MAP UNIT SUMMARY REPORT- PART 1 - BACKGROUND

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

PART 1 – OVERVIEW (Part 2 in a separate file titled "Map Unit Summary Report Part 2 – Instructions")

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 has been used to display the results in a user-friendly format.

Justification for a new map unit summary method and report

There are several reasons for transitioning to a new environmental property summary method and report format. Some of the more important reasons include: software upgrade issues, standardized statistical approaches, and access to new options for the analysis and display of data.

Software upgrade issues

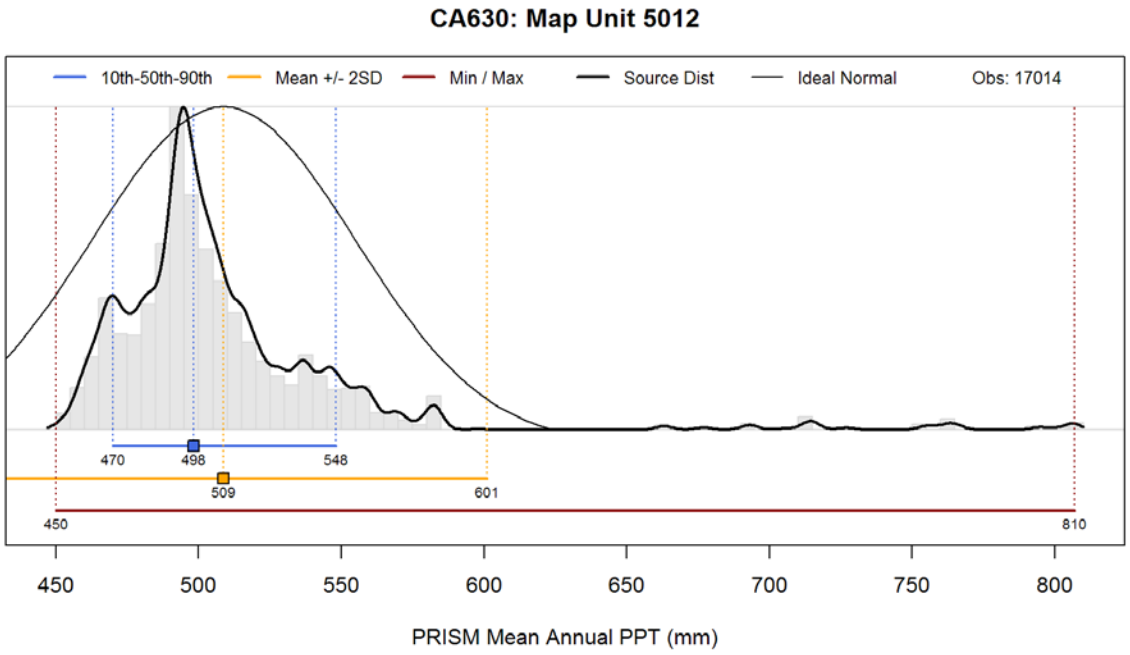
Lucas Wisely is not actively maintaining the scripts for this model for ArcGIS version updates. Because ArcGIS regularly creates new versions, and not all offices get the updates at the same item, the maintenance of the tools is cumbersome. Additionally, the Crystal Report software also has upgrade issues as well as now requiring the agency to purchase a copy of the software for each computer that will be using it.

Transition to standardized statistical approaches

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 vaguely 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. MLRA offices that are creating MLRA map units as a result of SDJR projects are generally 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.

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. Here is an example figure from that discussion 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 is from a related proposal by Tom D'Avello for edits to the NSSH 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."

And, this is from the NRCS National Water and Climate Center website:

"What is the median and how is it different from the average? Although average is a commonly-used and well understood statistic, median is also a common descriptor used to express a "middle" value in a set of data. This "middle" value is also known as the central tendency. Median is determined by ranking the data from largest to smallest, and then identifying the middle so that there are an equal number of data values larger and smaller than it is. While the average and median can be the same or nearly the same, they are different if more of the data values are clustered toward one end of their range and/or if there are a few extreme values. In statistical terminology, this is called skewness. In this case, the average can be significantly influenced by the few values, making it not very representative of the majority of the values in the data set. Under these circumstances, median gives a better representation of central tendency than average."

