VERIPATH
FARMLAND FUNDS

Soil Clustering for Farmland Portfolio Monitoring

Leveraging Recurring Agrology Data to Monitor Soil Health and Land Quality Across a Large-Scale Portfolio

April 2026

EXECUTIVE SUMMARY

Veripath Partners manages approximately 138,000 acres of non-operated Canadian farmland across multiple provinces, crop types and geographic zones. As a non-operator, Veripath does not farm the land directly but instead leases parcels to experienced farm operators under long-term agreements. This approach preserves farmland's characteristic low-volatility return profile while requiring robust systems to monitor the condition and productivity of each parcel over time.

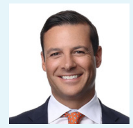
Taking a non-operating approach to farmland investing demands asset managers become experts in monitoring. A central element of Veripath's monitoring infrastructure is recurring agrology reporting. Professional agrologists conduct periodic soil sampling and analysis across the portfolio, generating a rich dataset of physical, chemical and biological soil indicators for each parcel. These indicators include organic matter content, macro and micronutrient levels, pH, electrical conductivity, cation exchange capacity, among others. Over multiple reporting cycles, this dataset grows into a multivariate time series spanning hundreds of parcels and dozens of measured variables.

The challenge for any farmland portfolio manager is translating this volume of agrological data into actionable intelligence. With potentially 20 to 40 measured soil parameters per parcel, and hundreds of parcels across diverse geographies, the raw data is too high-dimensional for simple visual inspection or threshold-based monitoring. Principal component analysis (PCA) offers an elegant statistical solution to this problem.

A combination of principal component analysis (PCA) and K-Means clustering offers an elegant statistical solution to this problem. This paper explains how Veripath applies this two-stage approach to its recurring agrology dataset within the TerraFIRST platform. PCA reduces the dimensionality of the soil data while preserving the dominant sources of variation, and K-Means clustering then partitions all observations into four discrete soil groups. This enables portfolio-level monitoring through a simple, actionable signal: tracking which parcels have changed cluster membership between agrology cycles. The framework also supports geographic and crop-type analysis, early identification of parcels experiencing quality changes, and integration with valuation and risk models.

The approach transforms what would otherwise be an unmanageable volume of agrological data into a structured, interpretable monitoring framework that supports both operational decisions and investor reporting.

Stephen Johnston
Director
sjohnston@omnigenceam.com



Matt Barr
Director
mbarr@omnigenceam.com

Keenan Viney
Data Scientist
kviney@omnigenceam.com

PRINCIPAL COMPONENT ANALYSIS: A PRIMER FOR INVESTORS

Principal component analysis is a well-established multivariate statistical technique that reduces a large set of correlated variables into a smaller number of uncorrelated dimensions, called principal components, that capture the majority of the variation in the original data. The technique has been widely adopted across scientific disciplines, from genomics to finance, wherever high-dimensional datasets require summarization without excessive information loss.

In the context of soil science, PCA has become a standard tool for constructing soil quality indices. Researchers at institutions including Cornell University, Oregon State University and the University of Alberta have used PCA extensively to identify minimum datasets of soil health indicators, reducing sets of 20 or more measured parameters to three or four composite indices that explain 70 to 85 percent of total variance. The technique is particularly well suited to soil data because many soil properties are intercorrelated. Organic matter, for example, is positively associated with cation exchange capacity, water holding capacity, aggregate stability and microbial activity. PCA captures these interrelationships efficiently.

In practical terms, PCA works by computing the eigenvectors and eigenvalues of the covariance (or correlation) matrix of the input data. The eigenvectors define the directions of maximum variance in the data space, while the eigenvalues quantify how much variance each direction captures. The first principal component (PC1) captures the largest share of total variance, PC2 captures the next largest orthogonal share, and so on. Typically, the first two to four principal components capture the substantial majority of meaningful variation, allowing the remaining components to be discarded as noise.

