**DSM Practitioner’s Discussion**

**3/3/2020**

**Discussion leaders: Suzann Kienast-Brown, Dave White**

**Topic: Covariate Selection**

Attendees: not captured due to issues in Adobe Connect

* **Welcome and Announcements:** Suzann Kienast-Brown
  + If you are interested in Remote Sensing course email Suzann and Shawn NLT this week.
  + Intro to DSM course will be offered in May.
  + Found on DSM Focus Team website
    - Request for mentors is open for applications
    - DSM Field Week requests are now open year-round
* **Presentation:** Dave White – *Covariate Selection Methods*
  + **Link to presentation** [**HERE**](https://new.cloudvault.usda.gov/index.php/s/4SJ2zqzD7BRoGQC)
  + Overview of Pedological and Statistical Methods
  + There is no magic number or combination of covariates
  + Utilize any available tool to aid in covariate selection
  + The modeling process is iterative
    - Select covariates
    - Build a model
    - Look at model performance and assess covariate importance
    - Select a new set of covariates
    - Build a model…
  + If using statistical methods, can you explain why the covariates are being selected?
* **Discussion**

**Stephen Roecker** – Can you expand on what you do if combination derived from data reduction method doesn’t make pedological sense?

**Dave White** – An example -- What might it mean if PC-6 is the highly predictive covariate?

**Suzann Kienast-Brown** - Sometimes what we can’t understand requires more study. In the example of PC-6 it may be that the combination of several covariates provide information that a single covariate does not.

**Stephen Roecker** - Sometime selection process is wrong. Example: Aspect as a predictor of surface rock fragment cover. Does this make sense?

**Stephen Roecker** - Sometimes the machine-learning model can find pattern in the noise that is not representative of the natural world.

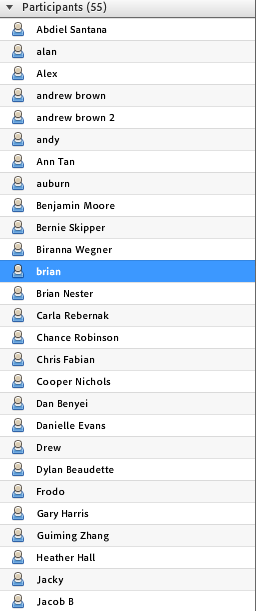
**Dave White** – If a covariate that you think is important isn’t represented in the selected covariate set, you may want to add it to the model and compare results. You could also utilize data exploration methods to compare and consider why that covariate might be missing from the model.

