sitetool: an application for field site selection and evaluation

Natalie Imirzian[1]*, David Simons[2], Christina Harden[3], Emilia Johnson[4], Madison Hewitson[1,5], Sagan Friant[2,6], David Redding[1]

[1]Science Department, Natural History Museum, London, United Kingdom;
[2]Department of Anthropology, Pennsylvania State University, State College, PA, USA;
[3]Center for Infectious Disease Dynamics, The Pennsylvania State University, University Park, PA, USA;
[4]School of Biodiversity, One Health & Veterinary Medicine, University of Glasgow, Glasgow, United Kingdom;
[5]University College London, Biosciences, Gower Street, London, United Kingdom;
[6]Huck Institutes of the Life Sciences, The Pennsylvania State University, University Park, PA, USA
*Corresponding Author: nimirzian@gmail.com

Abstract

Field studies are fundamental to ecological research, yet many studies rely on unspecified or convenience-based methods for site selection, potentially introducing bias that can compromise research results. Remote-sensing data provides a quantitative way to evaluate potential sites without expensive pilot visits, however, interacting with spatial data can be computationally complex. We present an R Shiny application that integrates geospatial data into the site selection process, helping researchers generate a list of potential field sites in a region of interest and ensuring sites fall along a gradient of variation relevant to their research questions. Through integration of remote-sensing data into an easy-to-use interface, this tool improves the ability of researchers to make quantitative site selection decisions, ultimately leading to more robust studies and research results.

Introduction

Field studies are essential to research across disciplines, as they allow observation and measure of natural systems in ways irreplicable in laboratory studies.

Researchers in ecology and evolution have a long history of optimizing sampling methods for statistical inference, resulting in extensive guidance on carefully selecting sampling intervals and structures (Andrew and Mapstone 1987; Anderson 2001; Stevens and Olsen 2004; Green and Plotkin 2007; Legendre and Fortin 1989; Albert et al. 2010). Despite this guidance, systematic reviews of field studies show many studies use unspecified, haphazard, or convenience-based sampling methods (Smith, Anderson, and Pawley 2017; Lewis 2004). Poorly specified sampling designs can bias input data and resulting parameter estimates and model outcomes, possibly leading to unrepresentative or even false conclusions (Elphick 2008; Albert et al. 2010; Hurlbert 1984; Conn, Thorson, and Johnson 2017). In studies on conservation and public health, false conclusions can have costly impacts, such as incorrect species abundance estimates altering protection prioritization (Reddy and Dávalos 2003) or missed detection of a disease vector thereby underestimating risk (Abad-Franch et al. 2014).

Sampling design involves three main components: *what* to sample (e.g. species, water, air, or soil), *how* to sample (the measurement procedure), and *where* to sample. A review of urban ecology studies demonstrated that while many studies reported on *what* they sampled and *how*, few studies provided details on their site selection methods (Dyson et al. 2023). A variety of R tools can help researchers with selecting sampling sites, such as *spsurvey* (Dumelle et al. 2023), *BalancedSampling* (Grafström, Prentius, and Lisic 2024), and *Spbsampling* (Pantalone, Benedetti, and Piersimoni 2022), but there is a need for an easy-to-use interface integrating geospatial data into the selection process to standardize methods in field site selection.

Additionally, many existing tools focus on sampling design for an already selected *sampling frame,* which is the overall pool from which samples are selected.

Ideally, a sampling frame is representative of the target population to ensure study results can be generalized. For instance, consider a study that intends to set out rodent traps to assess abundance of a pathogen in the rodent population. The rodents serve as the sampling units, or the "*what";* the "*how"* is through rodent trapping and molecular identification; the "*where"* is the location of the traps; and the "*sampling frame*" is the entire area where researchers plan to put traps. If the sampling frame is biased to only include forested areas, despite higher abundance of the rodent species in other land cover types, the rodent and pathogen abundance measures may be biased, as the sampling frame is not representative of the population. In this situation, it would be useful to have a tool to evaluate bias in land cover types in the sampling frame, so that the researchers could adjust their sampling frame to better represent the target population.

