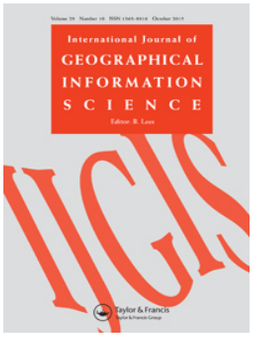

International Journal of Geographical Information Science

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

A citizen data-based approach to predictive mapping of spatial variation of natural phenomena

A-Xing Zhu, Guiming Zhang, Wei Wang, Wen Xiao, Zhi-Pang Huang, Ge-Sang Dunzhu, Guopeng Ren, Cheng-Zhi Qin, Lin Yang, Tao Pei & Shengtian Yang

To cite this article: A-Xing Zhu, Guiming Zhang, Wei Wang, Wen Xiao, Zhi-Pang Huang, Ge-Sang Dunzhu, Guopeng Ren, Cheng-Zhi Qin, Lin Yang, Tao Pei & Shengtian Yang (2015) A citizen data-based approach to predictive mapping of spatial variation of natural phenomena, International Journal of Geographical Information Science, 29:10, 1864-1886, DOI: 10.1080/13658816.2015.1058387

To link to this article: http://dx.doi.org/10.1080/13658816.2015.1058387

Published online: 24 Jun 2015.

International Journal of Geographical Information Science, 2015
Vol. 29, No. 10, 1864–1886, http://dx.doi.org/10.1080/13658816.2015.1058387

A citizen data-based approach to predictive mapping of spatial variation of natural phenomena

A-Xing Zhu[a,b,c,d,e], Guiming Zhang[d,c,f]*, Wei Wang[g], Wen Xiao[h], Zhi-Pang Huang[h], Ge-Sang Dunzhu[c,i], Guopeng Ren[h], Cheng-Zhi Qin[c], Lin Yang[c], Tao Pei[c] and Shengtian Yang[f]

[a]*Key Laboratory of Virtual Geographic Environment, Nanjing Normal University, Nanjing, China;* [b]*State Key Laboratory Cultivation Base of Geographical Environment Evolution, Nanjing, China;* [c]*State Key Laboratory of Resources and Environmental Information System, Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences (CAS), Beijing, China;* [d]*Department of Geography, University of Wisconsin-Madison, Madison, WI, USA;* [e]*Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing, China;* [f]*School of Geography, Beijing Normal University, Beijing, China;* [g]*Chinese Research Academy of Environmental Sciences, Beijing, China;* [h]*Institute of Eastern-Himalaya Biodiversity Research, Dali University, Dali, China;* [i]*Tibet Academy of Agricultural and Animal Sciences, Lhasa, China*

(*Received 22 November 2014; accepted 18 May 2015*)

The vast accumulation of environmental data and the rapid development of geospatial visualization and analytical techniques make it possible for scientists to solicit information from local citizens to map spatial variation of geographic phenomena. However, data provided by citizens (referred to as citizen data in this article) suffer two limitations for mapping: bias in spatial coverage and imprecision in spatial location. This article presents an approach to minimizing the impacts of these two limitations of citizen data using geospatial analysis techniques. The approach reduces location imprecision by adopting a frequency-sampling strategy to identify representative presence locations from areas over which citizens observed the geographic phenomenon. The approach compensates for the spatial bias by weighting presence locations with cumulative visibility (the frequency at which a given location can be seen by local citizens). As a case study to demonstrate the principle, this approach was applied to map the habitat suitability of the black-and-white snub-nosed monkey (*Rhinopithecus bieti*) in Yunnan, China. Sightings of *R. bieti* were elicited from local citizens using a geovisualization platform and then processed with the proposed approach to predict a habitat suitability map. Presence locations of *R. bieti* recorded by biologists through intensive field tracking were used to validate the predicted habitat suitability map. Validation showed that the continuous Boyce index ($B_{\text{cont}}(0.1)$) calculated on the suitability map was 0.873 (95% CI: [0.810, 0.917]), indicating that the map was highly consistent with the field-observed distribution of *R. bieti*. $B_{\text{cont}}(0.1)$ was much lower (0.173) for the suitability map predicted based on citizen data when location imprecision was not reduced and even lower (−0.048) when there was no compensation for spatial bias. This indicates that the proposed approach effectively minimized the impacts of location imprecision and spatial bias in citizen data and therefore effectively improved the quality of mapped spatial variation using citizen data. It further implies that, with the application of geospatial analysis techniques to properly account for limitations in citizen data, valuable information embedded in such data can be extracted and used for scientific mapping.

Keywords: citizen data; location imprecision; spatial bias; volunteered geographic information (VGI); *Rhinopithecus bieti*

*Corresponding author. Email: gzhang45@wisc.edu

1. Introduction

The rapid development of geospatial technologies in the last decade or so has empowered the general public to contribute geospatial data (Goodchild 2007a, Elwood 2008). With the popularization of the Internet, personal computers, and spatially enabled portable devices such as global positioning system (GPS)-equipped smartphones and personal digital assistants, citizens can now easily share georeferenced observations of the world (e.g., wildlife sightings) via interactive geovisualization interfaces (e.g., Google Maps, Google Earth, and Microsoft Virtual Earth), via social media (e.g., Twitter, Flickr, Instagram), or via citizen science projects (e.g., eBird) (Silvertown 2009, Sullivan *et al*. 2009, Dickinson *et al*. 2012, Guan *et al*. 2012). These technologies have created unprecedented opportunities for citizens to report their observations of the world. Such citizen-contributed observations have been referred to as volunteered geographic information (VGI) (Goodchild 2007a, Sui 2008, Sui *et al*. 2013). The technologies also create opportunities for scientists to conveniently solicit useful information from citizens on geographic features or phenomena of interest (e.g., Seeger 2008, Anadón *et al*. 2009, Haklay 2013). In this paper, we refer to the data contributed by citizens, whether voluntarily reported by citizens or deliberately solicited from citizens, as citizen data.

Citizen data have several advantages. First, they contain rich local information that spans a wide temporal spectrum because citizens, as local experts and sensors (Goodchild 2007a, 2007b), have long been sensing and accumulating knowledge of their respective areas. But citizen data also have the potential to provide information over large areas, given that seven billion networked human sensors are distributed across the globe. In addition, citizen data can provide timely updated information that cannot be obtained through remote sensing techniques but can be easily observed by citizens on the ground (Goodchild 2007b). Finally, collecting citizen data is much less expensive than traditional scientific data collection. In many cases, citizens contribute data purely voluntarily in the spirit of self-promotion and altruism without any hope of financial reward (Goodchild 2007a, 2007b), while traditional scientific data collection is costly and labor intensive and can only be conducted by professional scientists following strict protocols. This low cost might be of significant practical importance.

Interest in citizen data has grown rapidly, and many approaches have recently been proposed to assure their quality (see Goodchild and Li 2012, for a review; Elwood *et al*. 2013, Ali and Schmid 2014). Citizen data are now driving many successful ongoing applications. Among them, OpenStreetMap (Haklay and Weber 2008) is producing geographic data (e.g., road networks) for every corner of the world with voluntary efforts from Internet users. Such user-generated geographic data can be of quite high location accuracy, comparable to survey products of government mapping agencies (e.g., Haklay 2010). The eBird citizen science project (Sullivan *et al*. 2009) is documenting bird species (e.g., presence, abundance) with collective observations contributed by worldwide birders. For emergency management, citizen data are providing timely information for disaster monitoring and response such as wildfires and earthquakes (e.g., Goodchild and Glennon 2010, Zook *et al*. 2010). In the world's poor and remote areas, wildlife sightings can be solicited from subsistence farmers, herdsmen, and hunters who have long been living in the local area and whose livelihoods are closely linked to ecosystem services. For conservation programs with limited budgets in those areas, local citizens could serve as a cost-effective data source on wildlife distribution (e.g., Anadón *et al*. 2009).

Citizen data contain valuable information on the spatial distribution of geographic phenomena, and in some cases represent the only data that is available. The data thus have the potential to be used for mapping spatial variation of geographic phenomena of interest. Predictive mapping is a commonly used technique to map the spatial variation of geographic phenomena, which is particularly useful for mapping those physical geographic phenomena that cannot be observed and mapped using remote sensing techniques but that have spatial variation that is highly related with their environmental factors (covariates) (Franklin 1995, Scull *et al*. 2003, Zhu 2008). Examples of such geographic phenomena are forest composition (Franklin 1995), soil class and soil properties (Zhu *et al*. 2001, Scull *et al*. 2003, Zhu 2008), species richness (Pittman *et al*. 2007), and habitat suitability (Franklin and Miller 2009).

Under the framework of predictive mapping, one needs to first establish the relationship between the target geographic phenomenon and its environmental covariates, and then apply this relationship to an environmental database to create a predictive map of spatial variation of the target geographic phenomenon (Franklin 1995, Scull *et al*. 2003, Zhu 2008). The relationship is often derived from representative field samples of the target geographic phenomenon (each consisting of observation of the target geographic phenomenon at a particular location) using a wide range of methods (e.g., statistical modeling, machine learning) (Franklin 1995, Zhu 2008). Citizen data contain field observations of geographic phenomena and therefore could serve as 'samples' for predictive mapping. In combination with the vast accumulation of environmental data (e.g., Kerr and Ostrovsky 2003, Gillespie *et al*. 2008, Viña *et al*. 2008), the relationship between the target geographic phenomenon and its environmental covariates can be derived from citizen data and applied to predict spatial variation of the geographic phenomenon of interest.

