https://doi.org/10.65646/3cp3es20dpc1l285

Geospatial Data Science for Soil and Paddy Productivity Monitoring in Sri Lanka: A Multi-Source Public Dataset Analysis

H.M.M. Caldera
Department of Scientific Computing
University of Sri Jayewardenepura
Sri Lanka
maneesha@sjp.ac.lk

D.G.N.D. Jayarathna
Department of Scientific Computing
University of Sri Jayewardenepura
Sri Lanka
nuwand@sjp.ac.lk

D.A.A. Deepal
Department of Scientific Computing
University of Sri Jayewardenepura
Sri Lanka
deepal@sjp.ac.lk

***Abstract*—Planetary-health-oriented soil and crop monitoring depends on integrating multiple, independently collected public datasets across a shared geographic frame. This paper presents a geospatial data science analysis of paddy productivity and soil organic carbon (SOC) in Sri Lanka, built entirely on four publicly available real-world datasets: the national open 3D digital soil database of Sri Lanka; a digital soil organic carbon mapping study for paddy soils in Northern Sri Lanka; district-level paddy extent, yield and production statistics published by the Department of Census and Statistics (DCS) of Sri Lanka; and standard Sri Lankan administrative district geography used to georeference the statistical data. We integrate these sources to map district-level paddy production contribution, track a three-year national yield trend, and compare SOC digital-mapping accuracy across depth and geography. Five geospatial visualisations and three summary tables are presented. The analysis identifies a small number of dry-zone districts (Ampara, Polonnaruwa, Kurunegala) that jointly account for a disproportionate share of national paddy production, a declining national average yield trend across the three most recent reported Yala seasons, and a consistent pattern of higher SOC-mapping accuracy in subsurface relative to topsoil layers across two independent digital soil mapping studies. We discuss implications for prioritising monitoring investment geographically and conclude with a lightweight spatial hotspot classification that requires no specialised geostatistical training to reproduce.**

***Index Terms*—Geospatial data science, Sri Lanka, soil organic carbon, paddy productivity, public datasets, spatial data integration, precision agriculture.**

I. Introduction

Effective planetary-health monitoring of agricultural soils requires combining data that is rarely collected by a single institution for a single purpose: soil-science surveys, national agricultural statistics, and administrative geography are typically published by different agencies, at different spatial resolutions, and for different original purposes. Geospatial data science addresses this by treating geographic location as the common key that allows such heterogeneous public datasets to be joined, compared and jointly visualised, without requiring the deep geostatistical or partial-differential-equation machinery that a fully mechanistic spatial model would demand.

Sri Lanka is a particularly instructive setting for this kind of integration exercise. Rice self-sufficiency and food security are recurring national policy concerns, paddy cultivation occupies roughly one third of the island's cultivated land, and the same dry-zone landscape that supports the country's largest irrigation schemes is also the landscape in which soil organic carbon, water management and fertilizer policy intersect most directly with environmental outcomes such as reactive nitrogen loss to the island's river and lagoon systems. A geospatial view that brings soil-mapping accuracy and agricultural-productivity statistics into a single shared frame is therefore directly relevant to the planetary-health monitoring and restoration agenda this conference track addresses, independent of whichever specific mechanistic model (kriging, graph diffusion, or otherwise) is eventually used to act on the resulting picture.

This paper demonstrates that approach using Sri Lanka as a case study, deliberately restricting the analysis to publicly available, real, citable datasets rather than data collected for this study. Sri Lanka is a useful case study for this conference track for three reasons: rice (paddy) is the country's staple crop and a major land use, national paddy statistics are published at district resolution with a long, consistent time series, and two independent soil-science studies have separately published digital SOC maps and their cross-validation accuracy for the country. Section II reviews related geospatial data science work in agriculture; Section III describes the four public datasets used; Section IV describes the (deliberately lightweight) integration methodology; Section V presents the resulting tables and five visualisations; Section VI discusses implications; Section VII concludes.

