Chapter 17
Digital Summaries of Pedon Descriptions

Stephen Roecker, Jay Skovlin, Dylan Beaudette and Skye Wills

Abstract Soil scientists have been describing and analyzing pedons for over a hundred years. In the USA, a small portion of this data has been captured in the National Soil Information System (NASIS). While NASIS serves as a data repository, its analytical capabilities are limited, and the data are underutilized. In order to facilitate the analysis of soil horizon data in NASIS, we have used R to develop R Markdown (Rmd) reports. These Rmd reports are designed to provide numerical and graphical summaries of soil horizon data used for soil survey activities, such as the development of Official Series Descriptions and soil map unit components.

Keywords Soil series · Range in characteristics · NASIS · Pattern matching

17.1 Introduction

Pedon data consist of field estimates, observations, and laboratory measurements. Unlike the soil map unit polygons and their associated attribute data (component data), pedon data represent point data from individual soil observations. In support of soil surveys during the last 100 years, the National Cooperative Soil Survey (NCSS) has collected a substantial amount of pedon data. Since the introduction of the National Soil Information System (NASIS) in 1994 (Fortner and Price 2012), approximately 400,000 field pedons and approximately 63,000 laboratory pedons have been digitized (Ferguson, 2015, personal communication). Although

S. Roecker (✉)
USDA-Natural Resource Conservation Service, Indianapolis, IN 46278, USA
e-mail: stephen.roecker@in.usda.gov

J. Skovlin
USDA-Natural Resource Conservation Service, Missoula, MT 59808, USA

D. Beaudette
USDA-Natural Resource Conservation Service, Sonora, CA 95370, USA

S. Wills
USDA-Natural Resource Conservation Service, Lincoln, NE 68508, USA

A.E. Hartemink and B. Minasny (eds.), *Digital Soil Morphometrics*,
Progress in Soil Science, DOI 10.1007/978-3-319-28295-4_17

significant, this represents only a small portion of total field pedons ever described (Fig. 17.1). For digital soil mapping and updates to soil surveys, these pedon data are an invaluable resource.

In order to store soil data compactly and efficiently, NASIS has a hierarchical data structure (Fig. 17.2). One branch of the data structure stores point data—observations of site and pedon data, with soil horizons as the basic element. Aggregated data about soil map units and their soil components are stored in another part of the structure. Each aggregated soil component is made up of