The availability of remote-sensing data has unlocked new ways of comparing possible field sites. Publicly available high-resolution datasets exist on land-use and land cover, such as WorldCover (Zanaga et al. 2022), Copernicus Land Cover (Buchhorn et al. 2020), and Dynamic World (Brown et al. 2022), as well as for climate data, such as MODIS land surface temperature (Wan, Hook, and Hulley 2021) and CHIRPS precipitation (Hantak et al. 2021). Integrating such geospatial data into site selection can help researchers make data-driven decisions on where to sample, ultimately improving the robustness of research studies. Moreover, by plotting how sites fall along a gradient of climate and land cover characteristics, researchers can ensure their sampling frame and sampling sites exist across the full parameter range, thus avoiding a "truncated gradient" (Albert et al. 2010).

Here, we lay out a method using geospatial data to select and analyze potential field sites. The goal of this tool is twofold: (1) provide an easy way to select sites based on remote-sensing data and (2) evaluate already selected sites along a gradient of interest. The procedure can be applied to selecting individual sampling sites, as well as to finding the overall geographic and climatic area of the sites (the sampling frame). We describe the tool in the context of land cover data, although the method can easily be adapted to any dataset with a geographic resolution.

Methods

The two core steps in the procedure are generating a list of potential sites and calculating the landcover characteristics around each of the sites. Once the calculations are complete, a series of filtering steps and statistical comparisons can be used to narrow down to the final sites. There are three main input parameters in the workflow: (1) the **region of interest (ROI),** or the potential space where the study could take place; (2) the type of **geospatial data** and resolution; and (3) the **analysis distance** from the center of each site at which landcover data should be analyzed. The procedure is implemented in R (R Core Team 2023) and designed to be an iterative process, where the cycle is repeated until optimal field sites are found.

The Site Tool App

The procedure was developed into an R package featuring an easy-to-use R Shiny interface (Chang et al. 2024). A hosted version of the tool is available at (https://ecosyshealth.shinyapps.io/SiteTool/). A local version can be obtained from GitHub (https://github.com/BioDivHealth/sitetool).

The app divides the procedure into three steps. Users first select a region of interest by drawing on the map, entering bounding box coordinates, or uploading a shapefile (Figure 1). Users then add raster data to the map, either by uploading or using the default available data. Next, users specify parameters to generate a list of potential sites. Finally, users specify their desired analysis distance in meters, and the land cover characteristics are evaluated for each site. A series of plots are displayed with the results of the analysis. Users can interact with these plots to select sites, or download the site list to further analyze the resulting data in outside software.