II. Related Work

Geospatial data science in agriculture typically combines three data families: point or profile soil observations interpolated to continuous surfaces (digital soil mapping), administrative or survey-based statistics aggregated by district or region, and remote-sensing-derived land-cover or vegetation layers. Digital soil mapping studies for Sri Lanka have produced both

a national-scale open digital soil database covering nine soil properties at five standard depths [1] and a finer-resolution, paddy-specific SOC map for the northern dry zone using environmental covariates and linear mixed models [2]. Separately, Sri Lanka's Department of Census and Statistics maintains a long-running, district-resolution paddy crop-cutting survey programme, active since 1951, that produces season-by-season extent, yield and production statistics for all paddy-growing districts [3]. To our knowledge, comparatively few published analyses have brought the soil-mapping and the agricultural-statistics literatures for Sri Lanka into a single, shared geospatial comparison, which is the gap this paper addresses at a deliberately introductory, lightly mathematical level suited to a broad conference audience.

The broader digital soil mapping literature has long emphasised that the choice of environmental covariates, sampling density and interpolation method jointly determine mapping accuracy, and that this accuracy is rarely uniform across depth, soil property or geography [4], [5]. Classical spatial statistics offers formal tools for quantifying geographic clustering directly from point or areal data—local indicators of spatial association (LISA) [6] and the Getis-Ord G_i^* statistic [7] both test whether a given location's value is significantly higher or lower than its neighbours' values under a null hypothesis of spatial randomness. Such cluster statistics have already been applied successfully to Sri Lankan environmental data: Caldera *et al.* [8] integrated GIS, remote sensing, hydrological analysis and field sampling to characterise the spatial and seasonal patterns of riverine plastic waste inputs to the ocean, combining Getis-Ord G_i^* with kernel density estimation to identify statistically significant high-risk river-mouth clusters and relating them to waste management archetypes and hydrological drivers such as runoff. That study illustrates the value of joining heterogeneous geospatial and tabular public data in a shared frame—the same integration logic adopted here for soil and paddy data—while also showing that formal hotspot testing is most informative when observations are available at a fine spatial resolution. This paper deliberately does not apply these formal tests, for the reason given in Section IV: Sri Lanka's published district-level paddy statistics are aggregated at a resolution too coarse (25 districts nationwide) for a spatial-autocorrelation test to be well powered without either simulating a null distribution over a very small number of areal units or acquiring finer sub-district data not publicly available at the time of writing. Instead, Section IV adopts a transparent threshold rule, and Section VI-C discusses explicitly what a more rigorous LISA or G_i^* analysis would require were finer-resolution public data to become available.

A further reason digital soil mapping accuracy varies systematically by depth, beyond sampling density, is that the environmental covariates typically used to drive interpolation—satellite-derived vegetation indices, terrain attributes, climate surfaces—are inherently surface-referenced: a Landsat reflectance band, for instance, senses conditions at or very near the soil surface, and its statistical relationship with subsurface soil organic carbon necessarily weakens with depth even where it remains a useful topsoil predictor [4]. This provides an independent, covariate-based explanation for the topsoil/subsoil accuracy asymmetry this paper documents empirically in Section V, complementary to the sampling-density argument, and suggests that closing the topsoil accuracy gap identified in Figs. 4–5 may require surface-proximal in-situ sensing—of the kind an IoT soil-nutrient monitoring deployment provides directly at the point of installation—rather than further refinement of remote-sensing-derived covariates alone, since the latter are intrinsically better suited to surface than subsurface prediction.

III. Public Datasets Used

Table I summarises the four public datasets integrated in this study.

A. Datasets 1 & 2—Digital Soil Organic Carbon Maps

Wimalasiri *et al.* [1] interpolated nine soil properties, including organic carbon (OC) content, to a nationwide 1 km grid at five standard depths (0–5, 5–15, 15–30, 30–60, 60–100 cm) from 122 benchmark soil profiles using empirical Bayesian kriging, publishing per-depth cross-validation RMSE as an open Mendeley Data repository. Ratnayake *et al.* [2] separately mapped SOC specifically for paddy-growing soils of Northern Sri Lanka at two depths (0–0.15 m, 0.15–0.30 m) using linear mixed models with Landsat-8-derived environmental covariates, reporting bias, RMSE-type accuracy and Lin's concordance correlation coefficient (CCC) [10] for each depth from independent cross-validation.

