Soil Property Distribution Models from Point Data for Soil Survey

Howell, D [a]., Y. Kim [b], C. Haydu-Houdeshell [c], P. Clemmer [d], R. Almaraz [e],

[a] U.S.D.A. Natural Resources Conservation Service, Arcata, CA, U.S.A.
[b] Humboldt State University, Arcata, CA, U.S.A.
[c] U.S.D.A. Natural Resources Conservation Service, Victorville, CA, U.S.A.
[d] U.S.D.I. Bureau of Land Management, Denver, CO, U.S.A.
[e] U.S.D.A. Natural Resources Conservation Service, Lancaster, CA, U.S.A.

Abstract

Models estimating the distribution of soil properties were developed from soil profile descriptions and GIS landscape analysis to assist soil scientists with soil-landscape information prior to the completion of field soil investigations. Soil profiles and landscape features were described at 97 randomly located field sites within a 30,424 ha project area in the Mojave Desert. Explanatory variable information was also developed for each of these sites through GIS extraction from digital elevation model data, landform derivatives, band-ratio satellite images, and geomorphologic data. The models estimated selected soil characteristics continuously in a 30 m raster over the project area. The response variables that we modeled were soil genetic features that are used as diagnostic properties in Soil Taxonomy (e.g., presence or absence of argillic horizon).

1. Introduction

In the western United States the spatial and attribute resolution of digital soil-forming factor data is still quite coarse. The explicit relationships between these explanatory variables and the resulting individual soil properties is not well understood in most areas. Despite several decades of worldwide research and development of GIS soil modeling methods, the outputs from these models is rudimentary information.

The spatial resolution of input data is in the order of 10-30 m, while soil variation occurs at a scale as fine as 0.5 m. In addition, some of these data have been converted to raster format from large polygons, e.g., geology, which may to lead to an incorrect assessment of resolution. Attribute resolution is also out of sync, e.g., geologic attributes describe entire formations rather than individual rock types. It is beyond the resolution of the input soil-forming factor data and our explicit understanding of soil-forming relationships to attempt to create detailed, taxonomic (or even multi-property) soil maps directly from explanatory variable data. At this stage, we feel the appropriate goal for GIS soil-landscape modeling should be to produce maps showing estimates of important individual soil genetic features in order to increase understanding of soil-landscape relationships and to guide field data collection by soil scientists working on soil survey projects.

In this project we have attempted to develop models to estimate the spatial distribution of individual soil genetic features. These features are defined by objective criteria in Soil Taxonomy (Soil Survey Staff, 1999). These genetic features are commonly used to classify soil pedons using a soil taxonomic system. These genetic features also serve as markers in the stages of development, or genesis, of soils. We have made no attempt to model these genetic features as they would be combined in a taxonomic system. We feel that these individual genetic features are important measurable soil properties that influence soil use and management. Also, the relationships between each individual genetic feature and the soil-forming factors will be more direct than developing one relationship to model all of the soil properties together. Our models allow each feature to vary independently and continuously (at separate scales of variation) across the landscape as described by McKenzie, et al. (2000).

Soil survey depends on developing relationships between the soil-forming environment and the resulting soil properties. Dokuchaiev (Glinka, 1927), Hilgard (1914) and Jenny (1941) spoke of relationships between the soil-forming factors and soil properties. It is our interpretation that they were speaking specifically of soil properties rather than taxonomic classes based on combinations of soil properties. Jenny in particular spoke of developing quantitative relationships. Modeling soil properties directly seems more appropriate than modeling taxonomic classes when using quantitative statistical models based on physical soil-forming processes. There have been many papers published on the use of GIS and statistical inference used to develop these relationships with digital spatial data. We will not attempt to refer to other specific work except for the recent complete review and framework proposed by McBratney et al. (2003).

Our focus was on the development of GIS tools for production soil survey, not research. Our statistical modeling methods draw heavily on the work of others (Gessler et al., 1995)(McKenzie and Austin, 1993)(McKenzie and Ryan, 1999)(Webster and Burrough, 1972).