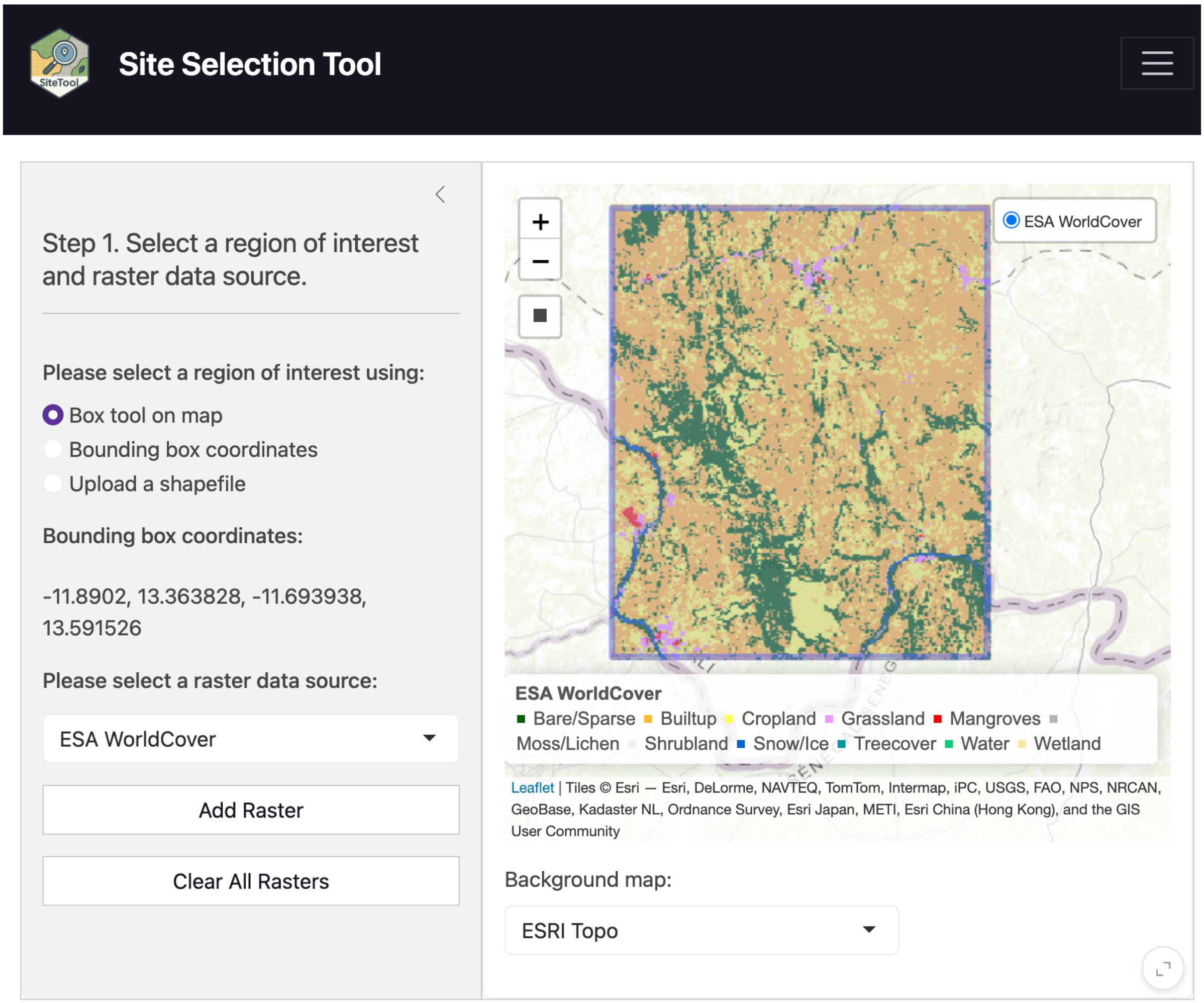

Figure 1. The application interface and preview of the workflow. The workflow involves three main steps, each in a separate box on the interface: (1) select an area and

type of input data; (2) generate a list of potential sites; and (3) analyze the landcover data surrounding each site. The figure shows a preview of step one, where a user has multiple options to select their region of interest and type of landcover data, after which the raster is displayed on the map.

Generating a list of potential sites

Depending on the study, a researcher may want field sites centered in a place of human habitation (“village” sites) or be randomly sampled throughout the region of interest (“random” sites). To obtain *village* sites, all areas tagged “city”, “suburb”, “village”, “town”, and “hamlet” are extracted from OpenStreetMap (OSM) using the R package *osmdata* (Padgham et al., 2017; OpenStreetMap contributors, 2017). Users have an option to return all of the villages within the region of interest, or user defined subset. These values can be used to further refine the list of sites.

To obtain *random* sites, points are randomly generated throughout the region of interest, up to a user specified number of points. To avoid sampling bodies of water, points are selected only from land areas, based on shapefiles derived from Natural Earth's 50m resolution land and lake boundaries (https://www.naturalearthdata.com/). After generating a shapefile of the sample space, the function *st_sample* from the *sf* package is used to select points (Pebesma 2018; Pebesma and Bivand 2023). As some studies may want to avoid human dominated areas or be near them for practicality purposes, users may refine the list of points by specifying proximity constraints to other points, roads and cities. The data on roads is downloaded from OSM (highways tagged as "motorway," "primary," "secondary," or "tertiary") along with the city data (any place tagged "city," "borough," "suburb," "quarter," "village," "town," or "hamlet"). Points within the specified distance are then filtered out. The sampling process repeats until a list of sites meeting all constraints and equal to the specified number is generated.

At this stage, users may also supply a list of selected sites or select potential sites on the map. These sites will be evaluated in comparison to the generated sites to assess whether the selected sites show bias in land cover characteristics or appropriately represent the variation that exists in the region of interest.