**DSM Practitioner’s Discussion**

**2/13/2020**

**Discussion leaders: Suzann Kienast-Brown, Tom D’Avello**

**Topic: Validation – developing qualitative means to express class data uncertainty**



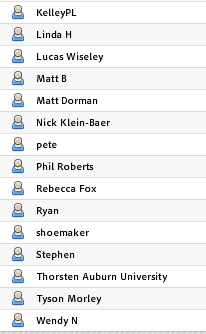
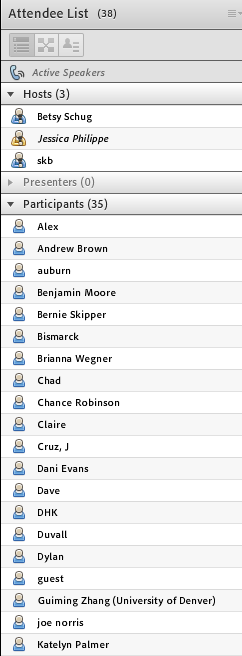
* Suzann
  + Raster Mastery topic
  + Webinar on Feb 20
  + March 26th FT meeting
  + Training update
    - Intro to DSM in OK
    - FL field week
    - ArcSIE March 16
    - RSSSA March 30
    - Intro to DSM May 4
* Tom D’Avello – qualitative assessment of uncertainty
  + Uncertainty calculations for continuous data is a done deal and taught in many of our training courses
  + Some classification methods have measures of uncertainty built in and some don’t
  + Incompatible nature of uncertainty
  + Qualitative uncertainty could be assigned by class (high, moderate, low)
    - Populated in raster attribute table
    - Displayed spatially as a layer with the map
  + Standard measure of uncertainty for RSS so similar info is being presented with map products
  + Tom is of the opinion that a standardizes quantitative metric is not possible.
  + Qualitative, ranked categories may offer a more useful solution than trying to devise a standard, quantitative metric
  + Developing a classification system to categorize uncertainty from various methods would provide a standardized procedure of assigning uncertainty class.
  + Possible description of a limited number of uncertainty categories would be useful to users
    - **Low** high degree of confidence named or similar class(es) to be found
    - **Moderate** moderate degree of confidence named or similar class(es) to be found
    - **High** moderate degree of confidence named or similar class(es) to be found
* Group discussion
  + How can we combine qualitative/quantitative measures?
  + How can we use uncertainty in delivery of our soil maps? We have to be able to decipher the information for users and present it in a meaningful way…through interpretations or risk or expression of complexity or ???
  + We don’t have to have all the answers right now and have the opportunity to figure it out as we go and as a team, which is fun 😊
  + Some resources
    - <https://gcc02.safelinks.protection.outlook.com/?url=https%3A%2F%2Fncss-tech.github.io%2Fmisc%2Faccuracy%2Faccuracy-uncertainty-soil-class-prediction.html&amp;data=02%7C01%7C%7Cfb6f7c71616f462b5d6408d7b0cc68e0%7Ced5b36e701ee4ebc867ee03cfa0d4697%7C0%7C0%7C637172264284377083&amp;sdata=FVbDyhVgumx6XbKflRT8UiVrrSsW8GrjSuU31RJCAw8%3D&amp;reserved=0>
    - <https://www.sciencedirect.com/science/article/pii/S2352009420300043>
    - weighted accuracy assessment <https://www.sciencedirect.com/science/article/pii/S0016706116303901?via%3Dihub> functions are now available in the aqp package in R

**DSM Practitioner’s Discussion**

**1/7/2020**

**Discussion leaders: Suzann Kienast-Brown, Jessica Philippe, Betsy Schug**

**Topic: Modeling and accuracy assessment**



**Update on Mille Lacs Uplands – Coarse-Loamy Basal Till - Modeling and Accuracy Assessment from Betsy Schug**

* Used ArcSIE and R
* RStudio
  + Develop statistics for determining ruleset for ArcSIE
    - Logistic regression to model physical properties
    - ArcSIE used modeled properties as covariates/environmental rasters to define rules and model soil classes
* Accuracy assessment design
  + Needed to maximize efficiency and see everything
  + Cluster sampling where possible
  + Independent targeting sampling in smaller areas
  + Poorly performed classes were correlated in with similar soils/classes
    - paired down 21 taxons to 14
  + Some poorly performing classes were retained in the model if they were important/highly variable

**Discussion**

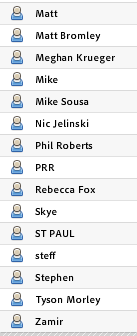
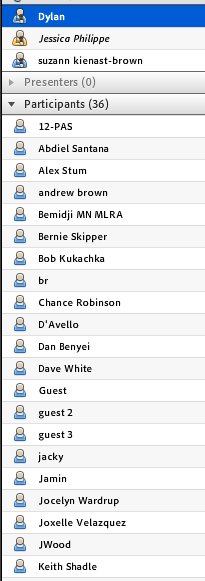
* Stephen – how much information is lost when targeting series rather than component phases? In this case not too much. Stoniness phases were combined, and it doesn’t necessarily mean lost information since the phasing was inconsistently applied in the past so it’s hard to know.
* Phased components that are important and have the data to support that (and make them modelable) they can certainly be part of the target
* Accuracy compared to SSURGO? How? Generate property map from both products, generate same depth interval property measurements for training data, and compare to both. Does the higher detailed product actually solve accuracy issues?
* There may be an assumption that DSM products are more accurate, but it’s really more about giving a more spatially explicit version of soils information; we don’t have accuracy measures for ssurgo but we can more easily generate them for raster products (doesn’t mean they are actually more accurate, just that we *can* report accuracy).
* If the raster is not actually more accurate, the added spatial detail is just noise. But can’t just assume that SSURGO is accurate, either.
* So focus is on providing multiple data types so the needs of the user can be addressed. Focus should be on making the best product possible.
* Maybe Stephen and/or Dylan will develop a way to compare SSURGO and RSS

