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

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

[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 being developed increasingly often to help guide both the agricultural development of native plant materials, and the selection of these materials for restoration projects. Despite the now widespread utilization of eSTZs standards for distributing these data are wanting, leading to inconsistent data products. In order to maximize the utilization of eSTZs we propose standards to guide the distribution of them, thereby increasing the focus of seed collection efforts and fostering the utilization and accessibility of the most appropriate commercially available seed sources. Further we propose that sharing metrics of model 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 ('easy rider') which implements our core suggestions for data dissemination.

IMPLICATIONS FOR PRACTICE:

- Developing a restoration plan in a short time period, as required after a natural disturbance, can be a stressful process. To decrease the chances of simple mistakes being propagated into plans, we develop standards to increase the consistency between eSTZ data products making there usage in GIS software more consistent.

*Correspondence: rbenkendorf@chicagobotanic.org

†Authors contributed equally, and are listed alphabetically.

- We implement these suggestions in an R package 'eSTZwritR' which should facilitate adherence to the guidelines for scientists developing eSTZ products, allowing for a rapid uptake of these conventions.
- We also suggest incorporating estimates of uncertainty for spatial eSTZ data products so practitioners have sufficient support for selecting material from non-target seed zones as is often required.

INTRODUCTION

[Figure 1 about here.]

Empirical seed transfer zones (eSTZs) are gaining popularity amongst restoration practitioners as a tool to help identify the most appropriate seed source for a species at a restoration site (McKay et al. 2005). eSTZs are popular for two primary reasons 1) they are based on empirical data - e.g. the phenotypes in a common garden, population genetics, or the correlation between occurrences of the species and environmental variables and 2) generally the zones are more coarse than provisional seed transfer zones thereby reducing the number of lineages requiring cultivation in agricultural settings. While popular, the development of eSTZs for a species is a costly and time consuming process, most often involving common garden or genetic studies, with many populations from across the species range incorporated as samples (Kramer et al. 2015).

In western North America, the majority of eSTZs have been developed by just a couple of lab groups, whilst the remainder have been developed as one-off's by other assorted groups. While standards for the best practices during the development of eSTZs are becoming more defined, no standards exist for *sharing* the results of eSTZs. Despite eSTZs being produced by a relatively small pool of lab groups and individuals, inconsistencies vary across the spatial data products used to report eSTZs, inconsistencies which we posit are associated with a combination of individual analysts preferences and to a lesser extent a natural evolution of the product reporting itself.

The success of a restoration project relies on the timely application of techniques which are suitable for the site at hand. Implementation of relevant techniques requires not only intrapersonal communication between a practitioner with themselves in time, e.g. avid note taking, but also interpersonal communication between practitioners. Hence the dissemination of ideas during and after a restoration project is our best opportunity to improve the outcomes of restorations (Figure 1). However, ideas have varying levels of complexity which may hinder their transmittal. For example seeding rates may be verbally communicated, while seed mixes are likely to require written documentation, whereas spatial data require both written and geographic data (e.g. coordinates and relations between them) in the form of spatial data products (e.g. rasters, shapefiles) to

accurately convey their meaning. Given the relative complexity of communicating precise spatial information standards should exist to ensure not only its accuracy and precision, but also the ease by which it can be interpreted and used.

Using 23 sets of eSTZs produced for 22 taxa, we show that most of the spatial data developed and disseminated, to share the results of an eSTZ, are inconsistent ((Doherty et al. 2017), (Erickson et al. 2004), (Johnson & Vance-Borland 2016), (Johnson et al. 2010), (Bradley St. Clair et al. 2013), (Johnson et al. 2015), (Johnson et al. 2013), (Johnson et al. 2012), (Horning et al. 2010), (Johnson et al. 2017), (Shryock et al. 2017), (Massatti 2020), (Massatti 2019), (Massatti et al. 2020)). We have already observed significant hindrances to the uptake of these data at the level of practitioners, and search for consensus within these data. 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 in making results more consistent.

Current Condition

[Figure 2 about here.]

