Analyzing Within-County Hydrogeomorphological Characteristics as a Precursor to Phosphorus Index Modifications

Tamie L. Veith,* Sarah C. Goslee, Doug B. Beegle, Jennifer L. Weld, and Peter J. A. Kleinman

Abstract

Phosphorus (P) site assessment is used nationally and internationally to assess the vulnerability of agricultural fields to P loss and identify high-risk areas controlling watershed P export. Current efforts to update P site assessment tools must ensure that these tools are representative of the range of conditions to which they will be applied. We sought to identify key parameters available in public GIS data that are descriptive of potential source areas in Pennsylvania and that ensure that modifications of the P Index span all feasible parameter combinations. Relevant soil and topographic variables were compiled for Pennsylvania at 30-m resolution, and areas within 90 m of permanent streams were extracted. Within each county, k-means and classification trees were used to identify and create classification rules for topoedaphic groups. Within counties, two to five groups adequately represented near-stream complexity, with available water capacity, hydraulic conductivity, and organic matter being the most important environmental variables. Discontinuities across soil survey boundaries made it impossible to develop clusterings beyond the county level. For county-scale research and management efforts, these groupings provide a manageable approach to identifying representative sites for near-stream agricultural lands. The full set of representative sites across the state enables evaluation of the P Index throughout the full hydrogeomorphic diversity of Pennsylvania. In future work, we can then combine a set of reasonable management practices with each of the main hydrogeomorphological regions resulting from this study and verify the revised P Index against expert knowledge and simulation results.

Core Ideas

- Hydraulic properties and organic matter are key for grouping near-stream soils.
- Discontinuities across soil surveys prevent modeling regions across counties.
- Two to five groups per county are sufficient to classify most near-stream soils.
- Cluster analysis of environmental data provides a focus for revising the P Index.

The Phosphorus (P) Index, a farmer-level tool for assessing risk of P loss from agricultural fields, is used throughout the United States. However, specific aspects of the tool and the degree of scientific processes supporting each state’s version varies based on data available, political goals, and, of course, the hydrogeomorphic characteristics that influence farming activities in each state. In Pennsylvania, for which 50% of the state’s area comprises 40% of the drainage area of the Chesapeake Bay Watershed, the Pennsylvania P Index was developed to support P management decisions by farmers, balancing production priorities with water quality protection objectives.

In response to US Revision 590 Nutrient Management Standards (Sharpley et al., 2013), the Pennsylvania P Index (Weld et al., 2007) is being evaluated and revised according to state-of-the-science experimental knowledge and nutrient management stakeholder input combined with water quality simulation models (Sharpley et al., 2012; Cela et al., 2016). Although individual Pennsylvania P Index parameters have been evaluated, statewide assessment of the aggregate P Index values is necessary. The ability to evaluate P Index recommendations on landscape and management properties that are representative of Pennsylvania agriculture is critical to ensure that the variability inherent in soils, topography, and management conditions is captured (Sharpley et al., 2001; Kleinman et al., 2002; Veith et al., 2005).

Evaluation and revision of the P Index necessitates a comparative process-based assessment of each of the two major P Index components: (i) source impacts driven by actual and potential rotational management practices, and (ii) P transport risks driven by the range of soil, topographic, and hydrological (i.e., hydrogeomorphic) characteristics throughout the state’s farmland. The large number of potential parameter combinations makes it impossible to efficiently use simulation modeling to assess all scenarios of environmental conditions and management practices; some way to reduce the environmental variability to a manageable number of classes is required (Goslee et al., 2014). Statistical clustering methods applied to the set of soils...
and topographic variables used in the P Index can be used to identify county-scale classes that share similar characteristics. In future work, we can then combine a set of reasonable management practices with each of the main groups resulting from this classification so that we can efficiently verify the revised P Index against expert knowledge and simulation results.

In this study, we develop and demonstrate a robust method for classifying agricultural land into a computationally feasible quantity of categories (<10) uniquely described by soil and topographic properties available from public geospatial data. We tested a variety of statistical clustering and classification tree methods on the near-stream land in Pennsylvania, at both the county and state levels, to develop a process that would reduce the environmental complexity of the relevant soils and topographic data into a manageable number of statistically similar groups to facilitate research and management. Specifically, we sought to categorize Pennsylvania riparian land into groups of soil property and topography that jointly represent major P transport drivers at the county and state levels.