There are two important requirements for samples to be used for predictive mapping. One is that the samples need to be representative of the study area in order to derive a relationship that can represent the covariation between the target geographic phenomenon and its environmental covariates. This is often accomplished by collecting samples following a well-designed spatial sampling scheme, such as random sampling, stratified random sampling, or systematic sampling (Haining 2003). The other requirement is that the precision of sampling locations needs to be high so that the locations can be used to accurately obtain the corresponding values of environmental covariates at these locations from environmental databases. However, data contributed by citizens sometimes do not meet these two requirements when used as samples for predictive mapping, though these limitations might not be of concern in some other applications. Spatial bias (Sullivan *et al*. 2009, Munson *et al*. 2010) in citizen data is one of the major challenges while other data quality issues are also under scrutiny (e.g., Flanagin and Metzger 2008, Bonter and Cooper 2012, Foody *et al*. 2013). The spatial coverage of citizen data is biased because observations made by citizens are opportunistic and 'ad-hoc' in nature. In the example of wildlife sightings, local citizens are not intentionally tracking wildlife of interest. Instead, they typically spot the wildlife en route to doing something else. In the sense of geographic sampling, the routes on which local citizens spotted the wildlife would be considered neither random nor regular by design but 'ad-hoc.' As a result, the spatial coverage of wildlife sightings elicited from local citizens is likely to be biased. Such spatial bias, if not appropriately accounted for, might adversely affect inferences made from citizen data (e.g., Ponder *et al*. 2001, Reddy and Dávalos 2003, Graham *et al*. 2004, Kadmon *et al*. 2004, Leitão *et al*. 2011).

In some cases, citizen data can be of quite high location accuracy (e.g., OpenStreetMap data) and location imprecision might not be a concern for these cases

(Girres and Touya 2010, Haklay 2010). For example, OpenStreetMap data on human tracks (e.g., roads, streets, buildings) and physical geographic features (e.g., rivers, lakes) are of high location accuracy because these stationary targets were digitized from accurately georeferenced high-resolution remote-sensing imagery. However, spatial location of citizen data might be imprecise depending on the geographic features or phenomena under observation, as well as on the availability of positioning technologies. In cases where reliable positioning technologies are in short supply and the geographic features under observation are on the move it is challenging to ensure the location accuracy of citizen data. For instance, unlike some biologists who can rely on GPS collars to pinpoint the location of wildlife occurrences, a farmer's descriptions of the location of the sighted wildlife are often imprecise or vague, particularly if a long time has lapsed since the actual sightings. It is not feasible for the farmer to report the exact location of the sighted wildlife with any level of confidence. In these situations, location imprecision might be a serious concern in the use of citizen data for predictive mapping (Graham *et al*. 2008, Fernandez *et al*. 2009, Osborne and Leitão 2009, Moudrý and Šímová 2012).

To the best of our knowledge, few studies have been conducted that focus on tackling spatial bias in citizen data for predictive mapping. Several methods related to reducing location imprecision in citizen data exist. Khalili *et al*. (2010) developed a method for disambiguating Flickr hometown location information. Jones *et al*. (2008) developed a method for improving the quality of information related to vague place names. Wieczorek *et al*. (2004) proposed the point-radius method for georeferencing locality descriptions in natural history museum specimens. However, these methods operate at very coarse spatial scale (e.g., place names, city names). Therefore, they might not help to further improve the location accuracy of citizen data at a finer spatial scale (e.g., if a farmer said, 'I saw monkeys over that area on the northern hill slope.'). The appropriate strategy for reducing location imprecision in citizen data should depend closely on the context of the data (what the data are about and how the data were generated).

This paper presents an approach to minimizing the impacts of spatial bias and location imprecision associated with citizen data for the predictive mapping of spatial variation of geographic phenomena using a monkey habitat-mapping example. The approach employs geospatial analysis techniques to reduce the location imprecision and to compensate for the spatial bias in citizen data, respectively. Section 2 introduces the basic idea and details of the approach. It is then illustrated with a case study that maps the habitat suitability of the black-and-white snub-nosed monkey (*Rhinopithecus bieti*) in Yunnan, China.

2. Methodology

2.1. The basic idea

When using data contributed by citizens as 'samples' for predictive mapping, the challenges raised by location imprecision and spatial bias in citizen data need to be addressed to make the data as representative as possible. The first challenge is to obtain representative locations of the observed geographic phenomenon. For example, local citizens often provide location information of their sightings of wildlife in the form of, 'I saw the monkeys over this area.' Clearly, 'over this area' can be depicted using a polygon, but this does not mean that the monkeys showed up at every location in the polygon area and certainly not at an equal probability within the polygon. Considering all locations in the polygon as wildlife sightings is thus not appropriate. The question then is how to obtain

locations that are representative of the actual presence location of wildlife in the polygon area outlined by local citizens.

Geographic phenomena that are highly related to their environmental covariates usually occur under certain typical environmental conditions (e.g., wildlife usually present in preferred habitats). Citizen data do contain valuable information on where the observed geographic phenomenon occurred, hence the polygons outlined by local citizens, though imprecise, do capture such typical environmental conditions under which the geographic phenomena would occur. Accordingly, the basic idea to reduce location imprecision assumes that locations at which values of environmental conditions are most frequent over the polygon area would best approximate, or represent, the actual location of the observed geographic phenomenon (e.g., wildlife presence) (Qi and Zhu 2003). This approach can serve as a first approximation given that additional information on the specifics of the habitat is not available. Otherwise, the approach can be fine-tuned to make use of the specifics of habitat. For example, if presence is more related to food source, the food source information should be weighted more heavily, even exclusively, in improving the precision of locations.

The second challenge is to find a way to remedy or compensate for the bias in the spatial coverage of citizen data. For instance, multiple sightings of monkeys at one location by many local citizens do not necessarily mean that the location is highly preferred by monkeys. This is because it might be that the location is easily visible from multiple routes. Thus, every time a monkey shows up at this location it is spotted by some local citizen(s). On the other hand, a monkey that shows up at locations that are preferred by monkeys but less visible to local citizens will have a lower chance of being spotted by local citizens. So, we must compensate for this spatial bias.

Citizen data are spatially biased because not every location on the landscape can be equally observed from the routes taken by the local citizens, given the irregular and nonrandom distribution of their routes and the variability of the terrain conditions. Accordingly, cumulative visibility, the frequency at which a given location can be seen by local citizens from the routes they take, was used as a weight to compensate for the bias due to unequal visibility over the landscape from the routes.

2.2. Detailed methodology

2.2.1. Collection of the citizen data

Citizen data were solicited from local citizens through a series of structured interviews using geovisualization techniques. The objective of each interview was to obtain from the local citizens sightings of the geographic phenomenon of interest (e.g., wildlife presence) and their activity routes in the study area. The connections between their experiences and the geographic phenomenon of interest were first reinforced by introducing them to materials relevant to the target geographic phenomenon (e.g., photos, videos, and audios of a wildlife species). The interview was then conducted using a 3D geovisualization platform, 3dMapper® (Burt and Zhu 2004; freely available at solim.geography.wisc.edu). 3dMapper® integrates digital elevation model (DEM) data with high-resolution satellite imagery to produce an intuitive 3D view of the study area, on which local citizens can rotate, pan, and zoom to any part of the study area. After being familiarized with the 3D view, it was not difficult for local citizens to locate areas where the geographic phenomenon had been sighted. Information on where and when they sighted the geographic phenomenon was recorded by drawing polygons on the 3D view and by filling in attribute

tables tied to polygons. The location of their activity routes and the timing and frequency of the use of the routes was recorded by drawing polylines and by filling in attribute tables tied to the polylines.

2.2.2 Refinement of the citizen data

2.2.2.1. Selection of representative locations. Under the assumption that those locations at which values of environmental conditions are most frequent over the polygon area would best approximate the location where a geographic phenomenon actually occurred (as discussed in Section 2.1), a frequency-sampling strategy proposed by Qi and Zhu (2003) was adopted to locate the representative locations within a polygon. A detailed discussion of this sampling strategy is beyond the scope of this paper, but an overview of this strategy is given here. The sampling strategy consists of three steps. First, within each sighting polygon solicited from local citizens for each environmental covariate taken into account (e.g., elevation, vegetation type, etc.), the frequency distribution of the environmental values was constructed. Here, a robust rule proposed by Freedman and Diaconis (1981, 1991) was adopted to determine the bin width for the construction of frequency distributions (histograms):

$$\text{bin}_x = 2(\text{IQR}_x)k^{-1/3}, \tag{1}$$

where bin_x is the bin width for environmental variable x for this polygon, which encompasses k pixels; IQR_x is the interquartile range of the k values of environmental variable x. Based on this frequency distribution, the locations (pixels) at which environmental values fall into the modal interval (the interval of environmental values with the highest frequency) were identified. Second, the pixels at which environmental values fall into the modal intervals of their respective environmental variables simultaneously were identified and these pixels were taken as representative locations for this polygon. Third, representative locations identified from all polygons were pooled to form a full set of presence locations.

2.2.2.2. Compensation for spatial bias. Cumulative visibility was used to compensate for spatial bias due to unequal visibility over the landscape from the routes taken by local citizens. Calculation of the cumulative visibility at every location across the landscape involves three steps. First, visibility at this location from each route was calculated by performing a viewshed analysis (Wang *et al*. 2000) from the corresponding route with a maximum visible distance (estimated based on field experience). Second, the corresponding frequency with which local citizens took the route was multiplied to the visibility from each route. Third, visibility from each route was summed up to obtain the cumulative visibility of this location. The spatial bias was compensated for by inversely weighting each presence location with the cumulative visibility at that location.