For a farmland portfolio, this means that a parcel described by 30 measured soil variables can be represented by its scores on two or three principal components without material loss of information**.** In TerraFIRST, PCA serves as a preprocessing step that makes subsequent clustering computationally tractable and more robust.

K-MEANS CLUSTERING: FROM CONTINUOUS SCORES TO DISCRETE GROUPS

While PCA reduces dimensionality, it still produces continuous scores that require interpretation. To transform these scores into actionable intelligence, TerraFIRST applies K-Means clustering to the PCA-reduced data, partitioning all soil observations into four discrete groups.

K-Means is an unsupervised machine learning algorithm that minimizes within-cluster variance. The algorithm iteratively assigns each observation to the cluster whose centroid (mean point) it most closely resembles, then recalculates centroids until assignments stabilize. The result is a partition of the data into k groups, where observations within each group are more similar to each other than to observations in other groups.

TerraFIRST employs the k-means++ initialization method, which selects initial cluster centers in a way that accelerates convergence and produces more stable, reproducible results than random initialization. A fixed random seed ensures that cluster assignments are identical across model runs, supporting auditability and backward comparability.

The choice of four clusters is derived using the elbow method; this value was distinct and stable in the dataset. Four groups provide enough differentiation to capture meaningful soil variation across prairie farmland – distinguishing, for example, high-fertility black soils from alkaline brown-zone soils – while remaining few enough that portfolio managers and agrologists can develop operational intuition for each profile.

Crucially, each cluster centroid is translated back to original soil units through inverse transformation of both the PCA and standardization steps. This produces a characteristic “soil profile” for each group expressed in familiar agronomic terms (percent organic matter, ppm phosphorus, pH units, and so forth). Portfolio managers can therefore describe Cluster 1 as “high-fertility, neutral-pH soils typical of black soil zones” and Cluster 3 as “moderate-fertility, alkaline soils common in southern brown zones” rather than working with abstract numerical scores.

INTEGRATION WITH TERRAFIRST

TerraFIRST is Veripath’s proprietary farmland operational and investment management platform. It encompasses a factor-driven portfolio construction model, a suite of investment screening tools, and a land management and reporting system that includes automated monitoring and farming practices verification. The platform channels data from satellites, AI-driven crop analysis, online farmer reporting and recurring agrology into a central data repository with a business intelligence layer containing over 200 key performance indicators and millions of data records with verified ground truth.

The agrology data pipeline within TerraFIRST collects soil sample results from certified agrologists, standardizes measurements across laboratories and sampling protocols, and stores the results in a structured format indexed by parcel, sampling date, depth and geographic coordinates. When a new round of agrology reports is received, the PCA model is refreshed to incorporate the latest observations.

Data Standardization

Before applying PCA, raw soil measurements must be standardized. Different soil parameters are measured on vastly different scales. Organic matter is expressed as a percentage (typically 2 to 8 percent for prairie soils), while phosphorus may be measured in parts per million (ranging from 5 to 80 ppm) and pH operates on a logarithmic scale between 4.5 and 8.5. Without standardization, variables measured on larger numerical scales would dominate the principal components regardless of their informational content.

TerraFIRST applies z-score normalization to each variable, subtracting the portfolio mean and dividing by the standard deviation, so that all parameters contribute equally to the analysis.

Each parcel in Veripath's portfolio is sampled at multiple points (typically ten) to account for within-field variability. Before cluster assignment, TerraFIRST aggregates these sample points into a single parcel-level centroid by averaging across all points for a given parcel and sampling year. This centroid represents the parcel's overall soil condition and is the unit of analysis for clustering. The aggregation ensures that cluster assignments reflect the parcel rather than being influenced by individual outlier samples.

Component Interpretation

The principal components generated from Veripath's agrology data are interpreted by examining the loading vectors, which indicate how strongly each original soil variable contributes to each component. In a typical prairie farmland dataset, the first component (PC1) tends to load heavily on organic matter, cation exchange capacity, total nitrogen and water holding capacity, and may be interpreted as an overall soil fertility index. The second component (PC2) often captures variation in pH, electrical conductivity and salinity-related parameters, reflecting the alkalinity and salt balance dimension that is particularly relevant in Alberta and Saskatchewan soils. A third component (PC3) may isolate micronutrient variation or soil texture differences**.** These two components form the feature space in which K-Means clustering operates.