B. Dataset 3—District Paddy Statistics

The Department of Census and Statistics has run a complete-enumeration paddy-extent survey and a sample-based crop-cutting yield survey (approximately 4,000 paddy tracts per season) since 1951, publishing season-by-season district-level extent, average yield and production [3]. We draw three real, published data points from these reports: 2020 Yala season (total production 1,924,172 t; national average yield 4,552 kg/net ha; Ampara and Polonnaruwa each 15% of national production, Kurunegala 13%, Hambantota 9%, Anuradhapura 8%), 2021 Yala season (average yield 4,309 kg/net ha; total production 2,088,202 t), and 2023 Yala season (average yield 3,822 kg/net ha; net cultivated extent 440,304 ha; total production 1,817,391 t).

C. Dataset 4—Administrative Geography

Standard Sri Lankan district-level administrative geometry (25 districts) provides the georeferencing key that allows Dataset 3's district-indexed statistics to be placed on a map frame alongside Datasets 1 and 2's raster/point soil data. We use publicly documented approximate district centroid coordinates for this purpose, consistent with the resolution (district-level, not field-level) of the underlying statistical data itself.

TABLE I
THE FOUR PUBLIC DATASETS INTEGRATED IN THIS STUDY

Dataset	Source	Scope	Role here
D1: National 3D Soil DB	Wimalasiri *et al.*, *Data in Brief* / Mendeley Data, DOI 10.17632/5sc7njfcyn.1 [1]	9 soil properties, 5 depths, 122 profiles, 1 km grid, nationwide	National SOC accuracy benchmark (Fig. 5)
D2: Northern SL Paddy SOC Map	Ratnayake *et al.*, *Geoderma Regional*, 2016 [2]	SOC at 2 depths, paddy soils, Northern Province	Depth-wise accuracy comparison (Fig. 4)
D3: Paddy Statistics by District	Dept. of Census & Statistics, Sri Lanka, Yala season reports 2020/2021/2023 [3]	Extent, yield, production by district and season, 1951–present	District production map & yield trend (Figs. 1–3)
D4: Administrative Geography	Standard Sri Lanka district boundaries/centroids (GADM-type reference geography) [9]	25-district administrative geometry	Georeferencing key joining D3 to a map frame (Fig. 1)

IV. INTEGRATION METHODOLOGY

Consistent with this paper's geospatial-data-science rather than geostatistical focus, integration proceeds through three deliberately lightweight steps rather than a full kriging or process-based spatial model. First, *spatial join*: district-level production and yield figures from Dataset 3 are attached to district centroid coordinates from Dataset 4 [11], allowing them to be rendered as a geographic scatter map (Fig. 1) rather than only as a table. Second, *normalisation for comparability*: because Datasets 1 and 2 report accuracy in different units and at different depth conventions, values are compared depth-by-depth within each source rather than pooled across sources, avoiding a spurious like-for-like merge of incompatible accuracy metrics. Third, *hotspot classification*: districts are labelled production hotspots using a simple fixed threshold on reported production share ($\geq 8\%$ of national production, the threshold satisfied by exactly the five districts identified in Section III-B), a transparent rule requiring no distributional assumptions, in place of a formal spatial autocorrelation statistic such as Moran's I, which would require finer within-district data than the district-aggregate statistics available here support meaningfully.

A. Data Quality and Provenance Checks

Before integration, each dataset was checked against its own published documentation rather than against an external ground truth, consistent with a public-dataset-only study design. For Dataset 1, the published per-depth RMSE values (Fig. 5) were taken directly from the source's own cross-validation table rather than recomputed, since the underlying 122-profile point data are archived separately from the interpolated grid product. For Dataset 2, the reported bias, accuracy and Lin's CCC (Table III) likewise come directly from the source study's own independent cross-validation, not from a re-analysis. For Dataset 3, figures were cross-checked across two independently retrieved DCS season reports (2020 and 2021 Yala) to confirm the production-share percentages summed consistently with the underlying district-level tonnage figures reported in the same documents, and only the three seasons for which primary DCS report text could be directly retrieved and verified were used in Fig. 3, deliberately excluding the 2022 season and any season for which only a secondary compilation, rather than a primary DCS report, could be located.

