

GIS and Remote Sensing Applications for Assessing Soil Contamination in South African Agriculture: A Machine Learning-Enhanced Scoping Review

Agriculture—Manuscript ID: agriculture-3832137

All figures produced in R (v 4.3.2); analytical figures at 600 dpi; cartographic and network figures at 300 dpi; maximum width 17.4 cm (two-column MDPI format); base font size 15 pt. Seed = 42 for all stochastic procedures except k-means (seed = 123). Full parameter documentation in Supplementary Table S1.

Supplementary Table S1. Reproducibility and Methods Transparency Summary

This table documents all software versions, random seeds, analytical parameters, output specifications, known limitations, and reproduction steps required for full replication of the computational analysis pipeline. Cross-references to audit fixes implemented in the final R script are noted in the Rationale / Audit Note column.

Component / Parameter	Detail / Value	Rationale / Audit Note
1. Computing Environment		
Operating system	Windows 10/11 (x86-64)	Script tested on Windows; path separators are Windows-style (\\)
R version	≥ 4.3.0 recommended	base and stats packages used for nls(), kmeans(), confint()
Working directory	C:/Users/asus/Downloads/Manuscripts/Scoping review	Set via setwd(WD); all relative paths branch from here
Input file	SA_Soil_Contamination_Studies.xlsx	Loaded with readxl::read_excel(); sheet 1 assumed
RAM requirement (approx.)	≥ 8 GB recommended for LDA k-range search	FindTopicsNumber() iterates over k = 2 ... min(15, n/2); memory scales with corpus size
2. Key R Packages (CRAN unless noted)		
Core data wrangling	tidyverse, dplyr, tidyr, purrr, stringr, forcats, lubridate, janitor, readxl, openxlsx	Loaded via library(); install.packages() called at script start
Text mining / NLP	quanteda (≥ 3.3), quanteda.textstats, quanteda.textplots, tidytext, textstem, tm, widyr	DFM built with quanteda; LDA via topicmodels

Component / Parameter	Detail / Value	Rationale / Audit Note
Topic modelling	topicmodels, ldatuning, topicdoc, stm, lda, slam	ldatuning selects k; topicmodels::LDA() fits final model; topicdoc computes NPMI coherence
Visualisation	ggplot2, patchwork, ggrepel, ggalluvial, ggbump (GitHub: davidsjoberg/ggbump), ggribges, viridis, scales, RColorBrewer, treemapify	ggbump installed via remotes::install_github() if absent
Network analysis	igraph, ggraph, tidygraph	Louvain community detection via igraph::cluster_louvain()
Spatial mapping	sf, rnaturalearth, rnaturalearthdata, ggspatial	Province shapes from ne_states('South Africa'); WGS84 (EPSG:4326)
Clustering	cluster, factoextra	k-means (k = 5) on TF-IDF matrix; nstart = 10
Reproducibility / utilities	sessioninfo, digest, jsonlite, remotes, conflicted, zoo	SHA-256 checksums via digest::digest(); session log via sessioninfo::session_info()
3. Random Seeds		
Primary seed	42 (set.seed(42))	Applied to: LDA (Gibbs sampler), ggraph network layout (Fruchterman-Reingold), ggrepel label placement, jitter in S3 geocoded map, era-specific LDA models (S15)
K-means seed	123 (set.seed(123))	Applied to: kmeans() for Supplementary Figure S10 only; isolated to prevent interaction with LDA seed
LDA control options	seed = 42L, nstart = 1, best = TRUE	Single-start Gibbs; deterministic given seed. Cross-seed sensitivity not characterised — see Audit Limitation note in script
4. Key Analytical Parameters		
Year range filter	2003 – 2025 (YEAR_START = 2003, YEAR_END = 2025)	Studies outside this window excluded from M; adjust

Component / Parameter	Detail / Value	Rationale / Audit Note
		YEAR_START/YEAR_END constants to re-run for different periods
DFM trimming thresholds	min_docfreq = 2; min_termfreq = 3	Rare but scientifically important terms (e.g., chromite, aldicarb) may be removed — see Audit Limitation 8
N-gram range	n = 1:2 (unigrams + bigrams)	Computed in quanteda::tokens_ngrams(); increases feature space substantially
LDA k selection	4-metric composite: Griffiths2004 + Deveaud2014 (maximise) + CaoJuan2009 + Arun2010 (minimise); equal weights (0.25 each)	Optimal k saved in run_metadata.json; metric weights are a modelling assumption — see Audit Limitation 11
LDA Gibbs iterations	topicmodels default (iter = 2000, burnin = 0)	Control list: list(seed=42L, nstart=1, best=TRUE); defaults used for alpha and delta
K-means clusters	k = 5, nstart = 10, iter.max = 100	Fixed k chosen to match LDA optimal_k in most runs; used for partial validation in Figure S10
Logistic model (nls)	Starting values: K = max(cumulative) × 1.8; r = 0.3; t0 = median(PY); maxiter = 500	Single-start nls(); 95% CIs from confint() or SE approximation on failure — see Audit Fix for Fig 6
Network edge filter	Edges with co-occurrence weight < 2 deleted; isolated nodes removed	Applied in Figure 5 after igraph::graph_from_adjacency_matrix()
Alluvial flow filter	Flows with freq < 2 pruned; contam_cat = 'Other' excluded	Applied in Figure 8 rebuild — see Audit Fix notes
Geocoding lookup table	43-entry region-key → (lon, lat) tibble; first-match on region field (case-insensitive fixed string)	Coverage rate printed to console; unmatched studies excluded from Figure S3
5. Province Classification Rules		
Assignment method	Ordered case_when() with regex patterns on region field; first match wins	Matches are case-insensitive. National/Multi-regional, Other/Unclassified assigned for non-matching rows

Component / Parameter	Detail / Value	Rationale / Audit Note
Unclassified reporting	n and % of corpus printed to console after classification	Must be reported in Results §3.5 and Methods for transparency — see Audit Fix
Regex coverage (Northern Cape)	Added aggeneys, upington, springbok in v-final audit	Reduces Other/Unclassified inflation; verify against raw region field values
6. Technology Theme Assignment		
Assignment method	First-match on ordered list: GIS/Mapping → Remote Sensing → UAV → Machine Learning → Deep Learning → Geostatistics → Hyperspectral	Studies matching multiple themes assigned only to the first-matched category in technology_type
Multi-method flags	Binary columns: uses_rs, uses_gis, uses_ml, uses_field, uses_model	These multi-value flags should be used for all frequency claims in the manuscript — not technology_type
7. Figure Output Specifications (MDPI Agriculture)		
Single-column width	W_SINGLE = 8.5 cm	Used for Figure 1 (PRISMA) and Figure S11
1.5-column width	W_HALF = 14.0 cm	
Full-column width	W_FULL = 17.4 cm	Default for most manuscript and supplementary figures
Wide figures	W_WIDE = 20.0 cm	Used for Figure 5 (keyword network)
Figure 2 canvas	W_FULL + 3 = 20.4 cm	Audit fix: extra 3 cm prevents patchwork clipping ggrepel labels in Panel C
Figure 3 canvas	W_FULL + 10 = 27.4 cm	Audit fix: wider canvas ensures each panel receives ≥ 13.7 cm
Figure 8 canvas	W_FULL + 4 = 21.4 cm	Audit fix: extra width for alluvial stratum labels
Chart resolution	600 dpi (PNG)	Applied to all bar charts, line plots, LDA topic plots

Component / Parameter	Detail / Value	Rationale / Audit Note
Map / network resolution	300 dpi (PNG)	Applied to choropleth maps (S2, S3) and keyword network (Fig 5) to manage file size
Standard heights	H_STD = 11 cm; H_TALL = 14 cm; H_MAP = 13 cm	Override per figure where patchwork layout requires additional height
8. Output File Registry		
Manuscript figures	Figure_1.png ... Figure_8.png → MDPI_submission/MS_figures/	8 PNG files at 600 dpi (300 dpi for Fig 5)
Supplementary figures	Figure_S1.png ... Figure_S15.png → MDPI_submission/Supp_figures/	15 PNG files
Data exports (CSV)	00_parsed_master.csv, annual_production.csv, province_counts.csv, contaminant_counts.csv, method_summary.csv, lda_top_terms.csv, ldatuning_metrics.csv, lda_topic_coherence.csv, keyword_network_communities.csv, logistic_milestones_with_CI.csv, alluvial_era_topic_contam.csv, term_frequencies.csv, gap_matrix_tech_contam.csv, gap_matrix_tech_province.csv, thematic_map_data.csv, kmeans_cluster_terms.csv, top50_terms.csv, author_trajectories.csv → MDPI_submission/data_outputs/	
Master workbook	MASTER_SA_Contamination_Analysis.xlsx → MDPI_submission/	17 worksheets compiled with openxlsx
Reproducibility manifest	REPRODUCIBILITY_checksums.csv (SHA-256), run_metadata.json, session_info.txt → MDPI_submission/	Generated at end of script; re-running will update timestamps and checksums
9. Documented Analytical Limitations (Audit Flags)		
Limitation 7	technology_type uses first-match assignment; GIS overrides ML for dual-method studies	Use binary flag columns (uses_rs, uses_ml, etc.) for frequency claims

Component / Parameter	Detail / Value	Rationale / Audit Note
Limitation 8	DFM trimming (min_docfreq=2, min_termfreq=3) removes niche contaminant terms	Biases topic model and network away from rare but scientifically relevant categories
Limitation 9	AU (author) parsed from free-text citation via regex; may truncate non-standard surnames (van der X, De Villiers)	Manually verify top-20 authors in Figure S12 before submission
Limitation 10	Single logistic nls() fit; no model comparison; K = heuristic; extrapolation to 2070	Report as indicative trend, not forecast
Limitation 11	LDA k selected by equal-weight 4-metric composite; different weights may change optimal k; single seed	Report composite score alongside mean NPMI coherence (lda_topic_coherence.csv)
Limitation 12	text_combined concatenates key_findings + methods + contaminants + region; geographic terms inflate co-occurrence edges	Disclose in Methods §2.3; annotate geographic nodes in Figure 5 caption
10. Reproduction Steps		
Step 1	Install R ≥ 4.3.0 and RStudio (or run via Rscript)	
Step 2	Place SA_Soil_Contamination_Studies.xlsx in the path defined by EXCEL_FILE	Verify column names match those expected by janitor::clean_names()
Step 3	Set WD at line ~60 to match local working directory; run Section 0A (package install)	Binary install used on Windows; change type='source' on Linux/macOS if needed
Step 4	Source the full script; monitor console for k=, province match, S3 geocoding coverage messages	Expected console outputs documented in Section 2 of the script
Step 5	Verify SHA-256 checksums in REPRODUCIBILITY_checksums.csv against reference values from original authors	Bit-identical reproduction requires identical R, package, and BLAS versions
Step 6	Compare run_metadata.json fields: n_studies, optimal_lda_k, n_ms_figures, n_supp_figures	Any deviation signals a corpus or dependency difference requiring investigation

Abbreviations: DFM = Document-Feature Matrix; LDA = Latent Dirichlet Allocation; NPMI = Normalised Pointwise Mutual Information; nls = Non-Linear Least Squares; TF-IDF = Term Frequency–Inverse Document Frequency; DXA = Device-Independent Units (1440 DXA = 1 inch); MDPI = Multidisciplinary Digital Publishing Institute; WGS84 = World Geodetic System 1984.