**Table 1: Illustrative PCA loading structure for prairie farmland soil parameters.
Actual loadings are portfolio-specific and recalculated with each agrology cycle.**

Soil Parameter	Typical PC1 Loading	Typical PC2 Loading	Measurement Unit
Organic Matter	High (+)	Low	% by weight
Cation Exchange Capacity	High (+)	Low	meq/100g
Total Nitrogen	High (+)	Low	ppm
Available Phosphorus	Moderate (+)	Low	ppm
Potassium	Moderate (+)	Low	ppm
Soil pH	Low	High (+)	pH units
Sulphate	Low	Moderate (+)	ppm

Cluster Profiles

Once K-Means clustering is applied to the PCA-reduced data, each of the four cluster centroids is inverse-transformed back to original soil units. This produces interpretable soil profiles that describe what each cluster "looks like" in agronomic terms. The table below provides an illustrative example of how cluster profiles might appear for a prairie farmland portfolio.

**Table 2: Illustrative K-means clusters with average soil values.
Actual clusters are portfolio-specific and recalculated with each agrology cycle.**

Cluster	Organic Matter (%)	pH	Available P (ppm)	CEC (meq/100g)	Interpretation
0	5.8	6.9	42	28	High-fertility black soils
1	4.1	7.6	31	22	Moderate-fertility dark brown soils
2	3.2	7.9	18	16	Lower-fertility brown soils
3	2.8	8.2	12	14	Alkaline, lower-fertility soils

These profiles enable agronomists and portfolio managers to understand each cluster in operationally meaningful terms, facilitating communication with farm operators and supporting targeted management interventions.

PORTFOLIO MONITORING APPLICATIONS

Cluster Migration Tracking

The primary monitoring signal in TerraFIRST's soil analysis framework is cluster migration: has a parcel changed its cluster assignment between agrology cycles? Because each parcel is assigned to one of four discrete groups at each sampling period, the system maintains a year-over-year cluster history for every parcel in the portfolio.

A parcel that moves from Cluster 0 (high-fertility) to Cluster 2 (lower-fertility) across successive cycles represents a clear signal warranting investigation. Conversely, a parcel moving from Cluster 3 to Cluster 1 may indicate successful soil improvement efforts by the farm operator.

This discrete signal is more actionable than continuous score monitoring. Rather than interpreting whether a 0.3-point decline in PC1 score is meaningful, portfolio managers receive a binary flag: the parcel either changed groups or it did not. Parcels that have changed groups are escalated for review; parcels that remain in the same cluster are presumed stable.

The cluster migration approach aligns with Veripath's operational model as a non-operator. Lease agreements include soil stewardship provisions, and cluster migration provides clear, defensible evidence for discussions with farm operators about land management practices.

Soil Health Trend Analysis

In addition to cluster migration, TerraFIRST tracks the underlying centroid values for each parcel over time. By computing each parcel's cluster-transformed centroid values at each agrology cycle, TerraFIRST generates a time series of soil condition indi-cators for every parcel in the portfolio.

A parcel that remains within the same cluster but shows directional movement toward the boundary of an adjacent cluster may be an early indicator of future migration. These secondary signals support proactive intervention before a parcel formally changes groups.

Because Veripath is a non-operator, the identification of negative soil health trends is operationally significant. Lease agree-ments typically include soil stewardship provisions, and recurring agrology data provides the evidentiary basis for enforcing those provisions. The clustering framework enhances this process by flagging parcels where the latent soil type has shifted in a statistically meaningful way, even when individual parameters remain within conventional ranges. A parcel may show organic matter at 4.2 percent (within normal range) and pH at 7.8 (within normal range), yet its PCA score may reveal a combination of subtle declines across multiple parameters that collectively indicate a negative trajectory.