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

[Figure 3 about here.]

In Figures 2 through 4, we present inconsistencies which we believe, or have observed to be, the most likely to interfere with practitioners' workflows. We encountered considerable inconsistency within file names (Figure 2), in directory structure and naming (Figure 3), and cartographic elements of maps (Figure 4). While some consensus existed around the use of USDA NRCS-Plants codes for denoting the taxon contained in the file (Figure 2), the lack of file names mentioning what attribute about the taxon they contained (e.g. 'zones', 'seed_zone', 'sz'), and the lack of specified geographic extents can make determining the specifics of the file difficult unless it is explicitly opened in a Geographic Information System (GIS) software. Unless all users have centralized directories on their networks for their STZ data products we propose that the current

approaches are a hindrance to a practitioner trying to find the relevant file within their file system using common searching functionality.

The naming of the fields (columns) within shapefiles likely presented the most problematic of all results (Figure 3), while many inconsistencies exist, here we focus on three. Different usages of polygon geometry were implemented 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 its own polygon. For most infrequent Geographic Information System (GIS) users, we have observed that multipolygons can be confusing and require them to use several moderately advanced spatial techniques to interact with. Surprisingly, within each shapefile the field denoting the seed zones was often ambiguously labelled, or entirely lacking any indication (Figure 3). In a number of instances it took us several minutes to determine which field was the seed zone by toggling through and visualizing many fields, despite us already having interfaced with all of these products multiple times.

[Figure 4 about here.]

Recommendations

Some consensus exists among the developers of eSTZs for a range of attributes related to distribution of data 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 following recommendations.

Directory Structure

[Figure 5 about here.]

eSTZs should be distributed using a predictable directory structure allowing users to be immediately familiar with where to find content (Figure 5). We recommend that all directories (folders) have two main subdirectories (Figure 5), one containing the essential data products, preferably in both raster and vector data formats (*see 'Data Formats'*). The second directory contains information relating to the product, including a formatted citation for data use, a map for quick reference, and any materials describing the development of the product both as a paper, and a text file of quick metadata attributes.

File Naming

[Figure 6 about here.]

The files within the directory should follow a naming convention which is easy for users to interpret and import to various software's, while also describing essential attributes 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, and the second is the method used to develop the STZ, and the final is the two main regions which the product overlaps. We recommend the use of the 12 Department of Interior regions as they cover considerable geographic expanses and reflect the ecology of North America.

Maps

Maps of the data product 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 of the popular spatial data models, vector and raster. For vector data we advocate for the continued usage of 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, and are supported by virtually all GIS software.

Vector Data Field Attributes

[Figure 7 about here.]

The order of the fields (or columns) of the vector data should follow a predictable pattern (Figure 7), allowing humans interacting with the data in a GUI to quickly detect their field of interest, and while it's bad practice – allow users code to subset columns by position rather than field name.

We recommend that each shapefile has at least 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. 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 numeric value which is hard to misspecify. 3) SZName (character) a human developed name for the zone which may refer to an axis 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. 4) AreaAcres (numeric - integer) of each polygon.

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 (see package website).

Estimating Uncertainty

[Figure 8 about here.]

We have observed consternation from seed collection crews, curators, and restoration practitioners alike, over the appropriate classification for a seed source, and the selection of a seed source for a restoration. Generally, these hesitations relate to a source which is on the border of multiple seed zones. We predict that with the increasing availability of fine resolution spatial data which more accurately reflect local ecological heterogeneity - seed zones will become more fine, increasing the perimeter to surface area ratio and the prevalence of this already common confusion (Gibson et al. 2019).

Currently eSTZs are distributed exclusively as polygon vector data (e.g. shapefiles). Vector data convey a sense of separation between the entities they represent, i.e. discrete classes with hard borders between them. Common examples of polygon vector data usage include: administrative units (e.g. zip codes, states, and countries), watersheds, and the geographic range of a species. Whereas raster data, or gridded surfaces, are used for representing continuous phenomena, i.e. gradients. Common examples of raster data usages include climate variables, land cover classes, and predictions of modelled species habitat suitability.