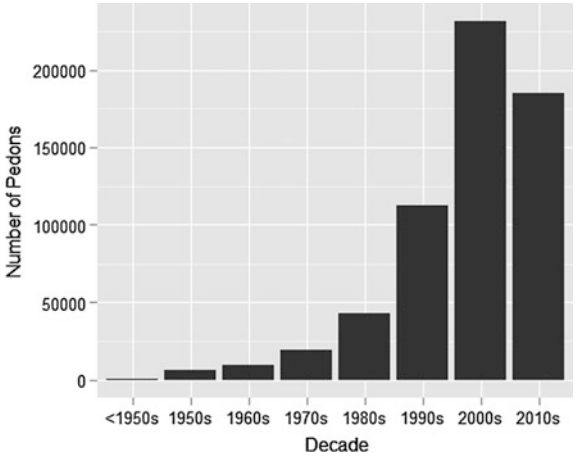

Fig. 17.1 Number of pedons sampled per decade recorded in NASIS

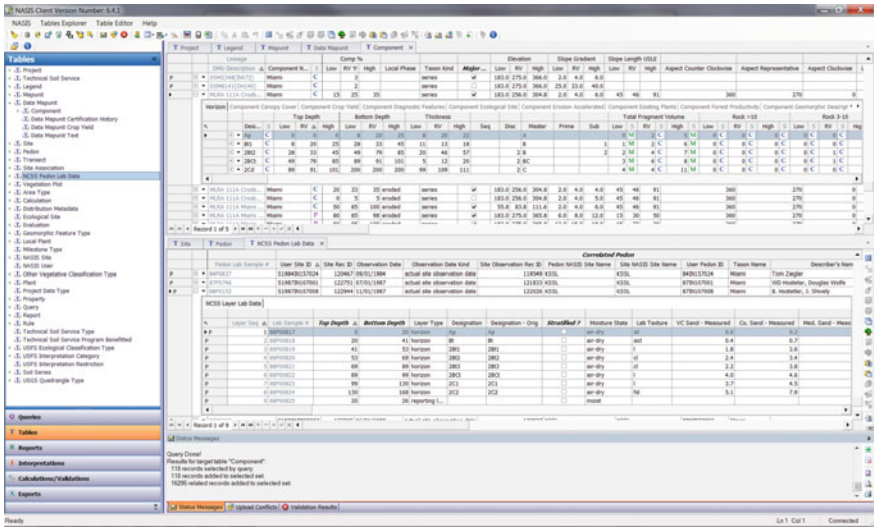

Fig. 17.2 Screenshot of the NASIS database interface, and the component and laboratory tables

generalized soil horizons based on a sample of pedon observations. Also linked to each horizon record are additional child tables. Each of these nested child tables may include several related child tables in order to capture heterogeneous soil conditions within each soil horizon. The dominant condition is specified as the representative value (RV). For numeric component data, it is also possible to specify a range with low (L) and high (H) values. This makes it possible to characterize the distribution or variation of a particular soil variable, such as clay content. Using this database structure, it is possible to capture soil horizonation, aggregate the data, and then generate spatial predictions by linking it to the soil polygons.

Soil mapping involves aggregating horizon descriptions from field and laboratory pedons into component horizon data. While there are standards that guide the process of describing individual sites and pedons in the Soil Survey Manual (Soil Survey Division Staff 1993) and the Field Book for Describing and Sampling Soils (Schoeneberger et al. 2012), there are no guidelines for the process of aggregating point/pedon observations into their component database elements. The NCSS guidelines either address developing Official Series Descriptions (OSDs) (USDA 2015), or how component ranges relate to the OSD (USDA 2013). Historically, the process of determining the ranges (L, RV, H) for various soil properties has been done with pencil and paper or spreadsheets and then selected by expert knowledge. This is a practice that continues today for a variety of reasons:

1. Familiarity with existing protocol,
2. Inconsistency among the existing data,
3. Additional workload involved in digitizing data,
4. Perceived or real software limitations,
5. Lack of training in new software and statistical methods.

Prior to the advent of NASIS, there were many early attempts at estimating low, RV, and high values for soil properties (Young et al. 1991; Jansen and Arnold 1976). These earlier attempts looked at estimates for portions of the soil profile, such as surface texture or subsoil clay content, and utilized parametric estimates (i.e., mean and confidence intervals). They also demonstrated the disconnect between the limits set for taxonomic units and those observed within map unit components. This issue is now addressed by Soil Survey Technical Note 4 (USDA 2003), which allows the range (i.e., low and high) of map unit components to extend beyond those specified by the OSD.

It is possible to manipulate and summarize pedon data directly in NASIS with reports and pivot tables, but the majority of summary functions within NASIS have been designed to analyze and evaluate component-level aggregate data. Data can be exported from NASIS to other software (Table 17.1), but these other software do not provide the same concise summary of data as do the reports designed for component data in NASIS. New reports can be added to NASIS, but complex reports are difficult to write because NASIS supports a limited implementation of the Structured Query Language (SQL) which has few functions for performing statistical analysis. Here, we advocate exporting pedon data to R (R Core

Table 17.1 Sample of tools for analyzing soil data sorted by user sophistication

Tabular analysis
1. Pencil and paper
2. Excel spreadsheets
3. PedonPC and AnalysisPC (microsoft access databases)
4. NASIS
5. R
Spatial analysis
1. SoilWeb
2. Web soil survey
3. Soil data viewer
4. SSURGO file geodatabases
5. R

Development Team 2015). R now supports R Markdown (Rmd) reports that provide access to report-writing capabilities (Xie 2014; Allaire et al. 2015) and user-contributed functions specifically designed for digital soil morphometrics, such as the aqp (Beaudette et al. 2012), soilDB (Beaudette and Skovlin 2015), and soil texture (Moeys 2015) packages.