**DSM Practitioner’s Discussion**

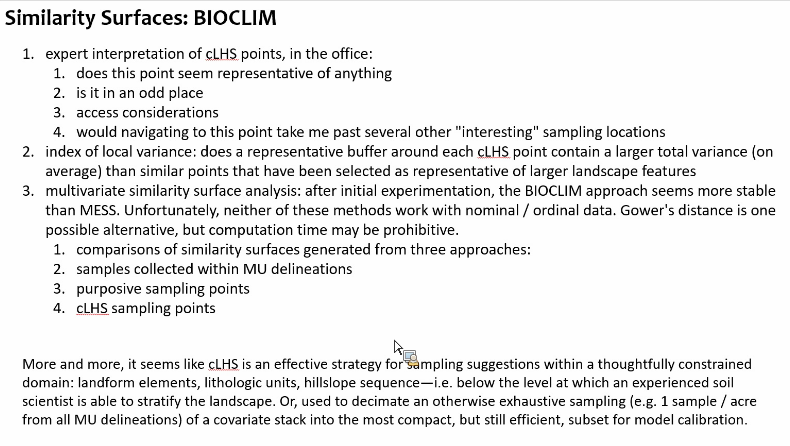
**12/3/2019**

**Discussion leaders: Suzann Kienast-Brown, Dylan Beaudette, Jessica Philippe**

**Topic: Incorporating similarity surfaces in sampling design**



* DSM Focus Team overview – Suzann
* Sampling design and similarity surfaces – Dylan
* Collecting data to…
  + Calibrate
  + Validate
  + Stratify
* Striking a balance between pure statistical sampling (I know nothing)
* Vs
* Pure subjective selection based on tacit knowledge (I have experience, I am confident in choosing a sample point)
* The nice thing about cLHS is that it takes subjective bias out. cLHS is good for smallest number of points to validate a statistical model
* BIOCLIM seems more stable than MESS for multivariate similarity surface analysis



* Discussion on what to do after using these techniques to evaluate points - key is a hybrid approach, allowing for adjustment/exclusion/addition of points based on expert knowledge.

**DSM Practitioner’s Discussion**

**10/1/19**

**Deeper Discussion of Sampling**

Discussion leaders: Suzann Kienast-Brown, Jessica Philippe, Dylan Beaudette



Preparing for Sampling (Kienast-Brown):

Sampling need to focus on scope and project objectives

Covariate selection needs to be carefully considered:

* + Represent important physical relationships with covariates
  + What covariates best represent physical features
  + Simplest Answer is the Best – Less Covariates is more
  + Just because you develop a model with 100 covariates, why would you
  + Must be able to explain covariate selection

In developing a sampling plan smaller feature space may render more logical results:

* More covariates require more training points to explain variability
* <10 covariates should be adequate
* Standard set of covariates may be considered (some literature suggests slope, curvature and wetness)
* List/standard set of covariates developed may be a good starting point for consideration

Alternative Ideas to Standard Set of Covariates/Standard Sampling Design

* Regional Specific Set of Standard Covariates

Dylan Beaudette suggested several that he’s found most effective in work west of the Rocky Mountains

Defining standard set could be a function of the Regional Office

* Strongly consider scale of processes

Scale of processes in relation to properties

Scale of mapping/resolution

* Do covariates have equal weight (cLHS is not weighted and distributes evenly based on feature space
* Does a standard set of covariates make sense when objectives are unique for every project.
* What are we trying to achieve: Property Mapping or Class Mapping
* Need to hone in on objectives before data wrangling

Stratification

* Recommended
* Soil scientists are uniquely qualified to do this by training
* Parent material/lithology/geology
* Difficult to explicitly represent in model
* We can cook up strata via multiple covariates
* Scale of processes considered – Landscape & Parent Materials
* Stratification should consider reassembly of domains
* Training points should be scattered across meaningful units
* Use pedologic tools for parameterization
* Strictly data driven approaches to sample design assume no prior pedologic knowledge
* Stratification again relates to objectives – National Scale vs Local Project Scale
* Is Geographic Stratification simply of method of clustering/sub-clustering?
* Political/Anthropogenic vs Natural Stratification
* Stratification imposes simple patterns
* Stratification is a tried and true method in soil survey, and in digital soil mapping/remote sensing studies
* Are we introducing bias