Geographic and Cluster-Based Analysis

Cluster assignments can be visualized geographically to reveal spatial patterns within the portfolio. Parcels are color-coded by their cluster membership and plotted on maps, allowing portfolio managers to identify whether certain geographic regions are dominated by particular soil groups.

These patterns inform both acquisition strategy and risk management, as parcels within the same cluster are expected to share similar return characteristics and weather sensitivities.

Cluster Change Detection and Early Warning

The most operationally valuable feature of the clustering framework is its ability to automatically flag parcels requiring atten-tion. Each time the model is refreshed with new agrology data, TerraFIRST compares each parcel's current cluster assignment against its assignment from the previous cycle. Parcels with changed assignments are automatically escalated for review.

This approach is particularly important at scale. With 138,000 acres across hundreds of parcels, it is impractical to review every agrological report in detail. Cluster-based screening allows the land management team to focus investigative resources on the parcels most likely to require intervention, significantly improving operational efficiency.

Cluster changes may indicate a range of underlying conditions: deterioration due to inadequate crop rotation or fertiliser application, improvement resulting from enhanced management practices, contamination, drainage problems, compaction from heavy equipment, or operator non-compliance with farming practices requirements. The cluster change flag identifies that something has shifted; diagnostic follow-up determines the cause.

Integration with Valuation and Risk Models

Soil health is a fundamental driver of farmland productivity and, by extension, farmland value. Veripath's valuation framework incorporates productivity metrics alongside comparable transaction data and income capitalization approaches. Cluster as-signments and cluster-level soil profiles provide a structured, quantitative input to this framework.

A parcel that migrates from Cluster 1 to Cluster 3 over successive cycles may signal potential impairment risk. This is relevant for net asset value reporting, where parcel-level valuations must reflect current and expected future productivity.

The integration of clustering with valuation models ensures that soil health information is reflected in portfolio valuations in a systematic, defensible manner.

METHODOLOGY AND WORKFLOW

The PCA and clustering-based monitoring workflow within TerraFIRST follows a structured sequence that begins with data collection and concludes with actionable reporting.

Step 1: Agrology Data Collection. Certified agrologists conduct soil sampling across the portfolio on a recurring cycle. Samples are taken at standardized depths (typically 0 to 15 cm and 15 to 60 cm) and submitted to accredited laboratories for analysis. The laboratory returns a panel of 20 to 40 measured soil parameters per sample. Typically, ten sample points are collected per parcel to account for within-field variability.

Step 2: Data Ingestion and Quality Control. Laboratory results are electronically ingested into TerraFIRST. Automated quality control checks identify missing values, out-of-range measurements and laboratory inconsistencies. Parcels with incomplete data are flagged for follow-up sampling. Missing values for numeric fields are imputed using the portfolio mean for each parameter.

Step 3: Standardization. All measured variables are z-score normalized against the portfolio distribution. This ensures that the PCA is not dominated by variables with larger numerical ranges and that all soil parameters contribute proportionally to the analysis.

Step 4: PCA Dimensionality Reduction. Principal component analysis is applied to the standardized data, reducing the feature space to two principal components. This fixed two-component reduction ensures consistency across model refreshes and provides a stable basis for clustering. The two components capture the dominant sources of variation in soil conditions while filtering out noise from minor or uncorrelated variables.

Step 5: K-Means Clustering. K-Means clustering with four clusters is applied to the PCA-reduced data. The algorithm uses K-Means++ initialization and a fixed random seed for reproducibility. Each soil observation is assigned to one of four clusters based on proximity to cluster centroids in the two-dimensional PCA space.

Step 6: Sample Point Aggregation and Parcel Assignment. For each parcel and sampling year, the multiple sample points are aggregated into a single centroid by averaging their positions in PCA space. This parcel-level centroid is then assigned to the nearest cluster, producing a single cluster assignment per parcel per year.