2. Materials and Methods

2.1 Study Area

The study site is located in the western Mojave Desert approximately 100 miles northeast of Los Angeles, California, U.S.A. The study site receives 76 to 127 millimeters of rain per year with the majority falling between November and March. Summer precipitation is common after convection storms. Elevation ranges from 180 to 1425 meters. Sampling sites were located on broad alluvial fans and associated landforms. Soil development varied from young soils with little or no soil development (Typic Torriorthents) to soils consisting of older well-developed fan remnants underlying younger, more recently deposited alluvial material (Argidic Argidurids). Vegetation communities are dominated by arid climate shrubs such as *Larrea tridentata* (Sessé & Moc. ex DC.) Coville (creosote bush) and *Ambrosia dumosa* (Gray) Payne (white bursage) with *Yucca schidigera* Roezl ex Ortgies (Mojave yucca) and *Yucca brevifolia* Engelm. (Joshua tree) occurring in some areas (U.S.D.A., 2004).

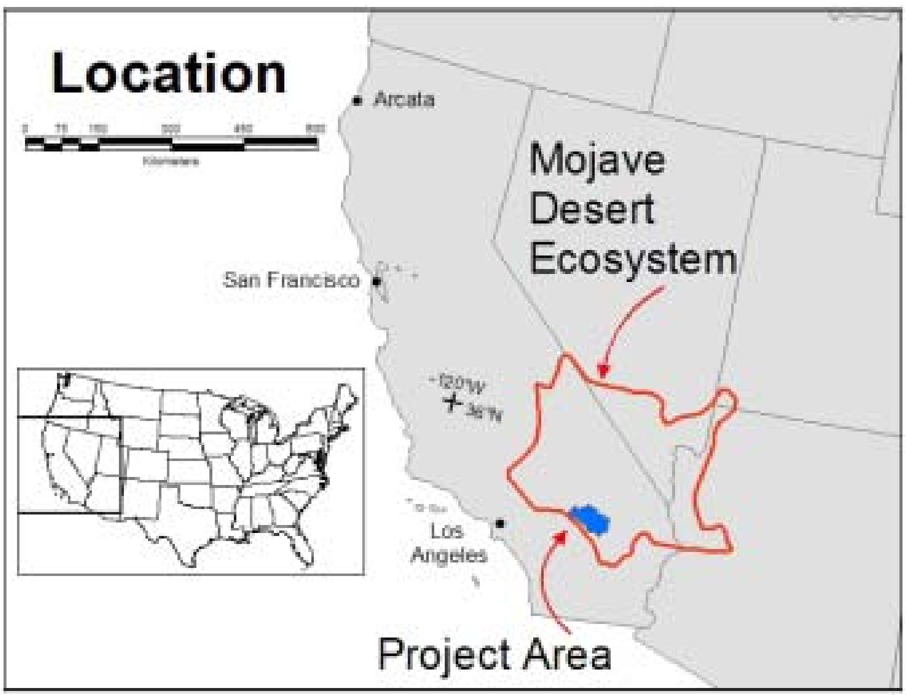

Figure 1 Location of the Project Area in the western United States

2.2 Attribute selection

The soil attributes we modeled were soil genetic features such as: presence or absence of argillic horizon, secondary carbonates, calcic horizon, durinodes, duripan, and separate (continuous) models estimating the depth to the occurrence of these features. We also estimated particle-size class.

The resolution of the model spatial input and output data is 30 m. The entire study area is approximately 288,000 ha.

2.3 Digital Spatial Data

The data layers used to represent the soil-forming environment were DEM and derivatives, band-ratioed Landsat Thematic Mapper (TM) imagery (Clemmer, 2003), and geomorphology (U.S. Army Topographic Engineering Center and Louisiana State University, 2000). See Table 1.

