RV definition Email discussion starting March 2016

If we make the decision that low-rv-high values will be based on percentiles, then the data we put into those fields will have to be calculated. The only way to calculate percentiles of (slope, MAAT, etc.) for components is to use pedon data--or some other down-scaling GIS layer that describes the locations of components.

I used the term "coarser-scale" (OK, a miss-spelled version of that term) because GIS summaries computed from map unit polygons represent that: map unit summaries. Manual adjustment of those values, based on expert knowledge, results in values that are no longer percentiles.

The statistical integrity that percentiles offer comes with a price tag: they need to be calculated. Calculations require data. It is our job as experts to decide which data are included in the calculation.

It is shovel time when sufficient data are not available.

Dylan

On map unit level summaries:

I didn't anticipate that this would be the case. Please give me a moment to take my foot out of my mouth.

So, MU-level summaries (someone's favorite quantiles) are the starting point for the component records, adjusted as needed? This is a nice mixture of mechanical + expert evaluation.

The only downside to this approach (there may be others) is that the low-rv-high values aren't really what we say they are: they are massaged versions of percentiles calculated at a courser scale.

Does that matter? I don't know, but this is probably one of the reasons for not writing specific guidelines.

Maybe the guidelines should adopt a loose interpretation of select percentiles. "The low value is approximately the 10th percentile of ....".

Dylan

Actually, all that gis/environmental data that we are generating from the Map Unit Summary Report that we just created is going into the component table because that is where they are stored in the data mapunit.  For slope and climatic variables they are likely to be the same for all components in the map unit.  But it very much depend on the landscape and mapping design that you are dealing with.

Jennifer Wood

A calculation that loads specific fields from a text file (linked via record id) would be the simplest. Henry has made a number of these to perform automated import of all kinds of data.

Population of component (environmental) data is an entirely different can of worms: I don't think anyone has suggested map unit level aggregates be dumped into the component records. The determination of those low-rv-high values should be derived from real data (pedons) when possible.

I don't have a clear idea of how this would work, but a standardized segmentation of the landscape (e.g. geomorphons) could someday be used to stratify the summary of environmental data by component.

We are back to the original discussion: picking low and high values--I think that everyone is on board with the median as the RV. I still suggest either the 5/95 or 10/90 percentiles.

One final note about basing low-rv-high values on percentiles: the values must be calculated from real data with a sample size of at least 3, better 5, and ideally >10. Do we have 5-10 points of real data for every component concept?

**Dylan**

What would it take to auto populate gis data in NASIS?

Create raster products for the whole country, then calculate the summaries for all the map units externally, then produce reports that compared the calculated with populated, then msso’s decide if the buttons should be pushed to autopopulate?

Another important thing to remember that John Fisher reminded me, that these calculations run for the whole map unit and do not differentiate by component.  There are plenty of instances where components should be distinguished in terms of slope, elevation, aspect, etc.  We don’t want to wipe out differences where they have been set up intentionally.

This is a pretty big undertaking we are talking about that I think is a great idea to start setting in motion.

Perhaps for now, for the Handbook, and for interim guidance, we go ahead and decide on a set of percentiles to recommend:  5th and 95th ?

Jennifer Wood

Definitely auto-population, we should value peoples time.

Tom D’Avello

Agreed, automated population of environmental data should be the new normal. Especially in an age of merging, splitting, and eventually line work changes. We don't want to accidentally change our jobs into database slaves.  
  
Dylan

Tom, I guess you are talking about replacing low-rv-high with those 7 parameters just for the GIS data elements?

Now would be the time since Teachman just sent out a notice asking for NASIS changes for the next version.

To be ready to ask msso’s to populate their databases with those parameters would be a major rollout of training.  Actually, I wonder if it could be automated?  Hand population of spatial data is starting to be pretty archaic.

Jennifer Wood

Regarding NSSH 618.55, I added the reference to percentiles to serve as a placeholder, solely to frame comments and elicit discussion from the reviewers.

The more I have thought and discussed this “problem” of populating lo-rv-hi, the more apparent it has become that too much thought and discussion has occurred. We just need to report standard distribution parameters. No muss, no fuss and users would always know what is what. An added benefit would be eliminating the rules and guidelines associated with making judgement calls.

A couple thoughts that would improve things:

1. Replace lo-rv-high in NASIS with a standard array of percentiles, e.g. – 5,10,25,50,75,90,95
2. Implement links to graphics from NASIS and SSURGO so users can see a pre-maid frequency distribution or box plot of respective variables

Option 1 could be implemented by adding a few more columns to NASIS

Tom

Agreed, we are getting into the realm of picking thresholds because they seem right.