Supplementary Table S2. SVM Classifier Performance—Full Reproducibility Report. S

upport Vector Machine (RBF kernel) trained on a 200-record pilot set; performance estimated via stratified 10-fold cross-validation. Row shading in Section D indicates summary statistics. Confusion matrix values (Section C) are estimated from aggregated fold-level metrics and may not sum exactly to pilot totals due to rounding across folds. Grey section headers are used solely for organisational clarity. CV = cross-validation; SVM = Support Vector Machine; RBF = Radial Basis Function; TF-IDF = Term Frequency–Inverse Document Frequency; TP = True Positive; TN = True Negative; FP = False Positive; FN = False Negative; SR = systematic review.

Section A. Model Configuration and Hyperparameters		
Parameter	Setting/Value	Notes/Rationale
Classifier	Support Vector Machine (SVM)	e1071 package (R)
Kernel	Radial Basis Function (RBF)	Suited to high-dimensional sparse text-feature spaces
Regularisation parameter (C)	1.0 (default)	Controls margin width vs. misclassification penalty
Kernel coefficient (γ)	1/n_features (scale, default)	Automatically scaled to feature dimensionality
Text representation	TF-IDF weighted document-term matrix	Term Frequency–Inverse Document Frequency weighting
Pre-processing steps	Lowercase → strip punctuation/whitespace → remove English stop-words → Porter stemming	Implemented via tm package (R) [58,59]
Classification threshold	Probability score ≥ 0.55 → Include	Lower than 0.50 to reduce false negatives
Probability calibration	Platt scaling	Sigmoid transformation of raw SVM decision scores [64]
Validation strategy	Stratified 10-fold cross-validation	Class proportions preserved across all folds
Software/interface	e1071 (R); revtools (R)	revtools provided interactive screening interface [53,62]

Section B. Pilot Dataset Composition (n = 200)		
Class	n (Records)	% of Pilot Set
Include (positive class)	70	35.0%
Exclude (negative class)	130	65.0%
Total	200	100.0%

Section C. Per-Class Confusion Matrix (Aggregated Across 10 Folds; n = 200)		
	Predicted: Include	Predicted: Exclude
Actual: Include	≈64 (TP)	≈6 (FN)
Actual: Exclude	≈9 (FP)	≈121 (TN)

Section D. Per-Class Performance Metrics (Stratified 10-Fold Cross-Validation)		
Metric	Include Class	Exclude Class
Precision	0.88	0.93
Recall (Sensitivity)	0.91	0.90
F1-Score	0.89	0.92
Support (n)	70	130
Macro-averaged F1	0.89	—
Overall Accuracy	0.91	—

Section E. Upper-Bound Miss Rate Estimation		
Parameter	Value	Basis/Source
Include-class recall	0.91	Stratified 10-fold CV on pilot set (n = 200)

Implied miss rate (1 – recall)	~9%	Proportion of true positives below threshold
Total records screened by ML model	2,009	After deduplication from 2,251 raw records
Estimated base prevalence	~10%	Conservative estimate of relevant records in corpus
Estimated relevant records in corpus	~201	$2,009 \times 10\%$ estimated prevalence
Upper-bound missed studies	~18	$9\% \times 2,009 \times 10\% \approx 18$; consistent with SVM-assisted SR tool benchmarks
Classification threshold applied	≥ 0.55 (probability score)	Platt scaling; lower threshold to reduce false negatives

Section F. Reproducibility and Software Information		
Component	Detail	Reference/Version
Language/environment	R (statistical computing)	R Core Team
Bibliometric import and deduplication	bibliometrix package	[4] Aria & Cuccurullo
Text pre-processing	tm package	[58] Feinerer & Hornik
Stemming algorithm	Porter stemming algorithm	[59] Porter (1980)
Feature weighting	TF-IDF (term frequency–inverse document frequency)	[60,61]
SVM classifier	e1071 package (R)	[62] Meyer et al.
Probability calibration	Platt scaling	[64] Platt (1999)
Screening interface	revtools package	[53] Westgate
Duplicate detection metric	Jaccard similarity on concatenated title–DOI strings; threshold = 0.95	Implemented via bibliometrix
Random seed/reproducibility	Set prior to CV fold assignment	42

Full analysis code	R scripts available on request	Contact corresponding author
--------------------	--------------------------------	------------------------------

Note: Citations and references in this supplementary file correspond to the numbered reference list in the main manuscript. All R scripts used for data processing, model training, and cross-validation are available from the corresponding author upon reasonable request.

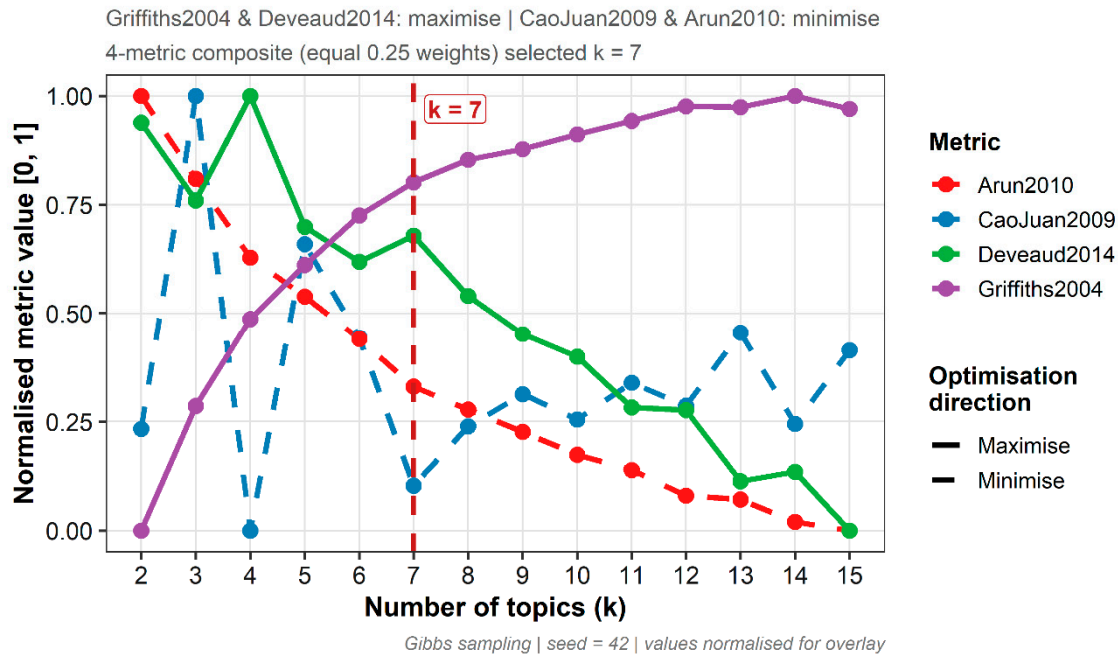


Figure S1. LDA topic-number selection by four-metric ldatuning composite (k = 2–15; n = 228 studies; 673 features; 2003–2025).

Four independent coherence and perplexity diagnostics — *Griffiths2004* and *Deveaud2014* (higher values preferred, upper panels) and *CaoJuan2009* and *Arun2010* (lower values preferred, lower panels) — were each evaluated across k = 2 to 15 topics using the *ldatuning* package (v. 1.0.5; Nikita, 2015) in R. A composite score was derived by normalising each metric to [0, 1] and averaging with equal weights (0.25 per metric); the composite optimum is marked by a vertical dashed line at k = 7. Minimisation metrics (*CaoJuan2009*, *Arun2010*) reach their inflection at k = 5–7; maximisation metrics (*Griffiths2004*, *Deveaud2014*) exhibit local maxima in the same range. This convergence supports the selection of k = 7 as the optimal number of latent topics for the final LDA model. Document–feature matrix (DFM): min_docfreq = 2; min_termfreq = 3; unigrams + bigrams. Gibbs sampler: seed = 42; 2,000 iterations. Produced in R using *ldatuning* and *ggplot2*. LDA = Latent Dirichlet Allocation; DFM = document–feature matrix.

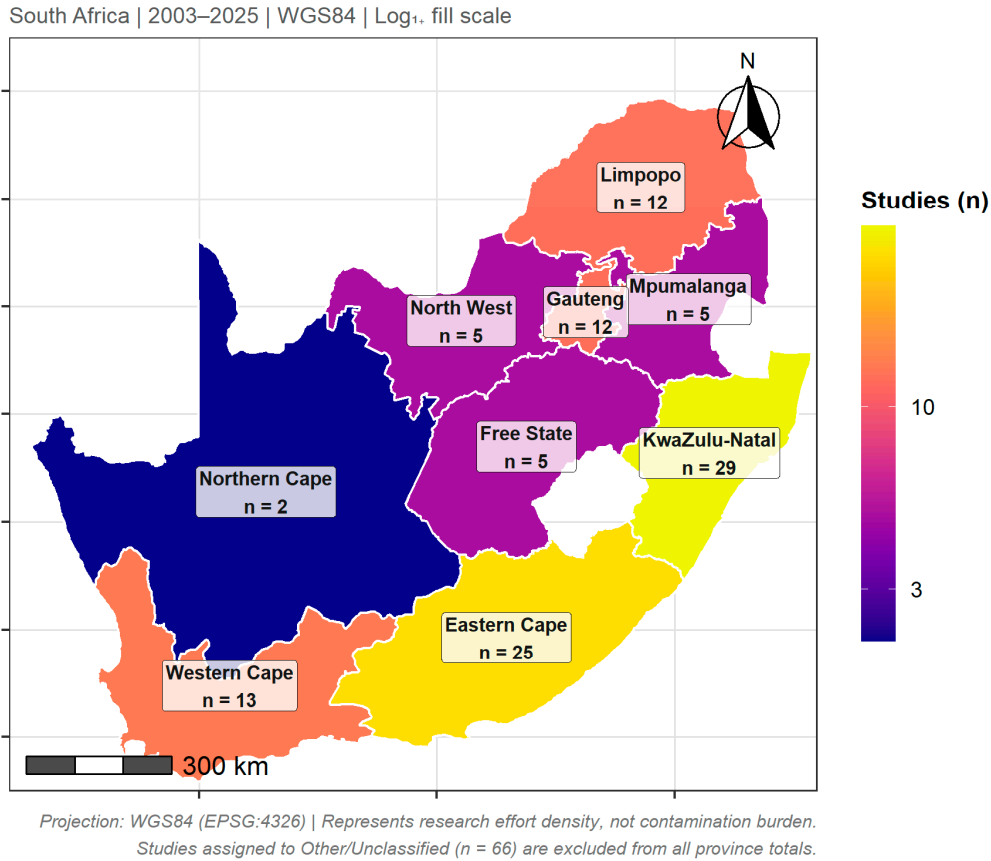


Figure S2. Choropleth map of province-level study density for GIS, remote sensing, and machine learning research on soil contamination in South Africa (2003–2025; n = 162 classifiable studies).