Calculating landcover characteristics around sites

Once a list of potential sites is determined, the landcover characteristics surrounding each site are calculated. The analysis procedure remains the same regardless of whether the sites are random locations or human settlements. Unless otherwise specified, calculations take place using the named functions in the *sf* and *terra* packages (Hijmans 2024) in R.

First, the scale for the analysis must be set by specifying the analysis distance. This distance represents how far from each potential site the input raster should be analyzed. The value depends on the question of interest and resolution of the input data. For example, a study on insects requires investigation of small-scale processes (a small analysis distance), while larger-roaming animals will need to consider a wider landscape area (a larger analysis distance). Ideally, researchers will conduct a thorough analysis to determine the appropriate scale for their study (*see* Faust et al. 2023; Andreo et al. 2021; Brock et al. 2019). Additionally, researchers must consider the resolution of their input raster, as the resolution of the input raster must be finer than the analysis distance; a raster resolution of 1 km^2 with an analysis distance of 1 kilometer would distinguish little of the variation surrounding sites.

After an appropriate analysis distance is set, the input raster is cropped at that distance from each potential site. The raster is cropped to be square surrounding each point, with the analysis distance, *d,* serving as the distance to each edge of the raster. As a result, the length of each edge of the cropped raster is $2d$, and the total analysis area for each site is equal to $4d^2$. Only entire cells are returned, and the raster is cropped to contain the entire area, so sometimes the cropped area is slightly larger depending on the raster resolution and analysis distance.

For categorical rasters (*e.g.,* landcover and land-use data), calculations take place among classes. For each class the metrics calculated are (1) class total area, (2) class proportion, and (3) class mean patch area. To calculate class total area, first the size of each raster cell (keeping the cell size the same as input data used) is calculated using the *cellSize* function in *terra*. Using the *cellSize* function allows close approximation of area whether rasters are in angular (longitude/latitude) or planar (projected) coordinate reference systems. We use the function *zonal* to sum the area cells for each class, providing us with a total area for each class. We divide this total class area by the total area of the raster ($4d^2$) to find class proportion. To calculate class mean patch area, the raster is split into separate layers for each class. We then detect all the patches in the layer, using the *patches* function and considering the 8 surrounding cells to be adjacent (Queen's case). We find the area of each patch by summing the cells from the cell area raster, then calculate the mean of the summed patch areas to get mean patch area.

For continuous rasters (*e.g.,* climate measures, NDVI, and elevation data), the cropped raster is considered as one layer and summary measures are calculated. Surrounding each site, the mean, standard deviation, and range of the values in the raster are calculated, providing an estimate on the variability of the measure. If the measure has high temporal variability, such as with climate measures, the analysis is best performed at multiple time points, or at critical points relevant to the research study. The tool allows upload of multiple rasters so the values from a series of rasters can be compared and visualized at one time.

Statistical comparison with selected sites

Following site generation and raster analysis, a series of plots are generated showing the distribution of land cover values across all sites analyzed. If selected sites are provided, a two-sided Mann-Whitney U test is performed to assess whether the distribution of the generated sites differs significantly from the distribution of selected sites. If the test is significant (*p-value* < 0.05), it suggests that the input sites may not adequately represent the full gradient range of the tested parameter. However, the statistical results should be interpreted with caution if there are land cover classes or areas irrelevant to the focus of the study. At this stage, sites from the generated list may be added to the selected list to narrow down to a short list of potential sites.

Example Use Case: Selecting across an agricultural gradient in Nigeria

The site selection tool was trialed as part of a multi-year field study on Lassa fever in Nigeria, aimed at integrating local-scale risk factors into broad-scale zoonotic spillover models. The study sought three sites across a gradient of postulated drivers of Lassa fever. Spillover is hypothesized to be driven by contact with infected rodents, with a key risk factor of agricultural land use (Redding et al., 2021). Sites had to be within a 3-hour drive of the project's base city and medium-sized villages (100–500 households).

