ERDC/EL TN-18-1
December 2018

US Army Corps of Engineers®

Taxonomic Soils Geomatics Investigation

by Austin Davis, John Furey, Cliff Morgan, and Jennifer Seiter-Moser

PURPOSE: This technical note describes initial investigative efforts to use soil classification data in a manner suitable for producing more accurate soil analogues for specific purposes. Those purposes include improving environmental modeling efforts and predicting complex biogeochemical processes affecting the fate and transport of contaminants and affecting spectral responses. This technical note also documents initial data analysis methods and the data structure and query system; further, this publication discusses the team's next steps.

Geomatics is the study of spatial properties, processes, and patterns inherent in existing spatial data. Soils data is therefore a suitable topic for geomatic research—and in fact, pedo-informatics is the hybrid discipline synthesizing information science, soils science, and geography. It is well known that soils mapping is an evolving science, constrained by the complex nature of soils, including geological heterogeneity, climatic and landscape variation, and anthropomorphic effects. These challenges create a ubiquitous and inescapable heterogeneity that confounds precise environmental modeling and prediction systems. Current modeling and geospatial tools cannot predict complex biogeochemical processes, because statistically accurate multivariate soil characteristics datasets do not exist. The lack of such datasets has been a limiting factor in the production of accurate soil analogues for predicting soil properties in austere and expeditionary environments. Here, the authors discuss the foundation upon which an evolutionary data system could yield better soil analogue suggestions over existing methodologies.

This research was performed to satisfy a component technology requirement for the Novel Taxonomic Approach to Predicting Soil Biogeochemical Processes work package. The U.S. Army (the Army) has lacked the ability to characterize soils and soil processes for militarily and environmentally relevant planning. The Army has relied on existing soil taxonomic systems to classify soils data; however, these systems do not by themselves suggest ideal soil analogues for any given purpose. Thus, the ability to predict relevant soil properties and soil processes (e.g., fate and transport of contaminants, spectral responses) at specific locations has been lacking. Here, the authors demonstrate the weaknesses of existing soil classification systems to suggest ideal soil analogues by comparing qualitative and quantitative data on like classified or analogous soils.

BACKGROUND: The U.S. Army Corps of Engineers (USACE) Maneuver Support Center of Excellence recently outlined a vision to provide "the Army organic capability to detect, assess, characterize, advise and mitigate all hazards." The argument can be made that when deployed into a true expeditionary environment, there would be extremely limited data concerning the environment and associated dangers. Furthermore, current methods to complete an all-hazards reconnaissance typically rely on the insertion of special teams to collect data and samples; a process that is often dangerous, slow, and not cost-effective. Therefore, the Army often lacks the ability to predict environmentally relevant soil processes and properties in data-scarce and

austere environments. The motivation for the project's approach is to utilize it to predict the properties that must be understood for specific operational concerns such as environmental liability, contaminant fate and transport prediction, and force maneuver during contingency operations.

Historically, the Army has lacked the ability to characterize soils and soil processes for militarily and environmentally relevant planning. Certainly, the Army has the capacity to quantitatively and qualitatively classify soils using existing soil taxonomic systems, but suggesting soil analogues that mimic relevant properties of soils at specific locations—both in the contiguous U.S. or outside the contiguous U.S.—where data is limited, has been lacking. Throughout the Global War on Terrorism, science and technology has made tremendous strides in the ability to rapidly and accurately monitor surface soil properties to assist in the decision-making process and mine large datasets for latent patterns. However, high levels of uncertainty and data limitations exist in current soil databases, remote sensing methodologies, and environmental modeling systems that limit the existing soil classification capability's accuracy and value. Therefore, an accurate and innovative way to suggest ideal soil analogues for specific military and environmental purposes is required.

Soils data has long been collected, structured, visualized, and analyzed for broader patterns. A majority of the soil taxonomy efforts over the past 100 years have focused primarily on agriculture. The authors want to utilize the inherent physicochemical properties (both qualitative and quantitative) found within these classification systems for military and environmental applications. Spatial analyses are also frequently conducted to observe regional trends and physiographic landforms. Existing soils databases of qualitative and quantitative data can become quite large (>100 GB) and often descriptive of the heterogeneity of soils. The aggregation of both qualitative and quantitative data on soil samples is used for taxonomic classification, which can also be mapped spatially. Each taxonomic designation has a corresponding Soils Series Description provided by the United States Department of Agriculture National Resource Conservation Service (NRCS). Taxonomic identification is also useful for discriminating soils in spatial datasets, including the Soil Survey Geographic database (SSURGO) and the National Cooperative Soil Survey (NCSS).