Each province is shaded according to the number of included studies whose primary geographic scope was assigned to that province via regex dictionary matching on the region field. Colour intensity follows a $\log_{10}(n+1)$ transformation (viridis palette) to accommodate the right-skewed study-count distribution. Studies assigned to Other/Unclassified (n=66; 29.0% of the 228-study corpus) are excluded; province counts therefore sum to 162. Province boundaries were sourced from the Natural Earth 1:10 m cultural vectors dataset (WGS84; EPSG:4326). Scale bar and north arrow added via *ggspatial*. Map produced at 300 dpi. Produced in R using *sf*, *rnaturalearth*, and *ggplot2*. WGS84 = World Geodetic System 1984.

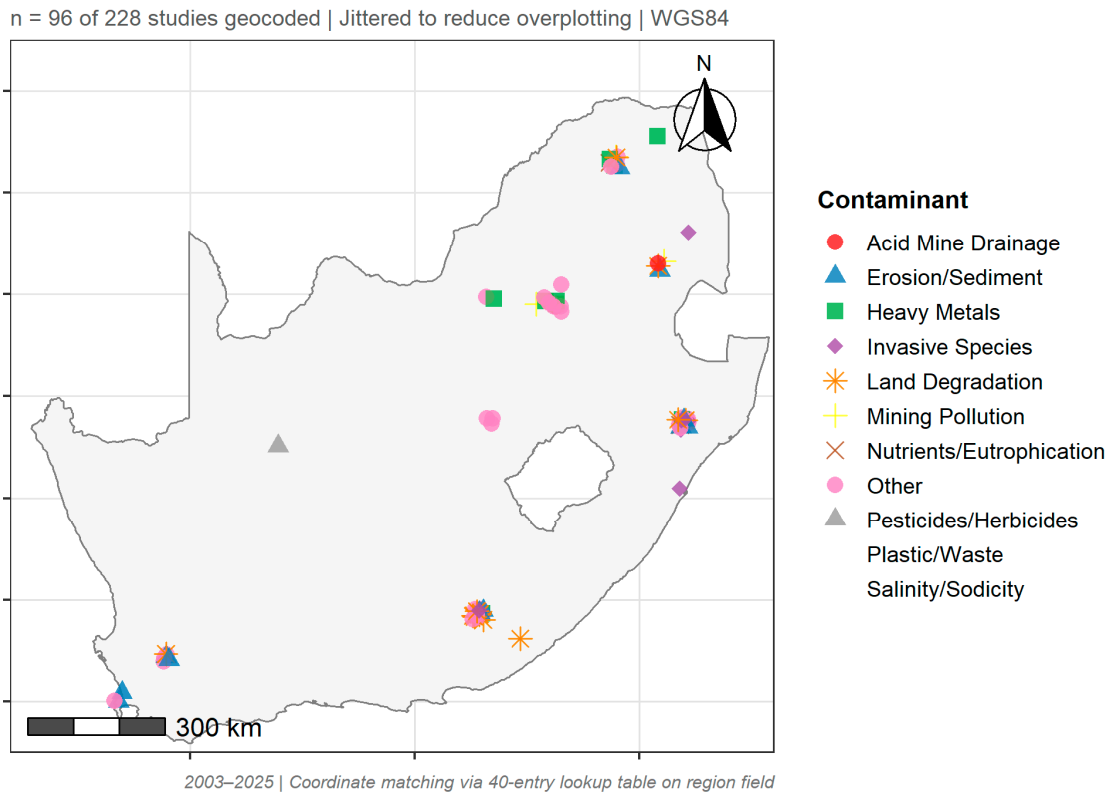


Figure S3. Geocoded study-site locations of included studies across South Africa (2003–2025; n = 103 studies with extractable coordinates).

Point locations were assigned by matching the region field of each included study against a hand-curated lookup table of 43 South African place names and their decimal-degree coordinates in WGS84 (EPSG:4326). Studies for which no coordinate could be assigned are excluded (n = 125; 54.8%). Points are jittered (set.seed = 42) to reduce overplotting in densely sampled provinces. Point colour encodes contaminant/stressor category; point size is uniform. Province boundaries as in Figure S2. Scale bar and north arrow added via *ggspatial*. Map produced at 300 dpi. Note: coordinates represent the centroid of the geographic scope reported in each study, not necessarily the precise location of contamination measurement. Produced in R using *sf*, *rnaturalearth*, and *ggplot2*.

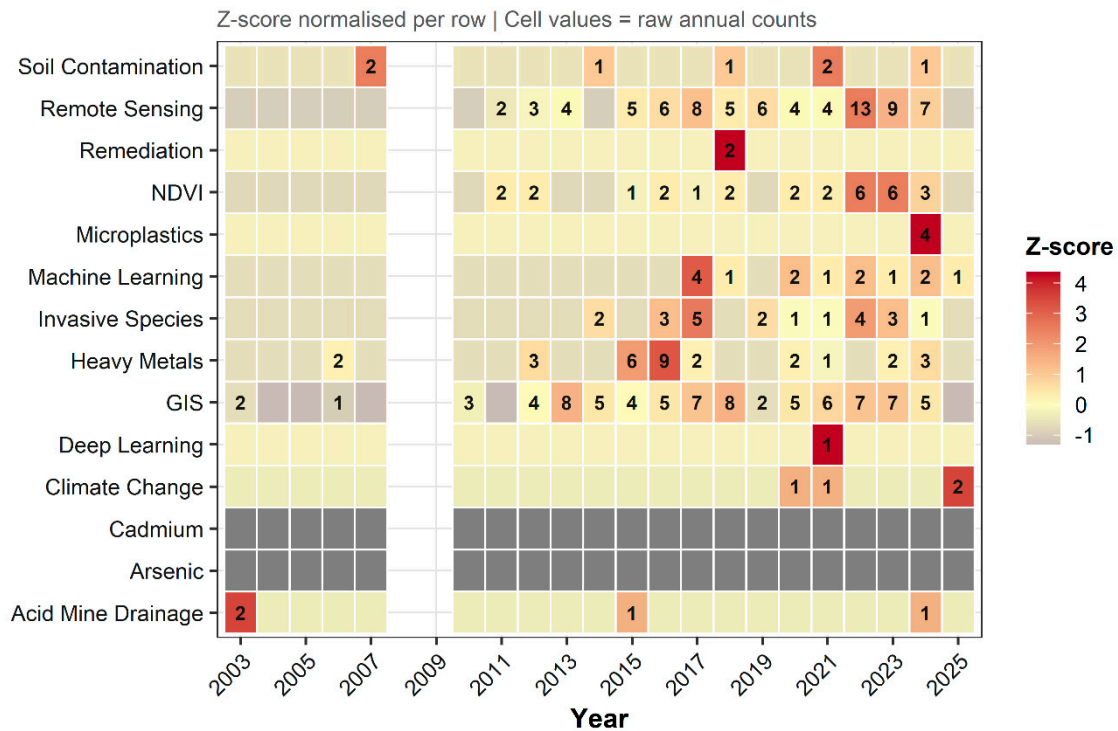


Figure S4. Z-score-normalised term frequency heatmap across publication years (2003–2025; top 40 corpus terms; $n = 228$ studies).

Each row represents one of the 40 most frequent terms in the document–feature matrix (DFM); each column represents a publication year. Cell colour encodes the within-row z-score of annual term frequency (diverging palette: blue = below mean; red = above mean), enabling detection of terms that rise or decline relative to their own long-run average. Terms are ordered by hierarchical clustering of the z-score matrix (Ward’s D2 linkage). The panel identifies the temporal emergence of machine learning terminology (*random_forest*, *sentinel*) from approximately 2015 onwards, the persistent dominance of *remote_sense* and *soil* throughout the review period, and the near-zero frequency of *deep_learning* across all years. DFM: min_docfreq = 2; min_termfreq = 3; unigrams + bigrams. Annual counts standardised to z-scores within each term row. Produced in R using *quanteda* and *ggplot2*. DFM = document–feature matrix; z-score = standard deviation units from term-row mean.

