Probabilistic Representation of Genetic Soil Horizons

D.E. Beaudette, P. Roudier and J.M. Skovlin

# Abstract

Published soil survey reports typically describe soil series concepts in the form of aggregate information: ranges in soil properties, interpretations, and limitations that derived from a collection of field-described soil profiles. While aggregate soil properties are readily estimated via standard statistical functions (mean, median, etc.), an aggregate representation of horizonation (e.g. genetic or functional horizon designation and depth) is typically difficult to construct. Variation in horizon designation among different soil scientists and different soil description systems, changes in horizon designation standards over time, variable depths at which horizons occur, and the various uncertainties associated with these are all factors that complicate the process of delivering an aggregate representation of horizonation. In this chapter we propose alternatives to the typical "representative profile", e.g. the selection of a single soil profile to represent a collection. Two possible methods for aggregating a collection of soil profiles into synthetic profiles are presented, describing depth-wise probability functions for each horizon. Both methods rely on an expert-guided description of generalized horizon designation (e.g. a subset of horizon designation labels that convey a reasonable "morphologic story") along with associated rules (regular expression patterns) used to correlate field-described to generalized horizon designation. The first method is based on (1-cm interval) slice-wise evaluation of generalized horizon designation; the second is based on a proportional-odds logistic regression model fit to depth-slices. These methods are demonstrated using USDA-NRCS soil survey data (USA).

# Introduction

Published soil survey reports typically describe soils in terms of *aggregate* information, *i.e.* soil properties, interpretations, and limitations that are based on a collection of field-described soil profiles. While aggregate soil properties are readily estimated via standard statistical functions (mean, median, etc.), an aggregate representation of *horizonation* (e.g. genetic horizon designation and depth) is typically difficult to construct (Beaudette et al., 2013). Variation in horizon designation "style" among different soil scientists, changes in horizon designation standards over time, variable depths at which genetic horizons occur, and the possible lack of a specific genetic horizon are all factors that complicate the process of delivering an aggregate representation of horizonation. The process of designating horizons by soil scientists can be somewhat subjective; even a second description of the same volume of soil can lead to a slightly different set of horizon designations and depths (Holmgren, 1988). In addition to human sources of variability, it is understood that most of the variation between profile descriptions is due to real differences between soils observed at different locations (Wilding et al., 1964).

This complex combination of variability in morphologic horizon designation and depths is rarely acknowledged at the series or component level: boundaries between horizons, expressed as horizon depths, are generally considered as "crisp" numbers, while in actuality they represent "fuzzy" numbers due to the varying distinctness of the horizon boundaries and how abruptly characteristics change at horizon boundaries.

Soil profiles and their corresponding soil horizons represent a record of soil formation and encapsulate significant information about soil morphology. Although new tools and technologies may make continuous-depth measurements of soils possible, horizon designations, have historically been the common pedological language used to annotate observations of changes in soil properties with depth (Hartemink and Minasny, 2014; Myers et al., 2011; Kempen et al., 2011).

The soil survey programs of many countries have historically used the "modal pedon" or “modal soil” concept to convey a reasonable example of morphologic central tendency. Several authors have expressed concern with this approach (Jones, 1959; Hudson, 1990) due to the loss of information on a complex natural body that exhibits continuous gradation in space. While the “modal pedon” concept fails as an aggregate representation of a collection, it does offer the user of soil survey a concrete example (of one possible realization) that can be visited and sampled as needed.

We demonstrate two possible methods for aggregating a collection of soil profiles into "representative synthetic profiles"; describing depth-wise probability functions for each genetic horizon. Both methods rely on an expert-guided description of generalized horizon designation (e.g. horizon designations that are deemed representative) along with associated rules (regular expression patterns) used to correlate field-described to generalized horizon designation. The first method is based on (1-cm interval) slice-wise evaluation of generalized horizon designation; the second is based on a proportional-odds logistic regression (McCullagh, 1980) model fit to depth-slices. Specialized classes for soil profile collections and depth-slicing algorithms are implemented in the [aqp](http://cran.at.r-project.org/web/packages/aqp/index.html) package for R.