17.2 Methods

To generate Markdown documents, RStudio was used. RStudio is an integrated development environment (IDE) for R and provides a minimalist graphical user interface (GUI) that organizes the R environment into four task-oriented windows. The initial start-up process of using RStudio and R to run the reports requires the user to install several R packages and their dependencies and setup an ODBC connection to NASIS. These steps are documented online at the NRCS Soils job-aid page, and readers are pointed to these reference documents for full details. R is an extendable environment and is in constant development, so installing additional packages is a common practice as packages are updated or new packages become available.

In order to access NASIS data for use in R, a user must first load a selected set of field or laboratory pedons in NASIS. A selected set is a view or virtual table that is created via a query, and serves as a working subset of a user's local NASIS database. NASIS has numerous queries to accomplish this. Once the data is loaded in NASIS, it can be imported into R via an ODBC connection using the fetchNASIS() function in the soilDB package. The user only needs to modify the report script by entering the name of the text file (e.g., "Miami") containing the GHL rules that correspond to the pedons loaded in the selected set. The report script is then run, and an HTML document is generated by pressing the Knit button in

RStudio. The necessary analysis steps are programmed into the report script, and the output is formatted to HTML using Rmd.

To develop a list of GHL, the user must specify which horizons are similar enough to be aggregated (Fig. 17.3). This is accomplished by mapping the existing horizon designations for each horizon and matching them to a generalized (i.e., simplified) horizonation sequence for each soil series or component. The assumption is made that the existing horizon designations accurately reflect the soil morphology and the corresponding soil properties of the horizons. For established soil series, the Official Series Description (OSD) can be used as a starting point for determining the appropriate GHL to assign to the horizons for the soil in question. The OSD provides a sample of likely horizons within either the typical pedon described or the range in characteristics (RIC) sections. For example, multiple Bt horizons might be aggregated or grouped together if it is determined that they are similar in clay content and other characteristics and that such an aggregation is not going to affect the use or interpretation of that soil. Also, Bw and Btk horizons might be aggregated if the development of the Btk horizons is incipient and does not meet the definition of an argillic or calcic diagnostic horizon. Another approach is to examine the frequency with which each horizon occurs (Fig. 17.4). Horizons that occur frequently are likely to be the most representative.

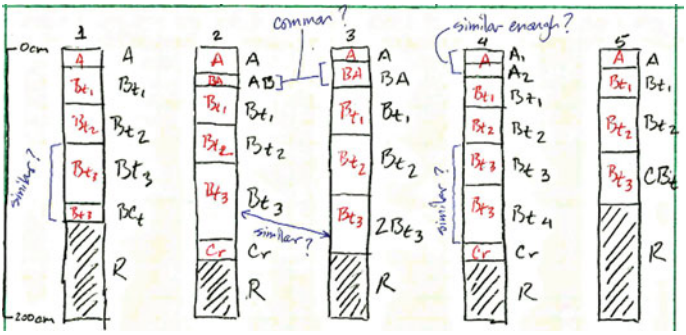

Fig. 17.3 Hand drawn illustration of the decision making (e.g., question asking) process soil scientists go through when determining the best selection of GHL for several similar soil descriptions

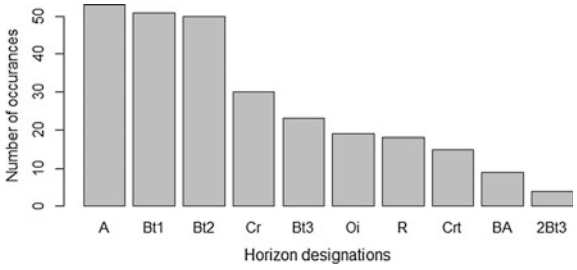

Fig. 17.4 Example of the original horizon designations sorted by frequency of occurrence for the Miami soil series

