Supplementary Information

Estimation of Water Balance and Nitrate Load in the Upper Basin of Aguascalientes, Mexico, Using SWAT

Víctor Hugo Santiago-Ayala [1], Arturo Corrales-Suastegui [2], David Avalos-Cueva [3], Saúl Hernández-Amparan [4], Cesar O. Monzon [5], Víctor Manuel Martínez-Calderón [6] and Lidia Elizabeth Verduzco-Grajeda

Complementary information on Potential Risk Index

Multivariate Analysis: MCA and HCPC

To identify latent configurations of agricultural management practices without imposing a priori assumptions, an unsupervised multivariate framework was implemented using the FactoMineR package v.2.11 in R. The analysis was conducted exclusively on nine categorical risk variables derived from fertilization practices, potassium management, crop type, irrigation system, and cultivated area. All variables were explicitly encoded as factors prior to analysis, and variables exhibiting a single category were automatically excluded to prevent degenerate structures.

Multiple Correspondence Analysis (MCA) was applied as a dimensionality-reduction procedure, projecting the categorical dataset into a continuous factorial space that preserves the principal associations among management categories. This transformation generated a metric representation suitable for clustering while avoiding direct reliance on arbitrary distance measures in the original categorical domain.

Hierarchical Clustering on Principal Components (HCPC) was subsequently performed on the individual factorial coordinates obtained from the MCA. The clustering procedure followed Ward's minimum inertia criterion, using Euclidean distances computed in the factorial space. The number of clusters was fixed at k = 3 to obtain balanced and interpretable agricultural typologies. No geographic coordinates or spatial weighting schemes were incorporated; therefore, the resulting clusters represent multivariate management configurations rather than spatial groupings. This sequential MCA–HCPC framework enabled the identification of internally homogeneous and externally distinct agricultural typologies defined exclusively by the intrinsic multivariate structure of management practices associated with potential nitrate leaching risk.

Robustness and Stability Assessment of the Multivariate Structure and IRPN Coherence

The robustness of the multivariate framework was evaluated through complementary procedures addressing dimensional adequacy, clustering quality, and partition stability. Dimensional retention in the MCA followed the Kaiser criterion, retaining components with eigenvalues above the average eigenvalue; cumulative explained inertia of the leading dimensions is reported in the Results section. Variable contributions, squared cosine (cos^2), and eta^2 coefficients were examined to assess representational fidelity and discriminative capacity.

Cluster quality was assessed using average silhouette width computed from Euclidean distances in the factorial space. The optimal number of clusters was independently examined across k values, supporting the selection of k = 3. Partition stability was quantified through bootstrap resampling (500 iterations, fixed random seed), recomputing MCA (first five components) and HCPC (k = 3) at each iteration. Agreement with the original partition was evaluated using the Adjusted Rand Index (ARI), and mean ARI with standard deviation was used to assess typological reproducibility. Pairwise chi-square tests among categorical variables confirmed significant dependency structures supporting the application of MCA.

A structured sensitivity analysis of the IRP_N was conducted to evaluate the effect of alternative weighting schemes on risk classification and typology-level interpretation. Three mathematically defined scenarios (moderate redistribution, partial hierarchical attenuation, and equal weighting) were implemented. For each scenario, the IRP_N was recalculated and producers were reclassified into predefined risk categories. Ranking stability was quantified using Spearman's rank correlation, and categorical robustness was assessed through concordance analysis. Cluster membership remained fixed, and mean IRP_N values were recomputed within each typology to verify preservation of the hierarchical risk gradient.

To assess coherence between typologies and risk levels, global differences in IRP_N among clusters were evaluated using a Kruskal–Wallis test, with effect size estimated via epsilon-squared (ε^2). Pairwise comparisons were conducted exclusively between the Low–Moderate Risk (n = 19) and Moderate–High Risk (n = 76) groups using the Wilcoxon–Mann–Whitney test with Holm correction; the Low Risk group (n = 4) was excluded from pairwise contrasts due to limited sample size. For this

group, medians and interquartile ranges were reported with 95% bootstrap confidence intervals (10,000 replicates). Effect sizes for pairwise contrasts were estimated using Cliff's delta (δ).