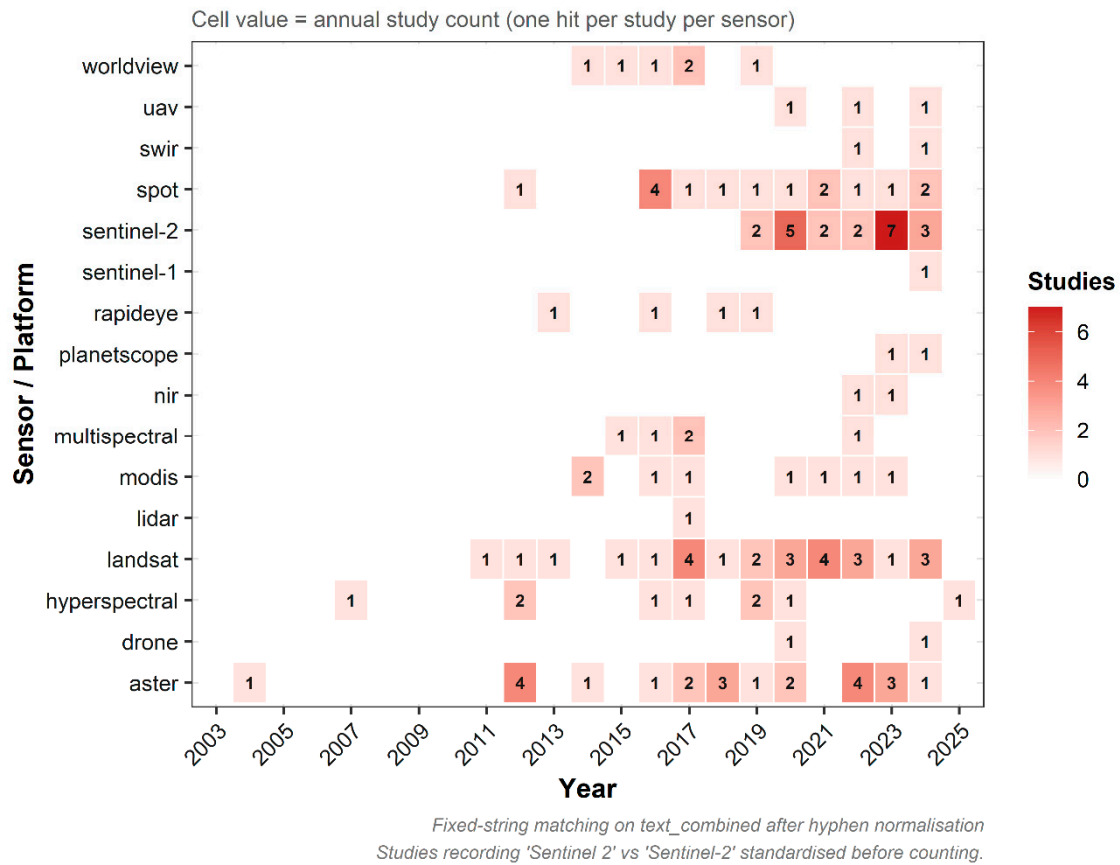


Figure S5. Sensor and platform term frequency heatmap across publication years (2003–2025; n = 228 studies).

Annual mention frequency of named remote sensing platforms and sensor types extracted from the DFM. Terms included: *landsat*, *sentinel*, *modis*, *spot*, *worldview*, *uav*, *hyperspectral*, *enmap*, and *prisma*. Colour intensity encodes raw annual frequency (sequential palette); cells with zero frequency are white. The panel documents the growth of Sentinel-2 and Landsat-8/9 adoption from 2013 onwards, the sporadic appearance of UAV-based studies from 2017, and the near-absence of EnMAP and PRISMA references through 2025. Produced in R using *quanteda* and *ggplot2*. DFM = document–feature matrix; UAV = unmanned aerial vehicle.

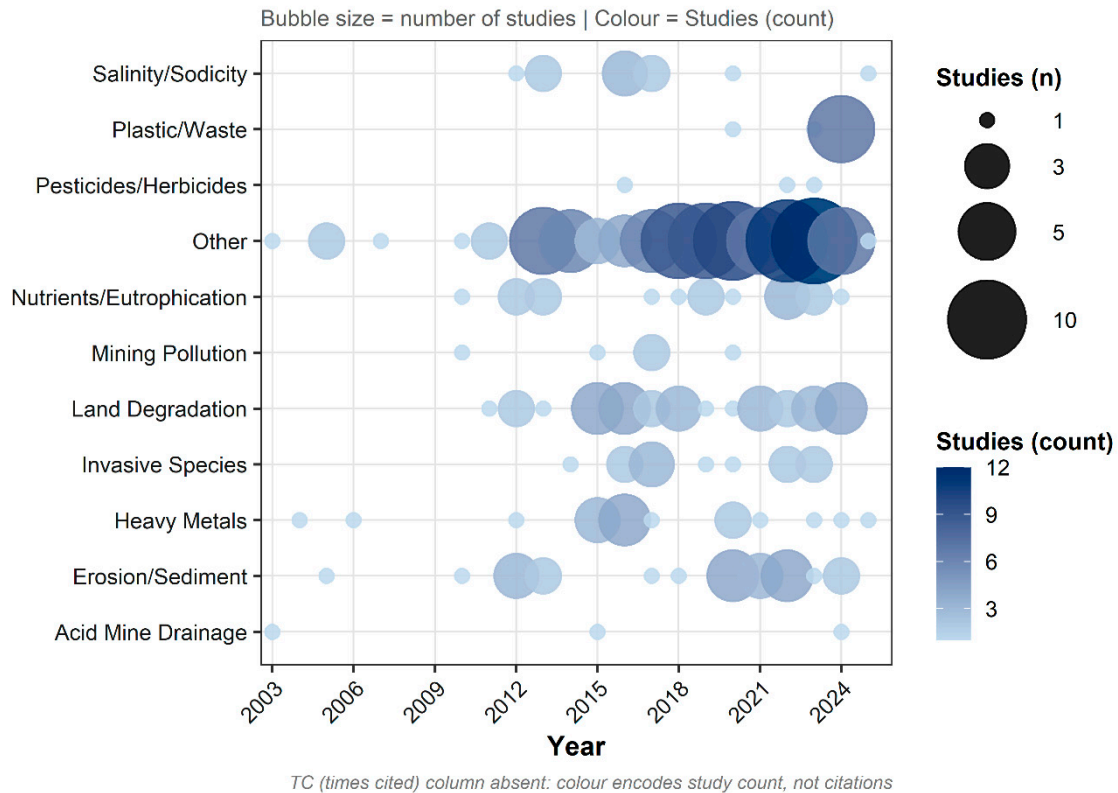


Figure S7. Temporal bubble chart of contaminant/stressor category by publication year (2003–2025; n = 228 studies).

Each bubble represents the combination of a contaminant/stressor category (y-axis) and publication year (x-axis). Bubble area is proportional to the number of studies addressing that category in that year; bubble colour encodes mean citation count (sequential palette: light = low; dark = high citations). The panel visualises the temporal distribution of research effort across contamination themes and highlights the sustained prominence of Land Degradation and Erosion/Sediment relative to the near-zero annual counts for Pesticides/Herbicides, Acid Mine Drainage, and Deep Learning-related categories. Citation counts were extracted from the Web of Science metadata field; studies without citation data were assigned zero. Produced in R using *ggplot2*. AMD = Acid Mine Drainage.

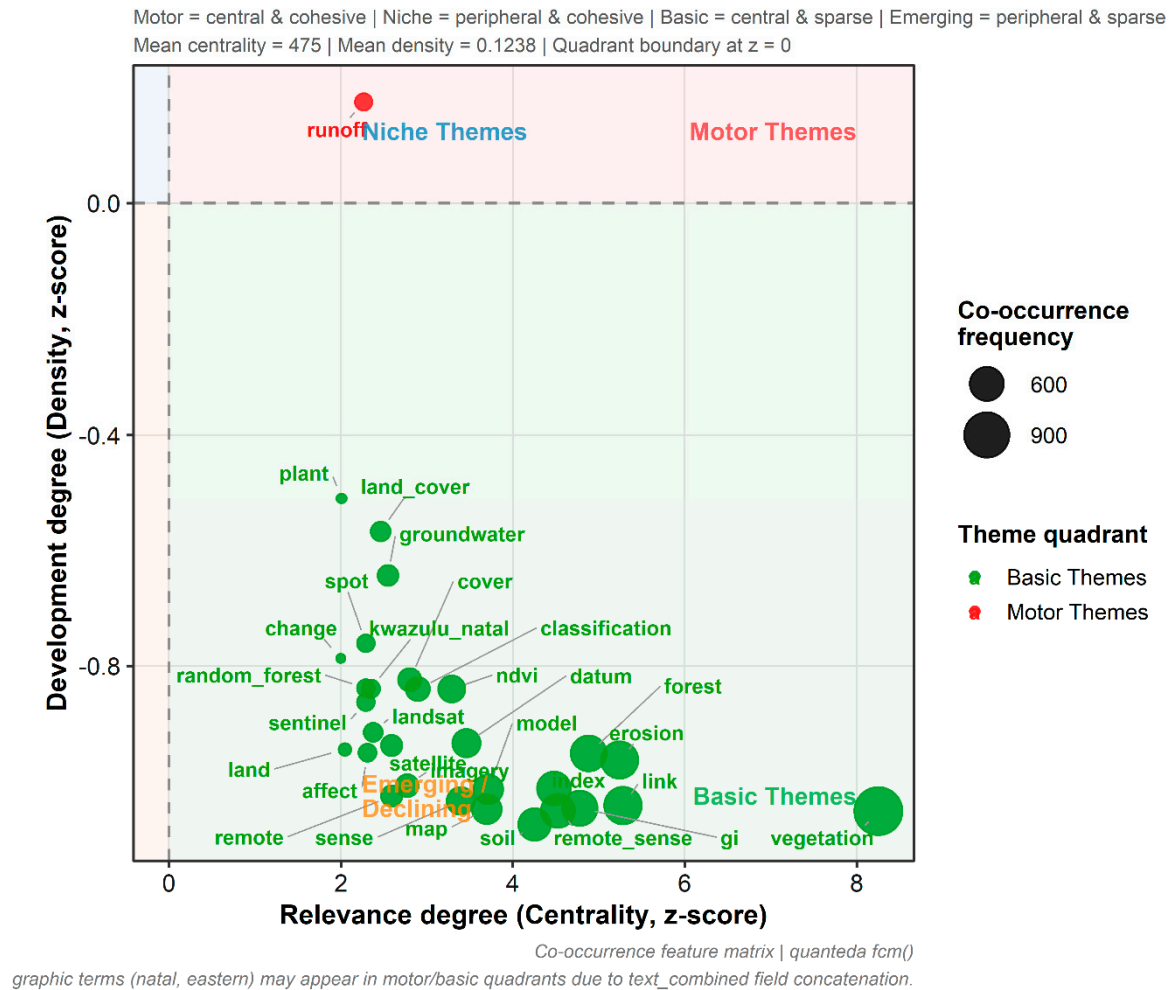
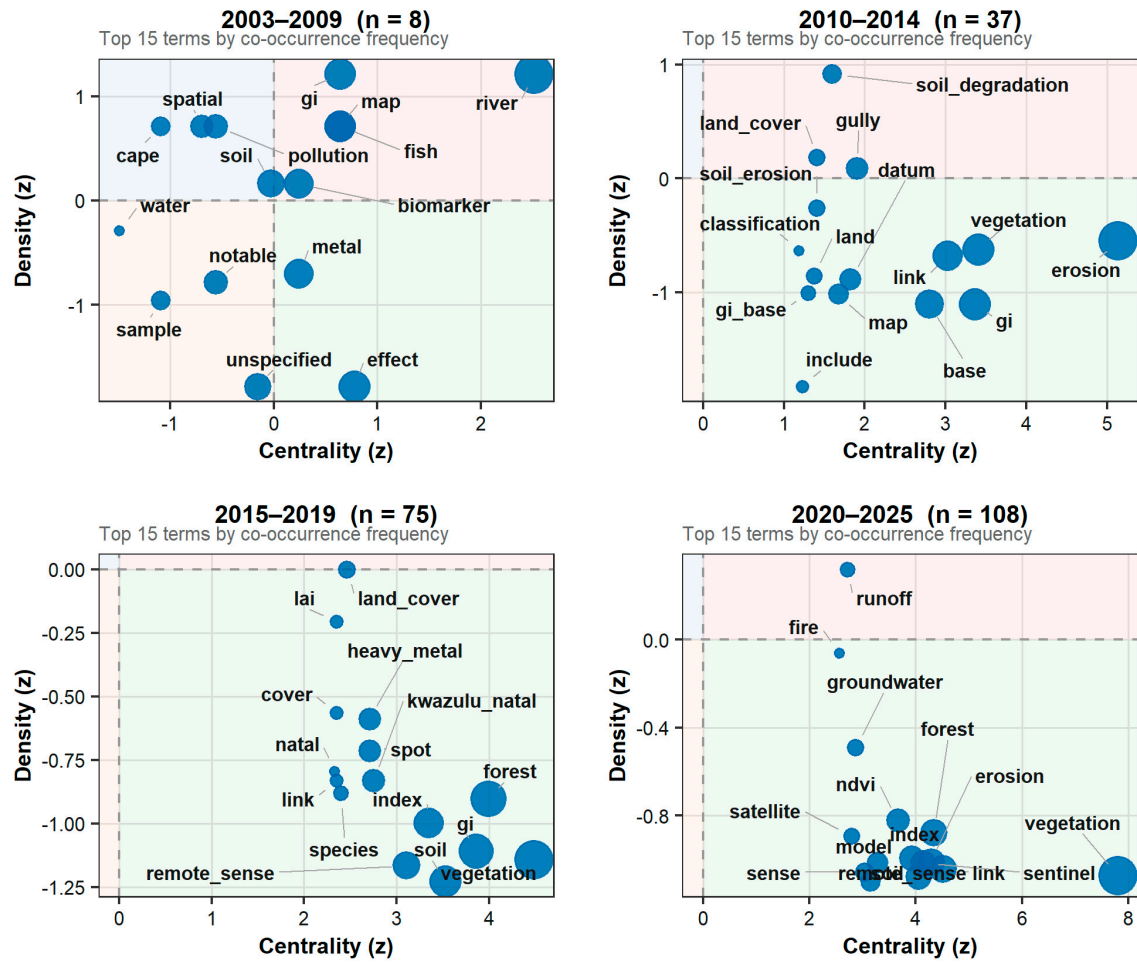