2.2.3. Use of citizen data for predictive mapping

With the refined citizen data, the relationship between the target geographic phenomenon and its environmental covariates can then be derived for predictive mapping of spatial variation of the target geographic phenomenon. The selection of the predictive mapping approach depends greatly on the characteristics of the 'samples' data, among other factors

(Franklin 1995, Scull *et al*. 2003). We adopted a fuzzy-based predictive mapping approach (Zhu 2008) over other alternatives (e.g., statistical approaches) for citizen data-based predictive mapping, because the use of citizen data as 'samples,' though refined, does not conform well to the requirements imposed by statistical methods. This is due to location imprecision and spatial bias. Fuzzy-based approaches, however, afford the ability to accommodate imprecision and uncertainty in spatial data (Altman 1994, Robinson 2003).

Further, wildlife habitat suitability was used as an example of the target geographic phenomenon to illustrate the proposed citizen-data-based predictive mapping approach. Habitat suitability at a spatial location indicates the degree of closeness between environmental conditions at this location and the optimal environmental conditions for species survival (Hirzel *et al*. 2002, Hirzel and Le 2008). For predictive mapping of wildlife habitat suitability, the relationship between habitat suitability and environmental variables that influence the wildlife need to be quantified first. This relationship can then be applied to an environmental database to create a predictive habitat suitability map (Guisan and Zimmermann 2000, Franklin and Miller 2009).

2.2.3.1. Quantification of suitability–environment relationships. Wildlife presents more frequently at locations where environmental conditions are better suited to their survival and reproduction (Hirzel *et al*. 2002, Phillips *et al*. 2006). Probability density functions (PDFs) of environmental conditions over the presence locations were, therefore, used to quantify the relationships between wildlife habitat suitability and environmental conditions (these PDFs were referred to as 'presence PDFs'). The presence PDF in respect to every environmental variable was estimated using kernel density estimation, a method that is capable of estimating continuous PDFs, which uses the equation (Silverman 1986, Porter and Reich 2012):

$$f(x) = \frac{1}{nh_x}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h_x}\right), \tag{2}$$

where $f(x)$ is the estimated PDF quantifying the relationship between habitat suitability and environmental variable x; x_i is the value of environmental variable x at presence location i; n is the total number of presence locations; K is a kernel density function and h_x is a bandwidth for environmental variable x. Here, we adopted the Gaussian kernel for K and used the 'rule-of-thumb' algorithm to determine h_x for each environmental variable in the equation (Silverman 1986):

$$K\left(\frac{x - x_i}{h_x}\right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-x_i)^2}{2h_x^2}}, h_x = \sigma_x\left(\frac{4}{3n}\right)^{1/5}, \tag{3}$$

where σ_x is the standard deviation of values of environmental variable x over the n presence locations.

To compensate for the spatial bias in citizen data, a weight was given to each of the presence locations when estimating PDF:

$$f'(x) = \frac{1}{h'_x}\sum_{i=1}^{n}\left[\frac{w_i}{\sum w_i} K\left(\frac{x - x_i}{h'_x}\right)\right], \tag{4}$$

where $f^{'}(x)$ is the estimated PDF for environmental variable x, and w_i is the weight to adjust the contribution of x_i to the estimated PDF; it was used to compensate for spatial bias. w_i is defined as inversely proportional to cumulative visibility (v_i) at presence location i:

$$w_i = \frac{w^{'}_i}{\max(w^{'})}, w^{'}_i = \frac{1}{v_i}, \tag{5}$$

a presence location with a higher cumulative visibility would have a smaller contribution to the estimated PDF and a presence location with a lower cumulative visibility would have a larger contribution. The bandwidth $h^{'}_x$ in Equation (4) was calculated with weights because $\sigma^{'}_x$ and $n^{'}$ were also calculated with weights:

$$h^{'}_x = \sigma^{'}_x \left(\frac{4}{3n^{'}}\right)^{1/5}, \sigma^{'}_x = \sqrt{\frac{\sum\left[w_i(x_i - \bar{x})^2\right]}{n^{'} - 1}}, n^{'} = \sum w_i, \bar{x} = \sum\left(\frac{w_i}{\sum w_i} x_i\right). \tag{6}$$

It is worth noting that there are weaknesses to this approach to accounting for spatial bias. In general, it is essential to incorporate information on spatial distribution of sampling/observation efforts to correct for spatial bias (Reddy and Dávalos 2003, De Solla *et al*. 2005, Kramer-Schadt *et al*. 2013). This approach used cumulative visibility as a proxy of observation efforts. In cases where observation efforts of local citizens are spatially incomplete, that is, there are 'gaps' in spatial coverage of their observation efforts, this cumulative visibility-based approach does not fully address spatial bias. This might be an issue for citizen data collected over large areas, but is less of a problem for small areas that would most likely be fully covered by the collective observation efforts of local citizens.

2.2.3.2. Mapping of habitat suitability. Based on these estimated presence PDFs, habitat suitability at each location (pixel) in the area was calculated in two steps through a 'rule-based' approach (Rüger *et al*. 2005, Van Broekhoven *et al*. 2006, Zhu 2008). First, habitat suitability was calculated with respect to every single environmental variable. Second, the overall habitat suitability was determined by aggregating habitat suitabilities to all environmental variables involved. The details of each step are as follows.

At the first step, with the assumption that wildlife habitat suitability was higher where observed presence probability density of the wildlife was higher, habitat suitability to every single environmental variable was computed by normalizing the corresponding presence PDF to [0, 1]:

$$S^x_j = \frac{f^{'}(x_j)}{\max(f^{'}(x))}, \tag{7}$$

where x_j refers to the value of environmental variable x at location j, $\max(f^{'}(x))$ refers to the maximum value of the PDF, and S^x_j refers to the habitat suitability in respect to environmental variable x at location j. It is worth noting that here S is not a probability of presence, but instead an index that indicates habitat suitability.

At the second step, with the assumption that wildlife habitat suitability was controlled by multiple environmental variables but determined by the least favorable factor (the

'limiting factor' approach), a minimum operator, which takes the minimum suitability value among the suitability values of individual environmental variables, was adopted to determine the overall habitat suitability (Zhu and Band 1994, Rüger *et al*. 2005, Li and Hilbert 2008, Zhu 2008):

$$S_j = \min\left(S_j^1, S_j^2, \ldots, S_j^m\right), \tag{8}$$

where m refers to the number of environmental variables.

2.3. Validation and evaluation

The continuous Boyce index (Hirzel *et al*. 2006), computed by comparing a suitability map against a validation data set (e.g., independently collected ground truth wildlife presence locations) was adopted to validate the accuracy of predicted habitat suitability map(s). The basic idea for computing the Boyce index (Boyce *et al*. 2002) was as follows. The habitat suitability range was partitioned into several classes (or bins). The Boyce index is the Spearman-rank correlation between area-adjusted frequency of validation points within individual bins and the bin rank. The continuous Boyce index $B_{\text{cont}}(W)$ overcomes the shortcomings of the Boyce index by calculating with a moving window (W is the window width, usually set to 0.1) instead of dividing into fixed classes. It is a robust and reliable accuracy measure of presence-only-based prediction (Hirzel *et al*. 2006). $B_{\text{cont}}(W)$ varies from −1.0 to 1.0. A positive value indicates a suitability map that is consistent with the distribution of presence locations in the validation data set, that is, the higher the suitability rank, the larger the proportion of presence locations in the validation data set that fall in areas of this suitability rank on the map. A value close to 0 indicates a suitability map that is not different from one predicted from a chance (random) model. A negative value indicates a suitability map that contradicts the distribution of presence locations in the validation data set (Hirzel *et al*. 2006). Details on how to compute $B_{\text{cont}}(W)$ for a habitat suitability map based on a validation data set can be found in Hirzel *et al*. (2006).

Wildlife presence locations collected by field biologists were used as the validation data set to evaluate the accuracy of the habitat suitability map predicted from citizen data. To evaluate the effectiveness of the proposed approach to addressing location imprecision and spatial bias in citizen data, the accuracy of suitability maps (validated against the validation data set) predicted from citizen data processed using the approach and maps predicted from citizen data processed without using the approach were compared.

3. Case study

3.1. Study area

The study area is located at Mt. Lasha (99°13′E–99°17′14″E; 26°17′52″N–26°20′58″N) in northwest Yunnan Province, China (Figure 1). The 20.3 km^2 study area is part of the Yunling Nature Reserve (established in 2006) and is an important habitat for a population of the black-and-white snub-nosed monkey (*R. bieti*). *R. bieti* is an endangered species endemic to the eastern Himalayas in northwest Yunnan and southeast Tibet, China, between the upper Mekong and Yangtze Rivers. Mt. Lasha is near the southern-most portion of the geographic range for the snub-nosed monkey (Long *et al*. 1994, Xiao *et al*. 2003). The Mt. Lasha group of *R. bieti* contains about 100 individuals (Huang *et al*. 2012).