Step 1: Identifying the region of interest and input raster data

First, a bounding box area was selected to encompass the project base city and areas within approximately a three-hour drive, with coordinates spanning from the southwest corner at 5.8° N, 8° E to the northeast corner 9.1° N, 7° E. These coordinates were entered in the tool and the raster data selected as the ESA WorldCover dataset (Figure 2A). This raster data source was selected as the main comparison of interest was land cover characteristics of potential sites.

Step 2: Generate a list of possible points and analyze land cover

Next, using the drop-down menu, “village” sites were selected, as the study intended to study disease risk around human settlements. The tool used OpenStreetMap to identify any settlements within the area and the 272 communities were display as blue points on the map (Figure 2B).

Step 3: Analyze and select sites

An analysis distance of 2 kilometers was selected based on the home range of the target rodent species and to fully encompass each village area. This was entered into the tool, and a series of plots were displayed showing the proportion of each land cover type contained within the analysis distance for each site (Figure 2C). Hovering over a point shows the value of each raster category for that site by turning the point yellow. Using this information, three sites that varied across cropland and grassland proportion were selected and added to the selected points list by clicking on the points. Sites with greater than 60% forest cover or 20% built-up were excluded, as the study intended to focus on rural agricultural areas. Visualizing the land cover for each site within the analysis area showed the final sites selected varied across the important land cover factors (Figure 2D). Additionally, the statistical comparison from the tool showed that the selected sites do not show bias relative to the other possible settlements in the area and appropriately reflect the land cover variation in the area.

