**Proposal for changes to the National Soil Survey Handbook**

New definitions of “low- representative value-high”

Proposed by**:** (Add your name if you concur with this proposal, include title and email address.)

Jennifer Wood, Soil Data Quality Specialist, Region 2 Soil Survey Office, [jennifer.wood@ca.usda.gov](mailto:jennifer.wood@ca.usda.gov)

Dylan Beaudette, Digital Soil Mapping Specialist, Region 2 Soil Survey Office, [dylan.beaudett@ca.usda.gov](mailto:dylan.beaudett@ca.usda.gov)

Tom D’Avello, Soil Scientist/GIS Specialist, NSSC-Geospatial Research Unit, [tom.davello@wv.usda.gov](mailto:tom.davello@wv.usda.gov)

Stephen Roecker – Soil Data Quality Specialist/GIS Specialist, NRCS Region 11, [stephen.roecerk@in.usda.gov](mailto:stephen.roecerk@in.usda.gov)

Russ Almaraz, GIS Specialist, Region 2 Soil survey office, [russ.almaraz@ca.usda.gov](mailto:russ.almaraz@ca.usda.gov)

Jason Nemecek, State Soil Scientist, Wisconsin, [jason.nemecek@wi.usda.gov](mailto:jason.nemecek@wi.usda.gov)

Skye Wills, Soil Scientist, National Soil Survey Center, [skye.wills@ca.usda.gov](mailto:skye.wills@ca.usda.gov)

**Proposal:** Establishment of a more precise definition of the terms “low, representative value, and high” (l-rv-h)in the National Soil Survey Handbook (NSSH) for the population of data elements in NASIS.

**Proposed definition for new section C. in Part 618.2**

1. For recent and newly populated information in NASIS, the representative values are meant to approximate the 50th percentile (median). The 50th percentile is the value where 50% of the data are both above and below that value. The low and high values are meant to approximate the 5th- 10th and the 90th-95th percentiles, respectively. For example, the 5th percentile is the value where 5% of the data are below that value and the 95th percentile is the value where 5% of the data are above that value. The low, high, and representative values for data populated prior to this version of the National Soil Survey Handbook were not guided by the percentile approach but generally approximate the current definition. Even where data used to populate the SSURGO database are not computationally derived, the populated values are designed to approximate the aforementioned percentiles for the data set being described.  
     
   The percentile approach is preferable to other measures of central tendency, such as the mean and standard deviation, because percentiles require no distributional assumptions and the percentile values fall within the bounds of the data set from which they are computed. This means that percentiles can provide benchmarks for the spread and central tendency for both normal and non-normal distributions, and the values will always fall within the minimum and maximum of the observed data. Consider a hypothetical data set for field-described clay content from the A horizon of the same taxa:  
     
   clay content: 11, 10, 12, 23, 17, 16, 17, 14, 24, 22, 14

clay content sorted: 10, **11**, 12, 14, 14, **16**, 17, 17, 22, **23**, 24

Low/10th percentile = 11

RV/50th percentile = 16

High/90th percentile = 23

**Justification:** Currently the terms low, rv, and-high are mentioned in sections [618.2 and 618.3](http://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/ref/?cid=nrcs142p2_054223) of the NSSH, but not defined. A loose definition has made sense because of the highly variable methods used to populate the low, rv, and high values in all of the various data fields in NASIS that represent a collection of many different vintages and sources of soil survey data. And, for much of the uses of soil survey information, this level of precision likely provides enough accuracy. Additionally, this new more precise definition actually does reflect how past data developers have thought about the meaning of low-rv-high. We have always tended to populate the rv as a median or a mode, rather than a mean. We have always thought about the low and high as values within the observed data set that approximated not the extreme, but commonly observed, ends of the data distribution.

For recent, ongoing, and future data collection efforts however, we aspire to populate NASIS fields using as much field-collected data as possible. As our data becomes more and more widely used, it has become increasingly urgent that we establish more precise definitions for these data ranges. While acknowledging that older data may not have been populated using this particular guidance, updated definitions are needed for current work that can accommodate more data-driven and modern approaches to computing soil survey data ranges. Clear definitions will more completely convey our best available knowledge to users and provide a standard to assure the correct uses and limitations of our data.

