Empirical seed transfer zones require conventions for data sharing to increase their utilization by practitioners

Brianna Wieferich[2,3], Reed Clark Benkendorf[1,2*]

[1] Northwestern University, Evanston, Illinois 60208, USA

[2] Chicago Botanic Garden, 1000 Lake Cook Road, Glencoe Illinois 60022 USA

[3] Dorena Genetic Resource Center, 34963 Shoreview Dr, Cottage Grove, Oregon 97424 USA

Abstract

Empirical seed transfers zones (eSTZs) are becoming increasingly developed to help guide both the development of native plant materials, and the most desirable seed sources for restoration projects. Despite the growing popularity of eSTZs, standards for distributing these data are nascent. In order to maximize the utilization of eSTZs we propose a set of standards to guide the distribution of eSTZs which should allow for making them easier to interact with, and increase the focus of seed collection efforts and foster utilization of the most appropriate seed source available. Further we propose that sharing of metrics of uncertainty for these data, which can help practitioners identify best alternatives for a seed transfer zone, should become common practice. Finally, we briefly introduce an R package eSTZwritR which implements our core suggestions for data dissemination.

IMPLICATIONS FOR PRACTICE:

- Developing a restoration plan in a short time period can be a stressful process. To decrease the chances of simple human induced mistakes we develop standards to increase consistency between eSTZs to make using them in GIS software more consistent.
- We implement these suggestions in an R package 'eSTZwritR' which should make adherence to the guidelines more simple for the scientists developing eSTZ products, allowing for a rapid uptake of these conventions.

*Correspondence: rbenkendorf@chicagobotanic.org

- We further suggest the incorporation of estimates of uncertainty for spatial eSTZ data products so practitioners can determine the most suitable alternative seed zones when material from a preferred zone is not available.

INTRODUCTION

[Figure 1 about here.]

Empirical seed transfer zones (eSTZ's) are gaining popularity among restoration practitioners as a tool to help ensure that the most appropriate locally adapted seed source is utilized at restoration sites (McKay et al. 2005). The development of empirical seed transfer zones (eSTZ's) for a species is a costly and time consuming process, most often involving common garden studies, or genetic studies, with many individual populations being represented as samples (Kramer et al. (2015)). To our knowledge, in Western North America, a couple research groups are regularly developing eSTZ's for multiple species at a time, while intermittently an eSTZ for a single species are being developed by various labs focused on ecology more broadly. These zones are being developed via funding from multiple agencies with multiple different final use cases and implementation strategies in mind. Accordingly, discord exists between the products generated by different labs, and for different funding agencies. While the suggestions for best practices during the development of eSTZs are becoming more available, the results of eSTZs need to be conveyed via spatial products, however standardization or guidance on the best practices for accomplishing this are wanting.

We argue that the dissemination of data during restoration is a vital portion of any projects success (Figure 1).... Given the relative complexity of interacting with spatial data, it is imperative that products which are inherently spatial, are shared in the most convenient fashion possible.

Here, using 23 sets of eSTZs produced for 22 taxa, we show that most of the data developed and disseminated, to share the results of an eSTZ, are incongruous Massatti et al. (2020). Subsequently, using any consensus (wisdom of the masses) from these data, combined with standard conventions of data sharing, we present a set of guiding standards for researchers to employ to make results more congruous.

Current Condition

[Figure 2 about here.]

We conducted a review of all eSTZs on the Western Wildland Environmental Threat Assessment Center (WWETAC) website (https://research.fs.usda.gov/pnw/products/dataandtools/datasets/seed-zone-gis-data, as of May 1, 2024). Each data product was analyzed for its file name structure (using 5 categories), metadata, naming conventions, and directory structure. All scoring was done by hand, and all analyses were carried out in R version 4.2.1.

[Figure 3 about here.]