Once appropriate GHL have been determined for the collection of pedons, pattern matching is used to assign the new GHL to each horizon. The process uses functions designed to parse the text from each horizon designation and match it to the new GHL. The function searches for any combination of characters before or after the specified pattern. Patterns that do not match any of the GHL are labeled "not used." Special meta-characters serve as anchors or anti-wildcards for the beginning (i.e., caret "^") and end (i.e., dollar sign "$") of the given pattern. For example, the GHL pattern "Bt" will match any permutation of Bt, such as 2Bt or Bt1. To exclude 2Bt horizons, a more specific pattern of "^Bt" would be necessary. Conversely, to exclude Bt1 horizons, a pattern of "Bt$" would be used. If a user wishes to match special character like the caret "^" symbol, which is also used for human-transported material, it is necessary to append it with two backslashes like so, "\\^." As the GHL rules are developed, they are stored in a text file and later referenced by the Rmd report. If the user is satisfied with the resulting GHL designations, they can upload it to the comp layer ID field in the horizon table in NASIS where it is stored for future use.

Example of the GHL rules for the aqp loafercreek sample data set:

- **A**: ^A$|Ad|Ap
- **Bt1**: Bt1$
- **Bt2**: ^Bt2$
- **Bt3**: ^Bt3|^Bt4|CBt$|BCt$|2Bt|2CB$|^C$
- **Cr**: Cr
- **R**: R

Embedded in the reports are numerical and graphical summarizes of the data elements typically collected and used to differentiate dissimilar soils. Numerical variables are summarized by percentiles (i.e., quantiles), instead of the mean and confidence intervals, because they provide nonparametric estimates of a distribution and are less influenced by skewness which is common for most soil properties. Also percentiles provide a neat and compact summary. The percentiles used can be adjusted by the user, but the default is set to the five number summary (i.e., 0, 25, 50 % or median, 75, and 100 %) (Tables 17.3 and 17.4). Additionally, the percentiles are appended with the number of observations (n) (e.g., (0, 25, 50 % or median, 75, and 100 %)(n)), to inform the user of the sample size. The standard graphics used are box plots which provide a similar summary and interpretation (outliers, ~5, ~25, 50 % or median, ~75, ~95 %, outliers) of the data (Fig. 17.5). To summarize categorical variables, frequency tables (i.e., contingency tables) are used which cross-tabulate the number of occurrences of matching pairs (Tables 17.5 and 17.6).