Table 1
Dependent and Independent Variables

Variable Names	Description [1]
Dependent	
Argillic	Argillic (clay accumulation) horizon in the soil: Yes=present, No=absent.
Argillic Depth	Depth to the top of argillic horizon
Calcic	Calcic (carbonate accumulation) horizon: Yes=present, No=absent.
Calcic Depth	Depth to the top of calcic horizon
Carbonates	Secondary carbonates: Yes=present, No=absent.
Carbonate Depth	Depth to the top of accumulation of secondary carbonates
Durinodes	Durinodes (silica masses): Yes=present, No=absent.
Durinode Depth	Depth to the top of accumulation of durinodes
Duripan	Duripan (silica cemented layer): Yes=present, No=absent.
Duripan Depth	Depth to the top of duripan
Taxpartsize	Particle-size class: 30, 33, 40, 44, 46, 50, 54, 59, 63, 69.
Independent	
Gisaspect	Slope direction: -1 to 360 DEM derivative
Giselev	Elevation above sea level (in meters) DEM derivative
Gisplan	Plan slope curvature (across the slope) DEM derivative
Gisprof	Profile slope curvature (up and down the slope) DEM derivative
Gisshape	Compound slope shape class (*categorical*) DEM derivative reclassification 9 classes for combinations of concave, linear, and convex
Gisslope	Slope steepness in percent DEM derivative
Ratio_band1	Reflectance value for band 1 TM Band Ratio Band 3/Band 2
Ratio_band2	Reflectance value for band 2 TM Band Ratio Band 3/Band 7
Ratio_band3	Reflectance value for band 3 TM Band Ratio Band 5/Band 7
landform1	Geomorphic landform general (*categorical*)

[1] See Soil Taxonomy (Soil Survey Staff, 1999) for soil definitions.

The compound slope shape data were derived from DEM data. The plan and profile curvatures were calculated as floating point numbers. These curvature numbers were evaluated against a digital raster graphic topographic map to assign these values to three classes each. Plan curvature was reclassed to concave, linear, and convex based on a subjective comparison to the contour lines. The same process was carried out for the profile curvature. The three shape classes for each direction were added together to form nine possible classes of compound curvature.

The soil enhancement band ratio product was processed using Landsat Thematic Mapper imagery acquired in August of 1993. There was some cloud contamination of the image, however this only appeared to effect the results in localized areas. The ratio composite was developed from research conducted previously in arid areas in Utah by the U.S.D.I. Bureau of Land Management, National Science and Technology Center.

Although extensive accuracy assessment has not as yet been accomplished, soil scientists in the Utah studies found this product to be useful in delineating and pre-mapping soil polygons. The product, along with other ancillary data, helped to plan field sampling, find discrete changes in soil make-up, and was very useful in helping to map remote and more inaccessible areas in difficult terrain. Although vegetation is directly linked to soil type and setting, this methodology appears to be most useful in arid areas where there is little interference from vegetation canopies and where more bare soil is exposed.

The indexing ratio uses bands 2,3,5, and 7 of the image and is usually displayed in the following color gun assignments: (Red) 3/2, (Green) 3/7, (Blue) 5/7. In this project the resulting digital number at each pixel was used as the value item in the models.

In the Utah studies, the 3/2 component was indicative of carbonate radicals (e.g., caliche, limestone); the 3/7 component seemed to indicate ferrous iron; while the 5/7 component was indicative of hydroxyl radical (e.g., clay).

The geomorphology and earth material data have been developed for the entire Mojave Desert region. This will form an important consistent data layer for modeling in this area. In some areas it appears to wander a little from the apparent landforms. It was developed at a smaller scale (1:100,000) than we are using it at for soil survey work (1:24,000). The models derived from these data have the same apparent misplacements in some areas.

2.3 Point Data

Three sets of soil point data were obtained from the project area. Field soil profile descriptions were used to characterize soil properties at each location. In order to simplify the models all sample points on mountain landforms were excluded and the models were not estimated for those areas.

Soil profiles were described at 97 randomized locations (randomly generated UTM coordinates) within a 30,424 ha portion of an ongoing soil survey project. The data from these randomly located points were originally used to fit models. We refer to these as random data points.