Here we present only the results which we consider to be the most consequential, and likely to interfere with practitioners workflows, in Figures 2 through 4. We encountered considerable inconsistency exists within file names (Figure 2), and even more so in directory structure names. While decent consensus existed around the use of USDA NRCS-Plants codes for denoting the contents of the file, the lack of file mentioning what they contained (e.g. 'zones', 'seed_zone', 'sz'), and the lack of specified geographic extents can make determining what the file contains, without opening it in a Geographic Information System (GIS), software difficult. Unless all users have centralized directories on their networks for all of their STZ products we propose that the current approaches offer considerable resistance to a practitioner trying to find a file within their file system using common GUI searching functionality.

The naming of the fields (columns) within shapefiles likely presented the most problematic of all results (Figure 3), while many inconsistencies existed, here we focus on only three. Different usages of polygon geometry types existed for representing the individual seed transfer zones, i.e. sometimes all portions of a seed transfer zone - when at least some components are disconnected - where stored within the same object or row (a multipolygon). Other times each discontinuous portion of the range would be stored as it's own polygon. For most infrequent Geographic Information System (GIS) users, multipolygons can be confusing... Surprisingly, within each shapefile the field denoting the Seed Zones were often unlabeled, or entirely lacking any indication; in a number of instances it took us several minutes to determine which field was the seed zone by toggling through visualizing each field.

[Figure 4 about here.]

Recommendations

Consensus exists among the developers of eSTZs for a range of attributes related to distribution of spatial products. Combining those opinions with our perceived best practices for data sharing, and experience as users of each of the existing empirical products, results in the recommendations below.

Directory Structure

[Figure 5 about here.]

eSTZ's should be distributed using a predictable directory structure allowing practitioners to be immediately familiar with where to find relevant contents (Figure XX). This predictable nature should decrease the time required to find particular attributes of the data.

We recommend that all directories have two main subdirectories (Figure 5), one containing the essential data products, preferably in both raster and vector data formats. The other directory contains information relating to the product, including a formatted citation for data use, a map for quick reference, and any materials describing the production of the product both as a paper, and a text file of quick metadata attributes.

File Naming

[Figure 6 about here.]

The individual files within the directory should follow a simple naming format which is easy for users of various softwares to interpret and readily import for use, while also containing key parameters of the data product. We recommend (figure 6) that each file name has three main components, in addition to the file extension. The first component is the USDA PLANTS code, the specific taxon, and the second is the type of data used to develop the STZ, the final is the two main regions which the product overlaps. We strongly recommend the use of the 12 Department of Interior regions as they cover considerable geographic expanses and reflect some degree of ecological patterns.

Maps

Maps should be included within the Information directory. Many questions about eSTZs can be answered quickly and simply from a practitioner consulting a map saved as a PDF with the essential cartographic components. We recommend that each map contains the following elements: north arrow, scale bar, state borders, geographically relevant cities, coordinate reference system information, sensible categorical color schemes for the seed zones, a legend, the taxons name as a title, and the maps theme ('Seed Transfer Zones') as a subtitle.

Data Formats

We recommend that the spatial data associated with an eSTZ be distributed using both popular spatial data models, vector and raster. For vector data we advocate for the continued usage of data using the shapefile format, while for raster data we propose the usage of geoTIFFs ('tifs', the .tif extension). In our experience tifs seem to be the most widely used of the raster data models in ecology for non-time series data, they are widely supported by a variety of geographic information systems, and generally seem to perform better than ASCII.

Vector Data Field Attributes

[Figure 7 about here.]

We believe that the fields, or columns, of the vector data should follow a predictable pattern. This will allow humans visualizing the data in a GUI to quickly visually detect their field of interest, and while it's bad practice – allow code to subset columns by position rather than field name. We further recommend the standardization of field names to allow for code and scripts to retrieve these values without more complicated coding techniques (Figure 7).