**Materials and Methods**

**Data Selection**

The Pennsylvania P Index (Weld et al., 2007) represents P transport through five factors: soil erosion loss, runoff potential, existence of subsurface drainage, contributing distance (i.e., distance from stream), and a modified connectivity factor to account for direct access to stream. The first P transport factor, soil erosion loss, is calculated as annual soil loss (t ha⁻¹) from the Revised Universal Soil Loss Equation (RUSLE) and varies yearly on the basis of crop management. Thus, to represent soil erosion loss, we considered the variables that comprise the RUSLE equation. Other than the rainfall erosivity factor of RUSLE, there are three management invariant factors: soil erodibility (K), field slope (S), and flow length (L) of uncontrolled flow. The K factor is important in calculating sediment detachment and movement within the field; S and L factors are both key drivers of sediment delivery from the field (Walling, 1983; Novotny and Olem, 1994). We selected soil erodibility factor (Kf), the US County-Level Soil Survey Geographical Database (SSURGO) soil survey representative of the RUSLE K factor, to estimate soil erodibility. We calculated percentage slope and flow length from the GIS elevation data to roughly estimate the RUSLE S and L factors. We recognize, however, that these 30-m estimations of percentage slope and flow length for use on a county- or statewide basis are highly uncertain representations of the measurements intended by the RUSLE S and L factors. In particular, although the uncontrolled flow length (L) can be measured in the field, it is difficult and controversial to estimate in GIS layers for the purpose of field erosion measurements, particularly when field size is less than the GIS resolution. It is not accurate to simply estimate slope length as the flow distance across the raster cell, especially with 30-m cells (Piechnik et al., 2012). Flow across the cell, particularly if across the diagonal, is likely to become concentrated over this distance in some of the raster cells.

The remaining four factors in the P transport section of the Pennsylvania P Index are categorical, estimated by the planner from tables based on general soil properties and site conditions. Although categories are useful in planning tools, we did not want to bias our classification results by using categorical values that were based on or could be represented by measurable physical properties. The second P Index transport factor, runoff potential, is based on the categorical SSURGO variables of surface runoff potential and drainage class. We represented these categorical variables by a total of five variables, which together indicate surface runoff potential and drainage class. Four of the variables, coming from SSURGO, were available water capacity (AWC), percentage clay particles, saturated hydraulic conductivity (Ksat), and percentage organic matter (OM). The fifth variable, a topographic wetness index (TCI), was calculated from GIS elevation data.

By restricting our analysis to a 90-m riparian area, we are considering the highest risk categories of the P Index “contributing distance” factor (i.e., factor values 4, 6, and probably a majority of value 2). We are excluding the land far enough away from the stream to have a 0 value for contributing distance. Because we do not have sufficiently detailed public data on the third and fifth P Index transport factors (subsurface drainage and modified connectivity), we are not considering manmade subsurface drainage or connectivity modifications, such as tile drains or grassed waterways.

**Data Source and Processing**

Statewide geospatial and associated attribute data, obtained from USDA-NRCS (2016), include 30-m resolution digital elevations (3D Evaluation Program digital elevation model, previously the National Elevation Dataset; USGS, 2016), county-level soil surveys (SSURGO; Soil Survey Staff, 2015), and blue-line stream networks (National Hydrography Dataset [NHD]; USGS, 2007). All data were processed and analyzed with open-source GIS and data analysis software GRASS and R (GRASS Development Team 2014; R Core Team 2016). Statistical significance was determined using t tests for two groups and ANOVA for three or more groups, and multiple comparisons were performed with Tukey’s all-pair comparisons (Hothorn et al., 2008). The weighted average values of the five soil variables (Kf, AWC, percentage clay, Ksat, and OM) were calculated for each county-level map unit based on component percentage. Values were calculated for the top 1 m of soil or to bedrock if the soil profile depth was <1 m and weighted by horizon thickness. Map unit values were rasterized at 30-m resolution to match the elevation grid. Percentage slope, flow length, and TCI were derived from the elevation data using GRASS routines r.slope.aspect, r.watershed, and r.terraflow, respectively (Toma et al., 2001).