Profile descriptions were subsequently obtained from the ongoing soil survey (n=313) (Haydu-Houdeshell, 2004) and another set (n=392) (Lato, 2002) was obtained from an adjacent, recently completed project. For some soil properties these were combined into a larger data set of 656 purposive sample points, after some points were eliminated due to missing data. The locations for these data were selected in a traditional manner by the judgment of the soil scientists to represent particular sets of soil-forming factors. These purposively located sample points are sometimes called judgment samples. We refer to these as purposive data points.

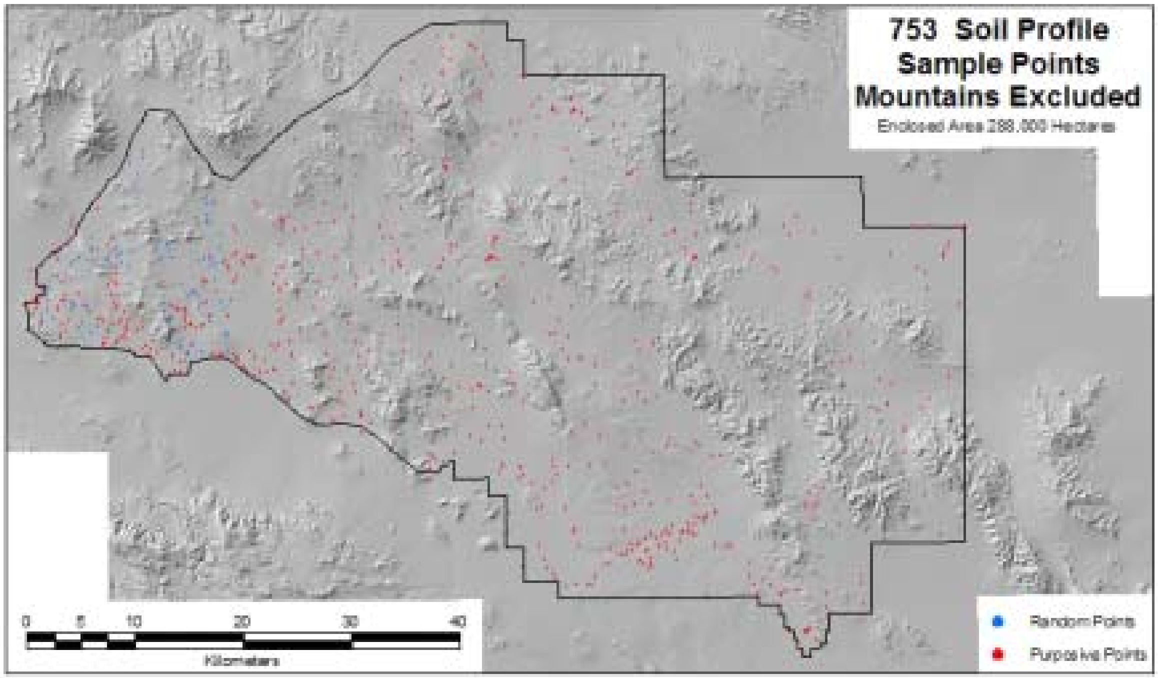

Figure 2 Location of Soil Profile Sample Points

The project area is sparsely vegetated and access is relatively unimpaired in most areas. We feel that these purposive samples represent the range of the soil-forming factors and that sample location bias will be low. Although this bias is not measurable. We wanted to see if it was possible to use these purposive data points to fit models for the entire project area. We wanted to make use of these extensive data. We refer to these models as models fit to the purposive data. We compared these to the models fit to the random data points.

We also used the UTM location coordinates of the random data points to extract the estimated soil property values from the models fit to the purposive data. A comparison was made of these extracted purposive model estimates to the actual measured values at the random data point locations (Environmental Systems Research Institute, 1998).

A range of methods was applied to develop optimal models. For continuous dependent variables, such as depth from surface to a feature, generalized linear models were used after thorough investigation of optimal Box-Cox transformations on variables and multicollinearity structures among variables. We then compared performance of models on randomly collected data set and purposively collected data set via maps, graphs, and summary tables. Various model selection criteria and diagnostic measures were used for these comparisons. We also compared the performance of logistic models on the presence-absence variables after transformation and multicollinearity checks.