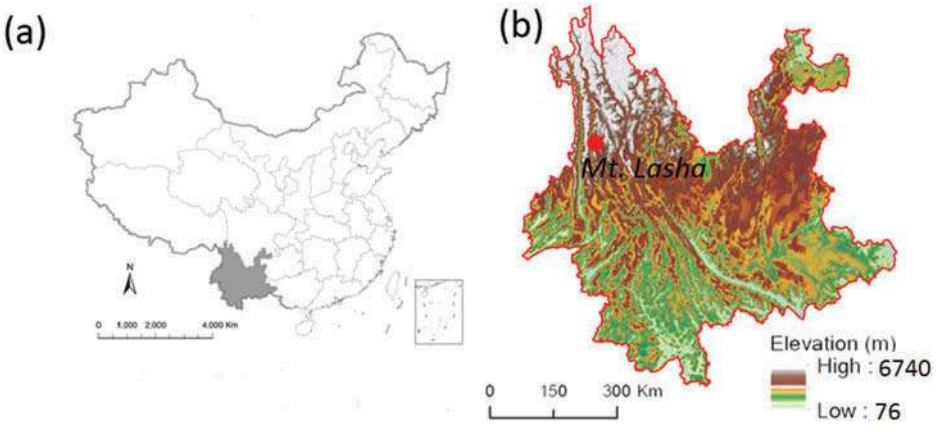

Figure 1. Location of study area. (a) Yunnan in China; (b) Mt. Lasha in Yunnan.

3.2. *Citizen data*

Citizen data were collected through interview sessions with local citizens during July and August 2010. Local citizens who had extensive experience in the field were interviewed. In total, 70 local citizens (including herdsmen, hunters and farmers) from all five nearby villages were interviewed. Local citizens in the area were more intensively active in the field in summer and therefore had more opportunities to spot the monkeys. Sightings from local citizens over the 2006–2010 summer months (June, July, and August) were then extracted (Figure 2).

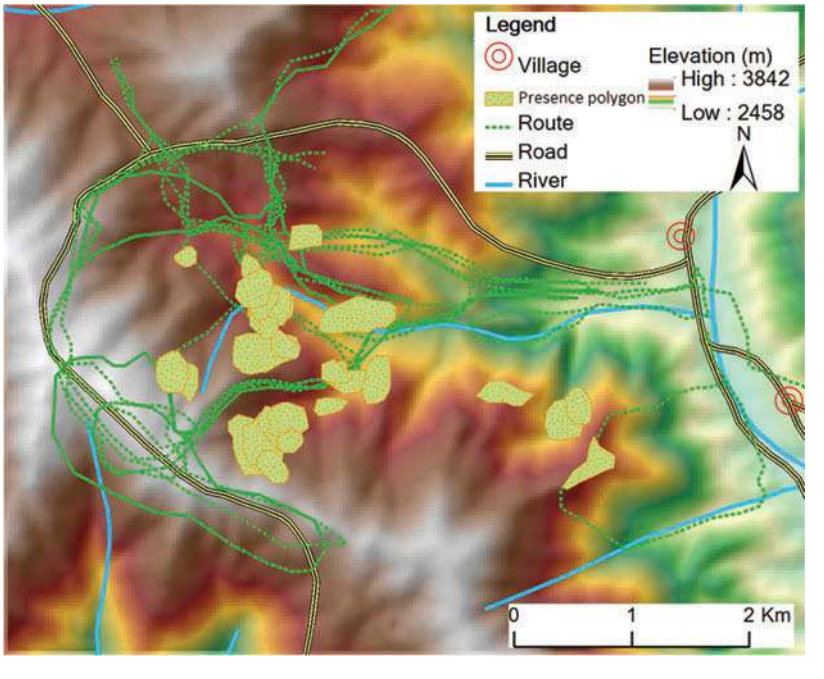

Figure 2. Sightings of *R. bieti* (27 presence polygons) and routes taken by local citizens over the 2006–2010 summer months.

Table 1. Environmental variables selected to characterize factors that influence *R. bieti* habitat suitability (the resolution of data is 30 m).

Category	Environmental variable	Value range
Terrain	Elevation	2460–3843 m
	Slope	0–52 degree
	Aspect	8 categories
Water source	Cost distance to river	0–628 m
Shelter and food	Vegetation type	10 types
Human impact	Cost distance to village or road	0–788 m

Notes: Eight slope aspect categories are: 1-North (0–22.5; 337.5–360) (in degree), 2-Northeast (22.5–67.5), 3-East (67.5–112.5), 4-Southeast (112.5–157.5), 5-South (157.5–202.5), 6-Southwest (202.5–247.5), 7-West (247.5–292.5), and 8-Northwest (292.5–337.5). In calculation of cost distances, tan(slope) was adopted as cost. Ten vegetation types are: 1-Evergreen coniferous, 2-Deciduous broadleaf, 3-Deciduous broadleaf and coniferous, 4-Evergreen broadleaf and coniferous, 5-Pasture, 6-Burning remains, 7-Charcoal remains, 8-Farmland, 9-Yunnan pine, and 10-Shrubberies.

3.3. *Environmental data*

Environmental variables on factors that influence *R. bieti* habitat suitability were selected to characterize the environmental conditions. For *R. bieti*, the factors include terrain conditions, water source, shelter and food, and human impact (Kirkpatrick 1996, Xiao *et al*. 2003, Huang 2009) (Table 1). A DEM at 30 m resolution was created from the contours digitized from 1:50,000 topographic maps of the area (20 m contour interval). Elevation, slope gradient, and slope aspect were derived from the DEM (Wilson and Gallant 2000) and used to characterize the terrain factor. Least-cost distances (ESRI 2010) to rivers and to villages or roads were computed to characterize the water source factor and human impact factor, respectively. A vegetation-type map of the study area derived from a field inventory (Huang 2009) was included to represent characteristics of forest structure and food resources.

3.4. *Habitat suitability map*

The relationships between *R. bieti* habitat suitability and the environmental variables (presented as normalized presence PDFs) are shown in Figure 3. The differences between derived suitability–environment relationships for which spatial bias is compensated and relationships for which spatial bias is not compensated were apparent, especially for the relationships with respect to elevation, vegetation type, and cost distance to river.

The habitat suitability map that was predicted based on derived suitability–environment relationships that compensated for spatial bias is shown in Figure 4 (cumulative visibility for bias compensation was computed with visible distance threshold 500 m, an estimation based on field experience in the study area). The following spatial patterns can be observed on this habitat suitability map: a large portion of the study area was predicted to have low suitability, mainly including areas over ridges, valleys, and southwest hill slopes; areas predicted to have relatively high suitability were located over middle elevation range and were distributed primarily on northeast hill slopes; small patches of relatively high suitability were isolated.

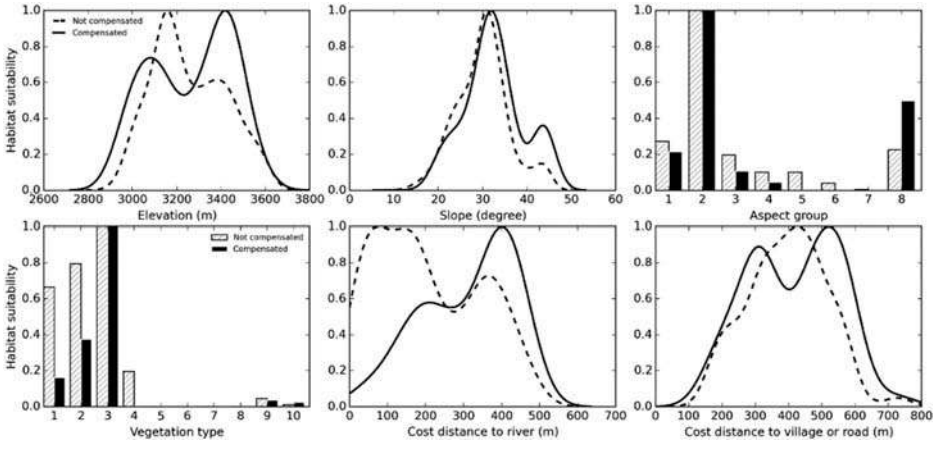

Figure 3. Suitability–environment relationships for *R. bieti*. Solid lines or bars are relationships derived with spatial bias compensated for. Dashed lines or bars are relationships derived with spatial bias not compensated for.

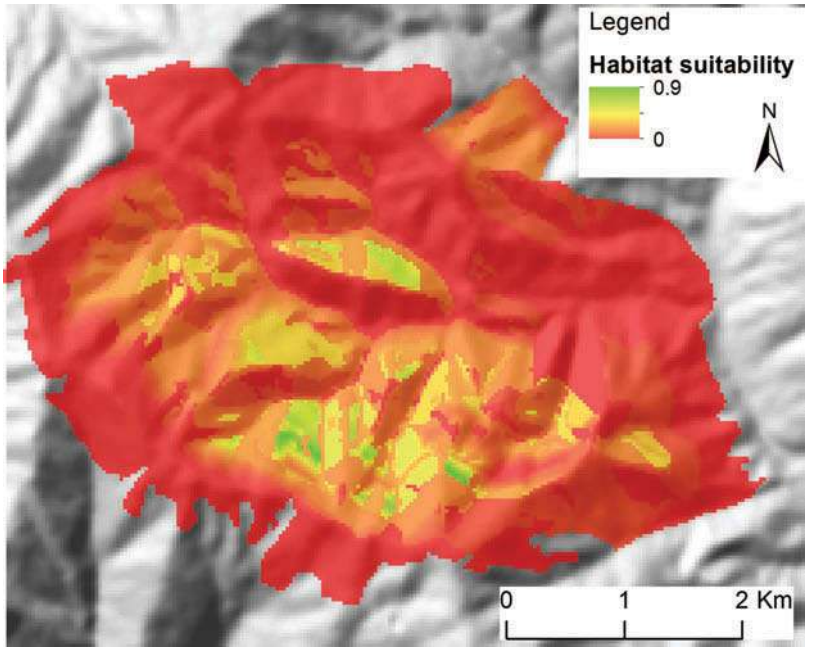

Figure 4. Habitat suitability map for *R. bieti* predicted using citizen data.