**Discussion:**

The topic of how to represent soils and landscapes in our tabular database is not new and many venerable soil scientist have grappled with this topic (see References section). The kinds of variables for which we provide numerical ranges in our database are as follows:

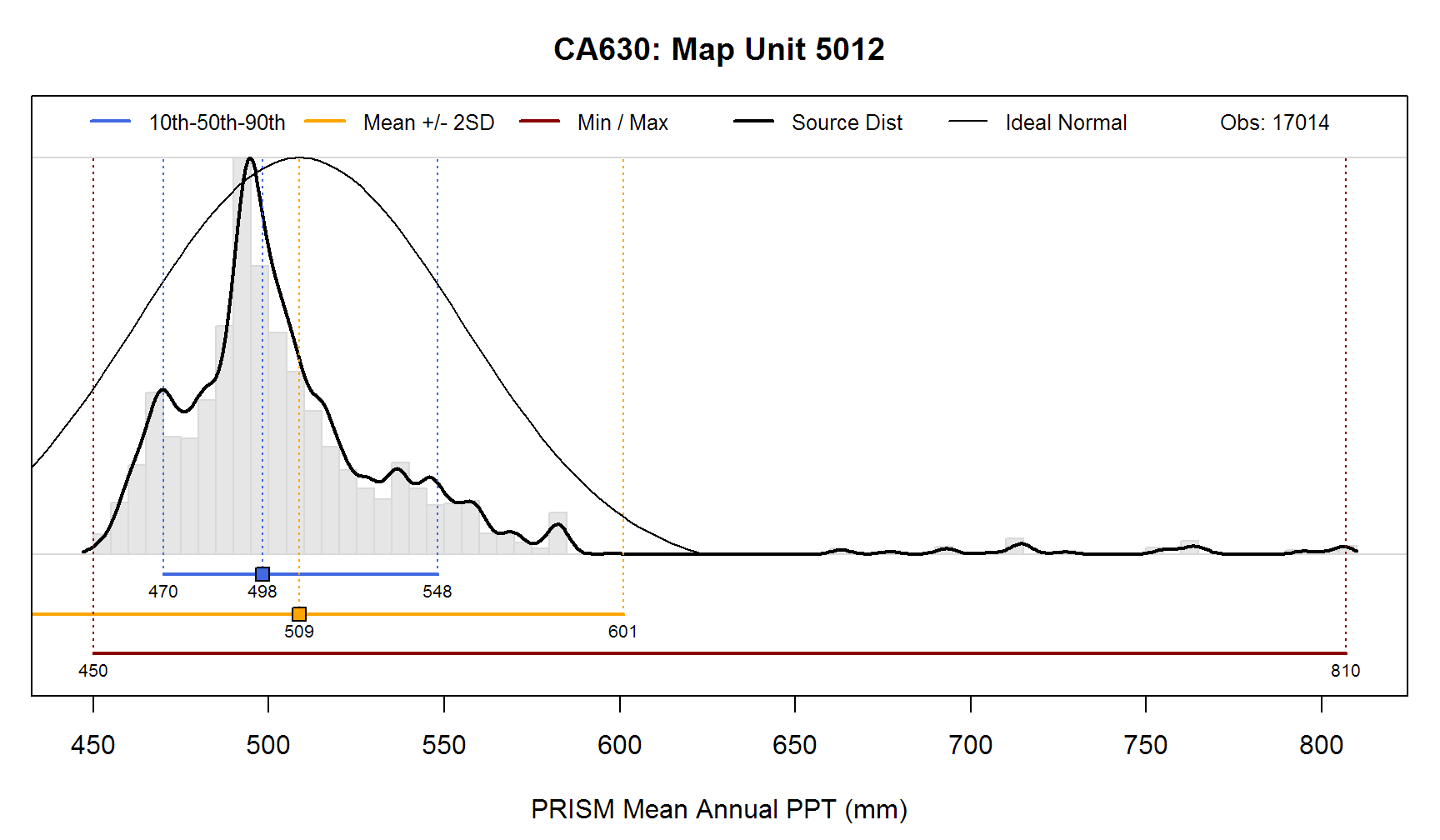
|  |  |
| --- | --- |
| Variable Type | Examples |
| Soil properties/derived values for component horizons in a map unit | Clay %, pH, Ksat, bulk density, %OM, EC, |
| Soil properties and environmental data for components in a map unit | Surface fragments, diagnostic features horizon depths, Slope, MAAT, MAST, |

Note - While l-rv-h columns are available in the Site and Pedon objects, they are generally not populated.