# Materials and Methods

## Soil Profile Data

A collection of (63) soil profiles from the Sierra Foothill Region of California were used to demonstrate two approaches for determining aggregate representation of genetic horizon boundaries. This collection of soil profile data represents the work of 13 different soil scientists, with described properties spanning ranges in physical properties (mostly related to bedrock composition) and are included within the **soilDB** package for R (Beaudette and Skovlin, 2015). These soils are associated with the [Loafercreek](https://soilseries.sc.egov.usda.gov/OSD_Docs/L/LOAFERCREEK.html) soil series (fine-loamy, mixed, super-active, thermic ultic haploxeralfs); moderately deep soils formed in colluvium and residuum from metavolcanic rocks (greenschist) (Figure 1). The climate is characterized by hot, dry summers and cool, wet winters. Mean annual air temperature is approximately 16 degrees C and mean annual precipitation is 760 mm. The native vegetation is blue oak and annual grass savannah. Land uses for this soil series include range, vineyards, recreation, and wild life habitat.



Figure 1. Eight photos of the Loafercreek soil series, collected in Tuolumne and Calaveras counties, CA, USA. How would you combine the wide range in morphology from these profiles into an aggregate concept?

The methods described in this paper are based on field descriptions: observations based on (experienced) visual and tactile investigation of the soil profile. Given sufficient laboratory characterization data, these same methods could be refined to use a combination of field and lab data.

## Horizon Generalization

Generalized horizon labels (GHL) represent an expert-guided selection of horizon designations that were consistently observed in the field, and meaningful in terms of soil morphology and management. These designations were determined to convey the "morphologic story" or conceptual framework of most-likely horizons typically observed in a suite of soil profiles associated with a specific soil series or map unit soil component. The Official Series Description or OSD (Soil Survey Staff, 2015) of the Loafercreek series typical pedon and range in characteristics defines this soil series concept. In this case, the OSD provided a useful GHL template, however, older OSDs or those based on a very limited set of data may not adequately convey an appropriate morphologic story.

Once a set of GHL have been determined (in the case of the sample dataset: Oi, A, BA, Bt1, Bt2, Bt3, Cr, R), it is necessary to create and apply a set of rules that map the field-described designations to corresponding GHL. When working with a set of pedons that have been described by a small number of individuals over a short period of time (i.e. consistency in both designation application and standards) it is possible to use a regular expression (REGEX) pattern matching to apply GHL. This process typically requires expert-guided review of: 1) regional patterns in horizonation style, 2) morphologic property differences by groups of field-described designation, and, 3) patterns of horizonation and properties with depth. We used a combination of field-described clay content, rock fragment volume, moist Munsell value, and horizon mid-point to evaluate GHL assignments and determine the final set of REGEX rules. Due to this iterative process, local experience with these soils and their properties are (mostly) preserved within the REGEX rules and corresponding GHL. It should be noted that there are some cases where pattern matching alone is not enough and manual adjustment of GHL on a horizon-by-horizon basis are needed. For simplicity, only REGEX-based assignment of GHL was used in this study.

At present there are limited means of capturing this type of soil horizon "micro-correlation" information developed in the application of GHL to soil horizon data. The authors suggest that future studies maintain a record of original horizon designations, generalized horizon labels suitable for aggregation, and the rules used to apply these labels. Such a record would be useful should more data on a soil be collected or laboratory data be included in the horizon data set. A convenient, quantitative evaluation of GHL assignments can be performed using the silhouette width metric (Rousseeuw, 1987). This metric, commonly used to assess clustering labels, provides a simple metric that can be used to address the basic question of GHL assignment: "given a set of data and labels, how well do these labels split differences within the data?". A more detailed description of this approach has been documented in chapter **(???)**.

### Aggregation of Generalized Horizon Labels

Aggregation of horizons as defined by GHL was performed using empirical probabilities, estimated along regular depth-slices from 0–150 cm (Beaudette et al., 2013). The "sliced" GHL data were then aggregated using proportional odds logistic regression (Figure 2). All computation was performed with the R package for statistical computing (R Core Team, 2013).