TECHNOLOGY DESCRIPTION: Preliminary Army research has shown that information on environmentally relevant soil properties are inherent in a limited set of taxonomic designations and taxonomic soil series descriptions. Further, contemporary data mining and natural language text processing have matured to a point that taxonomic data can be parsed, processed, and leveraged to extract needed information. Computational capabilities have also matured such that the large size and multi-dimensional aspects of taxonomic data can be processed in a reasonable period of time.

Conveniently, detailed information describing the factors contributing to inherent geochemical responses is contained with soil taxonomical classification systems. The purpose of soil taxonomical systems is to clearly and non-arbitrarily distinguish soils from different geographical areas or locales based largely on factors associated with their natural development. Thus, it is expected that environmental responses tied to inherent geochemical properties can be predicted if they could be calibrated with respect to soil taxonomic designations. Global soil taxonomic systems, including the NRCS (U.S. system), are largely based on individual assessments, including soil morphology observational descriptions within soil profiles. The soil taxonomic systems also contain chemical, physical, climatic, and topographical information related to overall soil

development over geological time. Most of this information is primarily contained in qualitative text descriptions, including reports and papers and books, and is not readily organized for use by a computer. However, with modern environmental information systems, it is possible to map existing semantic taxonomies into broader soil ontologies, with respect to specific purposes, by using text-based pattern recognition across different dimensions of taxonomic characteristics (e.g., fertility, aridity, moisture) and spatial features (e.g., shape, size, location).

Therefore, the NRCS taxonomic soils databases can be analyzed to derive environmentally relevant soil properties for the majority of soil regions in the United States. Once the taxonomic database is refined and reorganized it will be utilized to predict complex soil biogeochemical processes, through the soil properties evident in their taxonomy, to different regions of the world, especially in austere settings and environments. Soil data and taxonomy has been provided by the NRCS and will be used to develop a geographic database of environmentally relevant soil properties. The environmental relevancy of soil properties is determined by data demands of the U.S. Army Engineer Research and Development Center (ERDC) Environmental Quality and Installations Military Materials in the Environment, Military Research Program.

The scientific goal of the team's approach linking geomatics and pedoinformatics is to relate available qualitative and quantitative soils data of the contiguous United States in a manner suitable for analytical, mining, and learning algorithms. It is expected that once information is related spatially and aspatially, a viable dataset will emerge that is suitable for solving the previously mentioned problem: suggesting accurate soil analogues for specific military and environmental purposes.

The first step in the process is the collection of available datasets from the NSSC. The NSSC provided 3,262 Access Database files that each represent small sections of the much larger SSURGO database. The SSURGO database will largely function as the "Qualitative" dataset. Out of the 3,262 databases, 3,258 were built using Access Template macros contained within each database. Four databases representing sections of Alaska were excluded since they represent regions other than the contiguous U.S. Once all of the database files were populated with their associated data, the database data and schema were exported and merged to form a large SSURGO database representing the contiguous U.S., with a limited set of tertiary regions. This process was conducted using mdb-tools. Currently a local PostGIS implementation houses the merged SSURGO product.

At a minimum, the team's local implementation of the SSURGO product contains 36,401,756 unique spatial features, which have a many-to-one relationship with one of 950,438 soil components. There are 76 different tables containing different domains of information. Some tables have over 130 fields of data. The dimensional scale of this dataset is quite impressive at 58,897,037 different rows across all of the tables that can be related to each other through a variety of relational types.

The "Quantitative" database is the NCSS Soil Characterization Database, which was provided as an Access file by the NCSS and was also converted into a PostGIS database. This dataset represents two general types of data: Soil Sample Locations and Soil Sample Locations with Geochemical Data. The set of soils containing quantitative geochemical data is expected to be most useful within the NCSS. In spatial terms, the NCSS dataset is represented as point features,

which is in contrast to the polygonal SSURGO dataset. The NCSS dataset also contains more international data than the SSURGO.

The NCSS database contains 1,073,503 records, primarily as individual soil samples, including any sample metadata, such as sample designations and location, as well as sample analysis and measurement data. As constructed according to the header information, the database has up to 12 sample taxonomic descriptors per sample (generally having samp_tax in the header), including the sample mineralogy (samp_minalogy in the header), and 24 other similar taxonomic descriptors (having corr_tax and SSL_tax in the header). In addition to the text descriptors, there are up to 584 numerical entries per sample related to geochemical analyses and other physico-chemical measurements such as particle sizes, although most of this numerical data was sparsely populated.