All raster layers were clipped to a riparian buffer area extending roughly 90 m (three pixels) on either side of the streams defined by the NHD stream network. The decision was made to use all near-stream areas, rather than only those currently under agricultural management, because many nonagricultural areas nonetheless have the potential for that land use. However, the 7% of the near-stream raster cells that contained >20% soil organic matter (histosols) were excluded from the final dataset, as were very coarse sands with Ksat > 141.14 (0.04% of the near-stream raster cells), since neither soil type is generally suitable for agriculture. The final dataset for analysis contained 11.5 million 30-m raster cells and represents the entire population of near-stream areas in Pennsylvania.
Analysis Methods

Preliminary inspection of the soil variables showed substantial differences in value between soil survey regions. In Pennsylvania, soil survey regions correspond to county boundaries, the scale at which extension agents and resource professionals often operate. Thus, analyses proceeded by county (Drohan et al., 2003). Many statistical clustering methods require the calculation of a dissimilarity matrix, which is computationally prohibitive for large areas. After reviewing available statistical clustering methods, we chose two for this study: the fast and commonly used k-means method (Hartigan and Wong, 1979), and the slower Mclust, an implementation of model-based clustering for R (Fraley and Raftery, 2002). The two methods make different assumptions about the shapes of clusters in multivariate data, and little research has compared the usefulness of the two methods for toposequence data.

Data for each county were standardized to the range of each variable for that county, then randomly divided into 10,000-element subsamples. Both clustering methods were run for each county, with two to eight clusters per subsample. The partitioning methods, k-means and Mclust, require that the number of clusters to be fitted is specified in advance. The size of subsamples was chosen to be as large as feasible on the basis of computational overhead and computer capability. A typical cluster analysis metric, mean silhouette width, was used to determine the best number of clusters for each subsample within a county (Rousseeuw, 1987; Maechler et al., 2016). Cluster ensembles were created for the number of clusters that best fit the greatest number of subsets; the final ensemble clustering was the best grouping for that county (Hornik, 2005). The clusterings produced were different for each county.

Because the variable selection within these clustering algorithms is opaque, and so that we could compare clusterings between counties, we used classification trees to fit the final clustering results for each county to the original soils and topography dataset (Breiman et al., 1984; Therneau et al., 2015). The resulting tree distinguishes the importance of each of the eight variables included, making it possible to readily fit new samples into an existing classification, regardless of the source of that classification. Both the clustering methods and the classification tree approach are suitable for use with non-normal and potentially correlated variables.

The classification trees for each county were used to classify the entire state according to each county’s best clustering. The proportions of pixels that had two clusterings in common were used as the basis for a hierarchical clustering. The mean silhouette metric for this clustering was calculated for 2 to 25 groups. The dendrogram, pruned to the best number of groups, was used to combine very similar county-specific groups to produce the best merged statewide clustering. Similar groups of soils from different counties were combined, but the numbering of classes is arbitrary both between counties and for the state as a whole.

Results and Discussion

The faster and less resource-intensive k-means clustering worked better overall than Mclust in this study, based on silhouette values (0.395 for k-means and 0.348 for Mclust; t test p = 0.003), so only results from k-means are presented in further detail. Most individual counties were best fit with only two clusters (50 out of 67 counties, nine counties had three clusters, four had four clusters, four had five clusters). We selected one county from each of the three major physiographic provinces of Pennsylvania to present results: Lancaster from the Piedmont Province, Snyder from the Ridge and Valley Province, and Clarion from the Appalachian Plateau Province. The classification trees fit very well for each of these counties, with very low cross-validation error (Fig. 1a, 1b, and 1c, respectively by county). For two of the three counties, AWC was the primary variable defining the groups, with OM as the third. Where more than two splits were required, Ksat also defined groups. None of the other four variables evaluated (Kf, percentage clay, percentage slope, and slope length) contributed substantially to defining the groups. The spatial arrangements of the groups resulting from the clustering analysis generally followed patterns of topography and stream order (Fig. 2a, 2b, and 2c, respectively by county).