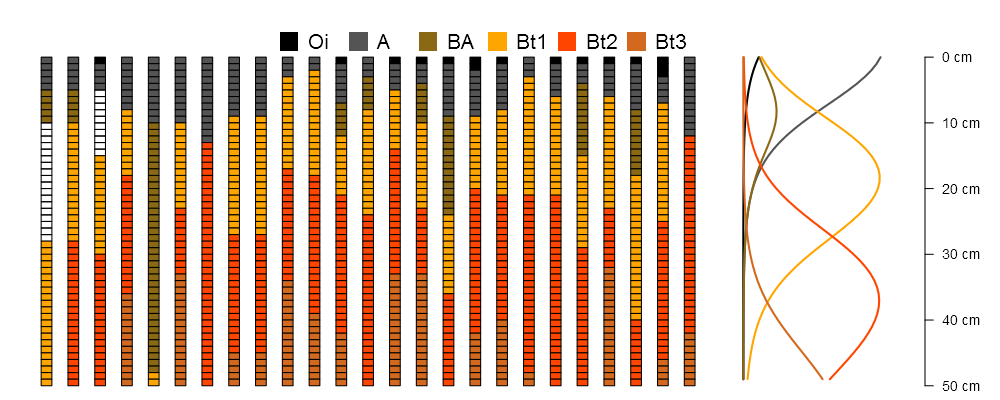


Figure 2. Demonstration of some Loafercreek soil profiles sliced into 1-cm chunks, colored by GHL, and associated probability estimates from the fitted PO-LR model.

A sequence of morphologic soil horizon designations can be modeled as an ordinal-scale variable: categorical by definition and ordered along a logical gradient, depth. Within the set of GHL associated with our sample data, "Bt2" horizons always occur after "Bt1" horizons and before "Bt3" horizons. The proportional odds logistic regression model (cumulative link model with logit link) (McCullagh, 1980) is a convenient framework for estimating the probability of encountering a GHL, as conditioned by depth. The proportional odds logistic regression (PO-LR) model can be defined as:

where is the estimated probability of encountering GHL , is a set of predictor variables, and a vector of fitted regression coefficients (Harrell, 2001). In this study, the PO-LR model was fit to "sliced" horizon data; 1-cm slices of GHL and slice top depth (Figure 2). Restricted cubic spline (RCS) basis functions (Harrell, 2001; Hastie et al., 2009) with 4 knots located at the 5th, 35th, 65th, and 95th percentiles of slice top depth were used to accommodate non-linearity. An empirical index of model stability was calculated by repeatedly re-fitting the PO-LR model to 25 randomly sampled profiles (out of 54 total), 250 times.

## Most-Likely Horizon Boundaries

Continuous estimates of GHL probability with depth are a convenient approach to communicating variability; however, there are still cases where discrete horizon depth information is required. For example, the USDA-NRCS Official Series Description pages are used by a wide range of individuals that may not need this level of detail. We used a simple strategy for converting these depth functions into a discrete set of "most-likely" GHL boundary depths. At each depth slice, the GHL with the highest probability is selected. Most-likely boundary depths are determined by locating upper and lower depths from contiguous sets of slices that share a common GHL. Within a collection of highly similar pedons, the most-likely boundary depths roughly correspond to crossings of the GHL probability depth functions.

## Quantification of Uncertainty

We used Shannon Entropy to quantify the relative amount of information present within GHL predictions at any given depth. Shannon Entropy was calculated according to (Kempen et al., 2009):

where is an index of entropy associated with predicted probabilities, , of encountering GHL through at any given depth. Values range from 0 (maximum information, minimum entropy) to 1 (minumum information, maximum entropy). Entropy values were computed along each 1-cm depth slice from predictions generated by the PO-LR model.

We used Brier scores (Harrell, 2001) to quantify agreement between assigned GHL and probabilities of predicted GHL:

where is an index of agreement between predicted probabilities, , and observed horizons,, over depth-slices through associated with a specific horizon. Larger values suggest less agreement between probabilities and observed horizon labels.

# Results

## Generalized Horizon Labels

A graphical representation of the association between field-described horizon designation and associated GHL is presented as a box and whisker plot in Figure 3. Assignment of GHL to the top (A) and bottom-most (Cr and R) genetic horizons by REGEX pattern matching resulted in the most internally-consistent groups of data. Transitional horizons near the surface (AB, BA, etc.) and lower Bt horizons (2Bt3, Bt4, BCt, etc.) were generally the most variable and thus difficult to place within a GHL by pattern matching. Further investigation of select soil properties and associated ranges (Table 1) made it possible to refine REGEX rules.