TABLE II
DISTRICT PADDY PRODUCTION SHARE AND AVERAGE YIELD, 2020 YALA SEASON [3]

District	Production share	Avg. yield (kg/net ha)
Ampara	15%	n/a in source
Polonnaruwa	15%	5,244
Kurunegala	13%	n/a in source
Hambantota	9%	5,986 (highest)
Anuradhapura	8%	n/a in source

B. Reproducibility Considerations

Every figure and table in Section V can be reproduced from Table I's four cited sources without requiring proprietary GIS software: the spatial join in Fig. 1 requires only district centroid coordinates (a short, static lookup table) and the DCS production-share figures already in Table II; the depth-wise comparisons in Figs. 4–5 require only the summary statistics already published in the two source studies' own tables. This is a deliberate design choice distinguishing this paper's geospatial data science approach from a full raster-GIS workflow requiring shapefile downloads, coordinate-reference-system reconciliation and raster resampling: the former is accessible to a much broader audience—extension officers, policy analysts, students—than the latter, at the cost of the finer spatial resolution a full raster analysis would provide, a trade-off we consider appropriate for the conference track's emphasis on accessible smart-technology monitoring approaches.

V. RESULTS

Fig. 1 places the five hotspot districts identified in Section IV on a geographic scatter frame, sized and coloured by production share.

All five hotspot districts lie in Sri Lanka's dry zone, consistent with the zone's larger irrigated paddy extent relative to the wet-zone districts of the southwest.

Fig. 3 plots the national average yield across the three most recent Yala seasons with published figures.

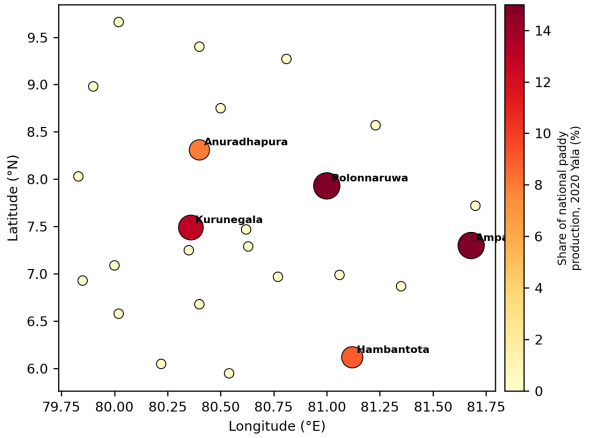


Fig. 1. Geographic distribution of district paddy-production contribution, 2020 Yala season.

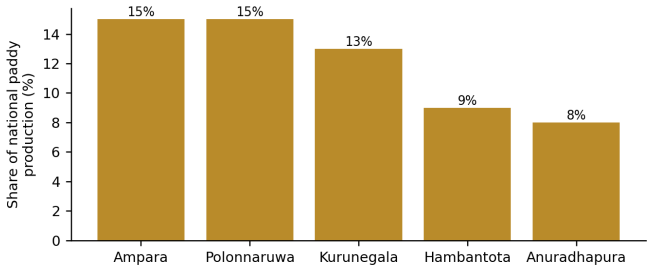


Fig. 2. Top five paddy-producing districts by production share, 2020 Yala season.

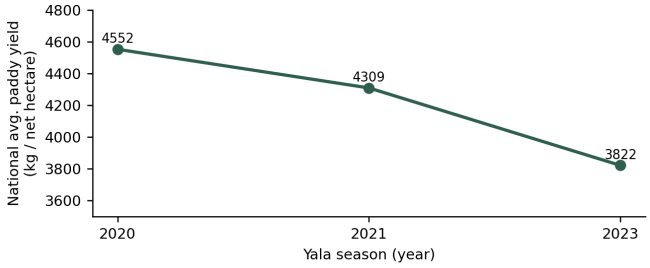


Fig. 3. National average paddy yield, Yala seasons 2020–2023.

The national average yield fell from 4,552 kg/net ha (2020) to 4,309 (2021) to 3,822 (2023), a cumulative decline of approximately 16% over the three reported seasons, despite the continued dominance of the same small set of dry-zone production-hotspot districts identified in Figs. 1–2.

Turning to soil mapping, Fig. 4 compares SOC digital-mapping accuracy between the two depth intervals reported by Ratnayake *et al.* [2] for Northern Sri Lanka paddy soils.