3.5. Validation

Presence locations of *R. bieti* recorded by field biologists in intensive field tracking were used as validation data to evaluate the accuracy of the predicted habitat suitability map. The field tracking of the Mt. Lasha *R. bieti* group (Huang *et al*. 2012) was conducted by one field biologist and two assistants. Nearly one year was spent in the field habituating

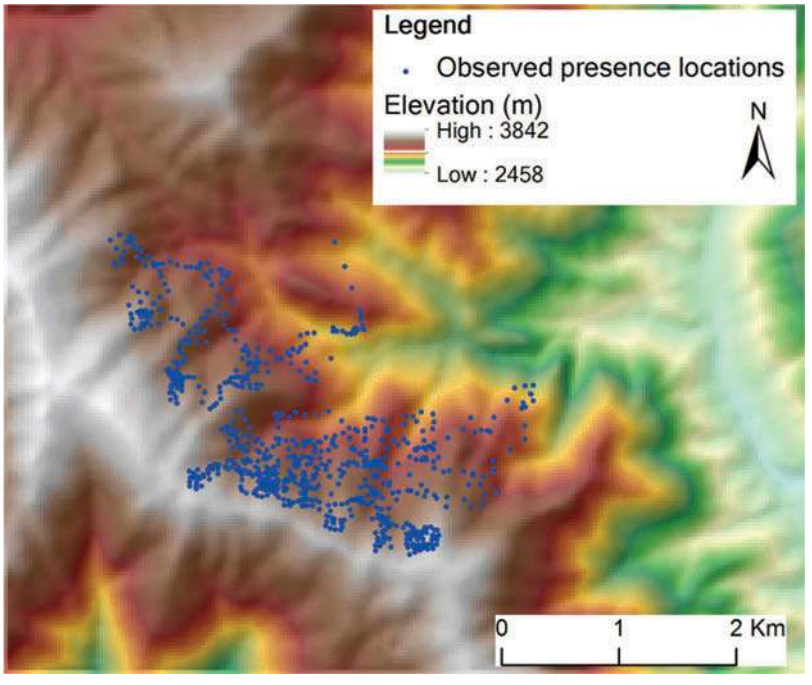

Figure 5. Presence locations of *R. bieti* observed in the field over summers in 2008 and 2009.

the monkeys. Tracking and direct observation of the group of monkeys were then conducted in 2008 and 2009 (primarily for behavioral studies). Locations of the group of *R. bieti* were recorded every 30 minutes during the field tracking periods. Due to the endangered status, small population size, and the secluded nature of *R. bieti*, the presence locations of *R. bieti* recorded in field tracking form the most detailed data set available reflecting the distribution of *R. bieti* in the study area. Presence locations of *R. bieti* over summers in these two years (853 presence locations in total) were selected as validation data to temporally match citizen data used to predict the habitat suitability map (Figure 5).

Based on these field-observed *R. bieti* presence locations, $B_{\text{cont}}(0.1)$ calculated for the predicted habitat suitability map was 0.873 (95% CI: [0.810, 0.917]) (Table 2), indicating that the habitat suitability map predicted from citizen data was highly consistent with the distribution of *R. bieti* presence observed in the field.

4. Discussion

4.1. *Effectiveness of the approach*

As shown in Table 2, if the frequency-sampling strategy was not applied to identify representative presence locations within sighting polygons (i.e., taking all locations within sighting polygons as representative presence locations to predict a habitat suitability map), $B_{\text{cont}}(0.1)$ for the predicted habitat suitability map was 0.173 (95% CI: [−0.048, 0.379]). Clearly, with the frequency-sampling strategy applied to reduce location imprecision, there was a significant improvement in accuracy of the predicted habitat suitability map. This indicates that the frequency-sampling strategy was effective to reduce location imprecision in citizen data.

Table 2. Evaluation of predicted habitat suitability maps.

	Not compensated		Compensated	
	$B_{cont}(0.1)$	95% CI	$B_{cont}(0.1)$	95% CI
Not sampled	−0.126	[−0.333, 0.092]	0.173	[−0.048, 0.379]
Sampled	−0.048	[−0.269, 0.178]	0.873	[0.810, 0.917]

Notes: 'Compensated' or 'Not compensated' means spatial bias was compensated for or not based on cumulative visibility calculated with a visible distance threshold of 500 m. 'Sampled' means applying frequency-sampling strategy to obtain representative presence locations from sighting polygons. 'Not Sampled' means simply taking all the pixels in presence polygons as representative presence locations. '95% CI' stands for 95% confidence interval for $B_{cont}(0.1)$ computed based on *Fisher's z-transformation*.

Table 2 also shows that, if spatial bias in presence locations was not compensated for, $B_{cont}(0.1)$ for the predicted habitat suitability map was −0.048 (95% CI: [−0.269, 0.178]). The accuracy of the habitat suitability map when it compensated for spatial bias was much higher than that of the map with no spatial bias compensation. This indicates that the cumulative visibility-based method was fairly effective at compensating for spatial bias in citizen data.

It is also worth noting that, without applying either frequency-sampling strategy or cumulative visibility to reduce location imprecision or to compensate for spatial bias, $B_{cont}(0.1)$ for the predicted habitat suitability map was −0.126 (95% CI: [−0.333, 0.092]) (Table 2). Clearly, there was a significant improvement in the accuracy of the predicted habitat suitability map with the two strategies applied, which indicates that the geospatial analysis techniques we adopted, as a whole, effectively mitigated the impacts of the two limitations of citizen data in predictive mapping.

4.2. *Sensitivity analysis*

4.2.1. *Estimation of visible distance threshold*

As discussed above, the cumulative visibility-based method effectively compensated for spatial bias in citizen data. In calculating cumulative visibility, visible distance threshold is a crucial parameter (initially estimated as 500 m based on field experience in the case study area). The estimation of this visible distance threshold might have an impact on the effectiveness of the spatial bias compensation method. To investigate this impact, cumulative visibility was calculated with variable visible distance thresholds to compensate for spatial bias in citizen data, and the accuracy of the predicted habitat suitability maps was evaluated.

Results (Figure 6) showed that, there was a significant increase in $B_{cont}(0.1)$ as visible distance threshold increased from 100 to 200 m; $B_{cont}(0.1)$, reaching the highest value of 0.917 when visible distance threshold increased to 300 m; and that it was around 0.9 when visible distance threshold was in the range of 200–500 m, but decreased dramatically to negative values after visible distance threshold exceeded 500 m. This suggests that suitability maps predicted based on citizen data can be of high accuracy with an estimation of visible distance threshold in the range of 200–500 m for the case study area. The optimal estimation of the visible distance threshold for the case study area is 300 m, though 500 m is already a reasonably good estimation.

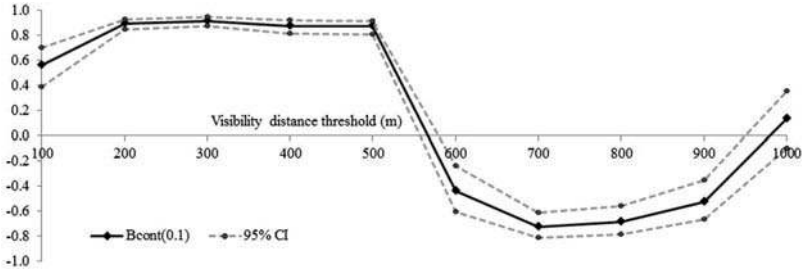

Figure 6. The impact of visible distance threshold on the accuracy of predicted habitat suitability maps.

It should be noted that in this paper cumulative visibility computed based on DEM is only an approximation of the realistic visibility over the landscapes. In fact, other than terrain condition, factors such as weather, vegetation coverage, and the eyesight of local citizens could also influence the visibility at a specific location from local citizens. Visibility under such complex conditions could be modeled if detailed information on these factors was available. For example, visibility in vegetated conditions can be modeled if spatial variation of vegetation structure and vegetation density is well known (Llobera 2007). However, for the study area, detailed information on these factors was unavailable. Therefore, the simplistic DEM-based viewshed analysis algorithm (Wang *et al*. 2000) was applied to compute visibility. In this sense, the visible distance threshold should be regarded as an estimation of the average visible distance threshold under the aggregated effects of all the factors influencing visibility for local citizens.

4.2.2. Impact of sample size

The number of pixels selected as presence locations (referred to as 'sample size') might have an impact on the accuracy of the predicted habitat suitability map. To assess this impact, presence location sets (referred to as 'sample sets') of variable sizes were obtained by randomly selecting a subset of the 542 presence locations (pixels). Accuracy of the habitat suitability map predicted from each sample set was evaluated. For each given sample size, this process was repeated 20 times to obtain a distribution of $B_{\text{cont}}(0.1)$.

Results (Figure 7) showed that, as sample size decreased, the median value of $B_{\text{cont}}(0.1)$ was consistently high (around 0.9), but the spread of $B_{\text{cont}}(0.1)$ widened slightly. For all sample sizes larger than 200, the minimum value of $B_{\text{cont}}(0.1)$ was greater than 0.8. As long as sample size was larger than 200, the accuracy was consistently high. For small sample sizes (e.g., 50), the accuracy deviated to a much lower value from the median. This might be because when sample size was small, there were no presence locations selected within some sighting polygons, and therefore only a smaller amount of information in citizen data was used for prediction. Overall, for the monkey habitat suitability mapping case study, accuracy of the predicted habitat suitability map was fairly robust when sample size was above 200. In general, the full set of presence locations identified with frequency-sampling strategy is recommended for prediction.