Need to emphasize context as opposed to bias

* Universal models require universal covariates/Non-stationary models commonly show positive spatial auto-correlation in residuals
* Need to continue to inform work with both pedologic knowledge and quantitative approaches

Evaluating Access Constraints

* Considering Accessibility

Trails

Access Points

Areas of High Opportunity

Areas of Constraint

Buffered Areas

Cost Distance from Accessible Features

* Evaluating Areas of Opportunity

Assure features are representative of the whole

Implementing Sampling

* Need to assure sampling plan/strategy is operational; projects are operational not academic, so flexibility is imperative
* Need to maximize field time
* Limit time on transitional points while still capturing underrepresented classes
* Achieve balance between free survey and data space
  + Data driven vs knowledge-driven
  + Balance somewhere in between with a hybrid approach
* Not fear some level of bias
* Allow crew to learn as they go, and afford the opportunity to adjust plans on the fly
* Compressed schedules for data collection require us to be dynamic
* How to select points that most represent covariate feature space
* In implementation practitioners realized that covariate space of a training set could be represented with a key subset of the total training base
* Look for balance between knowledge and data driven methods, what we know and what we think we know
* Visually assessing data collected and proposed data in covariate space is useful

Flexible Sampling Design will post-poned until next meeting to allow Dylan Beaudette to spend more time on demonstration tools for multivariate similarity analysis.

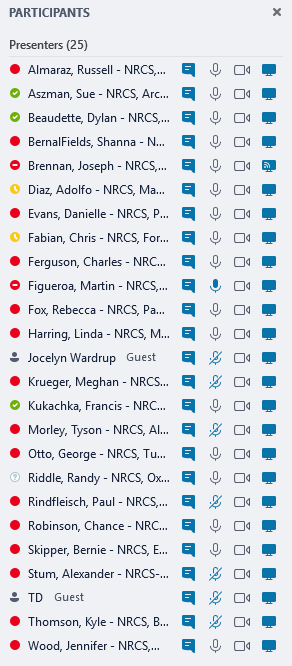
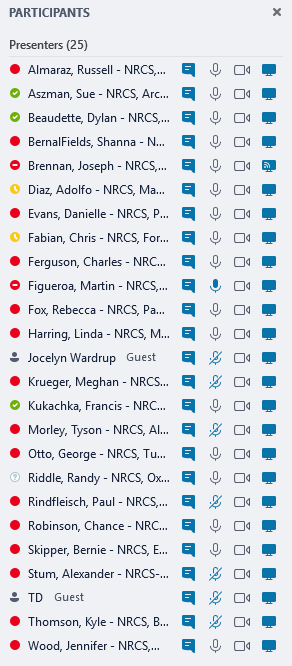
Developing Job Aids for Sampling and other components of digital soil mapping project work will be something the DSM Focus Team will be emphasizing soon and looking for contributors.

**DSM Practitioner’s Discussion**

**8/06/2019 – 11:00am to 12:25pm**

**This month’s topic:** DSM Update project design

**Roll Call (from Skype)**

****

**Introduction (Chance)**

1. Predictive modelling in soil survey
2. Scientific Method in Ag Handbook No. 18

**Making Digital Soil Mapping Operational: Defining the Modeling Domain (Chance)**

* Possible questions to ask in defining modelling domains/strata?
  + What is the landform pattern?
  + How many soil landscape units are differentiated within a single landform?
  + What is the order of the survey?
  + How many classes are to be modeled?