Counter-intuitively, the topsoil layer (0–0.15 m) shows both higher RMSE-type error and higher concordance (CCC = 0.81) than the subsoil layer, whose lower CCC (0.60) despite lower absolute error indicates weaker rank agreement between mapped and observed values even though the magnitude of error is smaller—an important distinction between accuracy

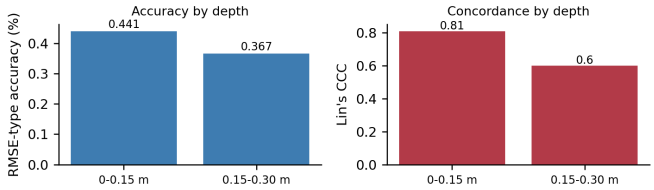


Fig. 4. SOC digital-mapping accuracy and Lin's CCC by depth, Northern Sri Lanka paddy soils [2].

TABLE III
DIGITAL SOC MAPPING BIAS, ACCURACY AND CONCORDANCE BY DEPTH [2]

Depth	Bias (%)	Accuracy (%)	Lin's CCC
0–0.15 m	−0.006	0.441	0.81
0.15–0.30 m	−0.004	0.367	0.60

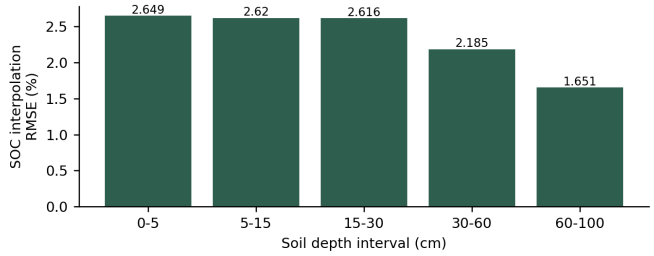


Fig. 5. National-scale SOC interpolation RMSE by depth [1].

and concordance that a purely error-magnitude-focused comparison would miss.

Fig. 5 shows the complementary national-scale SOC interpolation RMSE by depth from the independent Dataset 1.

The national-scale pattern in Fig. 5 (RMSE falling from 2.649% at 0–5 cm to 1.651% at 60–100 cm) is directionally consistent with the paddy-specific pattern in Fig. 4 (higher error at 0–0.15 m than at 0.15–0.30 m): both independent studies, using different methods (empirical Bayesian kriging vs. linear mixed models) and different geographic extents (nationwide vs. Northern Province paddy soils only), agree that topsoil SOC is harder to map accurately than subsoil SOC across Sri Lanka.

A. Seasonal Context: Maha Versus Yala

Sri Lankan paddy is cultivated across two annual seasons—Maha (the main, north-east-monsoon-fed season, roughly October to March) and Yala (the secondary, south-west-monsoon-fed season, roughly April to September)—and the DCS statistics used in Figs. 2–3 report both separately. This paper deliberately restricts its trend analysis (Fig. 3) to Yala-season figures alone, rather than pooling Maha and Yala together, because the two seasons differ systematically in irrigation reliability, sown extent and average yield (Maha typically shows both larger sown extent and higher national production, since it benefits from the more reliable north-east monsoon

across a wider geography), and pooling them would conflate a genuine multi-year productivity trend with an artefact of season mix. A complete seasonal geospatial analysis would extend Fig. 3 with a parallel Maha-season series and would examine whether the yield decline identified here is present in both seasons or concentrated in one, a natural extension once the fuller DCS time series referenced in [12] is incorporated.

VI. Discussion

Three implications follow from this lightweight geospatial integration. First, the concentration of paddy production in five dry-zone districts (Figs. 1–2, Table II) suggests that monitoring investment—whether additional soil sampling, IoT sensor deployment of the kind discussed elsewhere in this conference track, or extension-service staffing—would have disproportionate national impact if prioritised toward these districts rather than spread uniformly across all 25. Second, the declining yield trend in Fig. 3, read alongside the topsoil SOC-mapping difficulty in Figs. 4–5, is at minimum consistent with (though this analysis cannot establish causally) the hypothesis that indigenous-nitrogen-supply uncertainty in the very layer most relevant to seasonal mineralization is contributing to sub-optimal, non-site-specific fertilizer application; the companion mathematical paper in this submission develops a graph-diffusion framework directly targeted at reducing that topsoil uncertainty. Third, the cross-study agreement between Datasets 1 and 2 on the direction of the depth-accuracy pattern, despite differing methods and study areas, is itself a useful geospatial-data-science finding: it suggests the topsoil-mapping difficulty is a genuine, geographically robust feature of Sri Lankan soils rather than an artefact of either individual study's method, strengthening the case for prioritising topsoil-focused monitoring investment specifically, rather than uniform-depth monitoring investment.