Population of “low”, “RV”, and “high” values in the NASIS Component table is a critical component of initial mapping and update work. The definition of “low”, “RV”, and “high” values is not defined in the National Soil Survey Handbook and other National and Regional guidance is variable and even lacking. As a result, the population of these values varies across Regions, office areas, survey areas and even across map units within a survey area. In Region 2, many MLRA offices that are creating MLRA map units as a result of SDJR projects have been using the *mean* for population of the “RV” of elevation, MAAT, MAP, frost free days in NASIS. The “low” and “high” values are variously populated using the min, max, one or two standard deviations away from the mean, or some other method meant to capture the majority of the variation.

There have been discussions for a while, and current efforts are underway, to transition to a standardized approach to the population of “low”, “RV”, and “high” values in NASIS. See [this file located on GitHub](https://github.com/ncss-tech/soil-range-in-characteristics/blob/master/references/RV%20defnition%20email%20discussion.docx) for a 2016 email discussion, with additional comments, on this topic. Also see [this GitHub page](https://github.com/ncss-tech/soilReports) for the documentation of an R-based tool to use percentile-based statistical approaches to summarizing the environmental variables for map units.

A more in-depth discussion of the rationale for using the percentile approach, with examples using commonly described soil survey data, is presented at [this NCSS GitHub page](https://github.com/ncss-tech/soil-range-in-characteristics) and at the [NRCS National Water and Climate Center website](http://www.wcc.nrcs.usda.gov/normals/median_average.htm). Below is an example figure from the discussion on the GitHub page that demonstrates the problem of using the mean and standard deviation to represent the central tendency and spread of a data set. In this example, the Mean Annual Precipitation data has a long tail. Because the mean assumes normal distribution of the data if it is to represent the central tendency, the calculated mean in this example is higher than where the majority of the values are clustered in this data set.



This proposal is related to a proposed change to the NSSH, from Tom D’Avello for Part 618.55 in reference to the population of the “low”, “RV”, and “high” values for Component Slope Gradient in NASIS:

“These values may be determined by a statistical summary of the slope gradient layer for a given map unit layer. Slope gradient distributions are seldom normal, eliminating the use of conventional statistical parameters like mean and standard deviation as tools for determining the high, low a representative values. These values should be based on the robust parameters of percentiles. The representative value is based on the median. The low and high should be based on ranges that capture a majority of the area represented in a map unit. Using the 10th and 90th percentiles as the low and high, represents 80 percent of the area.”

We propose here to decide on a standard towards which we aspire for all values in NASIS. There is general agreement that there is justification for the central tendency approach, with RV as median/50th percentile.  For the low/high we can settle on a set of values, acknowledging any limitations. There is a lack of consensus on which percentiles to target for the low and the high. The 5th and 95th, 10th and 90th, and 25th and 75th have been suggested, see [this file located on GitHub](https://github.com/ncss-tech/soil-range-in-characteristics/blob/master/references/RV%20defnition%20email%20discussion.docx) for an extended email exchange on this topic, along with subsequent comments. In the currently proposed definition we suggest that the ‘low’ is intended to approximate the 5th to 10th percentile, and the high is intended to approximate the 90th to the 95th percentile. This acknowledges that the data developer has leeway in populating this value depending on the quantity of data, the quality of the data, and the kind of variable being described. For instance, the data source may be for the map unit, whereas the value being reported is for the map unit component. Or, data being summarized may be for component soil properties, where much less data might exist, and a wider range of the samples would like to be included ( for instance 5th to 95th percentiles, encompassing 90% of the range observed). This would be in contrast perhaps to modeled raster-based elevation or climatic data sets, where a narrow range of samples could be included due to the abundance of pixel values available to sample (10th to 90th percentiles, where capturing 80% of the range might be more appropriate).

Limitations

*Soil Property Values* We populate NASIS with l-rv-h for soil properties and some interpretations.  In some cases, when there are enough data (a judgement which depends on the philosophical approach to defining the population), applying the central tendency approach computationally, from field and lab data is possible. In many more cases, there are likely not enough data to compute values, but an approximation of the central tendency with approximately standardized percentiles is still the goal.

*GIS derived environmental and terrain values that are populated for the components* GIS summaries of central tendency values can be easily calculated for map units as a whole. When we want to constrain the ranges for these values by component, what techniques are available and/or how do we express to the user how populated values are derived?  It seems we have a choice of some kind of digitally derived, expert knowledge, or model-driven value – or some combination. Does the statement proposed adequately address the inability to computationally derive l-rv-h values in many cases?

**References**

[See dynamic list on this NCSS GitHub page and feel free to add more](https://github.com/ncss-tech/soil-range-in-characteristics)