Also agreed that the RV is more important than the low or high, and that the median is an ideal metric to standardize on.

The great thing about these reports is that it is relatively simple to adjust what is reported.

The concept of floating "low" and "high" values (e.g. the percentile used is variable) is better than our current framework (no standard), but seems that it would be very confusing to apply and document. The only way that this might work is if there are two suites of variables with their own thresholds for "low" and "high" values: GIS summaries use the 10th and 90th, pedon / lab summaries use the 5th and 95th... or something like that.

We may just have to pick something: limits associated with hypothesis testing sound nice, but our data rarely support any of the methods from that framework.

I like the idea of using the 25th-75th for comparisons of "central concepts"--series, components, map unit GIS data, etc.

I haven't seen much in terms of justification for any of these specific percentiles. I'll scan and send over some stuff from Fred Young and others who have previously wrote about this very dilemma.

Dylan

It would be preferable to see a single standard advised in the NSSH with allowance for circumstances.  But if a soil scientist desired to deviate from that, it is ok as long as supporting documentation is provided to explain any deviation from the NSSH standard. If the NSSH standard is followed then no supporting documentation is needed.

It could be that differing percentiles could be applied depending on the confidence of the mapping and the order.  Order 2 could have up to 5/95% depending on the complexity of the terrain.  Order 4 might have the 25/75 due to the scale and complexity of the terrain.  Size of polygons for a given mapunit might dictate as well, but this is usually in line with scale and order.

We need to remember that the NSSH is both a guide for the soil scientist and a base of documentation for the users.

**Kit Paris**

All thresholds are equally arbitrary, but if we want to be consistent with the current statistical default we should probably be using 2.5% and 97.5% as our thresholds. This would capture 95% of the data (i.e. a 1 in 20 chance), and coincides with a 95% confidence interval and the 5% exceedance threshold used in hypothesis testing. All of which are criticized in statistics, but none the less are still the common benchmark. Personally I’m less concerned about the Low and High thresholds, and more concerned about the RV, which I think we’re in agreement about.

I’ve always set the default in my reports to the min, 25th, 50th, 75th, and max because that is also the default given by the summary() and quantile() functions (i.e. Tukey’s five number summary). It seems that the 25th and 75th should constitute the central concept of a map unit or soil series. I’m not advocating them for L and H population, but I think they’re important conceptually.

Perhaps we should use different threshold for different properties. Thus we’d recognize that some properties are easier to quantify that others. Research has demonstrated this repeatedly. My observation in Region 11 has been, when it comes to slope gradient, we’re generally only capturing 50% of the data (i.e. the 25th and 75th percentiles). Even the reality of the 10th and 90th percentiles for slope gradient would probably be unsatisfying for most soil scientist.

All thresholds are equally arbitrary, but if we want to be consistent with the current statistical default we should probably be using 2.5% and 97.5% as our thresholds. This would capture 95% of the data (i.e. a 1 in 20 chance), and coincides with a 95% confidence interval and the 5% exceedance threshold used in hypothesis testing. All of which are criticized in statistics, but none the less are still the common benchmark. Personally I’m less concerned about the Low and High thresholds, and more concerned about the RV, which I think we’re in agreement about.

I’ve always set the default in my reports to the min, 25th, 50th, 75th, and max because that is also the default given by the summary() and quantile() functions (i.e. Tukey’s five number summary). It seems that the 25th and 75th should constitute the central concept of a map unit or soil series. I’m not advocating them for L and H population, but I think they’re important conceptually.

Perhaps we should use different threshold for different properties. Thus we’d recognize that some properties are easier to quantify that others. Research has demonstrated this repeatedly. My observation in Region 11 has been, when it comes to slope gradient, we’re generally only capturing 50% of the data (i.e. the 25th and 75th percentiles). Even the reality of the 10th and 90th percentiles for slope gradient would probably be unsatisfying for most soil scientist.

Stephen

PS – good work on the tutorial!