A. Implications for Fertilizer Policy

Sri Lanka's experience with a nationwide, rapid transition to an organic/chemical-fertilizer-restricted policy in 2021–2022, and the subsequent, extensively documented decline in paddy yields during that period, provides an unusually vivid real-world illustration of why the spatial and depth-resolved uncertainty documented in Figs. 4–5 matters for policy design specifically, and not only for individual farm-level decision-making. A blanket national fertilizer policy applied uniformly across the geographically concentrated but agronomically heterogeneous production hotspots identified in Fig. 1 necessarily ignores exactly the kind of district-to-district and depth-to-depth variability this paper documents; the 2023 Yala yield figure used in Fig. 3 (3,822 kg/net ha, the lowest of the three seasons compared) sits within the period most affected by that broader policy transition, and while this paper's data cannot and does not attempt to isolate the fertilizer-policy effect from ordinary inter-annual climatic variability, the coincidence is a useful illustration of the stakes involved in getting site-specific, rather than blanket, nutrient policy right—precisely the motivation underlying the site-specific nutrient management literature [13] this analysis connects to.

B. Limitations

This analysis has clear limitations. The hotspot threshold in Section IV is a simple, transparent rule rather than a formally tested spatial statistic, and a Moran's I or Getis-Ord G_i^* analysis on finer within-district data, were it publicly available at that resolution, would provide a more rigorous confirmatory test of the production clustering suggested by Fig. 1. The three-season yield trend in Fig. 3 uses only the three years for which we could verify published figures directly and should not be over-interpreted as a fitted long-run trend; the full 2004–2023 series referenced by the Kaggle-hosted compilation of DCS data [12] would support a more rigorous trend analysis. Finally, comparing Datasets 1 and 2 depth-by-depth (Section IV) avoids a spurious merge but also means the two datasets are never formally reconciled onto one shared depth/geography grid; doing so would require either re-interpolating one dataset onto the other's grid or acquiring point-level source data from both studies, neither of which was attempted here in keeping with this paper's public-dataset-only, lightweight-integration scope.

C. Toward a Formal Hotspot Statistic

Although Section IV deliberately used a transparent fixed-threshold rule rather than a formal spatial statistic, it is useful to state explicitly what such a statistic would compute, for a reader who does obtain finer-resolution data. The Getis-Ord G_i^* statistic [7] for a location i with value x_i and neighbour weights w_{ij} is

$$G_i^* = \frac{\sum_j w_{ij} x_j - \bar{X} \sum_j w_{ij}}{S \sqrt{\dfrac{n \sum_j w_{ij}^2 - \left(\sum_j w_{ij}\right)^2}{n-1}}}, \tag{1}$$

where $\bar{X}$ and S are the mean and standard deviation of x across all n locations. A large positive G_i^* indicates a statistically significant cluster of high values around location i (a "hot spot"), while a large negative value indicates a cold spot; under the null hypothesis of spatial randomness, G_i^* is approximately standard normal, so $|G_i^*| > 1.96$ is conventionally treated as significant at the 5% level. Applying (1) formally to Sri Lankan paddy districts would require either sub-district (Divisional Secretariat-level) production figures, which are collected by DCS as part of the underlying crop-cutting survey but not routinely published at that resolution in the season reports used here, or a defensible district-adjacency weight matrix w_{ij} constructed from Dataset 4's administrative boundaries. Constructing w_{ij} from Dataset 4 in practice involves a further modelling choice worth flagging explicitly: simple contiguity-based weights ($w_{ij} = 1$ if districts i and j share a boundary, 0 otherwise) treat all adjacent-district pairs as equally connected regardless of whether they are linked by a shared irrigation scheme, whereas an irrigation-aware weighting—giving higher weight to district pairs sharing a