Almost 61% of the samples have sample taxonomic order entries, which is the largest percentage of taxonomic entries. The distribution of entries among the twelve taxonomic soil orders is given in Table 1. The large variation among the orders probably results primarily from the agronomy interest driving the sampling, instead of just actual geographical coverage in the U.S. For example, much of the U.S. West is indeed arid, but there are 149 times as many aridisol samples in the database as there are gelisols.

Table 1. Distribution of the number of soil samples across the sample taxonomic order entries in the NCSS database, sorted alphabetically.

samp_tax order	number of entries
alfisols	86966
andisols	37711
aridisols	265908
entisols	33007
gelisols	1787
histosols	3124
inceptisols	48184
mollisols	94593
oxisols	2530
spodosols	26804
ultisols	42997
vertisols	9761

These 12 taxonomic order classifications are expected to be too broad to provide geochemistry-predictive capability by themselves; for example, merely knowing that a soil is an inceptisol may not provide enough information to usefully predict geochemical parameters such as pH, carbonates, cation exchange capacity, etc.

The similarly generated distributions of the 367 sample taxonomic Great Groups and 47 sample mineralogy classifications in the NCSS database are given in the Appendix. At these levels of groupings, the heterogeneity of sampling continues to prevail, but hundreds of these classification

levels contain more than the several dozen samples needed for good statistics. These more detailed classifications may provide more useful chemistry predictions than the broader levels.

Not every soil sample with taxonomic entries has physico-chemical entries, and vice versa. This could be partly due to this database aggregating contributions of many individuals and organizations, each with their own methodologies. Although in general most of the geochemistry is not well-populated—for example, less than 2% of the soil samples have an entry within the geochemistry header called "org_c," some of the geochemistry is better populated than the taxonomy; for example, over 83% of the soil samples have an entry within the geochemistry header called "k_nh4." Of most interest for further statistical analyses are those samples with both relevant taxonomy entries and relevant geochemistry (including other physico-chemical measurements) entries.

Slightly over 61% of the samples have at least some taxonomic entries and at least some geochemistry entries. Of those downselected samples, the average number of geochemistry entries is 130. To make further progress on the determining of which taxonomic classifications most usefully predict geochemistry, a reduction to the most relevant geochemistry entries is indicated. Due to the sheer size of the existing datasets, the number of potential covariance calculations that would have to be made for any given set of soils would be N factorial, where N is defined as the number of fields. To reduce N, expert judgment by a soils scientist at the ERDC Environmental Laboratory will select data fields considered to be the most important when describing or characterizing a soil. These fields are considered to have properties that do describe some inherent aspect of the soils and not just cosmetic or arbitrary values.

A portion of the table of pairwise correlations for these fields is shown in Table 2 expressed in terms of r^2. The vast majority (> 99%) of the pairwise r^2 values are < 0.3, when including the more than a hundred other columns not shown.

Table 2. The pairwise correlation r^2 values for the first 14 selected geochemical fields. The values that are > 0.3 are highlighted in red.

	n_tot	c_tot	s_tot	oc	cec_sum	caco3	h2o_satx	na_satx	k_satx	ca_satx	mg_satx	hco3_satx	cl_satx
ph_h2o	0.079	0.095	0.018	0.030	0.009	0.098	0.047	0.008	0.003	0.001	0.001	0.145	0.002
n_tot		0.725	0.000	0.700	0.250	0.011	0.611	0.000	0.000	0.000	0.001	0.000	0.001
c_tot			0.000	0.621	0.117	0.109	0.228	0.000	0.005	0.002	0.000	0.013	0.001
s_tot				0.011	0.453	0.000	0.000	0.034	0.034	0.171	0.073	0.005	0.017
oc					0.201	0.008	0.806	0.000	0.000	0.000	0.002	0.003	0.002
cec_sum						0.004	0.151	0.595	0.112	0.029	0.625	0.000	0.153
caco3							0.003	0.000	0.002	0.004	0.003	0.001	0.002
h2o_satx								0.001	0.000	0.000	0.002	0.002	0.001
na_satx									0.212	0.064	0.238	0.053	0.800
k_satx										0.018	0.067	0.026	0.197
ca_satx											0.038	0.002	0.135
mg_satx												0.001	0.129
hco3_satx													0.013