In the approach proposed here, the median, or 50th percentile, is generated, as well as the 5th, 10th, 25th, 75th, 90th, and 95th percentiles. The reports can also be edited to return the min, max, and any other percentile value that is desired.

New options for the analysis and display of environmental data are now available

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 – see Figure 1 in this document and Appendix 3 in the document titled "Map Unit Summary Report Part 2 – Instructions".

	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**	**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 **–** for complete instructions see "Map Unit Summary Report Part 2 – Instructions"

Setup

The mapunit summary report is provided as an R Markdown document (.Rmd file extension), configuration file (.R file extension), and basic documentation. For users who are completely unfamiliar with the R Studio environment, there is a basic R studio tutorial provided in Appendix 1 in the document titled "Map Unit Summary Report Part 2 – Instructions**".** The .Rmd document is opened in R Studio and the user is directed to ensure that the rasters required for analysis are in a specified location (defined in the file "config.R").

Inputs

Two kinds of inputs are required. The first is a shape file (or ESRI file 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. The set of raster data can be customized (regional 30 meter products, local 10 meter products, etc.) according to individual needs. Raster files need not have the same coordinate system, extent, or grid size.

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 to be analyzed. 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. For an explanation of why the raster pixels are sampled rather than using the whole population, please refer to the document linked at this site: http://ncss-tech.github.io/AQP/sharpshootR/sample-vs-population.html

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.

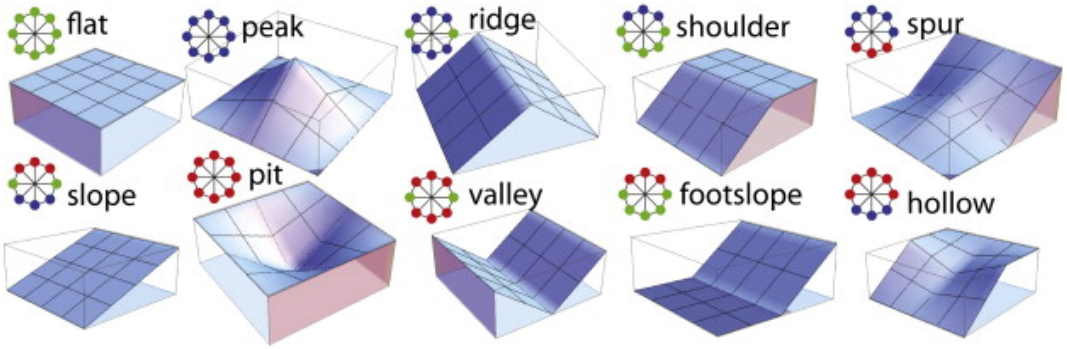
Output displayed:

- *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, 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 of Aspect:* 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 strongly directional pattern. Note: Summary of aspect data from landscapes with slopes < 3% are invalid.

- *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. This is a scale-independent method, in contrast to the classification of curvature approach. . Landform types are *approximately* correlated to standard slope positions and curvature classes:

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).

Caveats for use of the tool

1. The output generated is for the entire map unit, not the individual component. The individual data developer must evaluate the type of landscape being described and whether the output for the entire

map unit can be used to represent the component or if the component is represented by a subset of the data.

2. The output is dependent on the spatial delineation of the landscape in the geodatabase. Depending on the scale and tools available, the polygons may encompass areas that are meant to be associated with adjacent map unit concepts. The data developer must understand the source, scale, and accuracy of the linework that defines the data set.
3. The data developer must understand the source and limitations of the data set being summarized. All geospatial data is generated by a model of some kind. The data developer must have an understanding of the algorithms used to generate the input data sets.
4. This tool is not meant to generate values to automatically populate values in NASIS. All the previous caveats must be understood before using the output values to represent the component and map unit concepts to the users of the data.

