IEEE eScience Conference*, Stockholm, Sweden. Available from: http://itee.uq.edu.au/~eresearch/workshops/compcitsci2011/index.html [Accessed 24 Aug 2017].

Yenilmez, F., Düzgün, S., and Aksoy, A., 2015. An evaluation of potential sampling locations in a reservoir with emphasis on conserved spatial correlation structure. *Environmental Monitoring and Assessment*, 187 (1), 1–21. doi:10.1007/s10661-014-4216-5

Zuur, A.F., *et al*., 2009. *Mixed effects models and extensions in ecology with R*. New York, NY: Springer.