Lancaster County straddles the Conestoga Valley and Piedmont Upland physiographic subregions (Curran and Lingenfelter, 2016, p. 4). Soils typically have a silt loam texture, have a high AWC in the rootzone, are well drained with minimal rock fragments, and are capable of intensely productive agriculture. Lancaster near-stream land clustered into two groups. Class 1 had low AWC, OM, and Ksat (Fig. 1a) and occurred mainly in the ridges and slopes, where soils are prone to erosion and are expected to be relatively shallow (Fig. 2a). Lancaster’s second main hydrogeomorphic region (Class 2), occurring in the rich valley bottoms, was appropriately defined by high AWC, Ksat, and OM.

Snyder County (Fig. 1b and 2b), centered in the Ridge and Valley Province (Curran and Lingenfelter, 2016, p. 4), has distinct sandstone ridges with sandy loam soils that transition to a sandstone and shale mixture on the slopes. The clustering process captured this distinction, defining a near-stream cluster on the ridgetops by low AWC but high Ksat (Class 1), and defining a second cluster for near-stream areas on the slopes by both low AWC and low Ksat (Class 2). The k-means clustering also defined two distinct groups on the valley bottoms according to moderate or high levels of AWC (Classes 3 and 4, respectively, in Fig. 1b and 2b). Indeed, valley bottoms are either acidic and minimally productive shale-derived soils with lower AWCs or are deep, highly productive limestone-derived soils with high AWCs.

Clarion County (Fig. 1c and 2c) lies in the Allegheny High Plateau and upper portion of the Pittsburgh Plateau (Curran and Lingenfelter, 2016, p. 4). Soils are typically well drained with low to moderate AWC and substantial rock fragments. Agricultural productivity is further limited by the high elevation, mainly suited to pasture or potatoes (Solanum tuberosum L.). This county classified into two clusters on the basis of OM. Two-thirds of the near-stream areas in the county contained no more than 2.3% OM, whereas the soils adjacent to the largest river in the county contained >2.3% (Classes 1 and 2, respectively, in Fig. 1c and 2c). This provides a clear example of the ability of the study method to appropriately identify high phosphorus transport risk areas within the near-stream areas. An aerial close-up along a stream in Clarion County (USGS Digital Orthophoto Quadrangle, Fig. 3) shows an area where the low-OM Class 1 area along the stream is used primarily for agriculture, whereas the higher-OM Class 2 area has more tree cover. The southern stream branch in this figure shows an area where the classes split
Fig. 1. Classification trees for three counties in different physiographic regions of Pennsylvania: (a) Lancaster County, in the Piedmont Region; (b) Snyder County, in the Ridge and Valley Region; and (c) Clarion County, in the Appalachian Plateaus Region. The tree shows a dichotomous key to class types based on the most important input variables. Each node lists: the predicted class, the probability for each class member to be found at that node (an estimate of misclassification), and the percentage of total observations at that node. Class numbers and colors correspond to those in Fig. 2. awc, available water capacity; om, organic matter; ksat, saturated hydraulic conductivity.

Fig. 2. Mapped riparian clusterings for three counties in different physiographic regions of Pennsylvania: (a) Lancaster County, in the Piedmont Region; (b) Snyder County, in the Ridge and Valley Region; and (c) Clarion County, in the Appalachian Plateaus Region. County maps are shaded by elevation, such that darker shades indicate valleys and lighter indicate ridges. Class numbers and colors correspond to those in Fig. 1. Black inset box in Fig. 2c indicates close-up location depicted by Fig. 3.
along the north and south sides of the stream, both of which are agricultural fields.

Across all counties, it became clear that some variables were consistently more important in defining the county-level clusters, creating a significant difference in importance values among the variables (ANOVA $p < 0.001$). Three of the five soil properties—AWC, Ksat, and OM—were the most important, whereas the other two—percentage clay and Kf—were moderately important. The topographic variables of percentage slope, slope length, and TCI were rarely important, even though clustering tended to follow topographic boundaries.