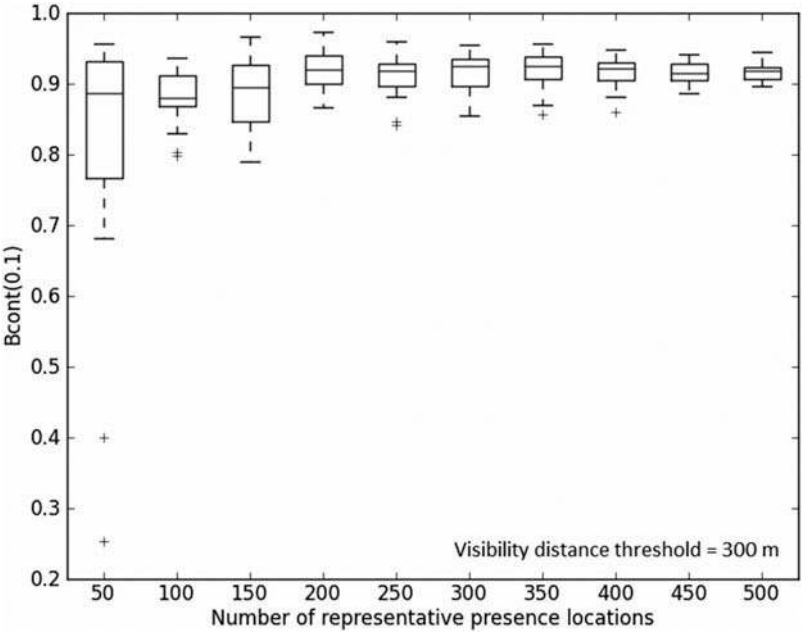

Figure 7. The impact of sample size on the accuracy of predicted habitat suitability maps. Boxplot at each given sample size was based on 20 values of $B_{cont}(0.1)$, each calculated on a suitability map predicted from a representative presence location set randomly selected from the 542 presence locations (pixels).

4.2.3. *Sensitivity to uncertainty in sighting polygon boundaries*

Wildlife presence data were elicited from local citizens by letting them draw polygons indicating where the wildlife had presented. When local citizens were drawing such polygons, though empowered by the intuitive 3D map of the area, uncertainties regarding where to draw polygon boundaries were unavoidable. For example, if a local citizen sighted monkeys in a large forested area, it is nearly impossible to draw a polygon that exactly delineates the extent of the monkeys' presence, due to the lack of easily recognizable references over such a homogenous area on the map. It was thus likely that the polygon drawn by the local citizen was larger or smaller than the actual size of the polygon. This uncertainty associated with polygon boundaries could possibly affect the accuracy of habitat suitability maps that are predicted based on citizen data. To assess this effect, we introduced different levels of area change into the sighting polygons to simulate such uncertainty. This was done by randomly enlarging or shrinking the sighting polygons proportionally while retaining their geometric centers and shapes. For example, for an area change level of 0.1, every sighting polygon was randomly either enlarged to 1.1 times the original area or reduced to 0.9 of the original area. Based on sighting polygons with introduced random area change, a habitat suitability map was predicted and its accuracy was evaluated. For each given area change level, this process was repeated 20 times to obtain a distribution of $B_{cont}(0.1)$.

Results (Figure 8) showed that accuracy of the predicted habitat suitability map decreased with the increase of area change level introduced in sighting polygons. However, as long as the area change level did not exceed 0.5 (50%), the accuracy of the predicted habitat suitability map remained consistently high with a minimum B_{cont}

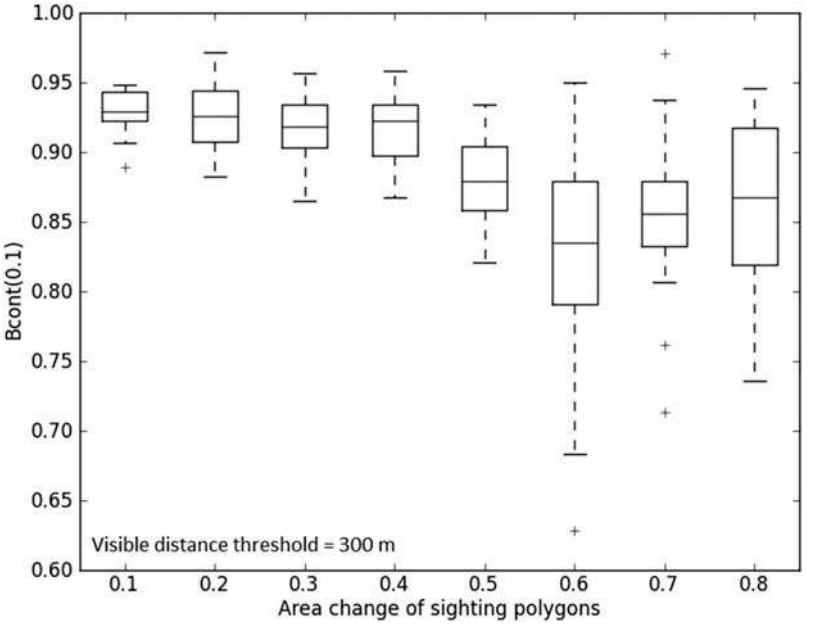

Figure 8. Impact of uncertainty in sighting polygon boundaries on the accuracy of the predicted habitat suitability map. Boxplot at each given area change level was based on 20 values of B_{cont} (0.1), each calculated on a suitability map predicted from sighting polygons introduced with random area change at this level.

(0.1) greater than 0.8. This indicates that, for the monkey habitat suitability mapping case study, the accuracy of the predicted habitat suitability map was fairly protected from the uncertainty associated with boundaries of sighting polygons. Even so, however, a general principle for soliciting sighting polygons from local citizens is to let them draw the polygons as small and accurate as possible to minimize uncertainties at the very beginning of generating citizen data.

4.2.4. *Impact of bandwidths in kernel density estimation*

Kernel density estimation was used to assess the PDFs of environmental conditions over the presence locations and so to quantify the relationships between wildlife habitat suitability and environmental conditions (Section 2.2.3.1). The bandwidth h'_x in Equation (4) is a critical parameter that determines the smoothness of the estimated PDFs and would therefore have an impact on the predictive mapping result. The 'rule-of-thumb' algorithm (Silverman 1986) was applied to determine bandwidth h'_x for each environmental variable x (Equation 6). Here the impact of h'_x on the accuracy of the predicted suitability map was further investigated by multiplying a coefficient a to h'_x (the bandwidth determined using the 'rule-of-thumb' algorithm):

$$H'_x = ah'_x, \tag{9}$$

where a varies from 0.1 to 2.0 with a 0.1 increment to decrease or increase the bandwidth H'_x (H'_x is equal to h'_x when a is 1.0). Figure 9(b) is an example of how bandwidth affects

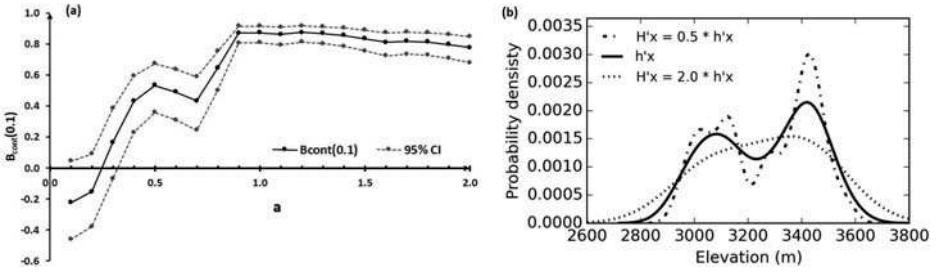

Figure 9. Impact of bandwidth in kernel density estimation on accuracy of the predicted habitat suitability map (a). The 542 presence locations identified with frequency-sampling strategy were used for prediction. Cumulative visibility computed at 500 m visibility threshold was used to compensate for spatial bias. (b) An example of how bandwidth H'_x affects the smoothness/shape of estimated PDF of elevation over the presence locations.

the smoothness/shape of the estimated PDF. The accuracy of suitability maps predicted with these various bandwidths were then compared. Results (Figure 9(a)) showed that when a is smaller than 0.9, prediction accuracy dropped dramatically ($B_{\text{cont}}(0.1)$ below 0.7). Prediction accuracy was consistently high when a was in a range from 0.9 to 1.4 ($B_{\text{cont}}(0.1)$ above 0.85), and then dropped slightly when a was larger than 1.4. This demonstrated that 'rule-of-thumb' is indeed a good algorithm to determine the bandwidth for kernel density estimation and achieve high prediction accuracy.

4.2.5. Suggestions on parameter selection

Several guidelines and suggestions related to parameter selection were presented for applying the proposed citizen-data-based predictive mapping approach. First, in soliciting citizen data, interviewers should let local citizens draw the sighting polygons as small and accurate as possible to minimize the uncertainty associated with polygon boundaries at the very beginning of generating citizen data, although the monkey habitat mapping case study showed that mapping accuracy holds up to a certain extent against such uncertainty. Second, it is recommended that all the representative locations identified with the frequency-sampling strategy from sighting polygons be used as samples for predictive mapping. Third, in computing cumulative visibility to compensate for spatial bias in citizen data the visible distance threshold should be estimated based on field experience in the respective study area (e.g., in the monkey habitat mapping case study it was estimated as 500 m according to field visibility in the study area). Last, in deriving suitability–environment relationships using kernel density estimation, the 'rule-of-thumb' (Silverman 1986) is a robust algorithm to determine bandwidth.