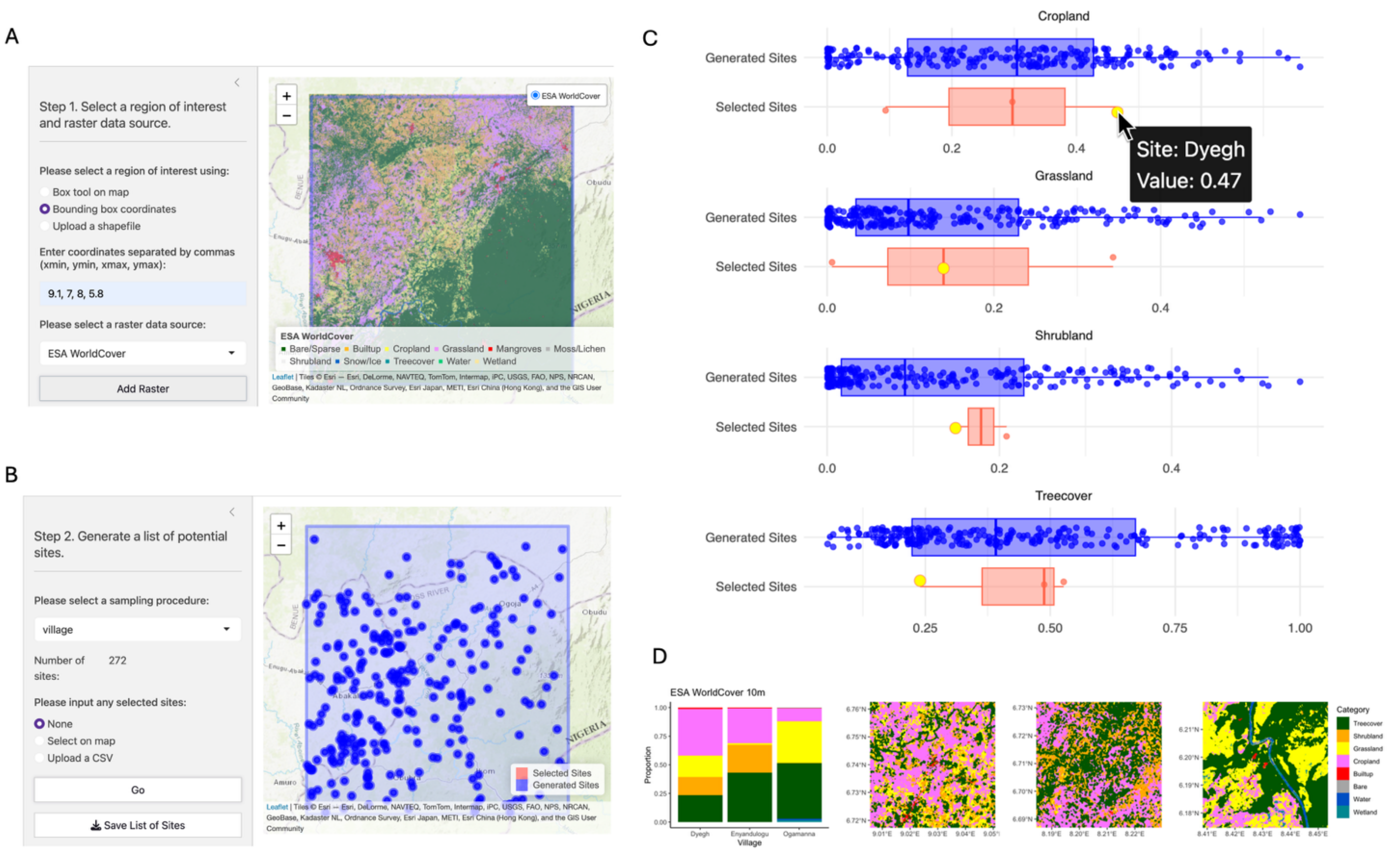

Figure 2. Example Use Case: Selecting field sites across a gradient in agricultural land-use. This example demonstrates the selection of three village sites that vary in land cover using the site selection tool. The bounding box area, visualized in the map from the tool in **(A),** resulted in 272 potential village sites (blue points in **B**). In Step 3 (visualized in **C**), the land cover characteristics of these sites were analyzed and compared. By hovering over a point, a user can see the characteristics of that site across the compared factors and click on the point to add to selected sites if it has suitable characteristics. Three points were selected that varied across the major land cover types. The bar plot in **(D)** compares the proportion of all major landcover types in the final selected sites. The individual rasters display the landcover raster surrounding each site, where the raster was cropped at the distance (2 km) land cover was analyzed for each site.

Discussion and Outlook

As the accuracy and resolution of remote-sensing products increases, there is a need for user-friendly tools incorporating geospatial data into the site selection process. We have developed such a tool that requires little manipulation of complicated raster files and geospatial formats, and helps researchers determine field study locations. We provide direct access to the tool via a web version, as well as the underlying code for easy adaptation for specialized use cases.

Drawbacks to this approach are potential errors in the input raster and OpenStreetMap data. OpenStreetMap relies on crowd-sourced data, and as a result, can have inaccurate place names or locations. This is particularly true in remote and rural areas, where data is not as regularly verified, so users should proceed with caution when using this data. However, the quality is consistently improving through user modification, and we encourage researchers to improve this resource if they encounter mistakes. Similarly, information inferred from satellites, whether for climate or land cover classification, has varying levels of accuracy for different times and locations. Researchers may be limited by the resolution of available geospatial products, which may not be detailed enough to compare possible sites, depending on the research question and target population of interest. The best use case is for researchers to use datasets they have verified as being relatively accurate in their region of study, or to use this tool for ground-truthing and testing of geospatial data accuracy. This process of site selection will be most effective when coupled with quantitative or qualitative assessments on the ground to further refine study sites.

The large number of studies that have used undisclosed site sampling methods serves as a call to action to improve and standardize site selection procedures and reporting. The advancement of code and data sharing methods, along with tools such as this one, should make it easier for researchers to report their methods, and improve research outcomes.

References

Abad-Franch, Fernando, Carolina Valença-Barbosa, Otília Sarquis, and Marli M. Lima. 2014. 'All That Glisters Is Not Gold: Sampling-Process Uncertainty in Disease-Vector Surveys with False-Negative and False-Positive Detections'. *PLOS Neglected Tropical Diseases* 8 (9): e3187. https://doi.org/10.1371/journal.pntd.0003187.

Albert, Cécile H., Nigel G. Yoccoz, Thomas C. Edwards Jr, Catherine H. Graham, Niklaus E. Zimmermann, and Wilfried Thuiller. 2010. 'Sampling in Ecology and Evolution – Bridging the Gap between Theory and Practice'. *Ecography* 33 (6): 1028–37. https://doi.org/10.1111/j.1600-0587.2010.06421.x.