We recommend that each shapefile has a bare minimum of four fields in the following order and of the following data types. 1) ID (numeric - integer) a unique number associated with each individual polygon in the file, we do not recommend combining polygons into multipolygon units, as individual polygons can retain information about their Area, and are easier for users to subset. 2) Seed Zone (numeric - integer) a unique identifier for each of the eSTZs delineated by the practitioners, these allow for quick filtering of the data based on a simple value which is hard to misspecify. 3) SZName (character) a human developed name for the zone this may refer to a components of a principal component analysis, e.g. LOW MEDIUM LOW, or be defined by the analysts. We opine that semi-informative names should be developed before data distribution to help practitioners more easily convey important attributes without having to rely on numeric values which may be more difficult to remember due to their nondescript nature.

In addition to these standard field naming and placement conventions, we further recommend a series of standards for the contents within these essential fields, and how to format any additional fields relevant to the project.

Recommendations for new products

Metadata

Metadata offers a way of rapidly distilling many parameters of the ESTZ development process for quick reference by practitioners. The inclusion of metadata may make it easier for a practitioner to find relevant details which they may otherwise have to peruse the literature files to glean. Given that the results of eSTZS are published in a variety of journals requiring different amounts of methodological reporting this is a fruitful format for sharing data in a standardized way.

Estimating Uncertainty

[Figure 8 about here.]

We have observed considerable consternation from seed collection crews, curators, and restoration practitioners alike over the 'proper' classification for both a seed source, and the selection of a seed source for a restoration. In most instances these hesitations relate to a seed source which is present from a population which 'straddles' two or more seed zones. We predict that with the increasing availability of fine resolution spatial data, and the wider availability of ecological relevant variables - which more accurately reflects local ecological heterogeneity - individual portions of seed zones will become more fine, increasing the perimeter to surface area ratio and thereby increasing the prevalence of this already common phenomenon (Gibson et al. 2019).

Based on our literature review we believe that currently eSTZs are distributed only as polygon vector data (e.g. shapefiles).

Vector data convey a sense of separation between the objects they represent, i.e. they are used to represent discrete classes with meaningful borders between them. Common examples of polygon vector data model usage include administrative units (e.g. zip codes, states, and countries), hydrologic basins, and the geographic range of a species. On the other hand raster data, or gridded surfaces, are used for representing continuous phenomena, i.e. gradients. Common examples of raster data model usages include climate variables, land cover classes, and predictions of species habitat suitability.

While we agree with the current prevailing census that in most applications, the use of the polygon vector data model is generally the preferred method of sharing data, we have witnessed enough scenarios where a population crosses multiple seed zones, that we believe the usage of raster data is warranted for all reports.

Raster data come with an enormous benefit in that they can readily incorporate multiple layers (individual raster files) for each pixel across a domain, thus allowing for a first layer of consensus predictions for each cell (the data conveyed in a vector data set), and a few other levels of prediction. For example, a raster with 4 layers could have 3 layers dedicated to raw model output and the final consensus layer between these products, in the case of regression type analyses these three layers would represent predictions at the specified lower and upper confidence intervals and the model prediction, and in the case of a classification algorithm the 3 classes with the highest predicted probabilities.

In the above examples the consensus layer would then be informed by the plurality of assignment between the predictions, and by the preferred prediction model (e.g. via the typical regression prediction) when no plurality exists.

We believe that conveying these uncertainties will allow data users to understand and explore the caveats with model predictions more. This practice is further grounded in best scientific practice as the spatial data used to develop the initial zones are imperfect, the study itself was imperfect, and the classification process is itself imperfect. Further on an ecological level we believe that a porosity exists between these populations of species – they are by virtue of being components of a species connected at least marginally via gene flow, and the expression of this continuity is the most appropriate course of action for data dissemination.

IMPLEMENTATION

The suggestions above may seem relatively onerous to carry out at the end of a multi -year study, especially when considering manuscripts are being prepared for publication and further funding opportunities are being applied for, and staff (e.g. postdocs) may be leaving the group at the end of the project. For these reasons we have created an R package, eSTZwritR ('pronounced easy rider'), which can implement all of them, less the statistical processing, with minimal user inputs. The package is installable from GitHub at https://github.com/sagesteppe/eSTZwritR, and a Github website (https://sagesteppe.github.io/eSTZwritR/) exists for users interested in better understanding it's functionality and which includes supplemental figures and more intricate details not discussed here.