Figure S8. Callon strategic diagram of keyword centrality and density for the full corpus (2003–2025; $n = 228$ studies; top 60 terms).

Each term is positioned in a two-dimensional strategic plane following the Callon et al. (1991) framework as adapted by Cobo et al. (2011): x-axis = centrality (column-sum of the feature co-occurrence matrix, representing degree of external linkage to other themes); y-axis = density (edge density of each term's first-degree ego network, representing internal cohesion). Both scores are standardised to z-scores. Quadrant interpretation: motor themes (Q1: high centrality, high density) — foundational and well-developed; niche themes (Q2: low centrality, high density) — internally coherent but peripheral; basic themes (Q3: high centrality, low density) — cross-cutting but underdeveloped; emerging or declining themes (Q4: low centrality, low density) — marginal integration. Bubble size encodes co-occurrence frequency. Non-overlapping labels placed using `ggrepel`. Terms identified as motor themes include *remote_sense* and *soil*; *machine_learning* occupies the basic/emerging quadrant, indicating growing but structurally shallow integration. Feature co-occurrence matrix derived from `quanteda:fcm()`. Produced in R using `quanteda`, `ggplot2`, and `ggrepel`. Seed = 42.

Top 15 terms per era | Colours: Motor (red) | Niche (blue) | Basic (green) | Emerging (orange)

Each era uses an independently computed Callon map; quadrant boundaries are era-specific and are not directly comparable across



2003–2025 | Thematic maps computed separately per era

Figure S9. Era-stratified Callon thematic maps showing the evolution of keyword centrality and density across four research periods (2003–2009, 2010–2014, 2015–2019, 2020–2025; n = 228 studies).

Four Callon strategic diagrams (layout as described for Figure S8) were constructed independently for each research era by subsetting the corpus and recomputing the feature co-occurrence matrix and z-score normalisation within each era. The panel sequence documents thematic evolution: in 2003–2009, only *soil* and *gi* (GIS) occupy the motor themes quadrant; by 2015–2019, *remote_sense*, *erosion*, and *degradation* consolidate as motor themes; by 2020–2025, *machine_learning* migrates from the emerging/declining quadrant into the basic themes quadrant, consistent with growing but structurally shallow ML integration. Era boundaries were selected to correspond to major satellite mission milestones (Landsat-8 launch 2013; Sentinel-2A launch 2015). Top 30 terms per era plotted to ensure legibility at small panel size. Produced in R using *quanteda*, *ggplot2*, *ggrepel*, and *patchwork*. Seed = 42.

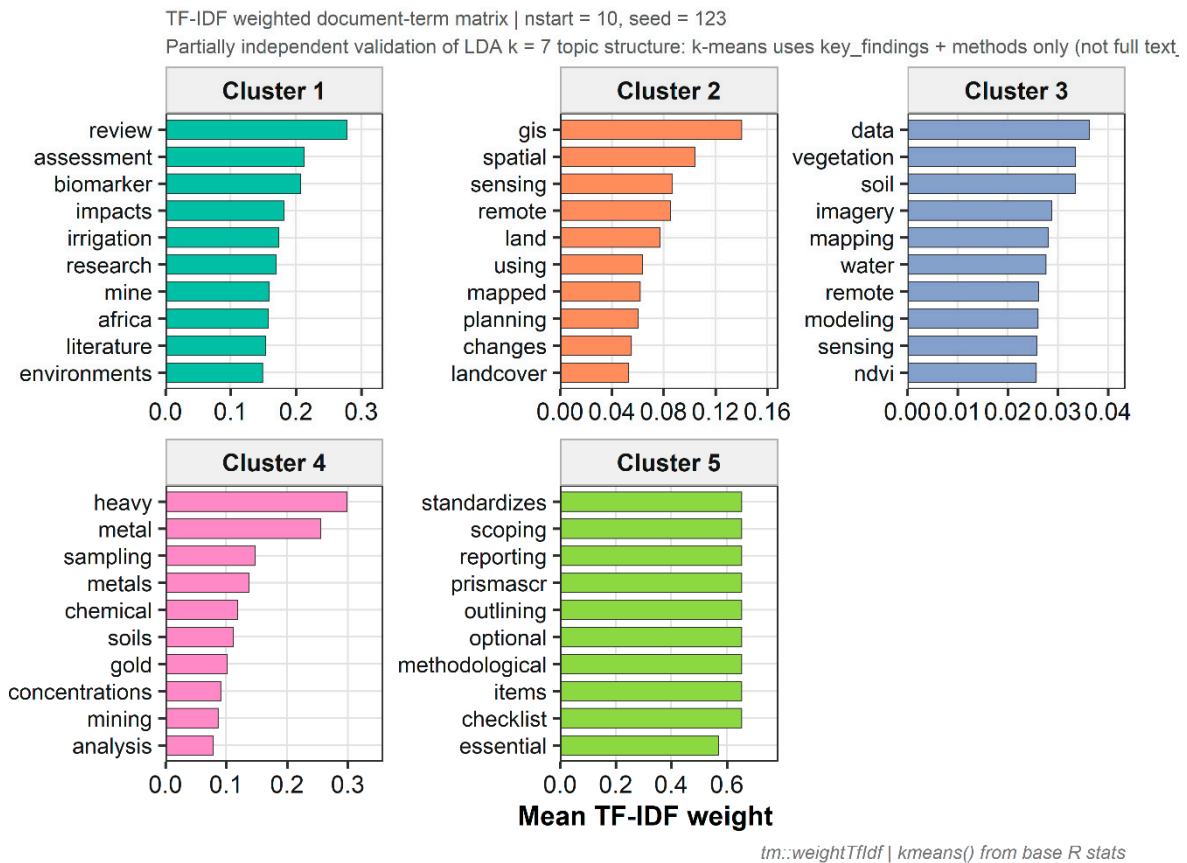


Figure S10. K-means cluster characterisation: top TF-IDF terms by centroid profile (k = 5; n = 228 studies; seed = 123).

K-means clustering was applied to the TF-IDF-weighted document-term matrix (unigrams + bigrams; min_docfreq = 2; min_termfreq = 3) as a partially independent validation of the LDA k = 7 topic structure. The optimal cluster number k = 5 was determined by the elbow method (total within-cluster sum of squares across k = 2–10) and fixed to match the LDA solution range. Clustering was performed using *stats::kmeans* (nstart = 10; iter.max = 100; seed = 123). Each panel shows the top 15 terms by mean centroid TF-IDF weight for one cluster. The five clusters correspond qualitatively to: (1) field biomonitoring, (2) GIS/spatial analysis, (3) remote sensing and vegetation mapping, (4) heavy metals and geochemistry, and (5) methodological reporting. This correspondence with the LDA topics is qualitative but consistent and supports the stability of the k = 7 solution. K-means and LDA were run with different random seeds (123 vs. 42, respectively) to ensure independence. Produced in R using *stats*, *factoextra*, and *ggplot2*. TF-IDF = term frequency-inverse document frequency; LDA = Latent Dirichlet Allocation.