Analysis of the Multivariate Analysis: MCA and HCPC and Robustness and Stability Assessment of the Multivariate Structure and IRPN Coherence

Multivariate typologies of agricultural management (MCA–HCPC)

Multiple Correspondence Analysis (MCA) was applied to nine categorical risk variables to identify latent configurations of agricultural management practices associated with potential nitrate leaching. The first five dimensions explained 51.78% of total inertia (Dim 1: 19.00%; Dim 2: 11.14%; Dim 3: 8.93%; Dim 4: 6.72%; Dim 5: 6.00%). According to the Kaiser criterion, ten dimensions exhibited eigenvalues above the average eigenvalue and were therefore retained for structural interpretation. Pairwise chi-square tests among categorical variables indicated significant dependency in 27 of 36 combinations (75.0%, $p < 0.05$), supporting the suitability of MCA for capturing multivariate association patterns.

Dimension 1 was primarily defined by low-risk categories of nitrogen and potassium management, including fertilizante_riesgo_0 (12.24%), cantidad_fertilizante_riesgo_0 (12.24%), aplicaciones_fert_riesgo_0 (12.24%), cantidad_potasio_riesgo_0 (11.47%), fuente_potasio_riesgo_0 (10.97%), and aplicaciones_potasio_riesgo_0 (10.68%), indicating that this axis represents a management intensity gradient. A moderate but significant Spearman correlation was observed between Dimension 1 coordinates and IRP_N values ($\varrho = -0.422$, $p < 0.001$), suggesting partial alignment between the factorial structure and the composite risk index.

Hierarchical Clustering on Principal Components (HCPC), using Ward's minimum inertia criterion and Euclidean distances in the factorial space, identified three distinct agricultural typologies. The average silhouette width was 0.40, indicating acceptable cluster separation. The optimal number of clusters ($k = 3$) was independently supported by silhouette-based evaluation across candidate k values. Final cluster sizes were 4 (4.04%), 19 (19.19%), and 76 (76.77%) localities, corresponding to Low Risk, Low–Moderate Risk, and Moderate–High Risk typologies, respectively.

Partition stability assessed through bootstrap resampling (500 iterations) yielded a mean Adjusted Rand Index (ARI) of 0.904 ± 0.210, with 100% successful convergence, indicating high reproducibility of the typological structure. Collectively, these results demonstrate that the identified agricultural

typologies represent internally coherent and statistically stable multivariate configurations of management practices linked to potential nitrate leaching risk.

To determine whether the multivariate agricultural typologies differed in their associated potential nitrate leaching risk, non-parametric comparisons of IRP_N values were conducted across clusters. Median IRP_N values increased progressively from the Low Risk typology (0.20; IQR: 0.17–0.22; 95% CI: [0.17, 0.22]) to the Low–Moderate Risk group (1.55; IQR: 1.49–1.66; 95% CI: [1.50, 1.65]) and the Moderate–High Risk group (1.89; IQR: 1.75–2.18; 95% CI: [1.80, 2.02]). The confidence interval for the Low Risk group median was derived from 9,999 bootstrap replicates due to its reduced sample size ($n = 4$).

A global Kruskal–Wallis test revealed statistically significant differences among typologies ($H(2) = 34.75$, $p < 0.001$), with a large effect size ($\varepsilon^2 = 0.355$), indicating that a substantial proportion of IRP_N variability is explained by the multivariate cluster structure. Post-hoc comparisons were performed exclusively between the Low–Moderate Risk ($n = 19$) and Moderate–High Risk ($n = 76$) groups using the Wilcoxon–Mann–Whitney test with Holm correction. This contrast was highly significant ($Z = 5.03$, $p < 0.001$) and yielded a large effect size (Cliff's $\delta = 0.75$; 95% CI: [0.46, 0.90]).

The Low-Risk typology was excluded from pairwise contrasts due to its limited sample size, which could compromise statistical power and effect size stability; however, its descriptive statistics were retained for interpretative completeness. Overall, these results confirm that the data-driven multivariate typologies correspond to statistically distinct levels of potential nitrate leaching risk, reinforcing the internal coherence between the composite index and the factorial cluster structure.

Scoring rubric and the category thresholds used for the variable's determinate in the IRPN.

Table S1. Scoring rubric and the category thresholds used for the variable's determinate in the IRPN. Risk scores represent ordinal categories reflecting relatively leaching susceptibility based on literature synthesis. The scoring system does not imply linear physical scaling. All weights sum to 1, and the index follows an additive weighted structure. Although assigned a score for multivariate characterization, this variable received weight = 0 in the IRP_N computation.