Anderson, David R. 2001. 'The Need to Get the Basics Right in Wildlife Field Studies'. *Wildlife Society Bulletin (1973-2006)* 29 (4): 1294–97.

Andreo, Verónica, Pablo Fernando Cuervo, Ximena Porcasi, Laura Lopez, Claudio Guzman, and Carlos M. Scavuzzo. 2021. 'Towards a Workflow for Operational Mapping of *Aedes Aegypti* at Urban Scale Based on Remote Sensing'. *Remote Sensing Applications: Society and Environment* 23 (August):100554. https://doi.org/10.1016/j.rsase.2021.100554.

Andrew, NL, and BD Mapstone. 1987. 'Sampling and the Description of Spatial Pattern in Marine Ecology'. *Oceanography and Marine Biology* 25:39–90.

Brock, Patrick M., Kimberly M. Fornace, Matthew J. Grigg, Nicholas M. Anstey, Timothy William, Jon Cox, Chris J. Drakeley, Heather M. Ferguson, and Rowland R. Kao. 2019. 'Predictive Analysis across Spatial Scales Links Zoonotic Malaria to Deforestation'. *Proceedings of the Royal Society B: Biological Sciences* 286 (1894): 20182351. https://doi.org/10.1098/rspb.2018.2351.

Brown, Christopher F., Steven P. Brumby, Brookie Guzder-Williams, Tanya Birch, Samantha Brooks Hyde, Joseph Mazzariello, Wanda Czerwinski, et al. 2022. 'Dynamic World, Near Real-Time Global 10 m Land Use Land Cover Mapping'. *Scientific Data* 9 (1): 251. https://doi.org/10.1038/s41597-022-01307-4.

Buchhorn, Marcel, Myroslava Lesiv, Nandin-Erdene Tsendbazar, Martin Herold, Luc Bertels, and Bruno Smets. 2020. 'Copernicus Global Land Cover Layers—Collection 2'. *Remote Sensing* 12 (6): 1044. https://doi.org/10.3390/rs12061044.

Chang, Winston, Joe Cheng, J. J. Allaire, Carson Sievert, Barret Schloerke, Yihui Xie, Jeff Allen, Jonathan McPherson, Alan Dipert, and Barbara Borges. 2024. *Shiny: Web Application Framework for R*. https://shiny.posit.co/.

Conn, Paul B., James T. Thorson, and Devin S. Johnson. 2017. 'Confronting Preferential Sampling When Analysing Population Distributions: Diagnosis and Model-Based Triage'. *Methods in Ecology and Evolution* 8 (11): 1535–46. https://doi.org/10.1111/2041-210X.12803.

Dumelle, Michael, Tom Kincaid, Anthony R. Olsen, and Marc Weber. 2023. 'Spsurvey: Spatial Sampling Design and Analysis in R'. *Journal of Statistical Software* 105 (January):1–29. https://doi.org/10.18637/jss.v105.i03.

Dyson, Karen, Emad Dawwas, Renata Poulton Kamakura, Marina Alberti, and Tracy L. Fuentes. 2023. ‘Say Where You Sample: Increasing Site Selection Transparency in Urban Ecology’. *Ecosphere* 14 (3): e4466. https://doi.org/10.1002/ecs2.4466.

Elphick, Chris S. 2008. ‘How You Count Counts: The Importance of Methods Research in Applied Ecology’. *Journal of Applied Ecology* 45 (5): 1313–20. https://doi.org/10.1111/j.1365-2664.2008.01545.x.

Faust, Christina L., Adrian A. Castellanos, Alison J. Peel, Peggy Eby, Raina K. Plowright, Barbara A. Han, and Nita Bharti. 2023. ‘Environmental Variation across Multiple Spatial Scales and Temporal Lags Influences Hendra Virus Spillover’. *Journal of Applied Ecology* 60 (7): 1457–67. https://doi.org/10.1111/1365-2664.14415.

Grafström, Anton, Wilmer Prentius, and Jonathan Lisic. 2024. ‘BalancedSampling: Balanced and Spatially Balanced Sampling’. https://cran.r-project.org/web/packages/BalancedSampling/index.html.