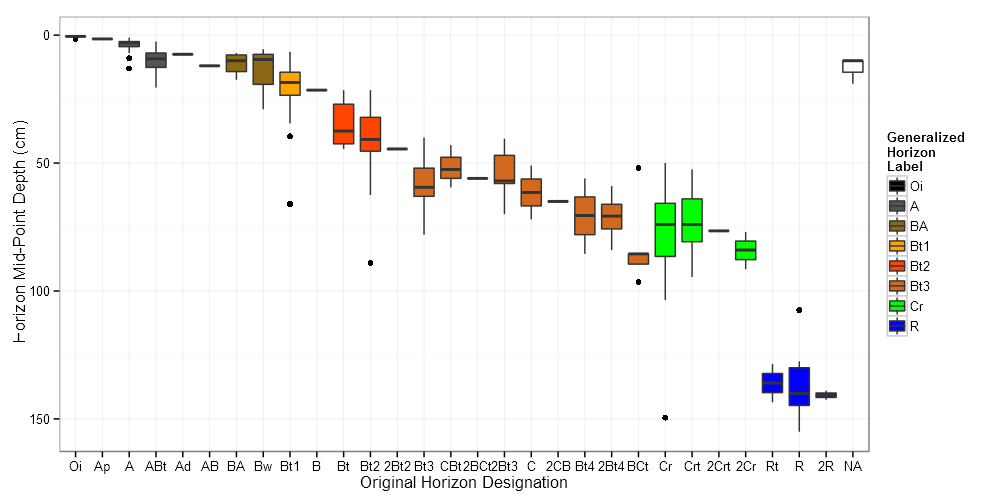


Figure 3. Original horizons designations (x-axis), GHL assignments (colors), and associated ranges in depth.

The degree of overlap in GHL concepts can be expressed in terms of measured soil properties (in this case a limited set of field-described properties), summarized by GHL (Table 1). The relatively low silhouette width values suggest that manual adjustment of GHL assignments may be required.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| GHL | Clay (%) | Horizon Mid-Point (cm) | Total RF\* Volume (%) | Moist Munsell Value | Silhouette Width |
| Oi | -- | 0.69 (0.37) | -- | -- | -- |
| A | 15.68 (3.34) | 4.17 (3.19) | 6.51 (7.29) | 3.28 (0.6) | 0.16 (0.17) |
| BA | 17.57 (3.74) | 12.07 (6.16) | 9.71 (6.5) | 3.57 (0.76) | -0.12 (0.1) |
| Bt1 | 21.43 (4.54) | 19.87 (9.29) | 12.67 (12.51) | 3.72 (0.56) | 0.02 (0.14) |
| Bt2 | 25.26 (4.98) | 39.63 (11.27) | 24.27 (21.76) | 4 (0.72) | -0.06 (0.14) |
| Bt3 | 28.61 (6.33) | 60.94 (13.56) | 35.02 (23.9) | 4.37 (0.6) | 0.06 (0.14) |
| Cr | -- | 76.96 (16.41) | -- | -- | -- |
| R | -- | 137 (11.54) | -- | -- | -- |

Table 1. Evaluation of GHL via field-described soil properties. Reported values are means with standard deviation in parenthesis. Values marked as "--" are the result of missing or insufficient data. \*RF: rock fragment percent by volume

## Probabilistic Representation of GHL

A graphical comparison of empirical and PO-LR predicted GHL probabilities is presented in Figure 4. The empirical probability curves are an exact representation of the 54 pedons within our sample data set, however, these curves are not likely a generalized representation of all possible soils correlated to the Loafercreek series. At the expense of a small amount of accuracy (as evaluated using the sample data set), the smoother and more generalized shape of the PO-LR derived GHL probabilities are better candidates for describing the central tendency of a soil series concept (Figure 5). When samples sizes are too small to support fitting a stable PO-LR model, the empirical probabilities can provide a reasonable alternative.

