**Dylan Beaudette**

Paul Benedict (regarding tracked changes in Word doc submittal):

I had some comments and suggestions on this proposal. A clear guide on low, RV and high values is probably needed, but I also wonder if the same rules need to apply to all data elements. I hope my comments don’t come across as too negative as I know this issue is important to those that proposed it and they are the ones dealing with this issue in their daily work and not me.

Ken Scheffe:

I read the proposal by Dylan and Jennifer regarding the population of low, rv, and high values in the NASIS database using percentiles.  It is a good proposal and would allow our scientists to increase the accuracy/precision in the database. I believe this is the correct way to move forward, but I have a few caveats and perhaps warnings that need to be considered.

The original population for low, rv, high values in NASIS most often was not with observed values, but actually represented the numeric limits of descriptive classes for the property or feature.  In some circumstances, the low and high actually represented the limits spanning two adjacent classes (soil reaction or pH was one of these).

*This is a valid concern and the proposed changes to {low, rv, high} could create a minor discontinuity between “new” and “old” data. While I wouldn’t want to add to our large list of discontinuities I do think that a more formalized definition of these concepts would improve consistency over the long term. The proposed concepts when applied via “update” work would iteratively smooth out these kind of discrepancies. I do think that class limits are a useful bracket when data are sparse—however, we should strive for something more robust when there are more than (for example) 5 observations. Honestly, we should be wary of series and component concepts that don’t have more than 5 observations per horizon—at least for the basic morphologic and derived properties.*

For example, if the texture was recorded in SSSD or the Soils-5 as ‘sandy clay loam’ the population of low, and high values for clay percentage in NASIS would have been low=20%, high=35% --- the class limits for clay for the texture ‘SCL’. We were permitted to modify the range with observed limits and the rv was to be the most commonly observed value based upon the experience or other supporting data.  However, most often, there was no one with field experience, nor time, nor desire to populate individual layers with tailored data.  We were ‘instructed’ by National Bulletin to populate l, rv, h – just get it done because it was needed now in order to generate new calculated interpretations (every property, every horizon, every component, every map unit, every survey in the state – and do it this year!)

*I think that those decisions made a lot of sense given the availability of resources and urgency of the request. Given sufficient data, it should be possible to derive the ranges and rv using percentiles as part of the update process.*

In NM, we were fully engaged in initial and update projects under reimbursable agreements during this time, so we had limited time/staff to research manuscripts, laboratory data, or field notes, so we developed routines to query for properties, such as texture = scl, and would populate in-mass using NASIS calculations,  the low, rv, and high values at the class limits, and the rv as an approximation of the middle of the class. I don’t believe NM was different than many other state, so much of the original population of l, rv, h in NASIS was by algorithms which generated class limits and a median value.

Not all is lost.  It has been 15 years since this was done.  Survey offices have done a lot of update work, and hopefully during this time, have adjusted the low, rv, and high values away from strict class limits with observations, lab data, and descriptive narratives gleaned from available sources to better approximate the real world. Kudos to those who tailored their data.  The problem is knowing which values for low, rv, and high have been tailored and which have not.  Perhaps the most telling thing would to be look for populated values that correspond to descriptive class limits as they likely have not been appropriately adjusted.

My concern, in addition to those expressed in the proposal by Dylan and Jennifer, as we go into this percentile representation of low, rv, and high would be a direct conversion, or simply a redefinition of data elements to the 5 or 10 percentiles and 90 or 95 percentiles.  I don’t believe it to be wise to simply redefine the terms low, representative value, and high to the percentiles without actually validating the numbers.  It would give the impression of a more statistically evaluated entry, without actually being true.  I didn’t see anything in the proposal by Dylan and Jennifer that indicating an evaluation method to assure appropriateness or validity would be performed as part of the redefinition.  I’m not worried about new data population based upon observations or measurements, but rather the legacy entries that simply represent the lower and upper limits of descriptive classes and the median value (rv).

*Fortunately, we are not proposing a change of existing data—rather, describing what “low, rv, high” aspire to be. Perhaps some language connecting past with present would help convey that goal.*

I think we should move forward on the proposal, but we probably need to talk with Dylan and Jennifer to see how this would be done.  It is not simply or only a redefinition of terms in the NSSH.  I believe the only honest way to do this would be to design and effort or initiative to systematically evaluate whether the current entries represent class limits, or are tailored values based upon observations, measurements, and documentation.  If the entries represent class limits, we need to evaluate and adjust them to more appropriate values which represent the reality on the ground for the appropriate percentile being used.

*It seems to me that this would only be necessary when/if there was discussion of re-calculating low, rv, high values for our entire collection of data. At this point in time that doesn’t seem like a wise decision.*

ALSO,  need to remember we populated map unit component data, which may have been wider, or narrower than the series named in the map unit.  Similarly, there were certainly real world occurrences, for example, where we had a soil with a fine-loamy argillic (upper 50cm for the family class) that very often was a sandy clay loam, however, it was known that on occasion in the ‘real world’ there were layers in the argillic that were actually heavy fine sandy loams at 18-20% clay.  This is still in the fine-loamy family, but we might have only entered the weighted average for he control section (sandy clay loam) to avoid potential confusion.   When the clay ‘low value’ was assessed and entered, we probably used 20% for the sandy clay loam class limit.

*I think that the language in the proposal leaves enough room for “adjustments” that may be required to make all of the RIC fit together into a coherent “story”. Removing subjectivity from the aggregation process was a large part of the motivation for this proposal. Perhaps we can adjust the wording to emphasize that (given sufficient data) percentiles are a scaffolding upon which objective RIC can be built and customized.*

As you stated, similar situations and accommodations were likely made for all the entries of low, representative value, and high.  And also, sometimes the value of zero is appropriate for such things as salinity in leached soils, slope on level soils,  etc. I don’t know that using the 5 or 10 percentile is appropriate when values of zero real.  You also mentioned in your track changes the need to avoid creating ‘gaps’ between classes if we restricted the current class limits to the 5th and 95th percentiles of that class.

*Given a large enough set of 0’s, it is quite possible to end up with 0 as the 5th percentile. Again, percentiles don’t replace critical thinking: they are the starting point for consistent RIC.*

Perhaps this is not so easy to do with existing data for (l, rv, h) as some would like to think.

Paul Benedict

Yes, your points were what I was thinking (although not in the detail you used) when I said I wondered if the same rules need to apply to all data elements. Sometimes for example, the lower value really should be a zero, even if the 10th percentile is a 0.1.

*Given a large enough set of 0’s, it is quite possible to end up with 0 as the 5th percentile. Again, percentiles don’t replace critical thinking: they are the starting point for consistent RIC.*