**Physiographic Regions/Soil Landscape Associations (Joe)**

* Cooperators had difficulty in mining parent material groups from SSURGO for [Soil Explorer](https://soilexplorer.net/)
* The NRCS MSSOs have the expert knowledge to effectively accomplish this task
* Developed SSURGO Based Map Unit Level Associations in conjunction with MLRA Legend Evaluations
* Hierarchical (1 to 3 levels based on complexity of MLRA)
* Developed through mining tabular data and expert knowledge
* Confined within MLRA and stored independent of NASIS
  + Text notes utilized to record information in NASIS

**Discussion**

* ***Tyson Morley*** comments about issues with the Soil Explorer in Texas
* ***Chris Fabian***
  + Economies of Scale
  + Many different geologic formations in foothills
  + Should each formation be a separate modelling domain, or should foothills be tackled as a single modelling domain?
  + Using gSSURGO is a helpful tool in project planning
* ***Dylan Beaudette***
  + Where to park simplified legend? – Parent Material Groups, etc.
    - Jay Skovlin has stored as specially formatted map unit text note.
  + Primary objection to POLARIS and other efforts is absence of stratification to model.
* Discussion of method for modelling across strata or modeling domains.
  + Dylan Beaudette, Colby Brungard, & others implementing modelling across strata
    - Develop modelling domains (i.e. MLRAs) then build and test predictive model of strata
    - Develop subset of soil forming factors that are general enough to describe differences between strata (modelling domains) (i.e. 5 to 10 indices of climate)
    - Using reduced set of covariates calculate distance between.
    - Create model to predict modelling domains
  + Possible future presentation from Dylan on topic of modelling across strata
* Perhaps in relation to discussion of modeling across strata
  + Anisotropy – ***Alex Stum***
    - [Soil Anisotropy](https://blogs.agu.org/terracentral/2013/07/01/soil-anisotropy-mechanisms-and-hydrologic-consequences/)

**Computer Configurations and Requirements**

* Bob, Alex, and Drew discussion regarding SBAG computing requirements. Working on spreadsheet for computers assigned to NRCS Soil Scientists
* With Windows 10 we now have major issues with network slowdowns and slower processing times
* Requests have been made for exemptions, but have not been approved by OCIO
* Alex shared various configurations which might be assigned depending on computing needs of soil scientist
* Dylan commented that if exemptions were approved everyone would see a tremendous improvement, but at present tasks that in the past took 10 seconds now might take 10 to 30 minutes due to network issues in the new Windows 10 computing environment.

**Closing Remarks**

Next call on September 3 as Suzann/Jess/Joe lead a deeper dive into sampling

**Practitioner’s Discussion**

**7/2/19**

Attendees:

1. Jess Philippe – R12 St. Johnsbury, VT
2. Suzann Kienast-Brown – R4 Bozeman, MT
3. Tom D’Avello – GRU, Morgantown, WV
4. Andrew Brown
5. Anthony C
6. Ben M
7. Bernie Skipper
8. Betsy Schug
9. Bob Kukachka
10. Brianna Wegner
11. Chad F
12. Dan Wing
13. Dave W
14. Duvall
15. George Otto
16. Greg S
17. Jamin
18. Jane
19. Jocelyn Wardrup
20. Joe B
21. Jennifer Wood
22. Katelyn
23. Kelley PL
24. Kyle Thompson
25. Linda Harring
26. Matt Bromley
27. Meghan Krueger
28. Nick Kozlowski
29. Phil Roberts
30. PRR
31. Rebecca Fox
32. T. Riebe
33. Chris
34. Alex Stum
35. DHK

* Suzann provided overview of covariate selection – importance in the DSM process and methods for selection
* Kyle Thompson
  + 2M acres of loamy till; flat; 6% slope on average, depressions, cropland;
  + 5m DEM
  + Relief derivatives (68 total)
  + Completed for two watersheds
  + Analyzed covariates for sampling/training data
  + Covariate reduction – Dave White
    - Scripts here: <https://github.com/ncss-tech/soil-pit/tree/master/sandbox/dave>
    - Near zero covariance filtering; correlation filter; IPCA
      * Calculate variance for each covariate; near zero gets removed
      * Correlation filter; correlation matrix then set threshold to remove highly correlated covariates
      * IPCA
        + PCA then computes loading factors; relates PCs to covariates with loading factors; set threshold ie. Keep everything that explains 95% of variance
      * 14-16 covariates for each watershed
      * List of covariates were different for each watershed
    - Random forests for recursive elimination (R caret package)
      * Training data
        + Summarize MU tabular data – 4 soil properties