From: Wood, Jennifer - NRCS, Davis, CA   
**Sent:** Thursday, March 31, 2016 9:40 AM  
**To:** D Avello, Tom - NRCS, Morgantown, WV <[Tom.Davello@wv.usda.gov](mailto:Tom.Davello@wv.usda.gov)>; Beaudette, Dylan - NRCS, Sonora, CA <[Dylan.Beaudette@ca.usda.gov](mailto:Dylan.Beaudette@ca.usda.gov)>; Roecker, Stephen - NRCS, Indianapolis, IN <[Stephen.Roecker@in.usda.gov](mailto:Stephen.Roecker@in.usda.gov)>  
**Subject:** Final answer on percentiles? FW: Map Unit Summary Report for Environmental Variables

Good morning fellas,

Region 2 office staff are working through the tutorial I put together from the map unit summary report that Dylan and I put together, attached (more complete versions than that ones I submitted for  my final project for the class).

I got a question, below, from Kit about potential conflicting guidance about the percentiles to use for low and high in nasis.  Currently Tom is proposing that the 10 and 90th be used for slope, in the NSSH.  I think Stephen has suggested 25th and 75th, and Dylan has always promoted 5th and 95th.

Perhaps an agreement can be reached?

Thanks, Jennifer

Jennifer Wood

**From:** Paris, Kit - NRCS, Davis, CA   
**Sent:** Wednesday, March 30, 2016 7:45 AM  
**To:** Wood, Jennifer - NRCS, Davis, CA <[Jennifer.Wood@ca.usda.gov](mailto:Jennifer.Wood@ca.usda.gov)>  
**Subject:** RE: Map Unit Summary Report for Environmental Variables

Jennifer,

Whey are the charts returning only the 5th, 25th, 75th, and 95th percentiles when the proposal for the NSSH Part 618.55 recommends the 10th and 90th  percentile?

**Kit Paris**

Appendix 2 – Region 2 R-based tool to summarize the environmental variables for map units

R-Based Map Unit Summary Reports

June 2016

Dylan Beaudette, Digital Soil Mapping Specialist, SSR-2, Sonora, CA

Jennifer Wood, Soil Data Quality Specialist, SSR-2, Davis, CA

Russ Almaraz, GIS Specialist, SSR-2, Davis, CA

# Objective

Provide quantitative summaries and comparisons of select environmental properties (as defined by raster data sources) according to map unit delineations.

# Background

Initial mapping and MLRA update work in the soil survey program require knowledge of the variation in the environmental properties across the spatial extent of a map unit.

Some of the environmental properties of interest include:

* Elevation
* Slope
* Aspect
* Surface curvature
* Mean annual air temperature (PRISM 800m)
* Mean annual precipitation (PRISM 800m)
* Frost free days (PRISM 800m)
* Growing degree days (PRISM 800m)
* Solar radiation (modeled from DEM)
* Land cover (NLCD, etc.)
* Derivatives and indexes based on any of the above such as:
  + Effective precipitation
  + Compound topographic index
  + Slope shape (curvature) classification
  + Geomorphon-based landform element classification

Soil scientists need quantitative descriptions of the central tendency and spread for these values in order to evaluate map unit concepts while actively mapping, performing update work, and addressing questions raised by SDJR projects.

NRCS stores and maintains manually populated summaries of the environmental properties of map units in NASIS (component and related tables). These data are provided to users of completed maps as part of Web Soil Survey reports and data downloads.

The environmental data in the NASIS Component tables associated with recently completed soil surveys were typically derived from some form of zonal statistical analysis. Those operations generally return a min, max, mean, and standard deviation. Information about the distribution of the data across classes such as in a table or histogram is often generated as well. For the last 10 years or so, Region 2 has been using a set of tools developed by Lucas Wisely, an NRCS GIS specialist currently in located in Denver, CO. He developed a system using an ArcGIS tool and Python script to run zonal statistics on raster data (for polygons associated with a single map unit) for MAAT, MAP, elevation, slope gradient, and slope angle. He also included a classification of curvature into slope shape classes for each map unit. The Crystal Reports program is used to display the results in a user-friendly format.

# Justification for a new map unit summary method and report

There are many ways to summarize, analyze, and visually inspect sets of data. The R computing environment offers access to a vast range of statistical methods, from very simple to very advanced. R also provides many ways to visualize data, which helps the user explore and understand the data better.

Scripts and reports can be created so that beginning users only have to point the report to the inputs required for the analysis. In the Methodology section below, we describe a report that does exactly that, and is designed to match or exceed the map unit summary functionality of the Lucas Wisely/Crystal Report method, see Table 1.