3. Results and Discussion

The resulting models and significant terms are listed below. Box-Cox transformation routines determined that the square root transformation for gisslope, carbonatesdept and calcicdept is optimal. We found that model fitting, model assumptions, and various diagnostics are better after the transformation.

Table 2 Models and Significant Terms

GLM Models

	Overall *F*	R^2	Significant terms
<u>Summary of GLM models for the ***randomly*** collected data set</u>			
Model 1: $\sqrt{\text{carbnates\_depth}}$	2.09**	29.8%	Giselev**, Gisplan**, Gisprof**, Ratio_band3**, Gisshape***, Landform1**
Model 2: durinodes-depth	1.02	45.4%	Ratio_band2*
Model 3: argillic-depth	1.38	35.9%	*None*
Model 4: $\sqrt{\text{calcic\_depth}}$	0.80	25.2%	*None*
Model 5: taxpartsize	1.81**	26.6%	Ratio_band1**
<u>Summary of GLM models for the ***purposively*** collected data set</u>			
Model 1: $\sqrt{\text{carbnates\_depth}}$	2.13***	19.3%	Giselev***, Ratio_band1***
Model 2: durinodes-depth	2.03**	28.1%	Ratio_band2**, Gisshape***
Model 3: argillic-depth	1.99**	25.9%	Giselev*, Ratio_band2**, Ratio_band3*, Landform1**
Model 4: $\sqrt{\text{calcic\_depth}}$	2.58***	37.4%	Gisshape**, Landform1***
Model 5: taxpartsize (*n*=656)	14.13***	36.9%	Giselev***, $\sqrt{\text{Gisslope}}$***, Ratio_band1***, Landform1***

Logistic Models

	Overall χ^2	% Concordant	Significant terms
<u>Summary of logistic models for the ***randomly*** collected data set</u>			
Model 1: calcic	12.08	69.2%	Gisprof*, Ratio_band1*
Model 2: argillic	17.93**	72.5%	Giselev*, Ratio_band1**
Model 3: duripan	15.73**	78.9%	Gisplan**, Ratio_band1**
Model 4: durinodes	21.16***	77.5%	Giselev**, Ratio_band3*
Model 5: carbonates	10.33	80.5%	Ratio_band2 **

(*continued on next page*)

Table 2 (*continued*) Logistic Models

	Overall χ^2	% Concordant	Significant terms
Summary of logistic models for the ***purposively*** collected data set			
Model 1: calcic (*n*=656)	51.76^{***}	67.6%	Giselev^{***}, Gisplan^{*}, Ratio_band1^{*}, Ratio_band3^{**}
Model 2: argillic (*n*=656)	51.54^{***}	66.4%	Gisaspect^{*}, Giselev^{***}, Gisprof^{*}, Ratio_band1^{***}, Ratio_band2^{**}
Model 3: duripan (*n*=656)	31.06^{***}	65.9%	Gisplan^{*}, Gisprof^{*}, $\sqrt{\text{Gisslope}}^{***}$, Ratio_band1^{***}, Ratio_band2^{**}
Model 4: durinodes	33.82^{***}	68.9%	Gisplan^{***}, Gisprof^{**}, Ratio_band1^{***}, Ratio_band2^{**}
Model 5: carbonates	11.86	62.3%	$\sqrt{\text{Gisslope}}^{***}$

*** *p*-value < 0.01
** *p*-value < 0.05
* *p*-value < 0.1

Table 3 Comparison of Model Estimates to Actual Measured Soil Properties

				Number of Classes Estimate Missed By					
Model Fit on Data Points	*n*	Compared to Actual Values at Points	Correct Class %	1 %	2 %	3 %	4 %	5 %	6 %
Particle-size Class									
Random	97	Random	31	40	26	1	2	0	0
Purposive	97	Random	27	43	25	3	2	0	0
Purposive	656	Purposive	33	37	22	5	3	0.3	0.2

(*continued*)

Table 3 (*continued*)