While we agree with the consensus that vector data are generally the best method of distributing data, given the number of times we have observed classification confusion we believe the inclusion of raster data is always warranted. Raster data have an additional benefit that they can intuitively incorporate multiple

layers (a 'raster stack') for each of their pixels. Thus allowing for a first layer of consensus predictions (the data conveyed in a vector data set), and other levels of raw predictions. For example, a raster with four layers would have three layers of raw model output while the final layer is a consensus of these products. In the case of regression type analyses two layers could represent predictions at the lower and upper confidence intervals and the final layer a model prediction, while in the case of a classification algorithm the three classes with the highest predicted probabilities and a consensus class would be present.

We believe that conveying these uncertainties will allow users to understand and explore the caveats with model predictions. 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.

IMPLEMENTATION

The suggestions above may seem onerous to carry out at the end of a multi-year study, especially when considering manuscripts are being prepared for publication, 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, lessen the statistical processing, with minimal user inputs. The package is installable from GitHub https://github.com/sagesteppe/eSTZwritR, and has a GitHub pages website (https://sagesteppe.github.io/eSTZwritR/) for users interested in better understanding it's functionality and which includes supplemental figures and details not discussed here.

FOR DEVELOPERS

The package requires only 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 non-gui GIS at the website above.

CONCLUSIONS

Seed based active restoration will always be a relatively expensive, yet necessary, option for terrestrial restoration. Here we present simple standards for the scientists developing eSTZs to use in order to standardize the data products they are developing to assist in their uptake. While these conventions should be easy to implement for a sufficiently motivated 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

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

List of Figures

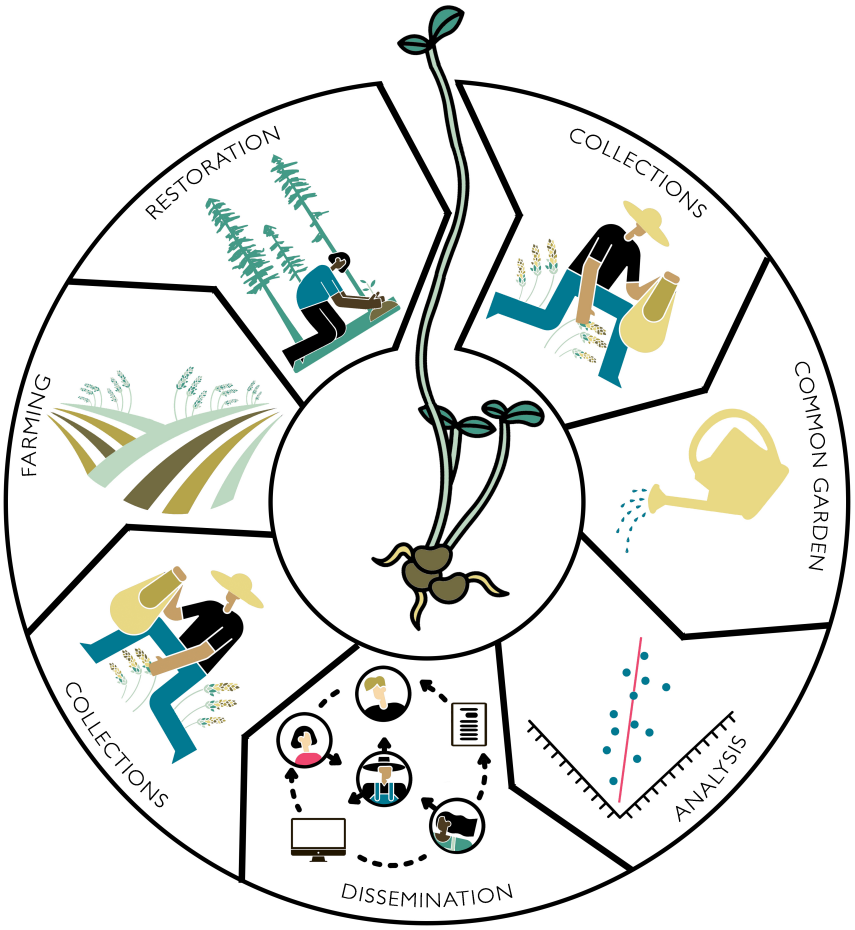

Figure 1: Dissemination. The first three panels indicate the process of developing an eSTZ, while the 'dissemination' panel showcases the need to share results so they can inform operational seed collections, agricultural increase, and selection of materials for a restoration. By Emily Woodworth