The final stage of the classification procedure was to combine the best $k$-means clusters from each county into a statewide classification, which resulted in seven groups in the best-fitting statistical result (Fig. 4) but had little practical success because soil classification differences across counties caused greater variance between even adjacent counties than existed within groups. Broadly, the classification followed physiographic province regions, but the discontinuities at county boundaries were more pronounced than any geographic trend. Thus, the seven clustering groups do not translate back into feasible combinations of the grouping variables. Discrepancies across county boundaries are clearly due to the classification of soils across survey areas. In many cases, a single soil unit that crosses a county boundary was classified differently on each side of this artificial border. Some of the cross-boundary quantitative differences were so extreme that one of the seven statewide groups was found only in a single county (Clinton).

Although many of the soil survey regions were sampled by different people and at different times, we were unable to find any clear patterns related to either date of initial survey or identity of survey staff (Eckenrode and Ciolkosz, 1999).

The potential difficulties of working at spatial scales that cross soil survey boundaries have been previously noted (Drohan et al., 2003). The discontinuities that occur at survey boundaries have been addressed at the watershed scale by using soil taxonomic information to aggregate groups into a continuous soil layer (Gatzke et al., 2011; Luo et al., 2012). This approach was effective at the watershed scale but may not be feasible at state or regional scales. A second approach incorporates machine learning methods to reclassify SSURGO data on the basis of environmental covariates (Heung et al., 2016). A recently developed soil series probability map, created using machine learning methods and intended to be internally consistent and eliminate artificial discontinuities, shows considerable promise for improving state and regional modeling and classification efforts (Chaney et al., 2016).

Realistic management decisions often require that the overwhelming environmental complexity of multiple continuous environmental variables be simplified into a small number of classes. Representative groups make it possible to target monitoring and modeling efforts most effectively, and to make management recommendations based on quantitative similarities (Goslee et al., 2014). The identification, in this study, of county-level groups that represent potential source areas for P loss enables evaluation of the P Index throughout the full hydrogeomorphic

---

*Fig. 3. Close-up of near-stream classification overlaying an aerial photo (USGS Digital Orthophoto Quadrangle), in which darker shades depict forests and lighter indicates less densely covered agricultural land. The region of Clarion County shown here is marked by the black inset box in Fig. 2c.*
diversity of Pennsylvania while still requiring consideration of a small number of combinations of relevant properties.

Tools for broadly relevant environmental problems require the use of large, complex datasets like the soils and topographic information collected here. These datasets are computationally challenging to work with and require substantial modification in methods that are effective with small datasets. Source datasets, notably the SSURGO soils data, are assembled from smaller units (i.e., county surveys) collected at different times and following different protocols, causing artificial boundaries that make it difficult to investigate state and regional attributes.

Conclusions and Implications

Developing and testing a site assessment tool such as the P Index has historically resulted in underrepresentation of certain conditions. We developed remote sensing and statistical clustering techniques to ensure comprehensive representation of field site conditions in P Index testing within Pennsylvania. Across the state, AWC, Ksat, and OM influenced the statistical clustering results the most strongly; Kf, percentage clay, percentage slope, slope length, and TCI had the least impact. The tested variables were chosen for their importance to the transport processes within the P Index, so these groupings provide a manageable and relevant starting point for modeling of management effects on P risk within all counties across Pennsylvania. Although we cannot currently conduct P Index assessment across multiple counties, the rapidly advancing improvements in digital soil mapping should soon make this possible. The methods we developed for clustering, classifying, and combining large environmental datasets are readily extensible to other regions and datasets. Although computational abilities are improving dramatically and rapidly, data reduction strategies such as chunking and recombining large datasets will continue to be important. Quality issues in the underlying data are more complex, and remain substantially more difficult to address.

Disclaimer

Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the USDA or the Pennsylvania State University. The USDA and the Pennsylvania State University are equal opportunity providers and employers.

References


Fig. 4. Statewide statistical merging (seven groups) of the 67 county-level classifications within Pennsylvania. Similar soils groups were combined across counties, producing seven statewide classes.