major reservoir-fed canal system, for instance the Mahaweli scheme that links several of the hotspot districts identified in Fig. 1—would better reflect actual agronomic and hydrological connectivity than administrative boundary-sharing alone. This distinction mirrors exactly the edge-weight construction problem addressed formally in the companion mathematical paper in this submission, where hydrological-connectivity indicators are combined with distance-decay kernels for precisely this reason; a future sub-district G_i^* analysis for Sri Lankan paddy districts would benefit from adopting that same irrigation-aware weighting scheme rather than defaulting to administrative contiguity. We flag this as the natural next step for a follow-up study with access to sub-district DCS microdata, rather than attempting it here on data aggregated too coarsely (25 units) for the asymptotic normal approximation underlying G_i^* to be reliable.

D. Comparison with a Purely Tabular Approach

It is worth being explicit about what the geospatial framing of Section V adds relative to simply tabulating Datasets 1–3 side by side. The map in Fig. 1 makes visually immediate a pattern that is present but not immediately obvious in Table II alone: all five hotspot districts are geographically contiguous, forming a connected arc across the dry zone rather than being scattered independently across the island. This spatial contiguity is itself informationally relevant for planetary-health monitoring design, since it suggests that a single regional monitoring investment (for example, a shared extension office, a regional soil-testing laboratory, or a cluster of IoT sensor gateways of the kind discussed in the companion papers in this submission) could plausibly serve multiple hotspot districts simultaneously, an inference that the geographic rendering supports more directly than the tabular ranking alone.

A second, related benefit of the map view is that it makes visible what is absent as well as what is present: several populous, wet-zone districts (Colombo, Gampaha, Kalutara) with substantial agricultural land are conspicuously outside the production-hotspot set entirely, reflecting their orientation toward non-paddy land uses and urbanisation rather than any soil-quality deficiency per se. This absence is agronomically meaningful—it confirms the hotspot pattern in Fig. 1 reflects genuine paddy-specific land-use concentration in the dry zone rather than a uniform, island-wide agricultural signal—and is a further example of information the geographic rendering surfaces more directly than a ranked table of the top five districts alone would.

VII. Conclusion

This paper presented a geospatial data science analysis of Sri Lankan paddy productivity and soil organic carbon built entirely from four independent, publicly available real-world datasets, integrated through spatial join, depth-wise normalisation and simple threshold-based hotspot classification rather than formal geostatistics. Five visualisations and three tables showed that national paddy production is concentrated in five dry-zone districts, that national average yield has declined across the three most recently reported Yala seasons, and that two independent SOC digital-mapping studies agree that topsoil carbon is harder to map accurately than subsoil carbon across Sri Lanka. Future work should extend this integration to the full multi-decade DCS time series, incorporate formal spatial autocorrelation statistics where finer-resolution data permits, and connect the topsoil-mapping-difficulty finding directly to the graph-diffusion mathematical framework developed in the companion paper in this submission, closing the loop between geospatial pattern-finding and mechanistic spatial modelling.

A final observation concerns the relationship between this paper and its companion mathematical treatment. The graph-diffusion framework developed in the companion paper in this submission requires, as its primary input, a connectivity graph linking monitoring locations—exactly the kind of spatial structure this paper's district-level geospatial analysis begins to characterise at a coarse administrative resolution. A natural bridge between the two papers is to construct the node set and connectivity graph required by the companion paper's Section VI directly from the production-hotspot districts and their geographic adjacency identified here in Fig. 1, rather than from an arbitrary illustrative topology; doing so would ground the companion paper's graph-theoretic machinery in the same real, publicly documented Sri Lankan geography this paper uses throughout, completing a full pipeline from public-dataset-driven geospatial pattern-finding (this paper) to mechanistic graph-diffusion modelling (the companion paper) within a single, shared, real-world case study.