4.3. Potential limitations of the approach

The proposed citizen-data-based mapping approach adopted specific geospatial analysis techniques to address the issues of location imprecision and spatial bias in citizen data and it is worth noting the potential limitations. The limitations for other applications are very much related to the two core assumptions used to address the location imprecision and spatial bias. The assumption in reducing location imprecision is that locations at which environmental conditions are most prevalent over the polygon area would best represent

the actual location of the observed geographic phenomenon. This assumption worked well as a first approximation for the monkey habitat mapping case studies. The suitability of this assumption for other cases very much depends on the nature of how the citizen data were generated and whether such an assumption would hold. In correcting for spatial bias it was assumed that the spatial bias was mostly due to the visibility of the observers. If this assumption does not hold, visibility from locations of observers will not be a good measure to compensate for the spatial bias. One important point to be made by this paper is that in working with citizen data one must understand the nature of how these data were generated and then use this understanding to preprocess the data so that the data meet the requirements of the application.

4.4. *Potential use of the approach*

The *R. bieti* habitat mapping case study shows that the proposed approach effectively minimized the impacts of the two limitations associated with citizen data (namely, location imprecision and spatial bias) and as a result, habitat suitability mapping based on citizen data achieved high accuracy. The approach presented in this paper also has implications for tackling data quality issues associated with citizen-contributed data in general (e.g., VGI and data produced in citizen science projects). It implies that, with the application of geospatial analysis techniques to properly account for the limitations of location imprecision and spatial bias in citizen data, valuable information embedded in citizen data can be extracted and used for scientific mapping.

The approach presented in this article thus has the potential to be used in a range of cases. For example, conservation programs in poor and remote areas where most of the world's biodiversity occurs (Myers *et al*. 2000) usually have only limited budgets, and might not be able to afford conducting traditional wildlife data collection protocols (e.g., field tracking, GPS collars). In such cases, data can be solicited from local citizens at low cost, processed with geospatial analysis techniques to increase its quality, and then used to support decision-making in conservation practices. In addition, citizens' observations include various geographic phenomena. This approach, therefore, can potentially be used for predictive mapping of geographic phenomena of interest other than wildlife habitat suitability.

5. Conclusions

This article presented an approach that employs geospatial analysis techniques to minimize the limitations of citizen data, location imprecision and spatial bias in particular, for predictive mapping of geographic phenomena. The approach minimizes the impact of location imprecision in citizen data by using a frequency-sampling strategy to identify representative locations from areas over which citizens observed the geographic phenomenon. It mitigates the impact of spatial bias in citizen data by using cumulative visibility (the frequency of a given location seen by local citizens) to compensate for the bias due to unequal visibility over the landscape.

As a case study and test of the principle, this approach was applied to *R. bieti* habitat suitability mapping in Yunnan, China. Citizen data (the sightings of *R. bieti*) were solicited from local citizens and then processed using the proposed approach. A habitat suitability map predicted from the processed citizen data was validated using *R. bieti* presence locations recorded by field biologists in an intensive field tracking. A validation revealed that the predicted suitability map was highly consistent with the distribution of *R.*

bieti presence observed in the field ($B_{cont}(0.1)$ = 0.873, 95% CI: [0.810, 0.917]). An evaluation of the approach showed that the proposed approach effectively minimized the impacts of the location imprecision and spatial bias associated with citizen data in predictive mapping.

The proposed approach has implications for tackling data quality issues of citizen-contributed data in general. It implies that, with the application of geospatial analysis techniques to properly account for the limitations in citizen data (e.g., location imprecision and spatial bias), valuable information embedded in citizen data can be extracted and used for scientific mapping.

Acknowledgements

The support received by A-Xing Zhu through the Vilas Associate Award, the Hammel Faculty Fellow, the Vilas Distinguished Achievement Professorship, and the Manasse Chair Professorship from the University of Wisconsin-Madison is greatly appreciated.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This study was funded by the National Natural Science Foundation of China (NSFC) [Project Numbers: 41431177, 30960085]; Natural Science Research Program of Jiangsu [14KJA170001]; Priority Academic Program Development of Jiangsu Higher Education Institutions; National Basic Research Program of China [Project Number: 2015CB954102]; 'One Thousand Talent' Program of China. Cheng-Zhi Qin thanks the National Science Foundation of China [Project Number: 41422109] for its support.

References

Ali, A. and Schmid, F., 2014. Data quality assurance for volunteered geographic information. *In*: M. Duckham *et al.*, eds. *Geographic information science*. Cham: Springer International Publishing, 126–141.

Altman, D., 1994. Research article. Fuzzy set theoretic approaches for handling imprecision in spatial analysis. *International Journal of Geographical Information Systems*, 8 (3), 271–289. doi:10.1080/02693799408902000

Anadón, J.D., *et al*., 2009. Evaluation of local ecological knowledge as a method for collecting extensive data on animal abundance. *Conservation Biology*, 23 (3), 617–625. doi:10.1111/cbi.2009.23.issue-3

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

Boyce, M.S., *et al*., 2002. Evaluating resource selection functions. *Ecological Modelling*, 157 (2–3), 281–300. doi:10.1016/S0304-3800(02)00200-4

Burt, J.E. and Zhu, A.X., 2004. *3dMapper*® [online]. Madison, WI: Terrain Analytics LLC. Available from: http://www.terrainanalytics.com [Accessed 14 March 2010].

de Solla, S.R., *et al*., 2005. Effect of sampling effort and species detectability on volunteer based anuran monitoring programs. *Biological Conservation*, 121 (4), 585–594. doi:10.1016/j.biocon.2004.06.018

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

Elwood, S., 2008. Volunteered geographic information: key questions, concepts and methods to guide emerging research and practice. *GeoJournal*, 72 (3–4), 133–135. doi:10.1007/s10708-008-9187-z

Elwood, S., Goodchild, M.F., and Sui, D., 2013. Prospects for VGI research and the emerging fourth paradigm. *In*: D. Sui, S. Elwood, and M. Goodchild, eds. *Crowdsourcing geographic knowledge: volunteered geographic information (VGI) in theory and practice*. Dordrecht, Netherlands: Springer, 361–375.

ESRI, 2010. Cost distance algorithm [online]. *ArcGIS Desktop 9.3 Help*. Available from: http://webhelp.esri.com/arcgisdesktop/9.3/index.cfm?TopicName=Understanding_cost_distance_analysis [Accessed 14 March 2010].

Fernandez, M.A., *et al*., 2009. Locality uncertainty and the differential performance of four common niche-based modeling techniques. *Biodiversity Informatics*, 6, 36–52. doi:10.17161/bi.v6i1.3314

Flanagin, A. and Metzger, M., 2008. The credibility of volunteered geographic information. *GeoJournal*, 72 (3–4), 137–148. doi:10.1007/s10708-008-9188-y

Foody, G.M., *et al*., 2013. Assessing the accuracy of volunteered geographic information arising from multiple contributors to an Internet based collaborative project. *Transactions in GIS*, 17 (6), 847–860. doi:10.1111/tgis.2013.17.issue-6

Franklin, J., 1995. Predictive vegetation mapping: geographic modelling of biospatial patterns in relation to environmental gradients. *Progress in Physical Geography*, 19 (4), 474–499. doi:10.1177/030913339501900403

Franklin, J. and Miller, J.A., 2009. *Mapping species distributions: spatial inference and prediction*. Cambridge: Cambridge University Press.

Freedman, D. and Diaconis, P., 1981. On the histogram as a density estimator: L2 theory. *Probability Theory and Related Fields*, 57 (4), 453–476.

Gillespie, T.W., *et al*., 2008. Measuring and modelling biodiversity from space. *Progress in Physical Geography*, 32 (2), 203–221. doi:10.1177/0309133308093606

Girres, J. and Touya, G., 2010. Quality assessment of the French OpenStreetMap dataset. *Transactions in GIS*, 14 (4), 435–459. doi:10.1111/tgis.2010.14.issue-4

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

Goodchild, M., 2007b. Citizens as voluntary sensors: spatial data infrastructure in the world of Web 2.0. *International Journal of Spatial Data Infrastructures Research*, 2, 24–32.

Goodchild, M.F. and Glennon, J.A., 2010. Crowdsourcing geographic information for disaster response: a research frontier. *International Journal of Digital Earth*, 3 (3), 231–241. doi:10.1080/17538941003759255

Goodchild, M.F. and Li, L., 2012. Assuring the quality of volunteered geographic information. *Spatial Statistics*, 1 (0), 110–120. doi:10.1016/j.spasta.2012.03.002

Graham, C.H., *et al*., 2004. New developments in museum-based informatics and applications in biodiversity analysis. *Trends in Ecology & Evolution*, 19 (9), 497–503. doi:10.1016/j.tree.2004.07.006

Graham, C.H., *et al*., 2008. The influence of spatial errors in species occurrence data used in distribution models. *Journal of Applied Ecology*, 45 (1), 239–247. doi:10.1111/j.1365-2664.2007.01408.x

Guan, W.W., *et al*., 2012. WorldMap – a geospatial framework for collaborative research. *Annals of GIS*, 18 (2), 121–134. doi:10.1080/19475683.2012.668559

Guisan, A. and Zimmermann, N.E., 2000. Predictive habitat distribution models in ecology. *Ecological Modelling*, 135 (2–3), 147–186. doi:10.1016/S0304-3800(00)00354-9

Haining, R.P., 2003. *Spatial data analysis*. Cambridge: Cambridge University Press.