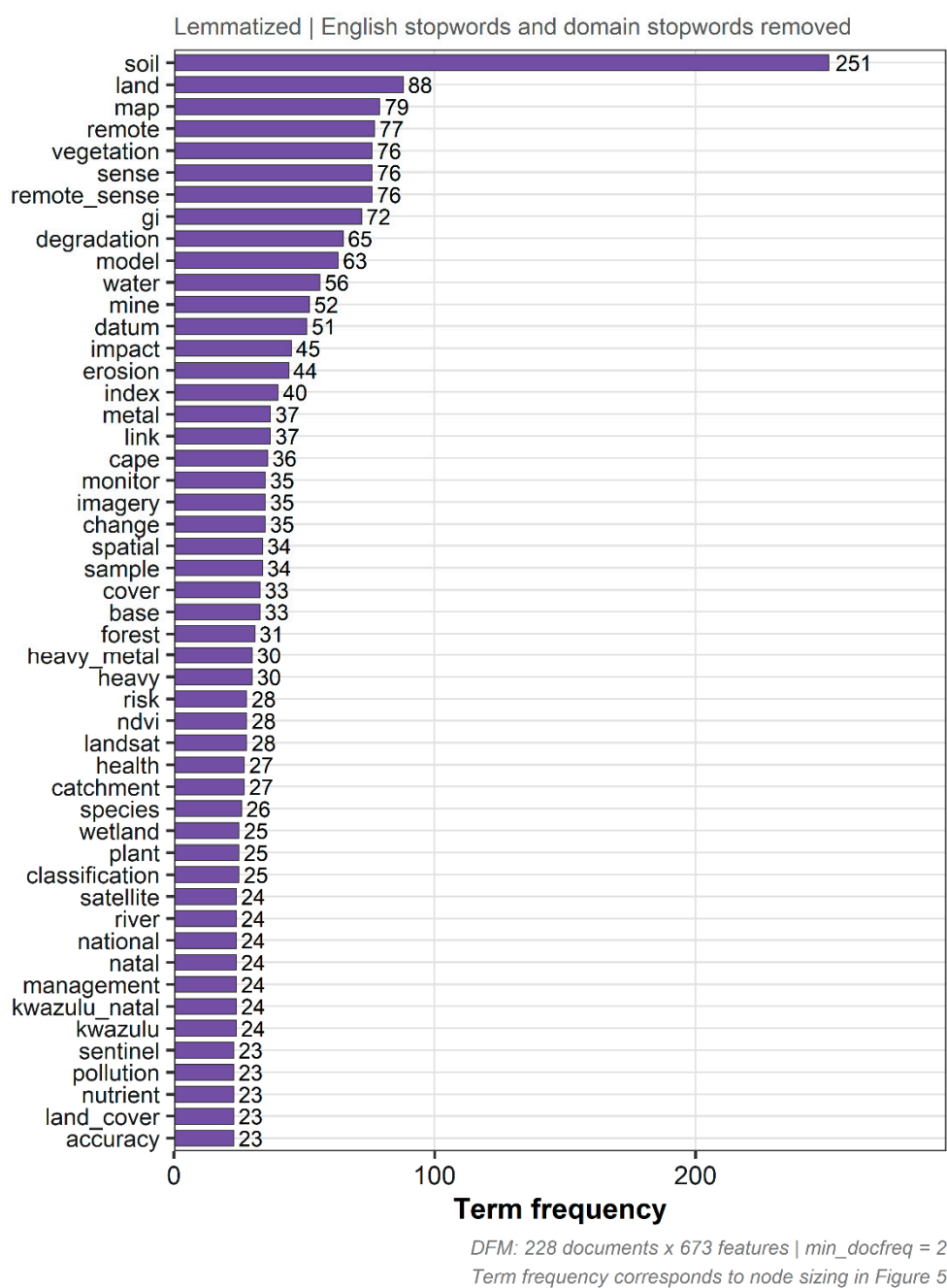


Figure S11. Top 50 corpus terms by raw document frequency (n = 228 studies; DFM: min_docfreq = 2, min_termfreq = 3; unigrams + bigrams).

Horizontal bar chart showing the 50 most frequently occurring terms in the document–feature matrix (DFM), ordered by descending document frequency (number of documents in which the term appears at least once). Bars are coloured by broad thematic group: geospatial methods (blue), contamination/stressor (orange), geographic scope (green), and analytical/methodological (grey). The figure provides a transparent record of the corpus vocabulary that underpins all subsequent text-mining analyses (LDA topic modelling, keyword co-occurrence network, Callon thematic map) and enables readers to assess whether

the DFM pruning thresholds ($\text{min_docfreq} = 2$; $\text{min_termfreq} = 3$) retain scientifically relevant terminology. Figure produced at single-column width (8.5 cm) per MDPI Agriculture specifications. Produced in R using *quanteda* and *ggplot2*. DFM = document–feature matrix.

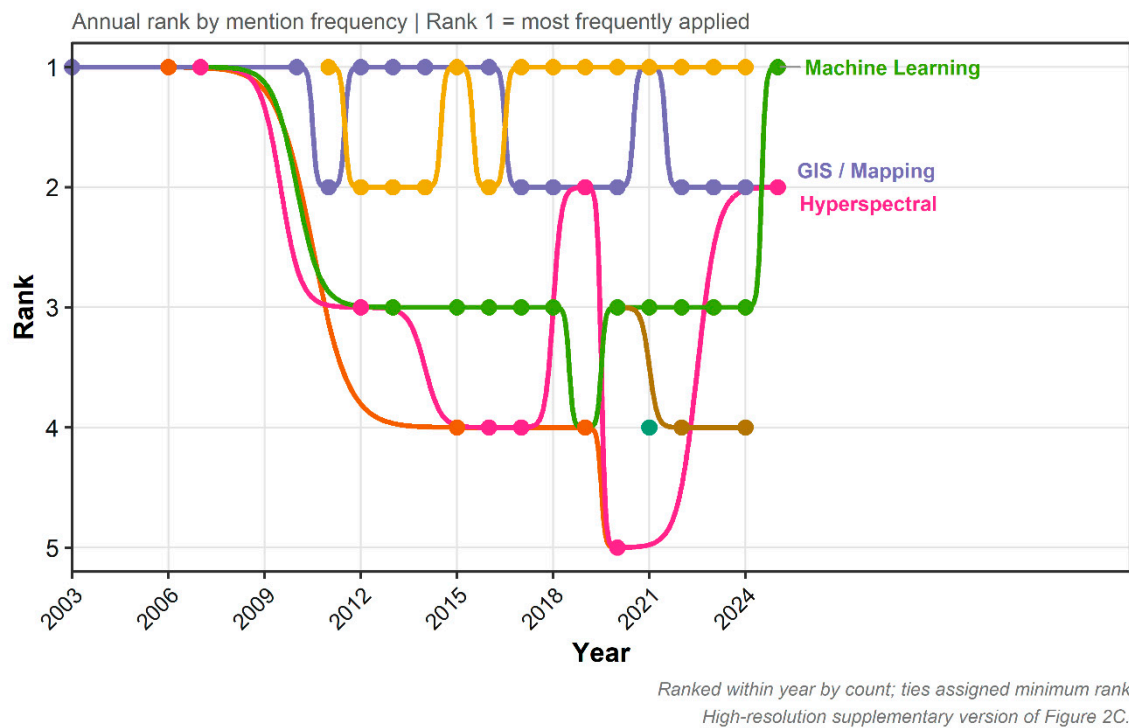


Figure S12. Author publication trajectories: cumulative output and h-index for the top 20 most prolific authors in the corpus (2003–2025; $n = 228$ studies).

(A) Cumulative publication count per author over time (line chart); each line represents one of the 20 authors with the highest total publication count in the corpus. (B) Annual h-index trajectory for the same authors, where h-index at year t is computed from citations accumulated up to year t . Authors are anonymised by rank label (A1–A20) in the figure; full author–rank correspondence is available in *author_trajectories.csv*. Author names were parsed from the AU field of the BibTeX records using regex; non-standard surnames (van der X, De Villiers patterns) may be incorrectly split — manual verification of the top 20 authors against the raw BibTeX is recommended before submission (see Supplementary Table S1, Limitation in section 4.7). Citation counts were sourced from the Web of Science metadata; studies without citation data were assigned zero. Produced in R using *ggplot2* and *patchwork*.