FOR DEVELOPERS

The package requires only 4-5 functions to produce a directory with the contents discussed above, with minimal data entry. Most importantly the entries are well outlined and easily entered without requiring close attention to detail, an omnipresent scenario when processing standards by hand.

FOR PRACTITIONERS

These results should allow for simple utilization of existing empirical seed transfer zone resources. We have re-processed all eSTZ data products we are aware of to follow these standards, with the exception of creating the uncertainty raster layers. We have provided some sample code which showcases loading these data into a Geographic Information System (GIS) which utilizes either R or python coding elements, as well as the freely available QGIS which is set up with an advanced graphical user interface (GUI), which allows a user to navigate via a computer mouse.

CONCLUSIONS

Seed based active restoration will always be a relatively expensive, yet necessary, option for restoration. Here we present simple standards for the scientists developing eSTZs to use in order to standardize the data products they are developing to ease their implementation. While these conventions should be easy to implement for a sufficiently detail oriented and interested individual, we also present an R package which can quickly achieve these results.

LITERATURE CITED

Bradley St. Clair J, Kilkenny FF, Johnson RC, Shaw NL, Weaver G (2013) Genetic variation in adaptive traits and seed transfer zones for pseudoroegneria spicata (bluebunch wheatgrass) in the northwestern united states. Evolutionary Applications 6:933–948

Doherty KD, Butterfield BJ, Wood TE (2017) Matching seed to site by climate similarity: Techniques to prioritize plant materials development and use in restoration. Ecological Applications 27:1010–1023

Erickson VJ, Mandel NL, Sorensen FC (2004) Landscape patterns of phenotypic variation and population structuring in a selfing grass, elymus glaucus (blue wildrye). Canadian Journal of Botany 82:1776–1789

Gibson A, Nelson CR, Rinehart S, Archer V, Eramian A (2019) Importance of considering soils in seed transfer zone development: Evidence from a study of the native bromus marginatus. Ecological Applications 29:e01835

Horning ME, McGovern TR, Darris DC, Mandel NL, Johnson R (2010) Genecology of holodiscus discolor (rosaceae) in the pacific northwest, USA. Restoration Ecology 18:235–243

Johnson RC, Cashman M, Vance-Borland K (2012) Genecology and seed zones for indian ricegrass collected in the southwestern united states. Rangeland Ecology & Management 65:523–532

Johnson RC, Erickson VJ, Mandel NL, St Clair JB, Vance-Borland KW (2010) Mapping genetic variation and seed zones for bromus carinatus in the blue mountains of eastern oregon, USA. Botany 88:725–736

Johnson RC, Hellier BC, Vance-Borland KW (2013) Genecology and seed zones for tapertip onion in the US great basin. Botany 91:686–694

Johnson RC, Horning ME, Espeland EK, Vance-Borland K (2015) Relating adaptive genetic traits to climate for sandberg bluegrass from the intermountain western united states. Evolutionary Applications 8:172–184

Johnson RC, Leger E, Vance-Borland K (2017) Genecology of thurber's needlegrass (achnatherum thurberianum [piper] barkworth) in the western united states. Rangeland Ecology & Management 70:509–517

Johnson RC, Vance-Borland K (2016) Linking genetic variation in adaptive plant traits to climate in tetraploid and octoploid basin wildrye [leymus cinereus (scribn. & merr.) a. Love] in the western US. PLoS One 11:e0148982

Kramer AT, Larkin DJ, Fant JB (2015) Assessing potential seed transfer zones for five forb species from the great basin floristic region, USA. Natural Areas Journal 35:174–188

Massatti R (2020) Genetically-informed seed transfer zones for cleome lutea and machaeranthera canescens across the colorado plateau and adjacent regions. Bureau of Land Management