The PO-LR probabilities were the least accurate within the very thin Oi (Brier Score of 1.40) and infrequently occurring BA horizons (Brier Score of 1.22). Accuracy was greatest in the most consistently defined horizons which were not surprisingly found at the “top” (A horizons) and “bottom” (R horizons) of the profiles. The degree of overlap in GHL concepts was greatest (as defined by Shannon Entropy) near the surface where A, BA, and Oi horizons spanned similar depth ranges, and lower in the profile where Bt3 and Cr horizons spanned large ranges in depth (Figure 5).

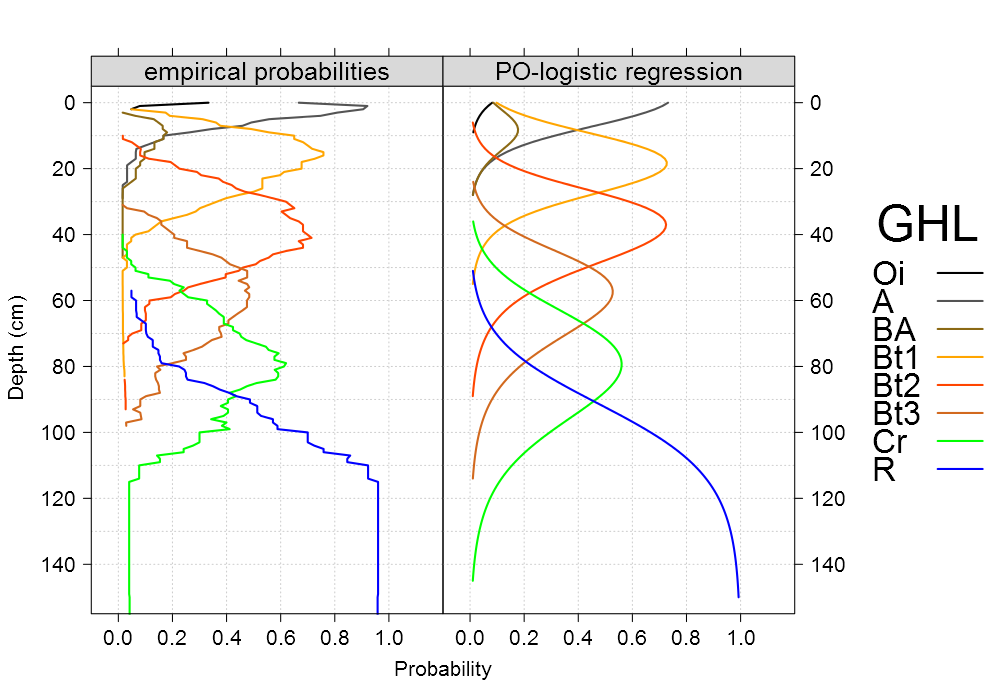


Figure 4. Comparison of empirical GHL probabilities, evaluated over 1cm depth slices, and predictions from the PO-LR model. Probabilities less than 0.01 have been removed for clarity.