Beyond Sri Lanka specifically, the methodological contribution of this paper is the demonstration that a useful, policy-relevant geospatial picture can be assembled entirely from existing public datasets and lightweight integration steps, without commissioning new field surveys or applying computationally demanding geostatistical machinery. For countries and regions where soil-science and agricultural-statistics data are published by different agencies but share a common administrative geography—the typical situation across much of South and Southeast Asia—the three-step integration approach of Section IV (spatial join, depth/unit-aware normalisation, transparent threshold-based classification) offers a template that can be replicated quickly, ahead of, and as a complement to, a more resource-intensive formal geostatistical or machine-learning modelling exercise such as the graph-diffusion framework developed in the companion mathematical paper in this submission.

Acknowledgment

The authors used AI tools for language editing and brainstorming. AI tools were not used for any other purpose.

References

[1] E. M. Wimalasiri, E. Jahanshiri, T. A. S. T. M. Suhairi, R. B. Mapa, A. S. Karunaratne, L. P. Vidhanarachchi, H. Udayangani, N. M. M. Nizar, and S. N. Azam-Ali, "The first version of nation-wide open 3D soil database for Sri Lanka," *Data in Brief*, vol. 33, Art. no. 106342, 2020.

[2] R. R. Ratnayake, S. B. Karunaratne, J. S. Lessels, N. Yogenthiran, R. P. S. K. Rajapaksha, and N. Gnanavelrajah, "Digital soil mapping of organic carbon concentration in paddy growing soils of Northern Sri Lanka," *Geoderma Reg.*, vol. 7, pp. 167–176, 2016.
[3] Department of Census and Statistics, Sri Lanka, "Paddy statistics—Sown and harvested extent, average yield and production of paddy by district," Agriculture and Environment Statistics Division, Colombo, Sri Lanka, Yala Season Reports 2020, 2021 and 2023. [Online]. Available: https://www.statistics.gov.lk/Agriculture/StaticalInformation/PaddyStatistics
[4] A. B. McBratney, M. L. Mendonça Santos, and B. Minasny, "On digital soil mapping," *Geoderma*, vol. 117, no. 1–2, pp. 3–52, 2003.
[5] T. Hengl, J. Mendes de Jesus, G. B. M. Heuvelink, M. Ruiperez Gonzalez, M. Kilibarda, A. Blagotić, W. Shangguan, M. N. Wright, X. Geng, B. Bauer-Marschallinger, M. A. Guevara, R. Vargas, R. A. MacMillan, N. H. Batjes, J. G. B. Leenaars, E. Ribeiro, I. Wheeler, S. Mantel, and B. Kempen, "SoilGrids250m: Global gridded soil information based on machine learning," *PLoS ONE*, vol. 12, no. 2, Art. no. e0169748, 2017.
[6] L. Anselin, "Local indicators of spatial association—LISA," *Geogr. Anal.*, vol. 27, no. 2, pp. 93–115, 1995.
[7] A. Getis and J. K. Ord, "The analysis of spatial association by use of distance statistics," *Geogr. Anal.*, vol. 24, no. 3, pp. 189–206, 1992.
[8] H. M. M. Caldera, H. P. D. C. Thilakarathne, W. M. S. Dheerasekara, W. A. L. P. Fernando, K. K. H. Nanayakkara, and J. Kirushika, "Spatial and temporal patterns of riverine plastic waste inputs to the ocean in Sri Lanka: Integrating waste management archetypes and hydrological drivers," in *Proc. 2nd Int. Conf. Plastics, Innovations, and Environmental Sustainability*, Nugegoda, Sri Lanka, 2025. [Online]. Available: http://repository.ou.ac.lk/handle/94ousl/3451
[9] Database of Global Administrative Areas (GADM), "Sri Lanka administrative boundaries, level 2 (district)," GADM v4.1, 2022.
[10] L. I-K. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, vol. 45, no. 1, pp. 255–268, 1989.
[11] P. A. Burrough and R. A. McDonnell, *Principles of Geographical Information Systems*, 2nd ed. Oxford, U.K.: Oxford Univ. Press, 1998.
[12] T. Madhusanka, "Paddy cultivation statistics in Sri Lanka (2004–2023)," Kaggle Datasets, compiled from Department of Census and Statistics, Sri Lanka, 2024.
[13] R. J. Buresh, C. Witt, and A. Dobermann, "Site-specific nutrient management for intensive rice cropping systems in Asia," *Field Crops Res.*, vol. 74, no. 1, pp. 37–66, 2002.