To relate the Quantitative dataset to the Qualitative dataset, a middleware program was used to build queries using a join relationship network model. Each database was cast into a hash-table data structure where the keys were table names and the value was a vector of table name and relation key pairs. An intermediate relational table between the two databases was produced spatially; such that a site location in the NCSS database was intersected by a SSURGO polygon. A Breadth First Search (BFS) query engine was developed in Python 3 to produce the least cost path through the relation networks linking all of the relevant tables to form a multi-database search query. This tool allows researchers the ability to query various soil properties from both databases simultaneously and relates those properties where they spatially intersect.

For the Qualitative data, the USDA documentation, NCSS manual, and SSURGO text entries have been parsed, and Latent Semantic Indexing completed at the paragraph level for individual words. These analyses implemented iPython calls to Gensim modules for topic modelling, and the software methodology allows for user exploration of the network of text relationships. A diagram of some of the relationships between software elements implementing database handling and textual analyses is shown in Figure 1. An illustration of the tree structure resulting from one of the semantic analyses is shown in Figure 2.

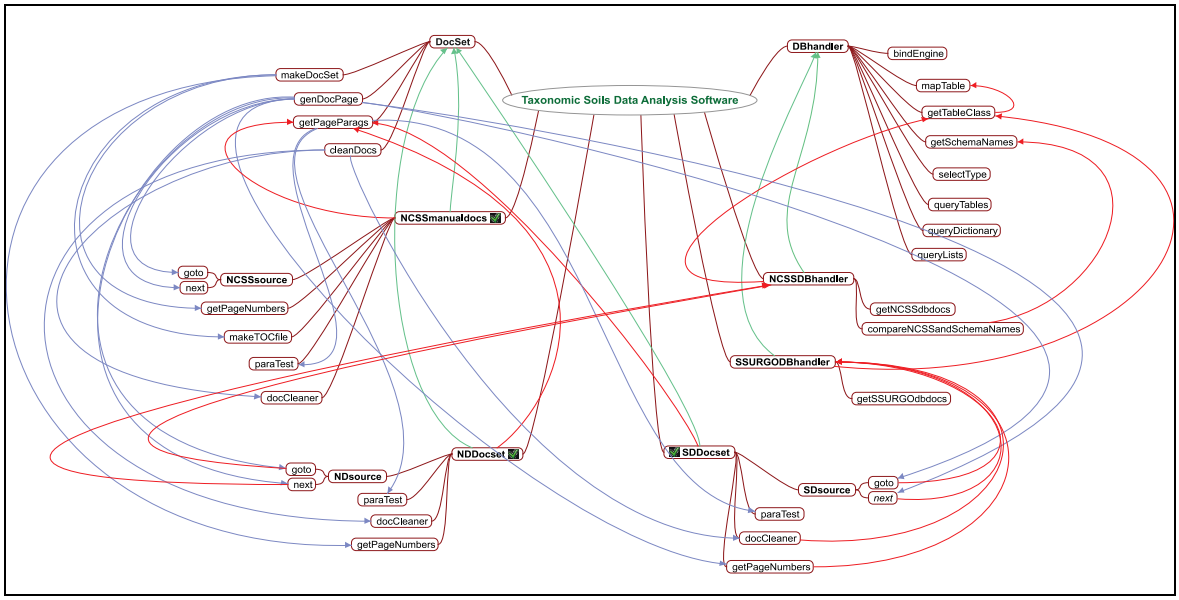

Figure 1. Relationship network showing the iPython handling of class types for use with the Latent Semantic Analysis software modules. Methods with no connected lines, such as selectType, are for use with other analysis methods; e.g., clustering.