Model Fit on Data Points	*n*	Compared to Actual Values at Points	% Estimates within 0 to 10 cm	% Estimates within 10 to 20 cm	% Estimates within 20 to 30 cm	% Estimates within 30 to 40 cm	% Estimates within >40 cm
<u>Argillic Depth</u>							
Purposive	46	Random	24	11	39	17	9
Purposive	116	Purposive	31	16	22	14	18
<u>Calcic Depth</u>							
Purposive	39	Random	18	39	26	15	3
Purposive	105	Purposive	33	26	13	10	17
<u>Carbonate Depth</u>							
Purposive	97	Random	39	33	15	4	8
Purposive	219	Purposive	39	35	12	3	12

These comparisons are made to the point data used for fitting a particular model and to the random point data.

The particle-size class comparison is for class assignment. The model estimate was assigned to the nearest class. The classes are indicated by a number code used in the National Soil Information System (U.S.D.A., 2004). The classes are: 30 sandy-skeletal, 33 loamy-skeletal, 40 sandy, 44 loamy, 46 coarse-loamy, 50 coarse-silty, 54 fine-loamy, 59 fine-silty, 63 clayey, and 69 fine. These class numbers are ordinal. Low numbers are coarse and high numbers are fine textures. The model assigned class was correct or within one class for approximately 70% of the sample points. This is true for both types of models and for comparisons to both types of sample points. The estimates were within two classes of the correct class for 92-97% of the sample points over an area of 288,000 ha (~712,000 acres). See Figure 3 for a map displaying the output from the purposive data model.

The comparisons for continuous estimates of depth to a certain genetic feature show a range of model performance. The model for depth to secondary carbonates performed best. The estimates were within 20 cm of actual measured values for 72% of the random sample points and 74% of the purposive sample points. The model estimates for calcic horizon depth were within 20 cm for 57% of the random points and 59% of the purposive points. The estimates for the depth to an argillic horizon were not as reliable.

The models based on logistic regression for the presence or absence of features are harder to evaluate. The model for probability of argillic horizon performed the best. The model fit to the random point data was much more sensitive, more accurate, and showed a greater range of estimated values than the model fit to the purposive data, even though there were many more purposive data points representing each level for each explanatory variable. Graphs of predicted versus actual values for each of the models showed a trend of agreement to actual values. Space limitations do not allow us to show graphs and maps displaying these results. More work is needed to improve these models.