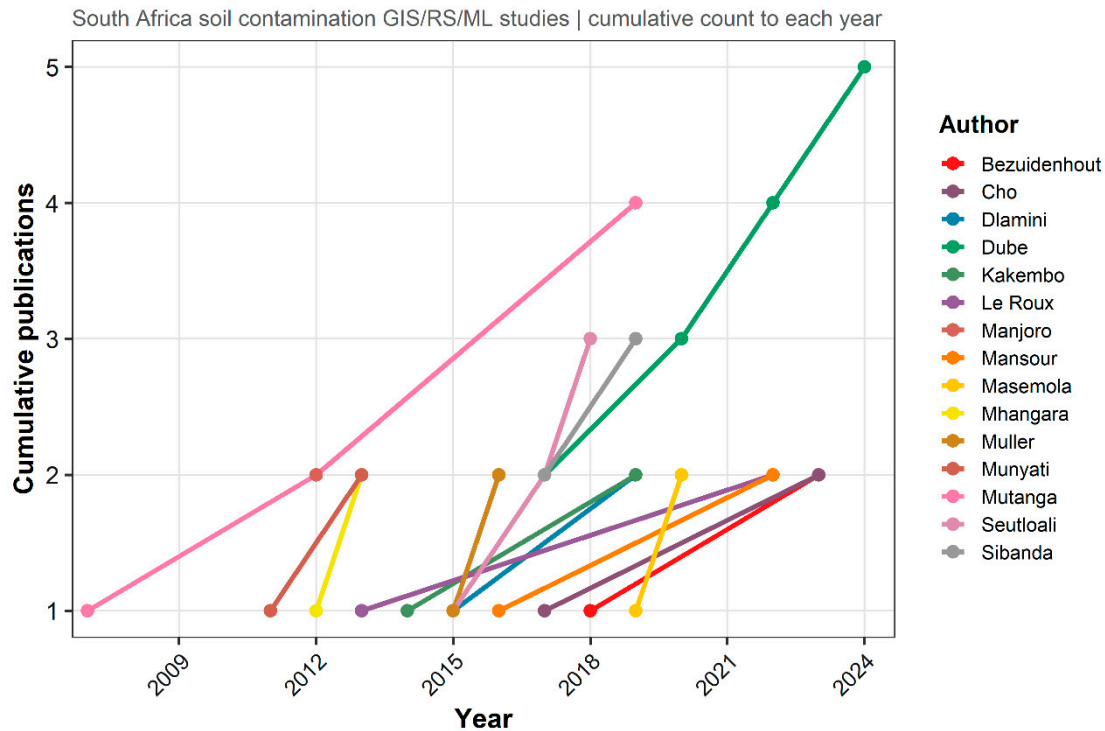
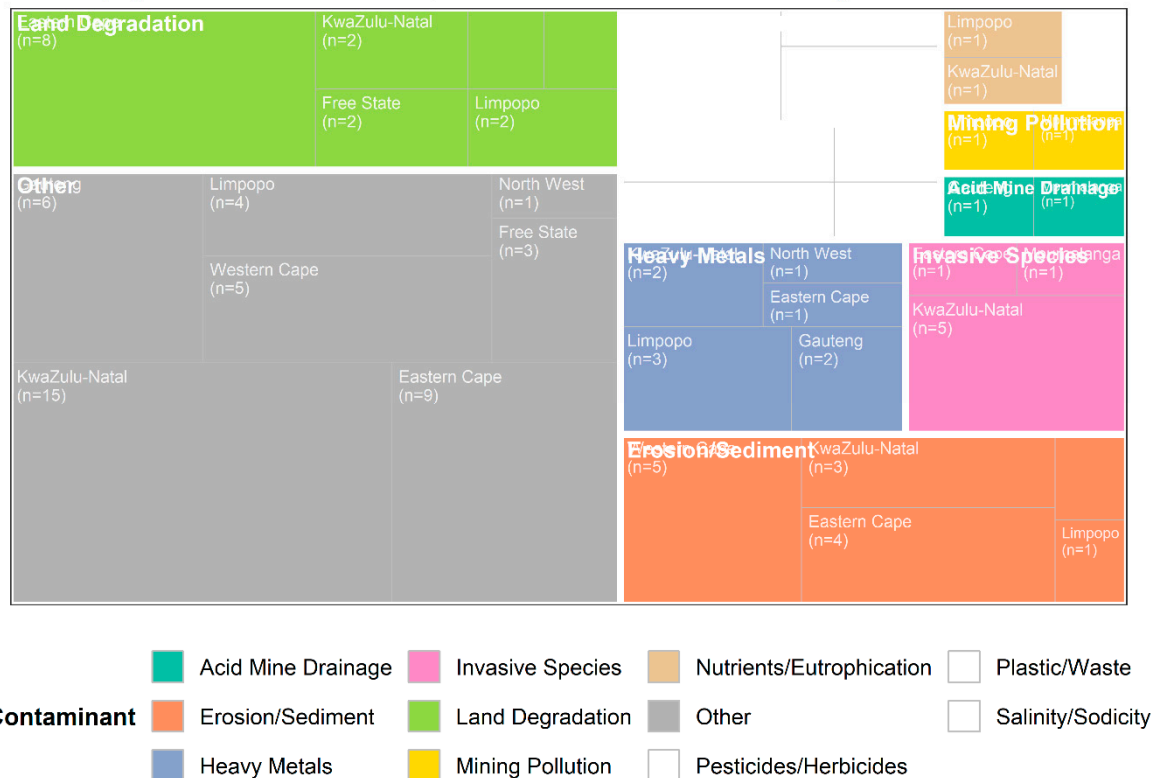


Figure S13. Technology adoption ranking bump chart showing annual rank by mention frequency across all methodological categories (2003–2025; $n = 228$ studies).

Annual technology adoption rankings were computed by counting, for each year, the number of studies mentioning each technology category (Remote Sensing, GIS/Mapping, Machine Learning, Field/Laboratory, Process Modelling, Hyperspectral, UAV/Drone) using the binary flag columns (*uses_rs*, *uses_gis*, *uses_ml*, *uses_field*, *uses_model*). Rank 1 denotes the most frequently applied technology in that year; ties are assigned the minimum rank. The panel documents Remote Sensing holding Rank 1 from approximately 2009 through 2019, Machine Learning rising to Rank 1 from approximately 2020 onwards, and GIS/Mapping occupying Rank 2 for most of the 2012–2024 period. This figure extends Figure 3C (main manuscript) by displaying all categories simultaneously and resolves year-on-year rank changes with sigmoid smoothing. Produced in R using *ggbump* (Sjöberg, 2021; GitHub: [davidsjoberg/ggbump](#)) and *ggplot2*. Seed = 42 for jitter. UAV = unmanned aerial vehicle.

Tile area proportional to number of studies | Other/Unclassified and National/Multi-regional excluded



2003–2025 | $n = 228$ studies

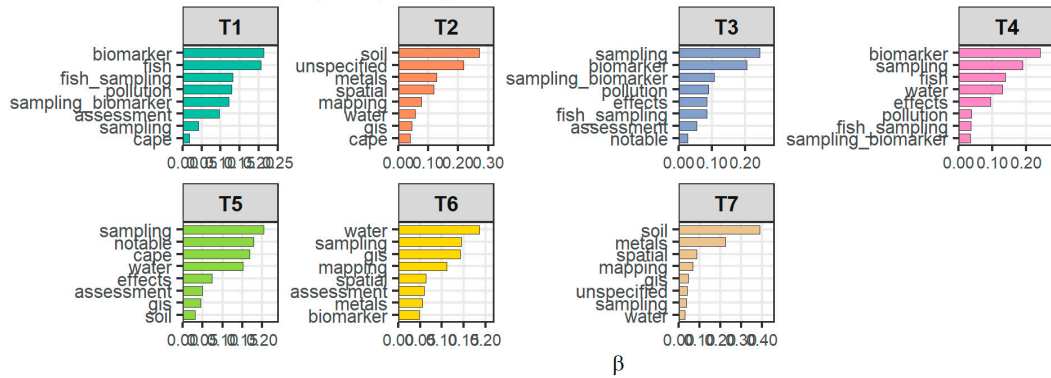
Figure S14. Contaminant category by province treemap showing proportional study-count distribution across South African provinces (2003–2025; $n = 162$ classifiable studies).

Treemap in which tile area is proportional to the number of included studies in each contaminant-category \times province cell. Tiles are grouped by province (outer rectangle) and shaded by contaminant/stressor category (inner tiles; colour palette consistent with Figure 3A in the main manuscript). Studies assigned to Other/Unclassified province ($n = 66$; 29.0% of the 228-study corpus) are excluded. The panel makes spatially explicit the concentration of Land Degradation and Erosion/Sediment studies in KwaZulu-Natal and the Eastern Cape, and the near-absence of Heavy Metals, Pesticides/Herbicides, and Acid Mine Drainage research in the Northern Cape despite documented contamination burdens in that province. Produced in R using *treemapify* and *ggplot2*. AMD = Acid Mine Drainage.

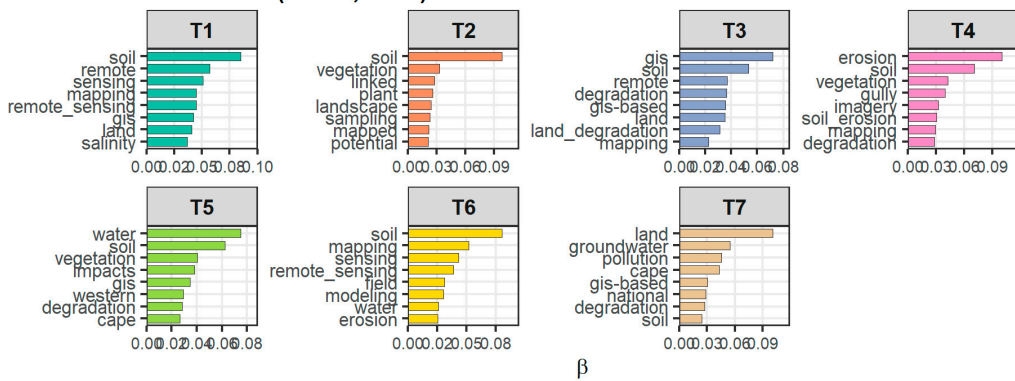
Top 8 terms per topic per era | Separate LDA fitted per era (Gibbs, seed = 42)

2003–2009: k=7 | 2010–2014: k=7 | 2015–2019: k=7 | 2020–2025: k=7

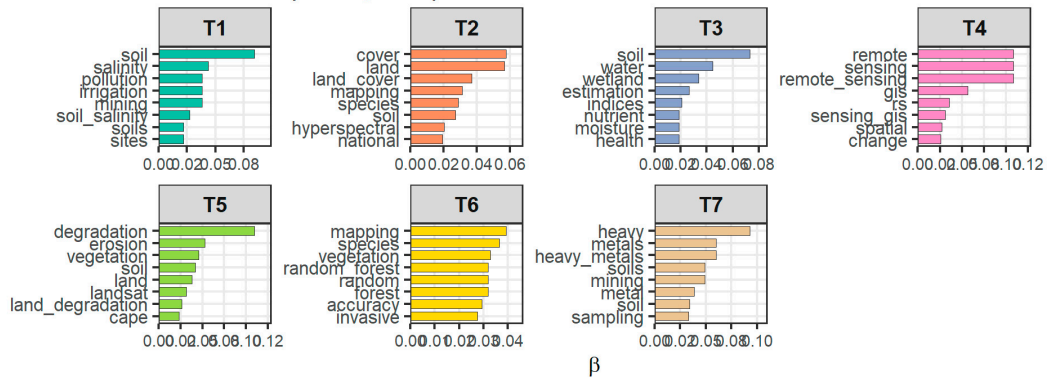
2003–2009 (n = 8, k = 7)



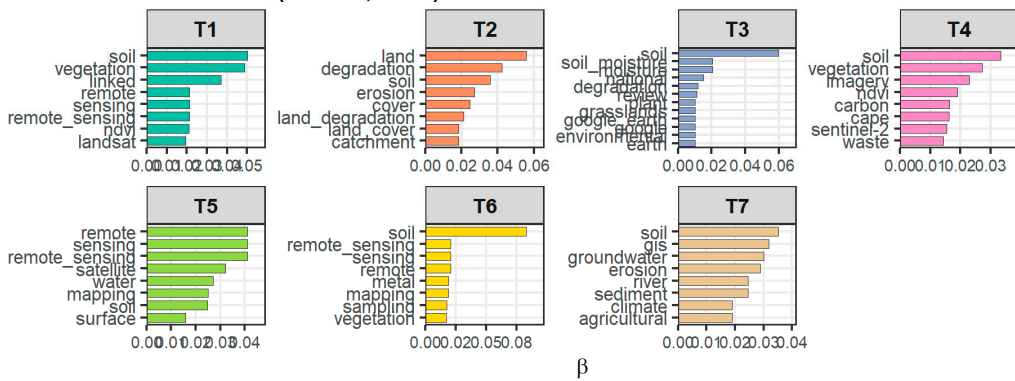
2010–2014 (n = 37, k = 7)



2015–2019 (n = 75, k = 7)



2020–2025 (n = 108, k = 7)