Massatti R (2019) Genetically-informed seed transfer zones for pleuraphis jamesii, sphaeralcea parvifolia, and sporobolus cryptandrus across the colorado plateau and adjacent regions. Bureau of Land Management

Massatti R, Shriver RK, Winkler DE, Richardson BA, Bradford JB (2020) Assessment of population genetics and climatic variability can refine climate-informed seed transfer guidelines. Restoration Ecology 28:485–493

McKay JK, Christian CE, Harrison S, Rice KJ (2005) ‘How local is local?’—a review of practical and conceptual issues in the genetics of restoration. Restoration Ecology 13:432–440

Shryock DF, Havrilla CA, DeFalco LA, Esque TC, Custer NA, Wood TE (2017) Landscape genetic approaches to guide native plant restoration in the mojave desert. Ecological Applications 27:429–445

List of Figures

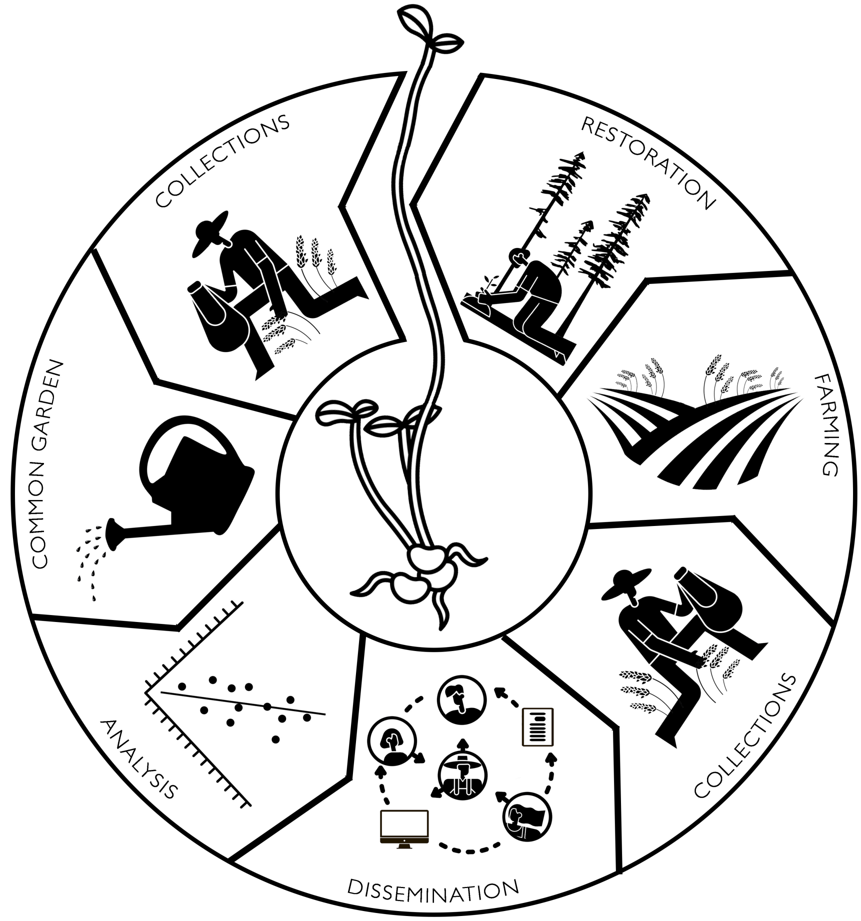


Figure 1: Dissemination... By Emily Woodworth

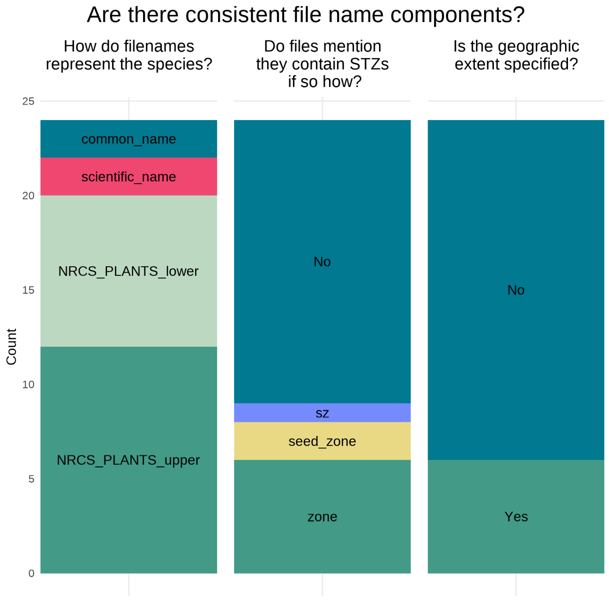


Figure 2: File Naming

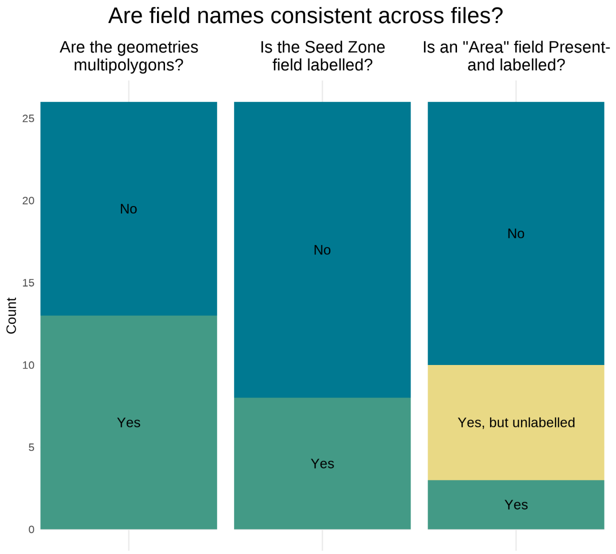


Figure 3: Field Names Shapefile

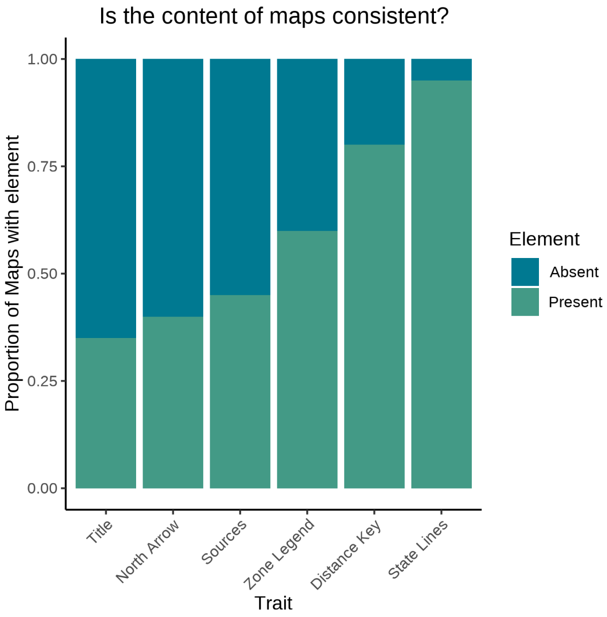


Figure 4: Map Components

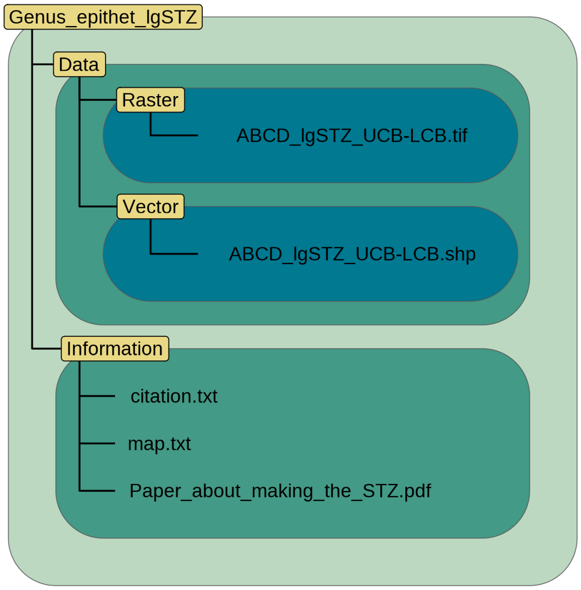


Figure 5: Directory Structure

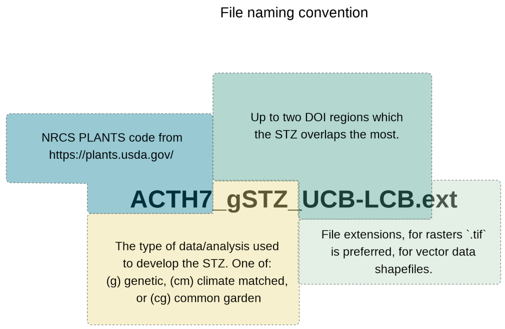
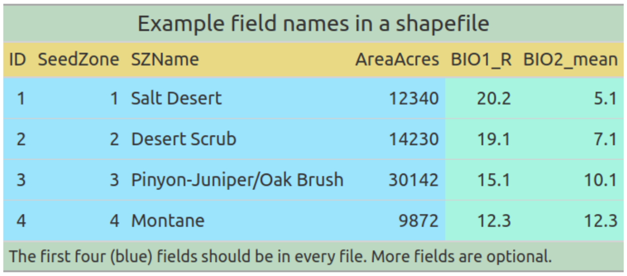

File naming convention
NRCS PLANTS code from https://plants.usda.gov/
Up to two DOI regions which the STZ overlaps the most.