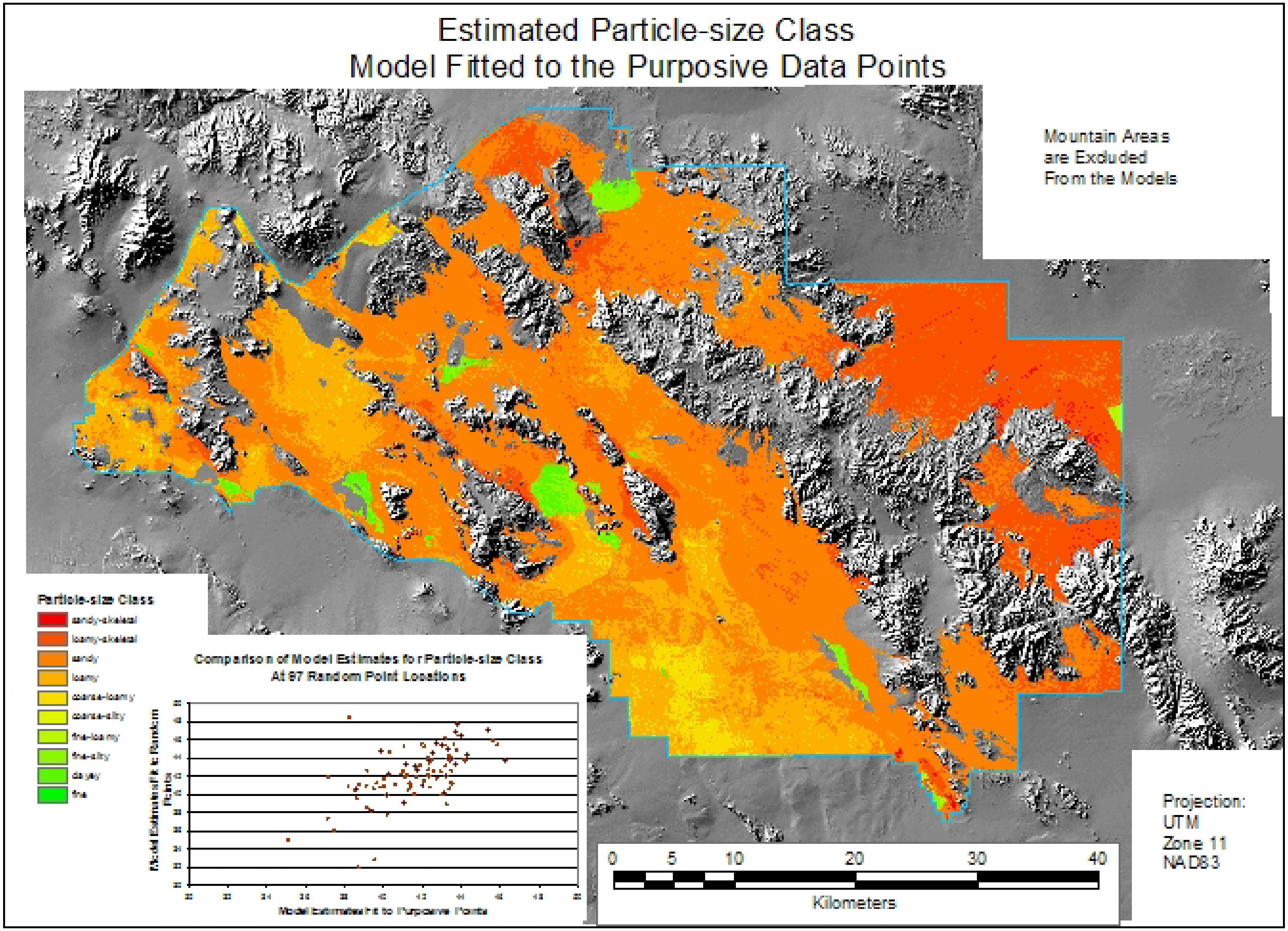

Figure 3 Model Estimate of Particle-size Class

Future work will focus on increasing the number of soil properties evaluated for the larger purposive point data set. We will also look into combining the binomial models (presence/absence) of features with the estimates of depth for those features, e.g., to produce a map estimating areas with 50% probability of the presence of an argillic horizon with estimates of depth in those areas. We also hope to improve the models for depth estimates so that several soil feature surfaces can be visualized in 3-dimensional perspective view draped over a land surface. These combinations of continuous estimations of soil features could form new soil survey products, when the models perform better. In this study each dependent variable was modeled separately, but future study needs to include modeling taking multiple dependent variables into account, i.e., from multivariate point of view.

Acknowledgements

This work was accomplished through the hard work of field soil scientists in difficult, remote conditions (who collected the 'real data') and by a statistician, a GIS specialist/soil scientist, and a remote sensing specialist in comparatively cozy offices. We appreciate the support of Dave Smith, State Soil Scientist, U.S.D.A. Natural Resources Conservation Service, California.

References

Clemmer, P. 2003. Band-ratio Landsat 5 Thematic Mapper Imagery. Personal communication. U.S.D.I. Bureau of Land Management.

Environmental Systems Research Institute. 1998. GridSpot70. [Online] Available: http://arcscripts.esri.com/details.asp?dbid=11037

Gessler, P. E., I. D. Moore, N. J. McKenzie, and P. J. Ryan. 1995. Soil-landscape modelling and spatial prediction of soil attributes. International Journal of Geographical Information Systems 9(4):421-432.

Glinka, K. D. 1927. Dokuchaiev's ideas in the development of pedology and cognate sciences. U.S.S.R. Academy of Science. Russian Pedological Investigations, I.

Haydu-Houdeshell, C. 2003. Soil profile descriptions Johnson Valley Off Highway Vehicle Area Soil Survey Project. Personal Communication. U.S.D.A. Natural Resources Conservation Service.