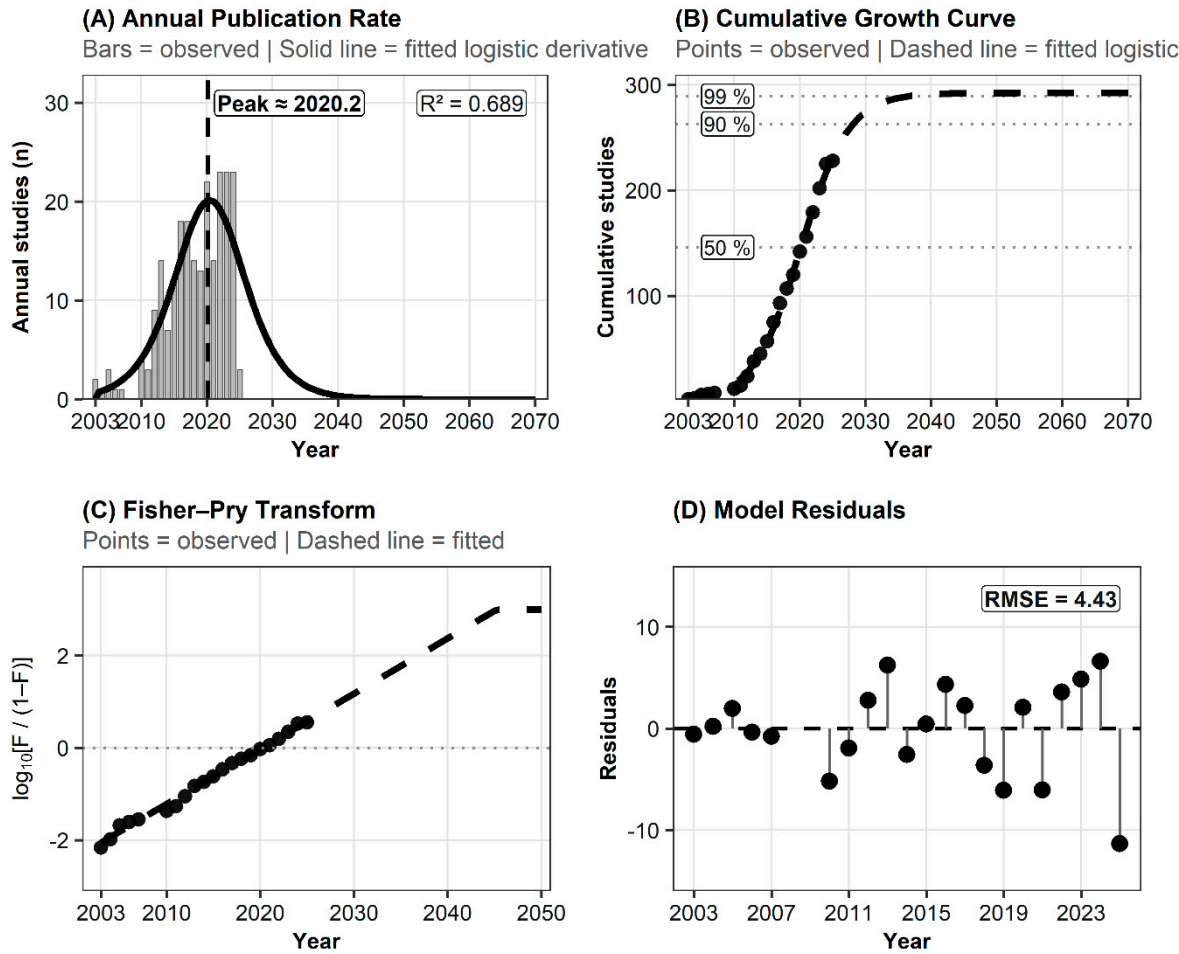


topicmodels::LDA | Gibbs | seed = 42

: numbers are era-specific and are NOT comparable across eras. Topic labels assigned post-hoc by authors from top-8 term distributions.

Figure S15. Dynamic LDA topic model: topic proportion trajectories across four research eras (2003–2009, 2010–2014, 2015–2019, 2020–2025; $k = 7$; $n = 228$ studies).

Logistic Growth Life-Cycle Model (Greyscale — Print Version)

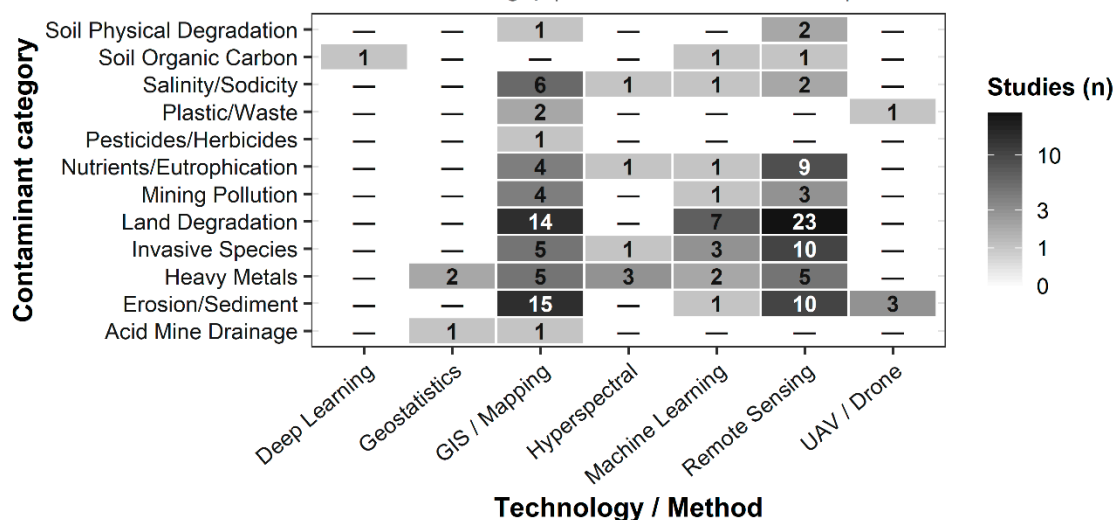


Greyscale version of Figure 6 | $K = 292$ [269–324] | $r = 0.275$ [0.251–0.302] | $t_{\square} = 2020.2$ [2019.4–2021.1] | 95% CI | 2003–2025
Line type distinguishes observed (points) from fitted (dashed) series

Evidence Gap Matrices (Greyscale — Print Version)

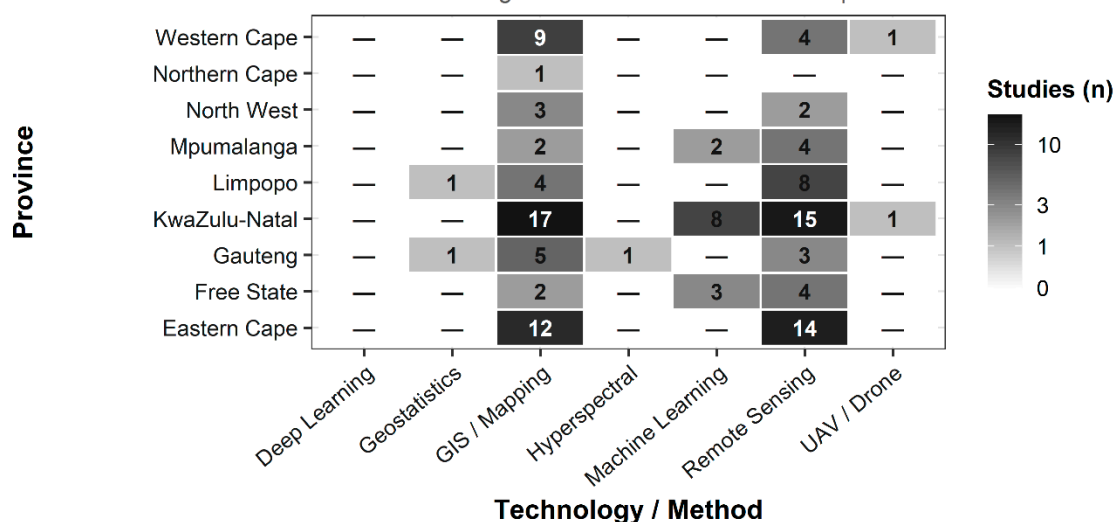
(A) Technology × Contaminant Evidence Matrix

White = evidence gap | Darker fill = more studies | Text confirms cell count



(B) Technology × Province Evidence Matrix

Excludes multi-regional and unclassified records | Other/Unclassified n = 66

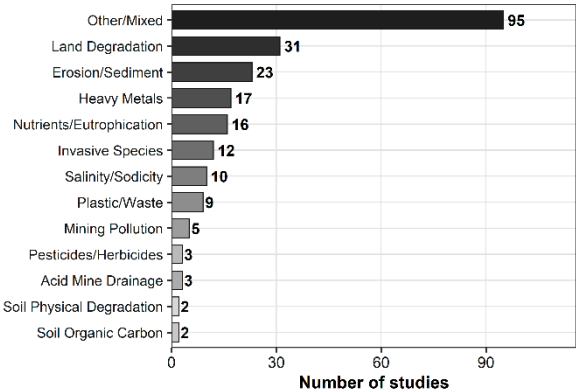


Greyscale version of Figure 7 | White = evidence gap | Grey intensity encodes study count (log₁₀ scale) | Numerical labels retained to disambiguate shading at low counts | Revised taxonomy: Other/Mixed excluded from Panel A | 2003–2025

Scope and Methodological Composition (Greyscale — Print Version)

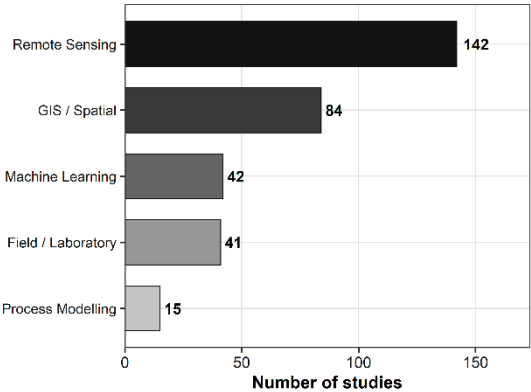
(A) Studies by Contaminant / Stressor Category

2003–2025 | n = 228 studies (revised taxonomy)



(B) Studies by Methodological Approach

Studies may use more than one method



Greyscale version of Figure 3 for print/accessibility compatibility | Revised contaminant taxonomy (Reviewer 1, Comment 10) | n = 228 studies | 2003–2025