First Page Of TOC From Manual

Contents

	Method code	Page
Preface		xix
I. Introduction		1
II. Primary Characterization Data		11
1 Sample Collection and Preparation	1	12
1.1 Field Sample Collection and Preparation	1A	13
1.1.1 Site Selection	1A1	13
1.1.1.1 Geomorphology	1A1a	13
1.1.1.2–4 Pedon, Water, and Biological Sampling	1A1b-d	13
1.1.2 Classification Systems and Soil Survey		13
1.2 Laboratory Sample Collection and Preparation	1B	23
1.2.1 Soil Samples	1B1	23
1.2.2 Water Samples	1B2	26
1.2.3 Biological Materials	1B3	26
2 Conventions	2	27
2.1 Methods and Codes	2A	27
2.2 Data Types	2B	28
2.3 Particle-Size Size-Fraction Basis for Reporting Data	2C	29
2.3.1 Particles <2 mm	2C1	29
2.3.2 Particles <Specified Size >2 mm	2C2	29
2.4 Sample Weight Basis for Reporting Data	2D	29
2.4.1 Air-Dry/Oven-Dry	2D1	29
2.4.2 Field-Moist/Oven-Dry	2D2	29
2.4.3 Correction for Crystal Water	2D3	29
2.5 Significant Figures and Rounding	2E	30
2.6 Data Sheet Symbols	2F	30
3 Soil Physical and Fabric-Related Analyses	3	30
3.1 Particle-Size Distribution Analysis	3A	31
3.1.1 Classification Systems		31
3.1.2 Particles <2 mm	3A1	33
3.1.2.1 Pipet Analysis	3A1a	38
3.1.2.1.1 Total Clay, <0.002-mm (<2 µm)		39
3.1.2.1.2 Total Silt, 0.002 to 0.05 mm		42
3.1.2.1.3 Total Sand, 0.05 to 2.0 mm		42
3.1.2.1.4 Fine Clay, <0.0002 mm (<0.2 µm)		43
3.1.2.1.5 Carbonate Clay, <0.002 mm (<2 µm)		44
3.1.2.1.6 Fine Silt, 0.002 to 0.02 mm		45
3.1.2.1.7 Coarse Silt, 0.02 to 0.05 mm		45

iii

Corresponding Tree Structure

```
(-1, 'root')
|___ (0, ['1', 'Sample', '12'])
|    |___ (1, ['1.1', 'Field', '13'])
|    |    |___ (2, ['1.1.1', 'Site', '13'])
|    |    |    |___ (3, ['1.1.1.1', 'Geomorphology', '13'])
|    |    |___ (2, ['1.1.2', 'Classification', '13'])
|    |___ (1, ['1.2', 'Laboratory', '23'])
|         |___ (2, ['1.2.1', 'Soil', '23'])
|         |    |___ (2, ['1.2.2', 'Water', '26'])
|         |         |___ (2, ['1.2.3', 'Biological', '26'])
|___ (0, ['2', 'Conventions', '27'])
|    |___ (1, ['2.1', 'Methods', '27'])
|         |___ (1, ['2.2', 'Data', '28'])
|              |___ (1, ['2.3', 'Particle-Size', '29'])
|              |    |___ (2, ['2.3.1', 'Particles', '29'])
|              |         |___ (2, ['2.3.2', 'Particles', '29'])
|              |___ (1, ['2.4', 'Sample', '29'])
|              |    |___ (2, ['2.4.1', 'Air-Dry/Oven-Dry', '29'])
|              |         |___ (2, ['2.4.2', 'Field-Moist/Oven-Dry', '29'])
|              |              |___ (2, ['2.4.3', 'Correction', '29'])
|              |___ (1, ['2.5', 'Significant', '30'])
|                   |___ (1, ['2.6', 'Data', '30'])
|___ (0, ['3', 'Soil', '30']) 'Phosphorus,', '159'])
|    |    |___ (2, ['4.3.2', 'Anion', '165'])
|    |    |    |___ (3, ['4.3.2.1', 'Two-Point', '165'])
|    |    |    |    |___ (4, ['4.3.2.1.1', '1-h,', '165'])
|    |    |    |    |    |___ (5, ['4.3.2.1.1.1', 'Phosphorus', '165'])
|    |    |___ (2, ['4.3.3', 'Aqueous', '166'])
|    |    |    |___ (3, ['4.3.3.1', 'Single-Point', '166'])
|    |    |    |    |___ (4, ['4.3.3.1.1', '1:10,', '166'])
|    |    |    |    |    |___ (5, ['4.3.3.1.1.1', 'Phosphorus', '166'])
|    |    |    |___ (3, ['4.3.3.1', 'Single-Point', '167'])
|    |    |    |    |___ (4, ['4.3.3.1.2', '1:5,', '167'])
|    |    |    |    |    |___ (5, ['4.3.3.1.2.23', 'Boron', '167'])
|    |    |___ (2, ['4.3.4', 'Bray', '168'])
|    |    |    |___ (3, ['4.3.4.1', 'Phosphorus', '168'])
|    |    |___ (2, ['4.3.5', 'Bray', '169'])
|    |    |    |___ (3, ['4.3.5.1', 'Phosphorus', '169'])
|    |    |___ (2, ['4.3.6', 'Olsen', '170'])
|    |    |    |___ (3, ['4.3.6.1', 'Phosphorus', '170'])
|    |    |___ (2, ['4.3.7', 'Mehlich', '170'])
|    |    |    |___ (3, ['4.3.7.1', 'Phosphorus', '170'])
|    |    |___ (2, ['4.3.7', 'Mehlich', '171'])
|    |    |    |___ (2, ['4.3.8', 'Citric', '171'])
|    |    |    |    |___ (3, ['4.3.8.1', 'Phosphorus', '171'])
|    |    |    |___ (2, ['4.3.9', 'New', '172'])
|    |    |    |    |___ (3, ['4.3.9.1', 'Phosphorus', '172'])
|    |    |    |___ (2, ['4.3.10', 'Nitrogen,', '172'])
|    |    |         |___ (2, ['4.3.11', '1', '173'])
|    |    |              |___ (2, ['4.3.12', 'Anaerobic', '174'])
|    |    |                   |___ (3, ['4.3.12.1', '2', '174'])
|    |    |                        |___ (4, ['4.3.12.1.1', 'Nitrogen', '174'])
|    |___ (1, ['4.4', 'Carbonate', '174'])
```