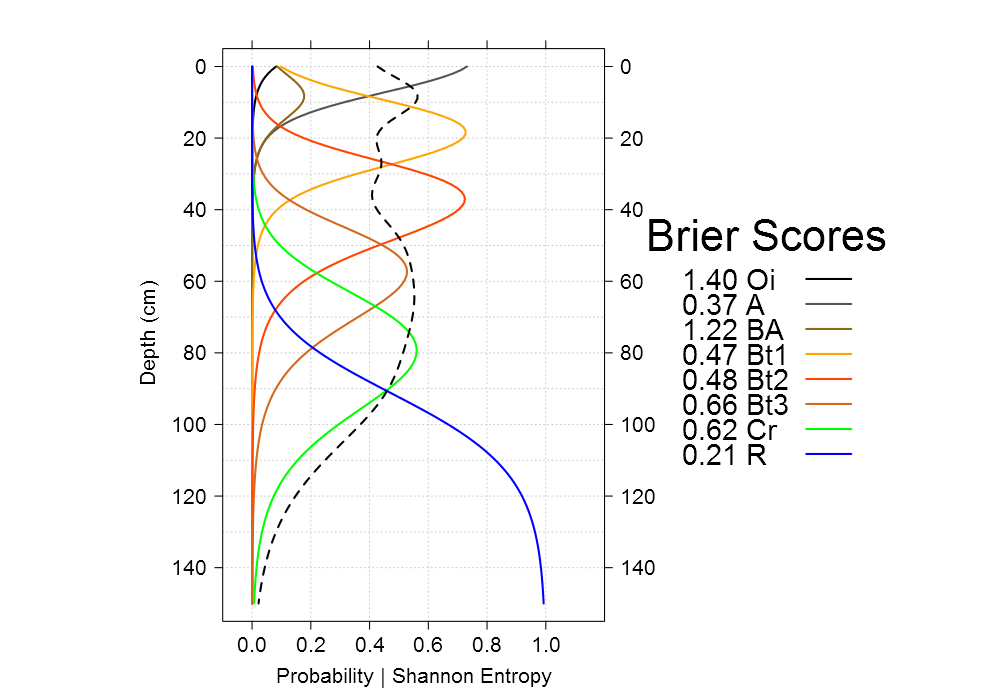


Figure 5. PO-LR predicted GHL probabilities (solid lines), Shannon entropy (dashed line), and associated Brier scores (values printed in legend).

## Model Fit and Stability

The fitted PO-LR model had a reasonably high coefficient of determination (). Removal of RCS basis functions from the model resulted in an of 0.79. Deviations between empirical and modeled probabilities were greatest in horizons near the surface and smallest in the lower-most horizons (Figure 6). Discrepancies between the two sets of probabilities can be attributed to two main factors: 1) lack of model fit, and 2) generalization (e.g. smoothing) of empirical probabilities by the PO-LR model.

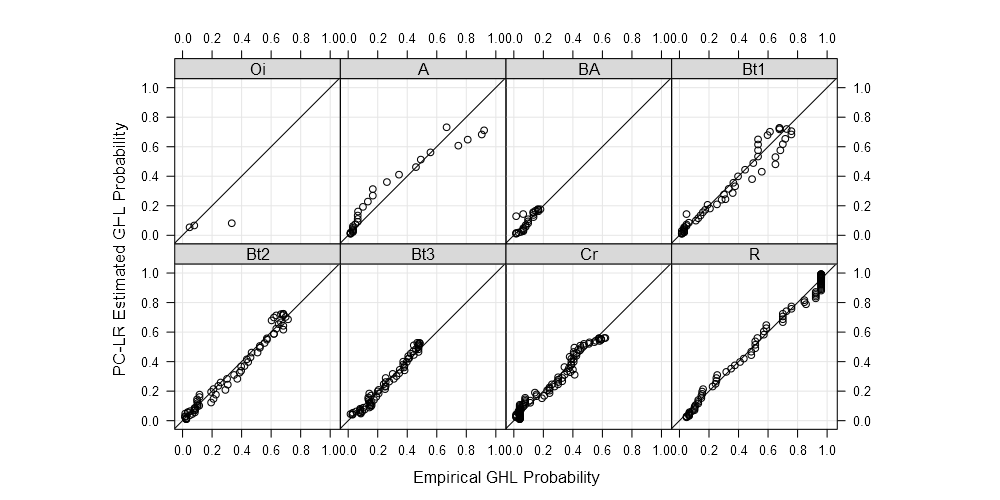


Figure 6. Scatterplot comparison of empirical GHL probabilities and PO-LR estimated GHL probabilities. Solid lines represent a 1:1 agreement.