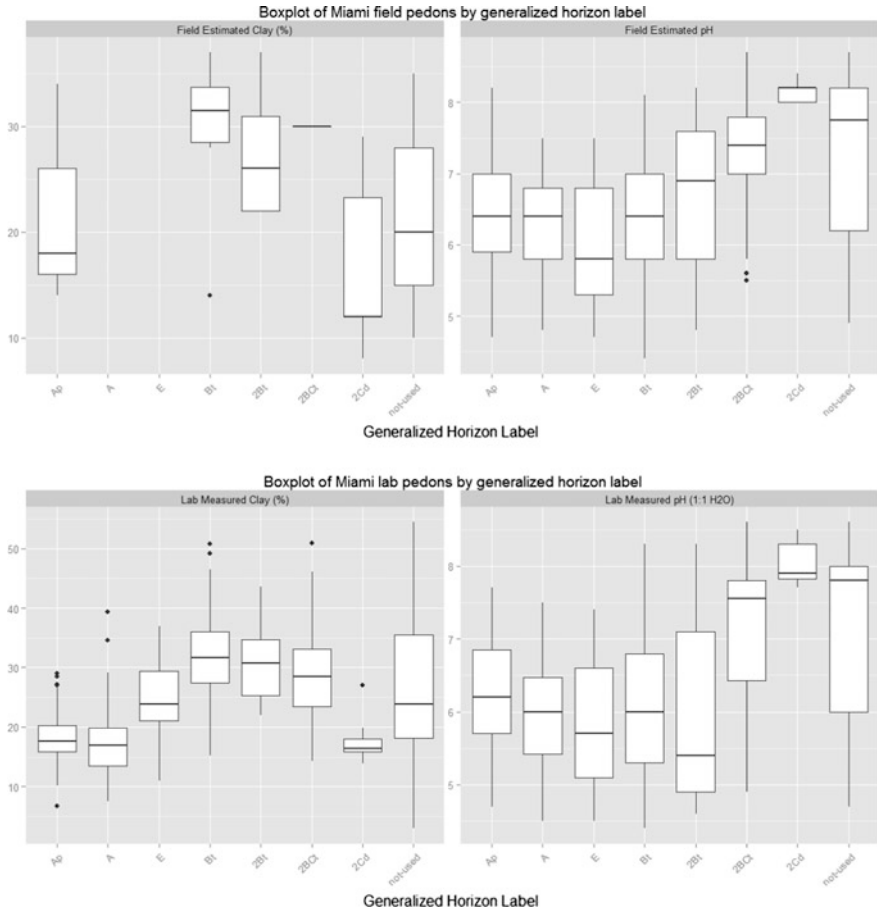

Fig. 17.5 Box plots of field (*top*) and laboratory (*bottom*) measurements for clay (%) and pH

17.3 Results and Discussion

The full field and laboratory reports are not shown here due to space limitations. The list below summarizes their content followed by sample excerpts and a discussion of the field and laboratory report content.

- Field pedon report content:
  - General map of georeferenced pedon locations overlaid on county boundary outlines;
  - Table of identifying information: pedon id, soil series, etc.,
  - Soil profile plots (Fig. 17.6),
  - Surface rock fragments,

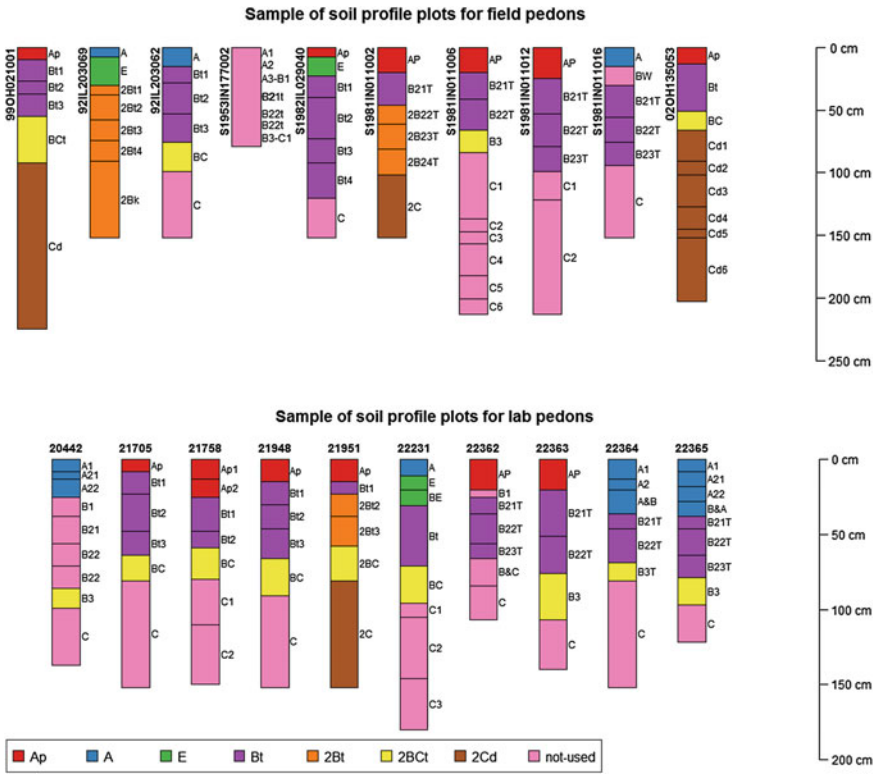

Fig. 17.6 Example of soil profile plots of the field (*top*) and laboratory (*bottom*) pedons for the Miami soil series. Horizons are colored according to their GHL