Moisture, OM, CaCO3, erosion class

Using dominant component, weighted average to define classes

* + - * + Zonal stats for properties for polygons and covariates; mean for covariates used in training data
        + 10% of each project map unit
      * Picked covariates based on pedological knowledge and quantitative approach
  + Biggest struggle was defining classes based on summarizing MU data
  + Questions for Kyle:
    - How did your knowledge of soil/landscape relationships play a role in covariate selection?
      * Most made sense
    - What would you do differently?
      * Look at MU data he was summarizing; changing depth intervals for OM and CaCO3
      * But overall very pleased with the process and the results
    - Probably can be done in a week or so first time around
    - Use mean or median value for summarizing covariates using polygons and zonal stats
      * Mean – may or may not be a value that occurs in the data
      * Median – will be a real value in the data
      * Arguments made for both
        + Depends on the modeling approach you’re using
      * This probably deserves more discussion/thought to provide better guidance

**Practitioner’s Discussion**

**6/4/19**

Attendees:

1. Jess Philippe – R12 St. Johnsbury, VT
2. Suzann Kienast-Brown – R4 Bozeman, MT
3. Tom D’Avello – GRU, Morgantown, WV
4. Stephen Roecker
5. Phil Roberts
6. Rebecca Fox
7. Adolfo Diaz
8. Anthony C
9. Bernie Skipper
10. Carl
11. Chad
12. Chris Gebauer
13. George Otto
14. Jay Skovlin
15. Jocelyn Wardrup
16. Jonathan Diaz
17. Kurt
18. Matt Bromley
19. Alex Stum
20. Talyor C
21. Wade Bott
22. Wendy Pierce
23. Jennifer Wood
24. Tim Riebe
25. Jacky
26. Tyson Morley

* Sampling discussion
  + Tom – Salmon-Challis NF
    - cLHS design for field sampling
    - cost layer – binary classification – 0 and 10,000
    - 1/4 mile buffer from trails
    - 1/3 mile buffer from roads
    - cLHS point in high cost areas – similarity index in cLHS package is good but only for points that they can actually access
    - clustering based on covariate values extracted at high cost points to then use in SIE to create fuzzy membership layers to identify similar areas that are accessible
    - touch base with Colby/Dylan to potentially improve cLHS package
    - number of clusters could have been adjusted if more time available – 1hr each to create fuzzy memberships in SIE
    - data management issues with similarity index as separate layer for each point
    - Tom will write up job aid and share scripts
    - Suzann shared Carla’s observations/considerations in implementing sampling design in the field
    - Discussion
      * overall, use primary cLHS point as primary sampling location and cluster around to capture landform position variability; such as a traverse
      * if covariates are chosen carefully to represent soil forming factors, then cLHS points should represent variability in soil-landscape relationships and when coupled with cluster/traverse sampling and provide adequate information to develop map unit concepts and map unit descriptions
  + Jess – sampling design evolution in Essex and WMNF
    - Essex Stratified grid sampling: extremely time & labor intensive for little return (highly variable area and points did not capture the variability)
    - Essex Random Catena/stratified random sampling: combined targeted & random approaches and better captured the catena variability targeted in the model, but didn’t necessarily meet statistical rigor of randomness
    - Essex -With SSURGO as an end goal, some traditional transects (makes traditional project leaders feel better and did yield some good data)
    - WMNF – stratified random based on parent material and some target components. Rushed process with resource limitations that impacted the usefulness for training data.
    - WMNF moving forward – cLHS for each parent material model
    - Discussion
      * sampling design and the stage at which it happens can really affect things like map unit composition (example of a transect designed to document raster map unit that “goes away” when raster is processed for ssurgo)
      * Unlikely to be perfect so documentation is key
  + Alex
    - transects served as unbiased approach to sampling landscapes in past
    - cLHS looks at cumulative distribution to sample feature space
      * you can use to represent map unit concept
      * add hoc sampling will risk this
    - many models are sensitive to sample size and purposive sampling risks oversampling a single class and affecting modeling

**Initial Sub-Team Meeting**

**4/2/19**

Attendees:

1. Jess Philippe – R12 St. Johnsbury, VT
2. Suzann Kienast-Brown – R4 Bozeman, MT
3. Jay Skovlin
4. Aaron Wells
5. Bernie Skipper
6. David Rand
7. Frodo
8. Jocelyn Wardrup
9. Matt
10. Matt Dorman
11. Nicholas Kozlowski
12. Tom D’Avello
13. Portland
14. Betsy Schug
15. Stephanie Shoemaker
16. Jacky
17. Kelley Paup-Lefferts
18. Sam Streeter
19. Nick Kozlowski
20. Pat O’Connell
21. Chad
22. Jennifer Wood
23. Kristi Mingus
24. Tim Riebe
25. Brian
26. Ben

* Project design overview – Jess
  + Discussion on nationwide covariate stack to support projects – hopeful for wide use and access via Fort Worth in CY19
* Bob Marshall project overview – Jay Skovlin
  + Use existing information – USFS, NRCS
  + Soil forming factors
  + Stratify landscape based on landform/parent material groups
    - Will ask Sonora SSO to share script on DSM FT Github
  + Use strata in sampling design – cLHS
    - Gives different scale of information than geomorphons
    - Geomorphons may still be useful – more component level information

**Initial Sub-Team Meeting**

**2/5/19**

Attendees:

1. Jess Philippe – R12 St. Johnsbury, VT
2. Suzann Kienast-Brown – R4 Bozeman, MT
3. Alex Stum – R9 Temple, TX; awareness for initial; loves dsm
4. Beth Rowley – Missoula, MT; bob marshall wilderness; glacier np; updates
5. Betsy Schug – Fergus Falls, MN; update and initial with dsm; learn new things; lessons learned
6. Brian Gardner – Moscow, ID; curiosity; notcom forested areas
7. Chris Fabian – Fort Collins, CO; USFS orphans; update projects
8. Kari Sever – Fort Collins, CO; current correlation
9. Chris Gebauer – Klamath Falls, OR; large traditional initial/update; notcom w/dsm
10. Chris Savastio – Mindon, NV; notcom w/dsm
11. Danny Wood – Powell, WY; notcom w/dsm
12. Dave White – Las Cruces, NM; mostly update w/dsm
13. David Rand – Salem, OR; initial for NF, lots mapped, leverage existing knowledge w/dsm
14. Jacqueline Vega – Kealakekua, HI; initial; improve knowledge
15. Jocelyn Wardrup – grad student DE; expanding knowledge
16. Joe Brennan – R10 St. Paul, MN; mostly update, some initial; fed lands liaison
17. Josh Paul – Fairbanks, AK; LOTS of notcom
18. Megan Krueger – Vale, OR; BLM initial survey; using ArcSIE
19. Mike Regan – R1 Portland, OR; awareness
20. Phil Goodin – Powell, WY; notcom w/dsm
21. Russ Almarez – R2 Davis, CA; initial mapping focus team
22. Sam Streeter – Alamosa, CO; update w/dsm; learning for initial
23. Shauna Bernal-Fields – Ontario, OR; some dsm training; use it more
24. Sue Aszman – Arcata, CA; initial NF, BLM
25. Tim Riebe – R13 AK; LOTS of notcom
26. Andy Oxford – Pierre, SD; assisting AK with Kodiak project
27. Bernie Skipper – DU; extracting linework from photography – digital
28. Bob Kukachka – ID; retiree; NF work; awareness
29. Whityn Owen – OR State GIS; awareness

* DSM Focus Team overview – Suzann
* Getting started with DSM – Jess
  + Visit the DSM focus team website – it’s all there!
    - Training curriculum
    - Available resources
  + Essex county project
  + White Mountain project
* Q/A
  + Properties
    - NASIS data? Yes, but w/QC. Project will start with KSSL, RCA data
    - Official data? Differing opinions – moving toward NRCS providing multiple data options – more discussion to come on this topic
      * Feel free to communicate with Suzann or leadership if you have opinions, concerns, etc to share
  + Time savings with DSM?
    - Depends really. Every project is different in scope.
    - Efficiencies are created in targeted field sampling, consistency, and ability to extrapolate across project area from smaller sampling area, if soil-landscape relationships are the same.
    - Initial investment and learning curve, especially for the first project. Like everything, efficiencies will develop over time.
    - More flexible product (raster-based) for users in the end may be one of the biggest gains for employing DSM.