Figure 2. Semantic analyses of individual words at the paragraph level for the NCSS manual (first page of table of contents, left) produced a tree structure (right).

DISCUSSION AND RECOMMENDATIONS: This investigation demonstrates the heterogeneity of soils and the inconsistency and uncertainty inherent in the existing soil taxonomic system. Pairwise correlations between soil properties within same or similar taxonomic groups were too weak (>99% of all relevant pairs had $r^2 < 0.3$) for predictive purposes in themselves. When considering the technology requirement to produce accurate soil analogues for specific purposes, this can be reduced to a two-step process: (1) prove that soil analogues can be determined repeatably and for many soils, and (2) produce a soil analogue for a specific purpose rather than an exact replica. Step 1 is accomplished by demonstrating that groupings of soils can have consistent properties across many fields. The team's initial efforts have unfortunately shown that while some consistency in soil data does exist, it is not sufficient for producing ideal analogues. This is why developing suitable soil analogues is difficult.

The authors propose that this is a surmountable problem by applying techniques and algorithms from machine learning. Since some parametric consistency does exist, iteratively grouping soils in a parametric vector space to achieve more consistent correlations is possible. The next effort for the Quantitative databases will rely on the employment of a genetic algorithm that will evolve groupings of soils toward a set of fitness metrics, where the fitness metrics represent the soil metrics of some unknown purpose. After several evolutionary steps, the genetic algorithm is expected to stabilize into a series of soil groups where the groups correspond to different fitness

levels of a particular soil's use as an analogue. Therefore, given an unknown set of soils **A,** we can determine an analogue set of soils **B** for some purpose *P*. Where a set of metrics and descriptors **M** is known about **A** and a set of metrics and descriptors **N** is known about **B** and a set of metrics and descriptors **O** is known about *P* by using an evolutionary technique.

The Qualitative and textual analyses will be expanded to phrasal analyses, which will also require expert judgment to down-select phrases. For example, "very fine" can have a definite, even if qualitative, soil science meaning depending upon the circumstances, but in other contexts isolated intensifiers such as "very" are usually ignored for textual analyses.

ADDITIONAL INFORMATION: This technical note was prepared by Austin Davis, Research Geographer, John Furey, Research Physical Scientist, Cliff Morgan, Research Physicist, and Jennifer Seiter-Moser, Soil Scientist, all of the Environmental Laboratory, U.S. Army Engineer Research and Development Center. The technology was developed as an activity of the Environmental Quality and Installations research program. This technical note should be cited as follows:

> Davis, A. V., J. Furey, C. Morgan, and J. Seiter-Moser. 2018. Taxonomic Soils Geomatics Investigation. ERDC/EL TN-18-1. Vicksburg, MS: U.S. Army Engineer Research and Development Center.

The below link will access Table A1, the distribution of soil samples across the sample taxonomic Great Group entries in the NCSS database, and Table A2, the distribution of the number of soil samples across the sample mineralogical description entries in the NCSS database.

http://dx.doi.org/10.21079/11681/27733