  - Depths and thickness of diagnostic horizons,
  - Comparison of GHL versus original horizon designations (Table 17.2),
  - Depth and thickness distribution of GHL,
  - Numeric variables: clay content, rock fragments, pH, etc., (Table 17.3)
  - Soil texture and texture class modifier summarized by GHL (Table 17.5),
  - Soil color hue summarized by GHL,
  - Elevation, slope gradient, and slope aspect,
  - Parent material versus landform,
  - Slope shape (down slope vs. across slope shape),
  - Drainage class versus hillslope position.

- Laboratory pedon report content:

  - General map of georeferenced laboratory pedon locations overlaid on county boundary outlines,
  - Table of identifying information: pedon id, soil series, etc.,
  - Soil profile plots (Fig. 17.6),

Table 17.2 Number of the GHL versus the original horizon designations for the Miami series

	Ap	Bt1	BC	Bt2	AP	B21T	B22T	B3	BE	A	E	A1	Bt3	A2	B23T	2Bt2	2Bt3	2C	BT3	1Bt1
Ap	54	0	0	0	29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A	0	0	0	0	0	0	0	0	0	23	0	16	0	14	0	0	0	0	0	0
E	0	0	0	0	0	0	0	0	24	0	17	0	0	0	0	0	0	0	0	0
Bt	0	36	0	35	0	27	26	0	0	0	0	0	16	0	14	0	0	0	7	6
2Bt	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	7	0	0	0
2BCt	0	0	35	0	0	0	0	26	0	0	0	0	0	0	0	0	0	0	0	0
2Cd	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	0	0
Not-used	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 17.3 Percentile summaries of field estimates of clay (%) and pH

genhz	Clay	phfield
Ap	(14, 16, 18, 26, 34)(3)	(4.7, 5.9. 6.4, 7, 8.2)(77)
A	(NA, NA, NA, NA, NA)(0)	(4.8, 5.8, 6.4, 6.8, 7.5)(54)
E	(NA, NA, NA, NA, NA)(0)	(4.7, 5.3, 5.8, 6.8, 7.5)(41)
Bt	(14, 28, 32, 34, 37)(6)	(4.4, 5.8, 6.4, 7, 8.1)(206)
2Bt	(22, 22, 26, 31, 37)(5)	(4.8, 5.8, 6.9, 7.6, 8.2)(30)
2BCt	(30, 30, 30, 30, 30)(1)	(5.5, 7, 7.4, 7.8, 8.7)(70)
2Cd	(8, 12, 12, 23, 29)(6)	(8, 8, 8.2, 8.2, 8.4)(17)
Not-used	(10, 15, 20, 28, 35)(29)	(4.9, 6.2, 7.8, 8.2, 8.7)(146)

Table 17.4 Percentile summaries of laboratory measurements of clay (%) and pH

genhz	Claytot	ph1to1h20
Ap	(7, 16, 18, 20, 29)(83)	(4.7, 5.7, 6.2, 6.9, 7.7)(83)
A	(7.5, 13.5, 17, 19.9, 39.4)(53)	(4.5, 5.4, 6, 6.5, 7.5)(54)
E	(11, 21, 24, 29, 37)(45)	(4.5, 5.1, 5.7, 6.6, 7.4)(45)
Bt	(15.2, 27.4, 31.6, 36, 50.7)(155)	(4.4, 5.3, 6, 6.8, 8.3)(155)
2Bt	(22, 25.3, 30.7, 34.7, 43.6)(13)	(4.6, 4.9, 5.4, 7.1, 8.3)(13)
2BCt	(14.3, 23.4, 28.4, 33.2, 50.9)(86)	(4.9, 6.4, 7.5, 7.8, 8.6)(86)
2Cd	(14, 16, 16, 18, 27)(10)	(7.7, 7.8, 7.9, 8.3, 8.5)(10)
Not-used	(3, 18.1, 23.8, 35.5, 54.4)(298)	(4.7, 6, 7.8, 8, 8.6)(299)