ACTH7_gSTZ_UCB-LCB.ext
The type of data/analysis used to develop the STZ. One of: (g) genetic, (cm) climate matched, or (cg) common garden
File extensions, for rasters `.tif` is preferred, for vector data shapefiles.

Figure 6: File Naming



Example field names in a shapefile

ID	SeedZone	SZName	AreaAcres	BIO1_R	BIO2_mean
1	1	Salt Desert	12340	20.2	5.1
2	2	Desert Scrub	14230	19.1	7.1
3	3	Pinyon-Juniper/Oak Brush	30142	15.1	10.1
4	4	Montane	9872	12.3	12.3

The first four (blue) fields should be in every file. More fields are optional.

Figure 7: Vector Data Field Attributes

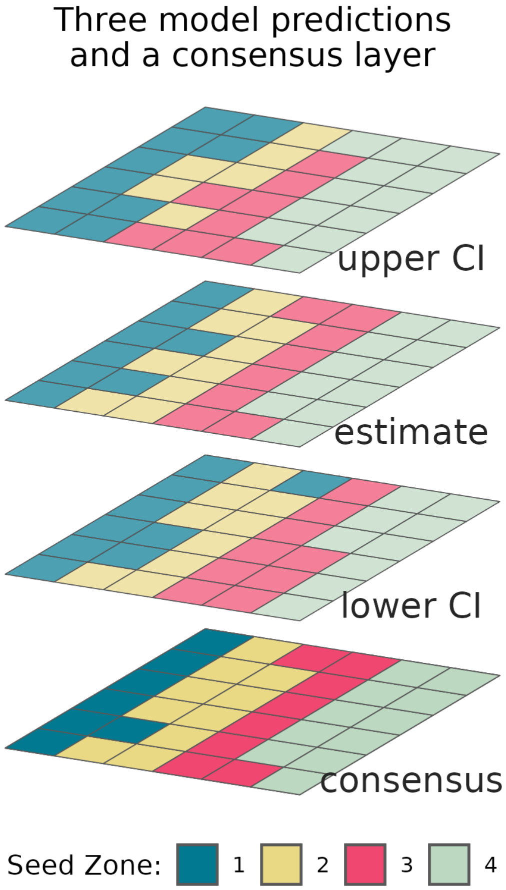


Figure 8: Using multiple predictions to create a consensus product