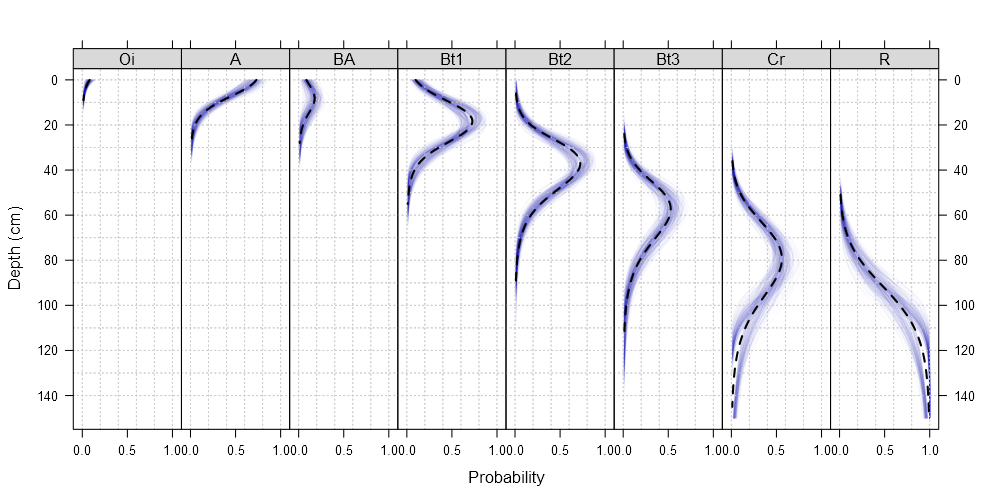
The stability of the PO-LR model was evaluated by iteratively re-fitting the model (250 times) using a random subset of 25 pedons (out of 54 total) within each iteration. The predictions from each iteration are presented in Figure 7. Mean model was 0.89 and ranged from 0.81 to 0.91. Variation between iterations appears to results in a range in predicted probabilities of about 0.2 probability units near the peaks associated with each generalized horizon label. The combination of predictions from the full model combined with many realizations of a reduced model could be a useful way to convey uncertainty in predictions of GHL probabilities at any given depth.

Figure 7. Predicted GHL probability depth-functions from 250 iterations of model fitting, based on a reduced training dataset. Dashed lines are predictions from the full model.

## ML Horizon Depths

The “most-likely” (ML) horizon depths extracted from empirical probabilities were quite similar to those extracted from PO-LR model predictions (Table 2). ML horizon depths represent one possible way in which probabilistic estimates of GHL occurrence can be simplified into a format that is more familiar to users of existing soil survey products. ML horizon depths could also serve as a template by which aggregate soil properties (clay, pH, CEC, etc.) are organized within soil survey reports. The Brier scores (Table 2) serve as an indication of how well each set of ML horizon depths fits the original collection of pedons. For example, predictions associated with the ML horizon depths for “A” horizons more consistently overlap with field-observed “A” horizons as compared to “Bt3” horizons. In all cases except “Bt3” horizons, Brier scores associated with the PO-LR model were equal to or greater than those associated with empirical probabilities; not surprising as predictions from the PO-LR model are much smoother than the over-fit empirical probabilities. The similarity in ML horizon depths and small differences in Brier scores suggests that the PO-LR model is a reasonable generalization of the GHL concepts defined for this collection of pedons.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Empirical Probabilities | | |  |  | PO Logistic Regression |  | |  |
| Horizon | Top (cm) | Bottom (cm) | | Brier | Horizon | Top (cm) | Bottom (cm) | Brier |
| A | 0 | 8 | | 0.20 | A | 0 | 9 | 0.26 |
| Bt1 | 8 | 28 | | 0.23 | Bt1 | 9 | 28 | 0.23 |
| Bt2 | 28 | 51 | | 0.25 | Bt2 | 28 | 50 | 0.23 |
| Bt3 | 51 | 68 | | 0.44 | Bt3 | 50 | 67 | 0.39 |
| Cr | 68 | 90 | | 0.36 | Cr | 67 | 91 | 0.36 |
| R | 90 | 203 | | 0.05 | R | 91 | 151 | 0.08 |

Table 2. Most-likely GHL boundary depths and associated Brier scores, computed from empirical probabilities and PO-LR predictions.

# Conclusions

Describing and sampling soil by genetic horizon designations represents an efficient approach that has provided a common pedological language used among soil scientists and classification systems. Processes for aggregating horizonation and deriving an aggregate representation across suites of similar soil profile descriptions of a soil series has been challenging. For this reason, soil series concepts have historically been defined using the modal profile; a single, field-observed pedon selected as a “representative” demonstration of central tendency. Advances in soil morphometrics are poised to change our understanding of what it means to describe soil profiles. Continuous depth functions of soil properties will further our understanding of how soil properties vary with depth, adding rich content to the existing genetic horizon framework.  
In this chapter we have outlined a simple approach for deriving continuous depth-functions of groups of field-described genetic horizon probabilities. Correlation of horizon designation to a subset of generalized horizon labels is fundamental to this approach and represents a series of “micro correlation” decisions that could support a wide range of soil data aggregation tasks. The two aggregation methods described in this chapter yield similar results; selection of an appropriate method depends on sample size, use of empirical probabilities is recommended for small collection of pedons and the PO-LR probabilities is recommended for large collections.