Notes on Methods

Curvature classification

The default data used by this report are based on a regional 10m resolution, integer DEM. Curvatures were calculated using a 5x5 moving window (Wood, 1996) and classified using a +/- 0.0001 m^{-1} threshold. A more accurate depiction of slope shape can be generated from 10m (or finer resolution) data when available. Window size and curvature class thresholds should be determined based on expert knowledge of local terrain and mapping scale. If you would like to use your own curvature data, be sure re-code curvature values using the standard notation of down-slope/across-slope as:

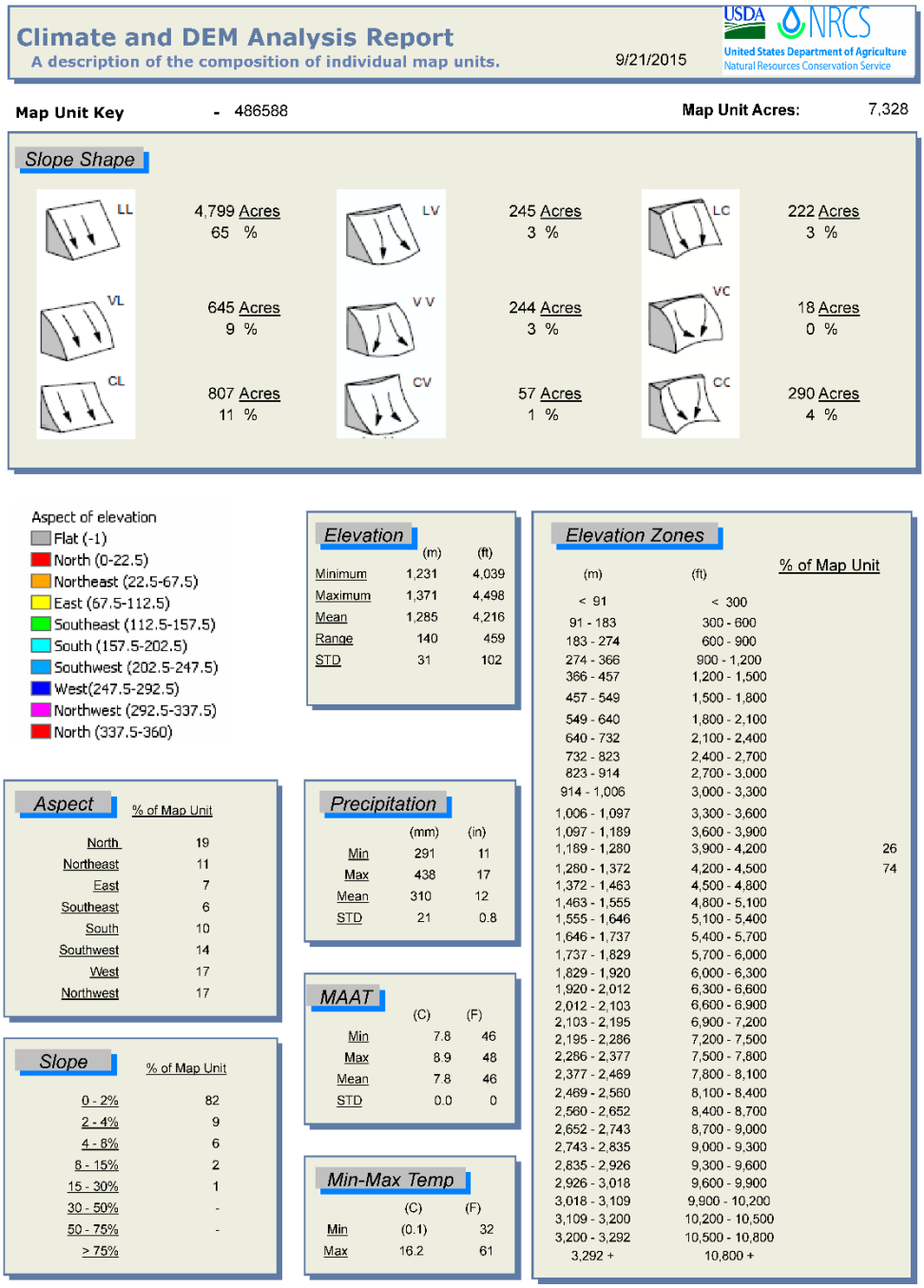
curvature classes			coded raster values		
L/L	L/V	L/C	22	32	12
V/L	V/V	V/C	23	33	13
C/L	C/V	C/C	21	31	11

Colors in the report are based on the approximate correlation of surface shape to moisture redistribution:

shedding positions ----------> accumulating positions
V/V, L/V, V/L, C/V, LL, C/L, V/C, L/C, C/C

Future Report Output and Functionality

Please contact us if there are other kinds of output that you need or would find useful.