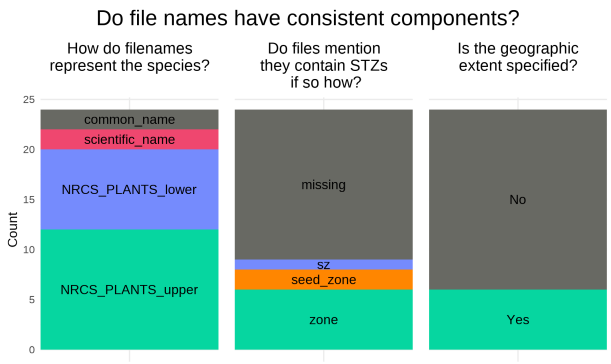

Figure 2: File Naming. Three inconsistencies in file names discussed here, with the advised format for data sharing in green, and the least desirable condition in grey.

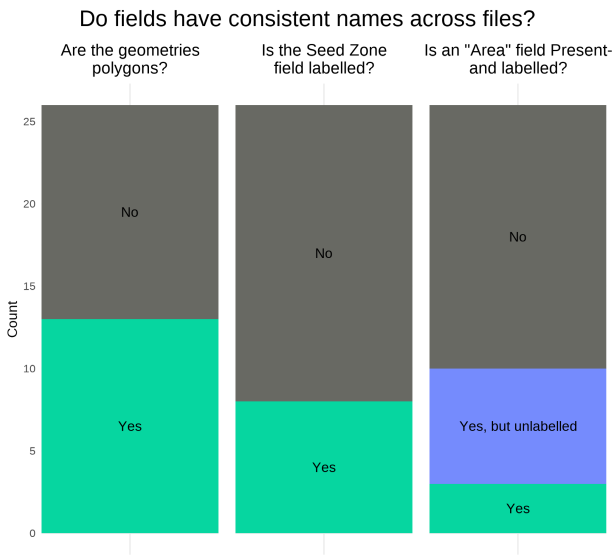

Figure 3: Field Names Shapefile. The three attributes of field names discussed here, with the most desirable condition in green, and the least desirable condition in grey.

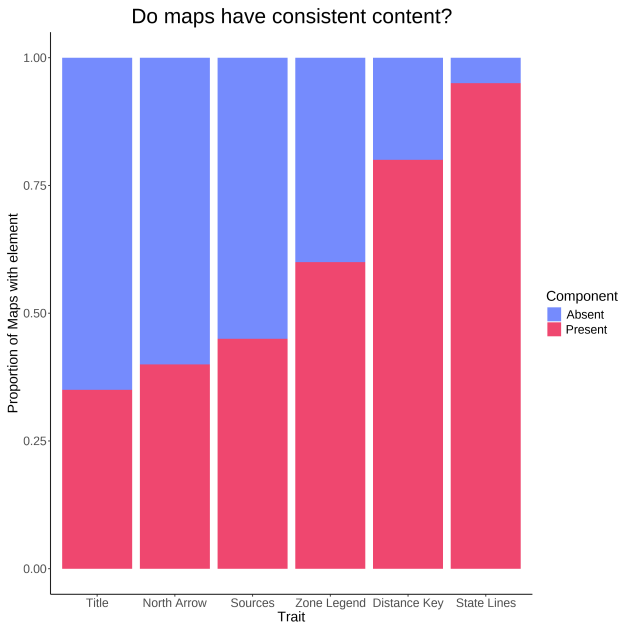

Figure 4: Map Components. Several essential cartographic elements - most notably a Title, a statement on data sources, and a legend for the seed zones, where missing from at least - or nearly half of the products inspected.