Fertilizer type				
Variable	**Operational Definition**	**Category**	**Score**	**Rationale**
Fertilizer type	Dominant nitrogen source	None	0	No direct N input
		Organic sources	1	Nitrogen supplied in organic forms requires mineralization
		Ammoniacal / ammonium	2	Nitrogen initially supplied as ammonium undergoes
		Soluble nitrates / NPK high-N	3	Fertilizers containing readily soluble nitrate provide

Fertilizer application rate (kg ha ^{1} cycle ^{1})		
Threshold (kg ha ^{1})	**Score**	**Rationale**
0	0	No N input
>0 – 400	1	Application rates within commonly reported agronomic
401 – 800	2	Elevated nitrogen inputs increase the likelihood of surplus
>800	3	High nitrogen application rates substantially exceed typical

Fertilizer application frequency (applications cycle ^{1})		
Frequency	**Score**	**Rationale**
0	0	No application
1–2	3	Low application frequency typically involves larger single-dose inputs, increasing the likelihood of nitrate accumulation in the soil solution and enhancing the risk of wash-off or deep percolation losses during subsequent rainfall or irrigation events.
3–5	2	Intermediate application frequency distributes nitrogen inputs across the cropping cycle, partially reducing pulse magnitude while still maintaining moderate potential for episodic nitrate mobilization under high soil moisture conditions.
>5	1	High-frequency split applications supply smaller nitrogen doses over time, improving synchronization with crop uptake and reducing the probability of instantaneous nitrate accumulation and leaching losses.

Potassium source (Potassium source_risk)		
Category	**Score**	**Rationale**
None	0	No potassium fertilizer applied during the cycle; this variable does not contribute
KCl / K_2SO_4	1	Potassium supplied in chloride or sulfate forms does not contain nitrogen
KNO_3	3	Potassium nitrate provides a direct source of nitrate (NO), immediately
NPK-00-00-49	0	No N component

Potassium application rate (kg ha ^{1} cycle ^{1})		
Threshold	**Score**	**Rationale**
0	0	No potassium input applied; no nutrient load contribution associated with this variable.
>0 – 150	1	Application rates within commonly reported agronomic ranges, representing low additional
151 – 300	2	Elevated application rates increase nutrient availability in the soil solution, potentially
>300	3	High application rates substantially increase soluble nutrient concentrations, elevating the probability of downward transport through percolation and runoff under irrigation or rainfall events.

Potassium application frequency		
Frequency	**Score**	**Rationale**
0	0	No potassium input applied during the cropping cycle; no contribution of nutrient pulses associated with this variable.
1–2	3	Low application frequency typically involves larger single-dose inputs, increasing the probability of nutrient losses through deep percolation or runoff when coinciding with irrigation or high-intensity rainfall events.
3–5	2	Intermediate frequency partially distributes nutrient inputs, reducing individual pulse magnitude relative to low-frequency regimes, but still maintaining moderate potential for episodic leaching losses.
>5	1	High-frequency split applications distribute nutrient inputs into smaller doses, improving synchronization with crop uptake and reducing the likelihood of percolation-driven nutrient losses.

Crop type (cultive_risk)		
Crop category	**Score**	**Rationale**
Legumes	1	Biological N fixation reduces synthetic
Cereals	2	High N demand
Leafy vegetables	3	High N demand, short cycles, multiple
Fruit vegetables	2	Moderate demand
Other	1	Variable but generally moderate

Irrigation system (irrigation_risk)		
Irrigation type	**Score**	**Rationale**
Drip irrigation	1	High water-use efficiency and localized application reduce deep percolation and nitrate leaching compared to surface irrigation systems.
Sprinkler	2	Intermediate control of water application; potential for percolation losses under mismanaged irrigation scheduling.
Rainfed	2	Leaching driven by high-intensity rainfall events; absence of irrigation control may induce episodic nitrate flushing.
Gravity irrigation	3	Low application efficiency and higher excess water application increase deep percolation and surface runoff, enhancing nitrate transport potential.

Cultivated area (Cultivated area_risk)		
Area (ha)	**Score**	**Rationale**
>15	0	Not used as a direct leaching driver; larger farm size is often associated with higher input-use efficiency (lower fertilizer use per ha at comparable yields), so this class is treated as the lowest-intensity proxy.
5–15	1	Intermediate proxy class: moderate management intensity expected; included for typology structure rather than mechanistic control.
1–5	2	Smaller holdings may reflect higher management pressure per unit area and lower input-use efficiency; treated as higher-intensity proxy.
≤1	3	Potential management-intensity proxy (smallholder scale), consistent with evidence linking smaller farm size with lower fertilizer-use efficiency / higher per-area input pressure in some contexts.