- Weighted averages for the particle size control section,
- Depths and horizon thickness for the particle size control section,
- Comparison of GHL versus original horizon designations (Table 17.2),
- Depth and horizon thickness of GHL,
- Numeric variables: particle size fractions, pH, base saturation, carbon content, etc. (Table 17.4),
- Laboratory soil texture summarized by GHL (Table 17.6).

Much of the information contained in the reports is used to summarize data for developing OSD and aggregated map unit soil components. Evaluating the graphics and tables within the reports quickly show where there are possible errors, narrow or wide ranges in values, or where data gaps exist due to insufficient data. One of the first outputs of the report that should be examined is the contingency table of the GHL versus the original horizon designations (Table 17.2). This shows the results of the pattern matching and should be examined to confirm whether the GHL assignments aggregate the soil horizons appropriately. For example, GHL that are labeled as "not used" did not match any of the given patterns and were not included in the data summaries. The user may in some cases wish to further examine these horizons and decide whether or not to refine the GHL rules to include/exclude them from the summaries.

Table 17.5 Number of GHL versus field textures

	cos	s	ls	lfs	si	fsl	l	sil	si	scl	cl	sicl	sc	sic	c	Sum
Ap	0	0	0	0	0	1	13	74	1	1	7	1	0	0	0	98
A	0	0	0	0	0	2	10	50	0	0	1	3	0	0	0	66
E	0	0	0	0	0	0	7	25	0	0	1	16	0	0	0	49
Bt	0	0	0	0	0	1	25	9	0	5	141	30	0	3	9	223
2Bt	0	0	0	0	0	0	8	0	0	0	22	6	0	0	0	36
2BCt	0	0	0	0	1	3	35	1	0	4	34	11	0	0	5	94
2Cd	0	0	0	0	0	0	27	1	0	0	1	0	0	0	0	29
Not-used	1	1	1	1	6	11	172	24	0	2	40	35	1	4	32	331
Sum	1	1	1	1	7	18	297	184	1	12	247	102	1	7	46	926

The values represent the frequency of occurrence (counts) for combinations of GHL and texture

Table 17.6 Number of GHL versus laboratory textures

	cos	si	fsl	l	sil	si	scl	cl	sicl	sc	sic	c	Sum
Ap	0	1	1	13	63	0	1	3	1	0	0	0	83
A	0	0	2	5	42	0	0	2	2	0	0	0	53
E	0	0	0	6	24	0	0	7	8	0	0	0	45
Bt	0	0	1	24	6	0	7	85	14	0	2	16	155
2Bt	0	0	0	4	1	0	0	5	2	0	0	1	13
2BCt	0	0	3	30	0	0	4	36	5	0	1	7	86
2Cd	0	0	1	8	0	0	0	1	0	0	0	0	10
Not-used	1	2	14	147	13	1	2	60	15	1	6	37	299
Sum	1	3	22	237	149	1	14	199	47	1	9	61	744

The values represent the frequency of occurrence (counts) for combinations of GHL and texture

As an example, the following tables and figures show excerpts from all the field and laboratory data labeled as the Miami soil series within NASIS (Tables 17.3, 17.4, 17.5, and 17.6) (Figs. 17.2, 17.4, 17.5, and 17.6). The example shows that the field estimates of clay content are missing for A horizons. Given the age of the data set, which ranges from 1951 to 2014, this is not surprising, as it has not always beencommon practiceto record field estimates for clay content. The laboratory data by comparison have numerous measurements of clay content. By examining the box plots, we can see a clay increase in the Bt and 2Bt horizons and a decrease in the 2Cd horizon. The box plots for pH show a wide interquartile range and a slight decrease in the median pH with depth. The subsoil (i.e., 2BCt and 2Cd) shows a much narrow interquartile range and higher median pH. Examining the contingency tables of GHL versus texture, we can see a greater frequency of silty textures in the A and E horizons (Table 17.5 and 17.6). The Bt horizon has a higher frequency of clay loam textures. If silty textures are indicative of the loess cap associated with the Miami soil series, numerous Bt horizons should be relabeled as 2Bt horizons. The report's summaries allow soil scientists to examine their data quickly particularly when the data are viewed in aggregate.

