Science of the Total Environment 633 (2018) 600–607

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

Representing a large region with few sites: The Quality Index approach for field studies

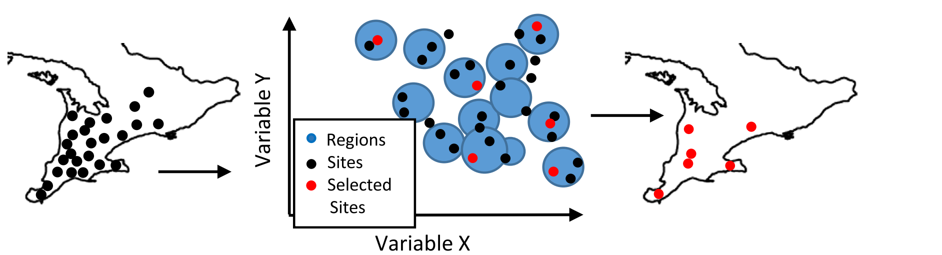
Madeline S. Rosamond [a,*], Christopher Wellen [b], Meguel A. Yousif [c], Georgina Kaltenecker [d], Janis L. Thomas [d], Pamela J. Joosse [e], Natalie C. Feisthauer [e], William D. Taylor [a], Mohamed N. Mohamed [d]

[a] Department of Biology, University of Waterloo, 200 University Avenue West, Waterloo, Ontario N2L 3G1, Canada
[b] Department of Geography and Environmental Studies, Ryerson University, 350 Victoria Street, Toronto, Ontario M5B 2K3, Canada
[c] Department of Chemical Engineering, University of Waterloo, 200 University Avenue West, Waterloo, Ontario N2L 3G1, Canada
[d] Environmental Monitoring and Reporting Branch, Ontario Ministry of the Environment and Climate Change, 125 Resources Rd., Toronto, Ontario M9P 3V6, Canada
[e] Science & Technology Branch, Agriculture and Agri-Food Canada, 174 Stone Road West, Guelph, Ontario, N1G 4S9, Canada

HIGHLIGHTS

- We present a novel way to select sites for field studies representing a larger region: the Quality Index.
- The Quality Index prioritizes sites that cover the range in variables of interest of a larger region and are well-distributed in variable space.
- A genetic algorithm allows for selecting optimal sites by maximizing the QI.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 29 June 2017
Received in revised form 9 March 2018
Accepted 10 March 2018
Available online xxxx

Editor: Ouyang Wei

Keywords:
Site selection
Field studies
Study design
Optimization

ABSTRACT

Many environmental studies require the characterization of a large geographical region using a range of representative sites amenable to intensive study. A systematic approach to selecting study areas can help ensure that an adequate range of the variables of interest is captured. We present a novel method of selecting study sites representing a larger region, in which the region is divided into subregions, which are characterized with relevant independent variables, and displayed in mathematical variable space. Potential study sites are also displayed this way, and selected to cover the range in variables present in the region. The coverage of sites is assessed with the Quality Index, which compares the range and standard deviation of variables among the sites to that of the larger region, and prioritizes sites that are well-distributed (*i.e.* not clumped) in variable space. We illustrate the method with a case study examining relationships between agricultural land use, physiography and stream phosphorus (P) export, in which we selected several variables representing agricultural P inputs and landscape susceptibility to P loss. A geographic area of 110,000 km^2 was represented with 11 study sites with good coverage of four variables representing agricultural P inputs and transport mechanisms taken from commonly-available geospatial datasets. We use a genetic algorithm to select 11 sites with the highest possible QI and compare these, post-hoc, to our sites. This approach reduces subjectivity in site selection, considers practical constraints and easily allows for site reselection if necessary. This site selection approach can easily be adapted to different landscapes and study goals, as we provide an algorithm and computer code to reproduce our approach elsewhere.

* Corresponding author.
E-mail address: msrosamo@uwaterloo.ca (M.S. Rosamond).

https://doi.org/10.1016/j.scitotenv.2018.03.113

1. Introduction

Environmental studies are frequently concerned with examining the effect of specific variables (*e.g.*, anthropogenic stressors) across a large, heterogeneous landscape, which may modify the effects of the variables. Often, a large region is of interest, but since it is frequently impractical to conduct detailed studies on the entire region, the selection of a small number of sites to represent the region is required. In many cases, sites are chosen primarily for logistical reasons (*e.g.* convenient access, existing datasets). However, since it is unlikely that a small number of randomly selected sites will cover the range in the variables of interest (Royall, 1970), systematic approaches to the selection of these representative sites are required (Kyllmar et al., 2014; Sharpley et al., 2015), and several approaches for a variety of study types have been developed (Coote et al., 1982; Danz et al., 2005; Fealy et al., 2010; Yates and Bailey, 2010).

Monitoring projects are often designed to capture the largest possible gradient across a variety of potential stressors. For example, sites were selected for a monitoring project in Laurentian Great Lakes (LGL) coastal zones using 201 geospatial datasets in six stressor categories and one soil category (Danz et al., 2005). The categories were previously identified as potential stressors in Great Lakes ecosystems (Environment Canada and United States EPA, 2003), though the stress gradients were not known *a priori*. Data in each category were compressed using principal components analysis (PCA). Sites were then grouped by performing cluster analysis on PCA values, with study sites selected from each cluster. Some approaches for designing monitoring projects focus on removing redundant sites; this is only useful when variables of interest have already been measured (*e.g.* Alameddine et al., 2013).