Hilgard, E. W. 1914. Soils. The McMillan Company, New York.

Jenny, H. 1941. Factors of soil formation. A system of quantitative pedology. McGraw-Hill Book Company, Inc., New York.

Lato, L. 2002. Soil profile descriptions Marine Corps Air Ground Combat Center at Twentynine Palms Soil Survey. Personal Communication. U.S.D.A. Natural Resources Conservation Service.

McBratney, A. B., M. L. Mendonca Santos, and B. Minasny. 2003. On digital soil mapping. Geoderma 117:3-52.

McKenzie, N. J. and M. P. Austin. 1993. A quantitative Australian approach to medium and small scale surveys based on soil stratigraphy and environmental correlation. Geoderma 57:329-355.

McKenzie, N. J., H. P. Cresswell, and M. Grundy. 2000. Contemporary land resource survey requires improvements in direct soil measurement. Communications in Soil Science and Plant Analysis 31(11-14):1553-1569.

McKenzie, N. J. and P. J. Ryan. 1999. Spatial prediction of soil properties using environmental correlation. Geoderma 89:67-94.

McSweeney, K., B. K. Slater, R. D. Hammer, J. C. Bell, P. E. Gessler, and G. W. Petersen. 1994. Towards a new framework for modeling the soil-landscape continuum. p. 127-145. *In* R. Amundson, J. Harden, and M. Singer (editors) Proceedings of factors of soil formation: a fiftieth anniversary retrospective symposium. Denver, Colorado. 28 October 1991. Soil Science Society of America Special Publication Number 33, Madison, WI.

SAS statistical software Release 8.02, SAS Institute Inc., Cary, NC, 1999-2001.

Soil Survey Staff. 1999. Soil Taxonomy, Second Edition, Agriculture Handbook No. 436. United States Department of Agriculture, Soil Conservation Service. Washington, D.C.

U.S. Army Topographic Engineering Center and Louisiana State University. 2000. Earth Materials and Landform Mapping Project. [Online] Available: http://www.mojavedata.gov/datasets.php?qclass=geo.

U.S.D.A. Natural Resources Conservation Service. 2004. The PLANTS Database, Version 3.5 [Online] Available: http://plants.usda.gov [June 4, 2004]. National Plant Data Center, Baton Rouge, LA 70874-4490 USA.

U.S.D.A. Natural Resources Conservation Service. 2004. National Soil Information System (NASIS) Choice list report. [Online] Available: http://nasis.nrcs.usda.gov/documents/

Webster, R. and P. A. Burrough. 1972. Computer-based soil mapping of small areas from sample data. I. Multivariate classification and ordination. II. Classification and smoothing. Journal of Soil Science 23(2): 210-234.

Authors:

David Howell, State Soil Survey GIS Specialist
USDA Natural Resources Conservation Service
1125 16th Street, Room 219
Arcata, CA 95521 707-822-7133 voice 822-7131 fax
David.howell@ca.usda.gov

Yoon Kim, PhD. Professor
Math Department
Humboldt State University
Arcata, CA 95521 (707) 826-5399 voice   (707) 826-3140
ygk1@humboldt.edu

Carrie-Ann Haydu-Houdeshell, Soil Survey Project Leader
USDA Natural Resources Conservation Service
Victorville Service Center
17330 Bear Valley Rd., Suite 106
Victorville, CA 92392 (760) 843-6882 x108 voice (760) 843-9521
Carrie-Ann.Houdeshell@ca.usda.gov

Pam Clemmer, Remote Sensing Specialist
U.S.D.I. Bureau of Land Management
Denver Federal Center
P.O. Box 25047
NSTC ST-134
Building 50
Denver, CO 80225-0047 (303) 236-0824 voice (303) 236-3508 fax
Pam_Clemmer@blm.gov

Russell Almaraz, Soil Scientist/GIS Specialist
USDA Natural Resources Conservation Service
Lancaster Service Center
44811 N. Date Ave. Suite G
Lancaster, CA 93534 (661) 945-2604 x113 voice (661) 942-5503 fax
Russell.Almaraz@ca.usda.gov