17.4 Conclusion

Here, we have presented an effort to efficiently analyze the large volume of soil horizon data present in the NASIS database. We have developed R Markdown reports that provide univariate summarizes of the data elements typically used to develop OSD and soil map unit components. Using the relational structure of the NASIS database combined with the extensible data handling and statistical analysis capabilities of R, it is possible to generate powerful graphical and tabular summaries for collections of pedon data bundled into one report. Summarizing pedon data by horizon is a critical and time-consuming step in the soil survey workflow. Because we can typically only investigate soil variability by examining several soil profiles and comparing multiple descriptions, viewing the data in aggregate allows us to approximate the representative values and ranges for soil horizons (i.e., polypedons), which are the building blocks of soil map unit components.

Acknowledgements The methodology presented here has benefited from the input from numerous individuals. Those that stand out include Alena Stephens and John Hammerly who assisted in testing different iterations of the reports, and Henry Ferguson, Paul Finnell, Carrie-Ann Houdeshell, and Samuel Indorante who provided valuable feedback.

References

Allaire JJ, Cheng J, Xie Y, McPherson J, Chang W, Allen J, Wickham H, Hyndman R (2015) rmarkdown: dynamic documents for R. R package version 0.6.1. http://CRAN.R-project.org/package=rmarkdown

Beaudette DE, Roudier P, O'Geen AT (2012) Algorithms for quantitative pedology: a toolkit for soil scientists. Comput Geosci 52:258–268

Beaudette DE, Skovlin JM (2015) soilDB: soil database interface. R package version 1.5–2. http://CRAN.R-project.org/package=soilDB

Fortner JR, Price AB (2012) United States soil survey databases. In: Huang PM, Li Y, Sumner ME (eds) Handbook of soil sciences: resource management and environmental impacts 2nd edn. CRC Press, pp 1–12

Jansen IJ, Arnold RW (1976) Defining ranges of soil characteristics. Soil Sci Soc Am J 40:89–92

Moeys J (2015) soiltexture: functions for soil texture plot, classification and transformation. R package version 1.3.3. http://CRAN.R-project.org/package=soiltexture

R Development Core Team (2015) R: a language and environment for statistical computing. R foundation for statistical computing. Vienna, Austria. ISBN3-900051-07-0

Schoeneberger PJ, Wysocki DA, Benham EC, Soil Survey Staff (2012) Field book for describing and sampling soils, Version 3.0. Natural resources conservation service, National Soil Survey Center, Lincoln, NE

Soil Survey Division Staff (1993) Soil survey manual. Soil conservation service. U.S. Department of Agriculture Handbook 18

U.S. Department of Agriculture (USDA) (2003) Natural resources conservation service. Technical note 4, populating map unit data: taxonomic classes and map unit componnents. http://www.nrcs.usda.gov/wps/portal/nrcs/main/soils/ref/. Feb 2013

U.S. Department of Agriculture (USDA) (2015) Natural resources conservation service. National soil survey handbook, title 430-VI. Available online. Accessed 24 June 2015

Xie Y (2014) knitr: a comprehensive tool for reproducible research in R. In Stodden V, Leisch F, Peng RG (eds) Implementing reproducible computational research. Chapman and Hall/CRC. ISBN 978-1466561595

Young FJ, Maatta JM, Hammer RD (1991) Confidence intervals for soil properties within map units. In: Mausbach MJ, Wilding LP (eds) Spatial variabilities of soils and landforms. SSSA Special Publication Number 28, Soil Science Society of America, pp 213–229