Step 7: Cluster Migration Detection. Each parcel's current cluster assignment is compared against its assignment from the previous agrology cycle. Parcels with changed assignments are flagged for review. The system maintains a complete cluster history for each parcel, enabling analysis of migration patterns over multiple cycles.

Step 8: Centroid Inverse Transformation. Cluster centroids are inverse-transformed through the PCA and standardization steps to produce soil profiles in original measurement units. These profiles describe what each cluster represents in agronomic terms.

Step 9: Reporting and Action. Summary reports are generated for the land management team, highlighting parcels that have changed clusters, portfolio-level cluster distributions, and cluster profile characteristics. Findings are incorporated into operator reviews, valuation updates and investor reporting.

ADVANTAGES OF THE PCA AND CLUSTERING APPROACH

The application of PCA combined with K-Means clustering to recurring agrology data offers several advantages over traditional threshold-based monitoring approaches.

Dimensionality Reduction. Reducing 20 to 40 soil variables to two principal components and four discrete clusters makes portfolio-level monitoring tractable. Without PCA, a portfolio manager would need to review hundreds of individual data points per parcel per cycle, an approach that does not scale to a portfolio of Veripath's size.

Actionable Discrete Signals. By assigning each parcel to one of four clusters, the framework produces clear, actionable outputs. Portfolio managers do not need to interpret continuous scores or determine whether small movements are meaningful. A parcel either changed clusters or it did not – a binary signal that drives clear follow-up actions.

Multivariate Sensitivity. The framework captures interactions between soil parameters that univariate monitoring misses. A simultaneous decline in organic matter, nitrogen and water holding capacity represents a qualitatively different signal than any one of those declines in isolation. PCA and clustering integrate these co-movements into a single cluster assignment.

Objectivity and Reproducibility. Both PCA and K-Means are mathematically defined transformations with no subjective weighting inputs. The principal components and cluster assignments are derived entirely from the structure of the data itself. The use of K-Means++ initialization and fixed random seeds ensures that results are identical across model runs. This makes the analysis reproducible, auditable and defensible in the context of institutional investor reporting.

Temporal Comparability. Because the PCA and clustering models are recalibrated with each cycle using the updated portfolio dataset, cluster assignments are directly comparable over time. This enables migration tracking that accounts for evolving portfolio composition as new parcels are acquired and others are disposed.

Interpretable Cluster Profiles. The inverse transformation of cluster centroids to original soil units produces human-readable soil profiles for each group. This enables meaningful communication with agrologists, farm operators and investors without requiring technical knowledge of PCA or clustering algorithms.

Scalability. The computational cost of PCA is trivial relative to the data volumes involved. The same framework that monitors 138,000 acres can scale to 300,000 or 500,000 acres with no change in methodology, a characteristic that aligns with Veripath's stated growth trajectory.

LIMITATIONS AND CONSIDERATIONS

PCA and K-Means clustering are powerful techniques**,** but they are not without limitations in the farmland context, and Veripath's implementation accounts for several known constraints.

First, PCA assumes linear relationships between variables. Soil chemistry includes some non-linear interactions, particularly at extreme pH values or in the presence of specific mineral imbalances.

Second, K-Means clustering assumes that clusters are roughly spherical in shape and similar in size. Soil data may contain irregularly shaped groupings or clusters of varying density that K-Means cannot capture optimally. The choice of four clusters is a modelling decision that balances interpretability with fit; the true underlying structure of soil variation may be more or less complex.

Third, both PCA and K-Means are sensitive to the composition of the input dataset. Adding or removing parcels changes the covariance structure and, consequently, the principal component loadings. As Veripath acquires new parcels, the PCA model is recalibrated, which can modestly alter historical scores. TerraFIRST maintains versioned PCA models to ensure backward comparability for reporting purposes.

Third, both PCA and K-Means are sensitive to the composition of the input dataset. Adding or removing parcels changes the covariance structure and, consequently, the principal component loadings and cluster boundaries. As Veripath acquires new parcels, the models are recalibrated, which can modestly alter historical assignments. TerraFIRST maintains versioned PCA and clustering models to ensure backward comparability for reporting purposes.