Haklay, M., 2010. How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and Planning B: Planning and Design*, 37 (4), 682–703. doi:10.1068/b35097

Haklay, M., 2013. Citizen science and volunteered geographic information: overview and typology of participation. *In*: D. Sui, S. Elwood, and M. Goodchild, eds. *Crowdsourcing geographic knowledge: volunteered geographic information (VGI) in theory and practice*. Dordrecht, Netherlands: Springer, 105–122.

Haklay, M. and Weber, P., 2008. OpenStreetMap: user-generated street maps. *IEEE Pervasive Computing*, 7 (4), 12–18. doi:10.1109/MPRV.2008.80

Hirzel, A.H., *et al*., 2002. Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data?. *Ecology*, 83 (7), 2027–2036. doi:10.1890/0012-9658(2002)083[2027:ENFAHT]2.0.CO;2

Hirzel, A.H., *et al*., 2006. Evaluating the ability of habitat suitability models to predict species presences. *Ecological Modelling*, 199 (2), 142–152. doi:10.1016/j.ecolmodel.2006.05.017

Hirzel, A.H. and Le, L.G., 2008. Habitat suitability modelling and niche theory. *Journal of Applied Ecology*, 45, 1372–1381. doi:10.1111/jpe.2008.45.issue-5

Huang, Z.P., 2009. Foraging, reproduction and sleeping site selection of black-and-white snub-nosed monkey (*Rhinopithecus bieti*) at the southern range. Thesis (MS). Southwest Forestry University, Kunming.

Huang, Z.-P., *et al*., 2012. *Seasonality of reproduction of wild black-and-white snub-nosed monkeys (Rhinopithecus bieti*) at Mt. Lasha, Yunnan, China. *Primates*, 53 (3), 237–245. doi:10.1007/s10329-012-0305-7

Izenman, A., 1991. Recent developments in nonparametric density estimation. *Journal of the American Statistical Association*, 86 (413), 205–224.

Jones, C.B., *et al*., 2008. Modelling vague places with knowledge from the web. *International Journal of Geographical Information Science*, 22 (10), 1045–1065. doi:10.1080/13658810701850547

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

Kerr, J.T. and Ostrovsky, M., 2003. From space to species: ecological applications for remote sensing. *Trends in Ecology & Evolution*, 18 (6), 299–305. doi:10.1016/S0169-5347(03)00071-5

Khalili, N., Wood, J., and Dykes, J., 2010. Analysing uncertainty in home location information in a large volunteered geographic information database. *In*: *Proceedings of the GIS research UK 18th annual conference*. London, UK: University College London, 14–16.

Kirkpatrick, R.C., 1996. Ecology and behavior of the Yunnan snub-nosed langur (*Rhinopithecus bieti, Colobinae*). Thesis (PhD). University of California, Davis.

Kramer-Schadt, S., *et al*., 2013. The importance of correcting for sampling bias in MaxEnt species distribution models. *Diversity and Distributions*, 19 (11), 1366–1379. doi:10.1111/ddi.12096

Leitão, P.J., Moreira, F., and Osborne, P.E., 2011. Effects of geographical data sampling bias on habitat models of species distributions: a case study with steppe birds in southern Portugal. *International Journal of Geographical Information Science*, 25 (3), 439–454. doi:10.1080/13658816.2010.531020

Li, J. and Hilbert, D.W., 2008. LIVES: a new habitat modelling technique for predicting the distribution of species' occurrences using presence-only data based on limiting factor theory. *Biodiversity and Conservation*, 17 (13), 3079–3095. doi:10.1007/s10531-007-9270-7

Llobera, M., 2007. Modeling visibility through vegetation. *International Journal of Geographical Information Science*, 21 (7), 799–810. doi:10.1080/13658810601169865

Long, Y.C., Kirkpatrick, C.R., and Xiaolin, Z., 1994. *Report on the distribution, population, and ecology of the Yunnan snub-nosed monkey (Rhinopithecus bieti*). *Primates*, 35 (2), 241–250. doi:10.1007/BF02382060

Moudrý, V. and Šímová, P., 2012. Influence of positional accuracy, sample size and scale on modelling species distributions: a review. *International Journal of Geographical Information Science*, 26 (11), 2083–2095. doi:10.1080/13658816.2012.721553

Munson, M.A., *et al*., 2010. A method for measuring the relative information content of data from different monitoring protocols. *Methods in Ecology and Evolution*, 1 (3), 263–273.

Myers, N., *et al*., 2000. Biodiversity hotspots for conservation priorities. *Nature*, 403 (6772), 853–858. doi:10.1038/35002501

Osborne, P.E. and Leitão, P.J., 2009. Effects of species and habitat positional errors on the performance and interpretation of species distribution models. *Diversity and Distributions*, 15 (4), 671–681. doi:10.1111/ddi.2009.15.issue-4

Phillips, S.J., Anderson, R.P., and Schapire, R.E., 2006. Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, 190 (3–4), 231–259. doi:10.1016/j.ecolmodel.2005.03.026

Pittman, S.J., *et al*., 2007. Predictive mapping of fish species richness across shallow-water seascapes in the Caribbean. *Ecological Modelling*, 204 (1–2), 9–21. doi:10.1016/j.ecolmodel.2006.12.017

Ponder, W.F., *et al*., 2001. Evaluation of museum collection data for use in biodiversity assessment. *Conservation Biology*, 15 (3), 648–657. doi:10.1046/j.1523-1739.2001.015003648.x

Porter, M.D. and Reich, B.J., 2012. Evaluating temporally weighted kernel density methods for predicting the next event location in a series. *Annals of GIS*, 18 (3), 225–240. doi:10.1080/19475683.2012.691904

Qi, F. and Zhu, A.-X., 2003. Knowledge discovery from soil maps using inductive learning. *International Journal of Geographical Information Science*, 17 (8), 771–795. doi:10.1080/1365881031000159604 9

Reddy, S. and Dávalos, L.M., 2003. Geographical sampling bias and its implications for conservation priorities in Africa. *Journal of Biogeography*, 30 (11), 1719–1727. doi:10.1046/j.1365-2699.2003.00946.x

Robinson, V.B., 2003. A perspective on the fundamentals of fuzzy sets and their use in geographic information systems. *Transactions in GIS*, 7 (1), 3–30. doi:10.1111/tgis.2003.7.issue-1

Rüger, N., Schlüter, M., and Matthies, M., 2005. A fuzzy habitat suitability index for populus euphratica in the Northern Amudarya delta (Uzbekistan). *Ecological Modelling*, 184 (2–4), 313–328. doi:10.1016/j.ecolmodel.2004.10.010

Scull, P., *et al*., 2003. Predictive soil mapping: a review. *Progress in Physical Geography*, 27 (2), 171–197. doi:10.1191/0309133303pp366ra

Seeger, C., 2008. The role of facilitated volunteered geographic information in the landscape planning and site design process. *GeoJournal*, 72 (3–4), 199–213. doi:10.1007/s10708-008-9184-2

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

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

Sui, D., 2008. The wikification of GIS and its consequences: or Angelina Jolie's new tattoo and the future of GIS. *Computers, Environment and Urban Systems*, 32 (1), 1–5. doi:10.1016/j.compenvurbsys.2007.12.001

Sui, D., Goodchild, M., and Elwood, S., 2013. Volunteered geographic information, the exaflood, and the growing digital divide. *In*: D. Sui, S. Elwood, and M. Goodchild, eds. *Crowdsourcing geographic knowledge: volunteered geographic information (VGI) in theory and practice*. Dordrecht: Springer, 1–12.

Sullivan, B.L., *et al*., 2009. eBird: a citizen-based bird observation network in the biological sciences. *Biological Conservation*, 142 (10), 2282–2292. doi:10.1016/j.biocon.2009.05.006

Van Broekhoven, E., *et al*., 2006. Fuzzy rule-based macroinvertebrate habitat suitability models for running waters. *Ecological Modelling*, 198 (1–2), 71–84. doi:10.1016/j.ecolmodel.2006.04.006

Viña, A., *et al*., 2008. Evaluating MODIS data for mapping wildlife habitat distribution. *Remote Sensing of Environment*, 112 (5), 2160–2169. doi:10.1016/j.rse.2007.09.012

Wang, J., Robinson, G.J., and White, K., 2000. Generating viewshed without using sightlines. *Photogrammetric Engineering & Remote Sensing*, 66 (1), 87–90.

Wieczorek, J.R., Guo, Q., and Hijmans, R.J., 2004. The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. *International Journal of Geographical Information Science*, 18 (8), 745–767. doi:10.1080/13658810412331280211

Wilson, J.P. and Gallant, J.C., 2000. *Terrain analysis: principles and applications*. New York: John Wiley and Sons.

Xiao, W., *et al*., 2003. Habitat degradation of rhinopithecus bieti in Yunnan, China. *International Journal of Primatology*, 24 (2), 389–398. doi:10.1023/A:1023009518806

Zhu, A.X., 2008. Rule-based mapping. *In*: J.P. Wilson and A.S. Fotheringham, eds. *The handbook of geographic information science*. Malden, MA: Blackwell Publishing, 273–291.

Zhu, A.-X. and Band, L.E., 1994. A knowledge-based approach to data integration for soil mapping. *Canadian Journal of Remote Sensing*, 20 (4), 408–418. doi:10.1080/07038992.1994.10874583

Zhu, A.X., *et al*., 2001. Soil mapping using GIS, expert knowledge, and fuzzy logic. *Soil Science Society of America Journal*, 65 (5), 1463–1472.

Zook, M., *et al*., 2010. Volunteered geographic information and crowdsourcing disaster relief: a case study of the Haitian Earthquake. *World Medical & Health Policy*, 2 (2), 6–32. doi:10.2202/1948-4682.1069