Climate and DEM Analysis Report

A description of the composition of individual map units.

9/21/2015

USDA NRCS United States Department of Agriculture Natural Resources Conservation Service

Map Unit Key - 486588 **Map Unit Acres:** 7,328

Slope Shape

LL	4,799 Acres 65 %	LV	245 Acres 3 %	LC	222 Acres 3 %
VL	645 Acres 9 %	VV	244 Acres 3 %	VC	18 Acres 0 %
CL	807 Acres 11 %	CV	57 Acres 1 %	CC	290 Acres 4 %

Aspect of elevation
- Flat (-1)
- North (0-22.5)
- Northeast (22.5-67.5)
- East (67.5-112.5)
- Southeast (112.5-157.5)
- South (157.5-202.5)
- Southwest (202.5-247.5)
- West(247.5-292.5)
- Northwest (292.5-337.5)
- North (337.5-360)

Elevation

	(m)	(ft)
Minimum	1,231	4,039
Maximum	1,371	4,498
Mean	1,285	4,216
Range	140	459
STD	31	102

Elevation Zones

(m)	(ft)	% of Map Unit
< 91	< 300	
91 - 183	300 - 600	
183 - 274	600 - 900	
274 - 366	900 - 1,200	
366 - 457	1,200 - 1,500	
457 - 549	1,500 - 1,800	
549 - 640	1,800 - 2,100	
640 - 732	2,100 - 2,400	
732 - 823	2,400 - 2,700	
823 - 914	2,700 - 3,000	
914 - 1,006	3,000 - 3,300	
1,006 - 1,097	3,300 - 3,600	
1,097 - 1,189	3,600 - 3,900	
1,189 - 1,280	3,900 - 4,200	26
1,280 - 1,372	4,200 - 4,500	74
1,372 - 1,463	4,500 - 4,800	
1,463 - 1,555	4,800 - 5,100	
1,555 - 1,646	5,100 - 5,400	
1,646 - 1,737	5,400 - 5,700	
1,737 - 1,829	5,700 - 6,000	
1,829 - 1,920	6,000 - 6,300	
1,920 - 2,012	6,300 - 6,600	
2,012 - 2,103	6,600 - 6,900	
2,103 - 2,195	6,900 - 7,200	
2,195 - 2,286	7,200 - 7,500	
2,286 - 2,377	7,500 - 7,800	
2,377 - 2,469	7,800 - 8,100	
2,469 - 2,560	8,100 - 8,400	
2,560 - 2,652	8,400 - 8,700	
2,652 - 2,743	8,700 - 9,000	
2,743 - 2,835	9,000 - 9,300	
2,835 - 2,926	9,300 - 9,600	
2,926 - 3,018	9,600 - 9,900	
3,018 - 3,109	9,900 - 10,200	
3,109 - 3,200	10,200 - 10,500	
3,200 - 3,292	10,500 - 10,800	
3,292 +	10,800 +	

Aspect

	% of Map Unit
North	19
Northeast	11
East	7
Southeast	6
South	10
Southwest	14
West	17
Northwest	17

Precipitation

	(mm)	(in)
Min	291	11
Max	438	17
Mean	310	12
STD	21	0.8

MAAT

	(C)	(F)
Min	7.8	46
Max	8.9	48
Mean	7.8	46
STD	0.0	0

Slope

	% of Map Unit
0 - 2%	82
2 - 4%	9
4 - 8%	6
8 - 15%	2
15 - 30%	1
30 - 50%	-
50 - 75%	-
> 75%	

Min-Max Temp

	(C)	(F)
Min	(0.1)	32
Max	16.2	61

Figure 1: Example output from the "Lucas model" reports.