Green, Jessica L., and Joshua B. Plotkin. 2007. ‘A Statistical Theory for Sampling Species Abundances’. *Ecology Letters* 10 (11): 1037–45. https://doi.org/10.1111/j.1461-0248.2007.01101.x.

Hansen, Matthew C, Peter V Potapov, Rebecca Moore, Matt Hancher, Svetlana A Turubanova, Alexandra Tyukavina, David Thau, et al. 2013. ‘High-Resolution Global Maps of 21st-Century Forest Cover Change’. *Science* 342 (6160): 850–53.

Hantak, Maggie M., Bryan S. McLean, Daijiang Li, and Robert P. Guralnick. 2021. ‘Mammalian Body Size Is Determined by Interactions between Climate, Urbanization, and Ecological Traits’. *Communications Biology* 4 (1): 1–10. https://doi.org/10.1038/s42003-021-02505-3.

Hijmans, Robert J. 2024. *Terra: Spatial Data Analysis*. https://rspatial.org/.

Hurlbert, Stuart H. 1984. ‘Pseudoreplication and the Design of Ecological Field Experiments’. *Ecological Monographs* 54 (2): 187–211. https://doi.org/10.2307/1942661.

Legendre, Pierre, and Marie Josée Fortin. 1989. ‘Spatial Pattern and Ecological Analysis’. *Vegetatio* 80 (2): 107–38. https://doi.org/10.1007/BF00048036.

Lewis, John B. 2004. ‘Has Random Sampling Been Neglected in Coral Reef Faunal Surveys?’ *Coral Reefs* 23 (2): 192–94. https://doi.org/10.1007/s00338-004-0377-y.

OpenStreetMap contributors. 2017. ‘Planet Dump Retrieved from Https://Planet.Osm.Org’.

Padgham, Mark, Bob Rudis, Robin Lovelace, and Maëlle Salmon. 2017. ‘Osmdata’. *Journal of Open Source Software* 2 (14): 305. https://doi.org/10.21105/joss.00305.

Pantalone, Francesco, Roberto Benedetti, and Federica Piersimoni. 2022. ‘Spbsampling: An R Package for Spatially Balanced Sampling’. *Journal of Statistical Software* 103 (August):1–22. https://doi.org/10.18637/jss.v103.c02.

Pebesma, Edzer. 2018. ‘Simple Features for R: Standardized Support for Spatial Vector Data’. *The R Journal* 10 (1): 439–46. https://doi.org/10.32614/RJ-2018-009.

Pebesma, Edzer, and Roger Bivand. 2023. *Spatial Data Science: With Applications in R*. Chapman and Hall/CRC. https://doi.org/10.1201/9780429459016.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Reddy, Sushma, and Liliana M. Dávalos. 2003. 'Geographical Sampling Bias and Its Implications for Conservation Priorities in Africa'. *Journal of Biogeography* 30 (11): 1719–27. https://doi.org/10.1046/j.1365-2699.2003.00946.x.

Smith, Adam N. H., Marti J. Anderson, and Matthew D. M. Pawley. 2017. 'Could Ecologists Be More Random? Straightforward Alternatives to Haphazard Spatial Sampling'. *Ecography* 40 (11): 1251–55. https://doi.org/10.1111/ecog.02821.

Stevens, Don L, and Anthony R Olsen. 2004. 'Spatially Balanced Sampling of Natural Resources'. *Journal of the American Statistical Association* 99 (465): 262–78. https://doi.org/10.1198/016214504000000250.

UNEP-WCMC. 2024. 'Protected Area Profile for Africa from the World Database on Protected Areas'. www.protectedplanet.net.

Wan, Z., S. Hook, and G. Hulley. 2021. 'MODIS/Terra Land Surface Temperature/Emissivity Daily L3 Global 1km SIN Grid V061'. NASA EOSDIS Land Processes Distributed Active Archive Center. https://doi.org/10.5067/MODIS/MOD11A1.061.

Zanaga, Daniele, Ruben Van De Kerchove, Dirk Daems, Wanda De Keersmaecker, Carsten Brockmann, Grit Kirches, Jan Wevers, et al. 2022. 'ESA WorldCover 10 m 2021 V200'. Zenodo. https://doi.org/10.5281/zenodo.7254221.