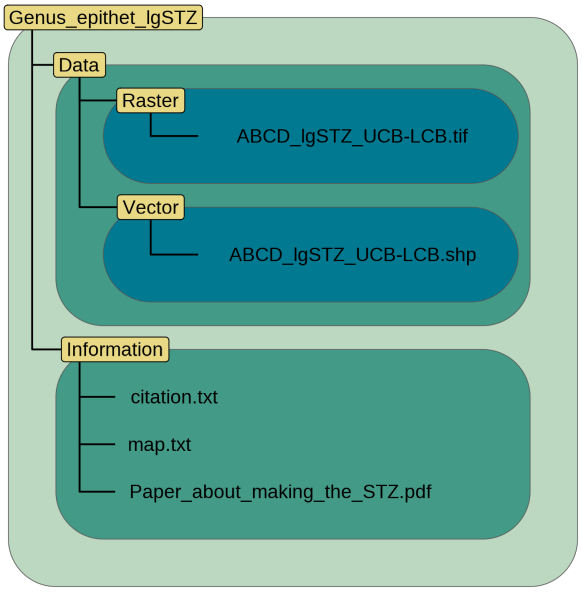

Figure 5: Directory Structure. Each directory is named in yellow, and spans the extent of variously coloured polygons. Individual files (or a set of files in the case of a shapefile) are depicted in black text within these directories.

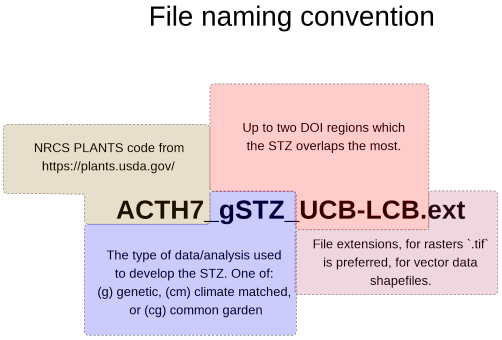

Figure 6: File Naming. The four proposed components of a filename are highlighted in different colours, and with appropriate cases.

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. The proposed field names for distributing vector data.

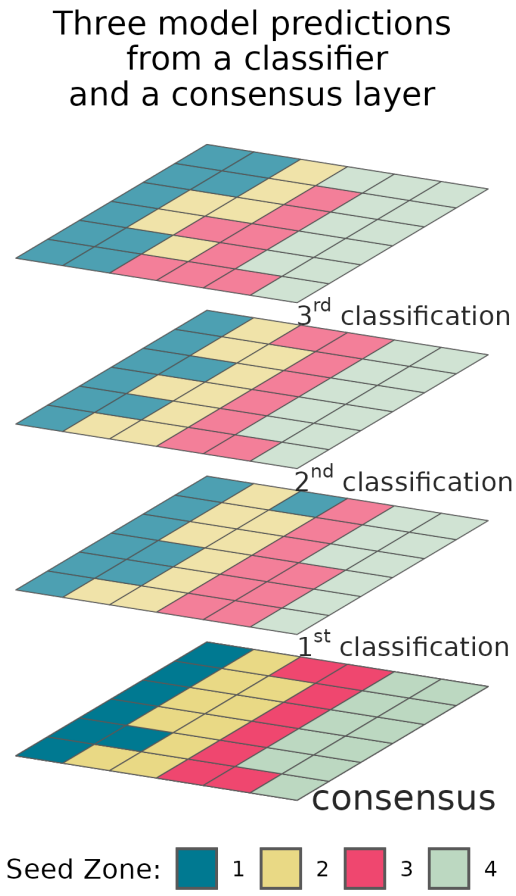

Figure 8: Many common classification algorithms can output probabilities for each class in the model. Reporting a subset of these allows users to interact with prediction probabilities directly.