**Table 1. Comparison of Lucas Wisely/Crystal Report vs R-Based report**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
|  | **Inputs** | **Software, scripting** | **Summary Statistics Provided** | **Landform classification** | **Report Format** |
| **Lucas Wisely/Crystal Reports** | **Shapefile with one or more map units, environmental data rasters** | **ArcGIS, Python script, Crystal Reports** | **Min, max, mean, standard deviation** | **Classification of curvature –scale/window size dependent** | **Crystal Report for each map unit** |
| **R-Based Report** | **Shapefile with one or more map units, environmental data rasters** | **R studio, R scripting, R markdown document** | **User-defined percentiles** | **Geomorphon approach – scale independent and,**  **Curvature classification with fixed window size (5x5 for region 2 DEM)** | **HTML Report with summaries of all map units by variable** |

Methodology

Setup

The proposed R-based mapunit summary report is provided as an R Markdown document (.Rmd file extension), along with a tutorial document. For users who are completely unfamiliar with the R Studio environment, there will be a link to a basic R studio tutorial. The .Rmd document is opened in R Studio and the user is directed change the path and file name to the location of their shape file. The user is directed to ensure that the rasters required for analysis are in a specified location.

Inputs

Two kinds of inputs are required. The first is a shape file (or geodatabase) containing polygons associated with one or more map units. The second is a set of rasters that contain relevant environmental and terrain-shape properties. These rasters can be customized (regional 30 meter products, local 10 meter products, etc.) according to individual needs.

Analysis

The script is directed to sample a specified number of points per polygon (e.g. 1 point per acre by default) in the shape file provided. These points are used to extract the value of each of the rasters provided at each of the points. Various analyses are performed on this set of sampled values.

Output

In R Studio, on the toolbar for the .Rmd file, the user clicks on the Knit HTML button. R begins running the script which could take several minutes, depending on the size of the area being analyzed. The results are displayed in an HTML report that opens automatically and is also saved automatically to the working directory.

Output displayed – For examples see [here (map unit summaries)](http://ncss-tech.github.io/example-reports/mu-comparison/CA630-mu-comparison.html) and [here (MLRA summaries)](http://ncss-tech.github.io/example-reports/mu-comparison/MLRA-comparison-report.html)

* *Input Data*: variables and file path
* *Area Summaries*: map unit total acreage and 5th, 10th, 25th, 50th, 75th, 90th, and 95th percentiles of polygon acreage
* *Box and Whisker Plots:* Whiskers extend from the 5th to 95th [percentiles](https://en.wikipedia.org/wiki/Percentile), the body represents the 25th through 75th percentiles, and the dot is the 50th percentile
* *Density Plots:* These are equivalent to a smoothed histogram, and visually display the distribution of the raster values across the range of those values in the map unit. They are more appropriate than histograms for continuous data, which can be sensitive to ‘bin size’ or size of class.
* *Tabular Summaries:* 5th, 10th, 25th, 50th, 75th, 90th, and 95th percentiles, by variable, based on the sampled values
* *Circular Summaries:* A graphical summary of aspect statistics on a circular diagram, based on the sampled values. Spread and central tendency are depicted with a combination of circular histogram and kernel density estimate. The circular 50th percentile is shown with a red arrow and the 10th and 90th percentiles are shown with gray arrows. Arrow length is proportional to the directionality of the data: longer arrows suggest a more directional pattern.
* *Slope Shape (Curvature) Summary:* Table and graphical summary of slope shape, based on a classification of surface curvature values into “concave”, “linear”, and “convex” groups. Combinations of down-slope and across slope shapes are given as proportions.
* *Geomorphon Landform Classification:* A table of values and graphical representation of landform types, based on the sampled values, expressed as a proportion of the map unit. The Geomorphons algorithm is [a new approach to classification of landforms](http://geomorphometry.org/StepinskiJasiewicz2011). This is a scale-independent method, in contrast to the classification of curvature returned in the Lucas Wisely report. Landform types are:

|  |  |
| --- | --- |
| Geomorphon Name | Approximate Slope Shape Equivalent |
| Flat | LL |
| Summit | VV |
| Ridge | LV |
| Shoulder | VL |
| Spur | LV |
| Slope | LL |
| Hollow | LC |
| Footslope | CL |
| Valley | LC |
| Depression | CC |

* *Multivariate Summaries:* This summary is the result of an “ordination” analysis, which quantifies the similarity of all of the raster variables across the sampled points in each of the map unit polygons. An optimal subset of raster data are generated via conditioned Latin Hypercube Sampling (clhs). (additional explanation / references pending)