Fourth, the framework does not incorporate external information about what constitutes desirable soil conditions. A parcel assigned to Cluster 3 may be low-quality in absolute terms or simply different from the portfolio average (for example, a sandy soil in a portfolio dominated by clay loams). Domain expertise from Veripath's agrology partners is essential for interpreting cluster assignments in context.

Finally, the framework captures variance and groupings, not causation. A cluster migration identifies that soil conditions have changed but does not diagnose why. Diagnostic follow-up, potentially including site visits, operator discussions, and supplementary testing remains necessary for parcels flagged by the cluster monitoring system.

CONCLUSION

For institutional farmland investors, the challenge of monitoring soil health at portfolio scale is non-trivial. Recurring agrology reporting generates valuable data, but without a structured analytical framework, that data risks being underutilized or reduced to overly simplistic threshold checks that miss important multivariate signals.

Veripath's application of principal component analysis combined with K-Means clustering to its agrology dataset addresses this challenge directly. PCA reduces high-dimensional soil data to two principal components, and K-Means partitions observations into four interpretable soil groups. This two-stage approach enables systematic cluster migration tracking, early identification of parcels experiencing soil condition changes, geographic cluster analysis and integration with valuation models – all within the TerraFIRST platform.

The framework produces actionable, discrete signals: each parcel is assigned to one of four groups, and parcels that change groups between agrology cycles are automatically flagged for review. This simplifies portfolio monitoring from an overwhelming volume of continuous measurements to a manageable set of binary flags, while preserving the multivariate sensitivity that threshold-based approaches lack.

The approach is scalable, objective and well suited to the requirements of institutional investors who expect rigorous, data-driven asset management.

As Veripath's portfolio continues to grow and agrology datasets deepen over successive sampling cycles, the informational value of the PCA framework will compound. Longer time series improve trend detection reliability, larger portfolios enable finer clustering resolution, and the growing dataset provides an increasingly robust foundation for linking soil health metrics to productivity outcomes and, ultimately, to land values. This represents a meaningful operational advantage in a farmland investment market where most participants rely on rudimentary monitoring approaches.

REFERENCES

1. Andrews, S.S., Karlen, D.L. and Cambardella, C.A. (2004). *The Soil Management Assessment Framework: A Quantitative Soil Quality Evaluation Method*. Soil Science Society of America Journal, 68(6), pp. 1945-1962.
2. Arthur, D. and Vassilvitskii, S. (2007). *K-means++: The Advantages of Careful Seeding*. Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 1027-1035.
3. Moebius-Clune, B.N., Moebius-Clune, D.J., Gugino, B.K., Idowu, O.J., Schindelbeck, R.R., Ristow, A.J., van Es, H.M., Thies, J.E., Shayler, H.A., McBride, M.B., Wolfe, D.W. and Abawi, G.S. (2016). *Comprehensive Assessment of Soil Health: The Cornell Framework*. Cornell University, Third Edition.
4. Matthew, C. (2025). *Notes on the use of principal component analysis in agronomic research*. Grassland Research, Wiley Online Library.
5. Singh, K. et al. (2023). *A weighted soil health index approach for refined assessment of soil health in cropping systems*. Frontiers in Soil Science.
6. T. Hastie, R. Tibshirani, and J.Friedman (2001). *The Elements of Statistical Learning*. Springer Series in Statistics
7. Valverde-Arias, O. et al. (2017). *Homogenisation of a soil properties map by principal component analysis to define index agricultural insurance policies*. Geoderma, 311, pp. 149-163.

OMNIGENCE ASSET MANAGEMENT

Toronto Office:
TD Canada Trust Tower, 161 Bay St.
27th Floor, P.O. Box 508
Toronto, ON, M5J 2S1

Calgary Office:
Suite 300, 4954 Richard Road SW
Calgary, AB, T3E 6L1

Montréal Office:
3 Place Ville Marie, Suite 3190
Montréal, QC H3B 2E3
www.omnigenceam.com