The methods presented in this chapter represent the first steps towards a quantitative description of soil morpology data in aggregate. Future application of these methods will depend on development of guidelines related to: minimum sample sizes, PO-LR model fitting parameters, model diagnostics, and recommendations on pedogenic interpretation of model coefficients. In addition, more work needs to be done on incorporating *depth-wise* correlation structure into the PO-LR model to support more realistic estimates of coefficient standard errors.

# References

Beaudette, D.E., and J.M. Skovlin. 2015. *soilDB: Soil Database Interface*. <http://CRAN.R-project.org/package=soilDB>.

Beaudette, D.E., P. Roudier, and A.T. O’Geen. 2013. “Algorithms for Quantitative Pedology: A Toolkit for Soil Scientists.” *Computers & Geosciences* 52: 258–268.

Harrell, Frank E. 2001. *Regression Modeling Strategies*. Springer Series in Statistics. New York, NY: Springer.

Hartemink, A.E. and B. Minasny. 2014. “Towards Digital Soil Morphometrics.” *Geoderma* 230–231: 305–317. doi:[http://dx.doi.org/10.1016/j.geoderma.2014.03.008](http://dx.doi.org/http://dx.doi.org/10.1016/j.geoderma.2014.03.008). <http://www.sciencedirect.com/science/article/pii/S0016706114001177>.

Hastie, T., R.Tibshirani, and J. Friedman. 2009. *The Elements of Statistical Learning*. Springer.

Holmgren, G.G.S. 1988. “The Point Representation of Soil.” *Soil Sci. Soc. Am. J.* 52: 712–716.

Hudson, B.D. 1990. “Concepts of Soil Mapping and Interpretation.” *Soil Survey Horizons* 31: 36–72.

Jones, T.A. 1959. “Soil Classification–a Destructive Criticism.” *J. Soil Sci.* 10: 196–200.

Kempen, B., D.J. Brus, and J.J. Stoorvogel. 2011. “Three-Dimensional Mapping of Soil Organic Matter Content Using Soil Type Specific Depth Functions.” *Geoderma* 162: 107–123. doi:[http://dx.doi.org/10.1016/j.geoderma.2011.01.010](http://dx.doi.org/http://dx.doi.org/10.1016/j.geoderma.2011.01.010).

Kempen, Bas, Dick J. Brus, Gerard B.M. Heuvelink, and Jetse J. Stoorvogel. 2009. “Updating the 1:50,000 Dutch Soil Map Using Legacy Soil Data: A Multinominal Logistic Regression Approach.” *Geoderma* 151: 311–326. doi:[10.1016/j.geoderma.2009.04.023](http://dx.doi.org/10.1016/j.geoderma.2009.04.023).

McCullagh, P. 1980. “Regression Models for Ordinal Data.” *Journal of the Royal Statistical Society, Series B* 42: 109–142.

Myers, D Brenton, Newell R Kitchen, Kenneth A Sudduth, Randall J Miles, E John Sadler, and Sabine Grunwald. 2011. “Peak Functions for Modeling High Resolution Soil Profile Data.” *Geoderma* 166 (1): 74–83.

R Core Team. 2013. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <http://www.R-project.org/>.

Rousseeuw, P.J. 1987. “Silhouettes: a Grapical Aid to the Interpretation and Validation of Cluster Analysis.” *Journal of Computational and Applied Mathmatics* 20: 53–65.

Soil Survey Staff. “Official Soil Series Descriptions.” Edited by Natural Resources Conservation Service, United States Department of Agriculture. <https://soilseries.sc.egov.usda.gov/OSD_Docs/L/LOAFERCREEK.html>.

Wilding, L.P., G.M. Scafer, and R.B. Jones. 1964. “Morley and Blount Soils: A Statistical Summary of Certain Physical and Chemical Properties of Some Selected Profiles from Ohio.” *Soil Sci. Soc. Proc.* 28: 674–679.