Field projects comparing "reference" and "impacted" sites may require a different kind of selection method for reference sites. Yates and Bailey (2010) demonstrate a selection process for reference headwater catchments in southern Ontario. Potential headwater sites were divided into soil type categories, and a gradient of human impact was calculated by linearly combining indicators of human impact (*e.g.* field tile drainage, septic systems, livestock) with PCA. Reference sites for each soil class were selected from watersheds in the lowest 25% of the impact gradient.

Site selection for field studies with specific hypotheses regarding relationships between variables (*e.g.* potential drivers) and responses of interest may focus on only a few variables and aim for a good range in these variables. These variables may be targeted for study *a priori* if they have been previously identified as relevant, or their relationships to environmental responses is of interest. For example, the Pollution from Land Use Activities Reference Group (PLUARG) study in the 1970s attempted to represent a large area (southern Ontario, Canada) with 11 small agricultural watersheds (<100 km^2) with the aim of quantifying and predicting sources of agricultural nutrients to the LGL. First, the agricultural region of interest was classified into five major soil types representing a gradient in the potential to transport nutrients and other pollutants. Next, a gradient of nutrient inputs (N and P) was determined by estimating nutrient contributions from manure production and fertilizer inputs (inferred from crop types). These two gradients were overlaid graphically and from them, 21 zones of similar combination of these two gradients were identified. Eleven headwater sites were then selected to cover a range of these 21 zones (Coote et al., 1978, 1982). As computerized geospatial analysis was not yet available, considerable professional judgement was used to weigh inputs, determine the gradients and to define the resultant zones (Coote, pers. comm.).

In a more recent study, Fealy et al. (2010) used a multi-criteria decision analysis (MCDA) approach to select agricultural watersheds as part of the Irish Agricultural Catchments Program. They first divided agricultural catchments into grassland and cropland categories. Within these two categories, several criteria deemed to be important to the potential for nutrient loss potential (*e.g.* % area in forage, livestock housing density) were measured and relative weights were assigned. These categories and their weights were determined by expert judgement and stakeholder engagement, which Fealy et al. (2010) described as a distinct advantage of the MCDA approach.

While these approaches have certain strengths, none provides a mathematical check on the distribution of sites, or attempt to avoid statistical leveraging by making sure sites are evenly distributed. Additionally, some methods include subjective weighting of input variables. Therefore, we developed a site selection process that combines the strengths of previous approaches for selecting monitoring and reference sites - the use of geospatial datasets to capture the regional range of variables of interest, reduction of expert judgement in ranking these variables - with a novel Quality Index (QI) value, which assesses the selected study sites' coverage of a range of interest (here, the range in a larger region) and the distribution of sites. Because we illustrate all potential sites in mathematical variable space, selecting new sites, if any are rejected for logistical reasons, is also easy with our approach.

We illustrate the method with a case study on phosphorus (P) export and cycling in small, agricultural watersheds in southern Ontario, Canada. We represent a large region (~110,000 km^2) with 11 study sites using four variables of interest relating to P sources (soil P and P inputs *via* fertilizer and manure) and P transport (runoff, soil texture, and slope). The study aims to examine relationships among agricultural activity, landscape variables and phosphorus (P) export. While P export from agricultural watersheds has been well-studied over several decades, the drivers of P export, and therefore appropriate management methods, are not well understood in many landscapes because of the complexity of P sources, storage, transport, and transformation in agricultural landscapes (Sharpley, 2016; Withers et al., 2017; Withers and Jarvie, 2008). P export results from these watersheds are not yet collected and will be reported elsewhere. The method has wider applicability to any studies where gradients (such as source and transport) of independent variables such as stressors are hypothesized *a priori* to be important drivers of a measure or outcome of interest.

2. Methods

2.1. Site selection approach and Quality Index

Our overall approach was to describe a large regional area (R) on a coarse spatial scale (*i.e.* divide it into subregions, R_s) using at least two independent variables of interest, and then describe many potential study sites ("candidate sites", population S) in the same way, mapping them both in mathematical space. We then selected n candidate sites (A_i) from S so that the n sites selected can represent the region R from which S was chosen (Fig. 1). We illustrate the method with a case study described in Section 2.4. Our specific procedure is as follows:

1. Break the region R down into subregions (R_s) and select potential study sites (S).
2. Map the population of subregions R_s and sites S into a mathematical space, the axes of which are defined by critical variables that represent the population. Selecting the critical variables inevitably involves some expert judgement, though the critical variables may be chosen to complement the hypothesis of the field study. We illustrate the method with four variables of interest (W, X, Y, Z); extending or contracting our method into spaces of arbitrary dimensionality to meet the needs of other projects is trivial. Once the candidate sites have been mapped into a mathematical space, we posit that representing the region R from which S was taken may be approached as a problem of geometrically representing the mathematical space defined, in our case, by X and Y.
3. We define the critical geometric attributes of the space in X and Y over which all of the candidate sites A_i co-vary. Specifically, we define the *Range* of the candidate sites A_i: