Comparative Evaluation of Spatial Indexing Methods Applied to the Georeferenced Characterization of Agricultural Units and Productivity in Peru during the year 2024

Ilma Magda Mamani Mamani[1] [https://orcid.org/0009-0002-8605-0086]

Faculty of Statistical and Computer Engineering
National University of the Altiplano, Puno, PERU
im.mamani@est.unap.edu.pe

Abstract. The efficient analysis of large volumes of georeferenced data is essential for modern agro-statistical management. This study compares the efficiency of four spatial indexing methods—R-Tree, Quad-Tree, KD-Tree, and Grid—applied to the microdata of the National Agricultural Survey (ENA 2024) of Peru. Coordinates and productive variables (area, production, and losses) were integrated into a spatial database processed in R using the libraries `sf`, `terra`, `spatstat`, and `FNN`. The results show significant contrasts in performance: the Grid method achieved the lowest query times (4–6 ms) and greater stability in heterogeneous regions; KD-Tree was superior in neighborhood queries (100 QPS), while R-Tree excelled in complex geometries at the cost of higher memory consumption (61.8 MB). These findings confirm that partitioning structures offer substantial advantages for the dynamic analysis of large volumes of agricultural data, providing a replicable basis for the modernization of statistical systems and agrarian territorial management.

Keywords: Spatial indexing, Geographic queries, Precision agriculture, Agricultural productivity

1 Introduction

The use of geospatial data constitutes a strategic component for sustainable agriculture and the formulation of policies based on territorial evidence. In Peru, the National Agricultural Survey (ENA 2024) of the INEI offers detailed information on georeferenced agricultural units, including coordinates, administrative codes, and productive variables. However, conventional statistical systems present limitations for executing efficient spatial queries, proximity searches, or detection of overlaps between units. Spatial indexing structures—such as R-Tree, Quad-Tree, KD-Tree, and Grid—optimize access to multidimensional data and improve analytical response. Previous studies show variations in their performance according to the density, distribution, and geometric complexity of the dataset [1, 5, 7, 8]. Recent research has demonstrated the potential of hybrid indexes to improve performance in large-scale agricultural applications [15, 19].

2 Objectives and Specific Objectives

The general objective of this study is to comparatively evaluate the efficiency and applicability of four spatial indexing methods (R-Tree, Quad-Tree, KD-Tree, and Grid) applied to the microdata of the ENA 2024 of Peru, with the purpose of optimizing geospatial queries and the territorial characterization of agricultural productivity.

2.1 Specific Objectives

1. Analyze the construction time, memory consumption, and query latency of each spatial indexing method.
2. Evaluate the efficiency of the indexes for range queries (buffers) and nearest neighbors (KNN).
3. Identify variations in the performance of the indexes according to the natural region (Coast, Sierra, and Jungle).
4. Apply the selected methods in two real scenarios: pest detection and drought impact.

3 Methodology

3.1 Data Acquisition and Preparation

During the integration stage of the ENA 2024 microdata, spatial and productive variables with analytical relevance and national coverage were selected (Table 1).

Table 1. Main variables used

Variable	Description	Type	Analytical Use
LATITUDE, LONGITUDE	Geographic coordinates (WGS84)	N	Spatial location
REGION	Natural classification (Coast, Sierra, Jungle)	C	Regional analysis
P217_SUP_ha	Harvested area (ha)	N	Agricultural density calculation
P219_CANT_1	Main production (kg)	N	Productive indicator
P224B	Crop losses (kg)	N	Vulnerability/efficiency
P223A	Presence of climatic impact (binary)	B	Risk identification
RENDIMIENTO_kg_ha	Production per hectare (derived)	N	Productivity evaluation

3.2 Analysis and Processing

Each agricultural unit was represented as a point in two-dimensional space, associated with its productive attributes. The analysis was implemented in R (version 4.3) with the packages `sf`, `terra`, `spatstat`, and `FNN`. The methodological flow included:

1. Massive data loading and cleaning (valid coordinates, imputation of missing values).
2. Construction of spatial indexes (R-Tree, Quad-Tree, KD-Tree, Grid).
3. Execution of spatial queries (by range, proximity, and overlap), with recording of performance metrics (response time, disk accesses, number of nodes, memory used).

3.3 Theoretical Description of the Models

R-Tree: hierarchical structure based on *minimum bounding rectangles* (MBR), where each node N_i groups a set of spatial objects S_i within a minimum area such that:

$$MBR(N_i) = \min_{R}\{\, R \supseteq \bigcup_{p \in S_i} p \,\}.$$

where R is the minimum rectangle that contains the union of points in S_i [1]. Searches are performed by verifying the intersection between the MBRs and the queried regions, with expected complexity $O(\log n)$. Its efficiency in multidimensional queries is high, although it can degrade due to overlaps when the data is heterogeneous [4]. Recent studies have proposed significant improvements in R-Tree optimization for agricultural data [16]. **Quad-Tree**: recursively divides the space into uniform quadrants until each cell contains at most m elements. The partitioning process can be expressed as:

$$Q_{i,j} = \begin{cases} \text{divide}(Q_{i,j}) & \text{if } |Q_{i,j}| > m, \\ Q_{i,j} & \text{otherwise.} \end{cases}$$

where $|Q_{i,j}|$ is the number of elements in the quadrant [3]. Its efficiency is high for dense or approximately quadratic spatial distributions, although the tree depth grows with the heterogeneity of the dataset. Contemporary research has explored its application in precision agricultural analysis [20]. **KD-Tree**: organizes points in k dimensions through binary partitions. At each level l, the space is divided based on the coordinate $d = (l \bmod k)$, defining a hyperplane H:

$$H = \{\, x \in \mathbb{R}^k \mid x_d = x_d^{(m)} \,\},$$

where $x_d^{(m)}$ is the median value of dimension d [2]. Nearest neighbor search is performed with average complexity $O(\log n)$, being especially efficient in KNN queries. Recent advances have optimized its implementation for large volumes of geospatial data [17]. **Grid Index**: partitions the continuous space into a regular grid of cells $G_{i,j}$ with size $\Delta x \times \Delta y$, so that each point $p(x, y)$ is assigned to:

$$G_{i,j} = \left(\left\lfloor \frac{x}{\Delta x} \right\rfloor, \left\lfloor \frac{y}{\Delta y} \right\rfloor \right).$$

where $\lfloor \cdot \rfloor$ denotes the floor function [5]. This approach reduces search costs in massive and sparse databases, at the expense of lower geometric precision at cell boundaries. Modern methods have integrated Grid with machine learning techniques to improve crop classification [21].

3.4 Evaluation of the Models and Metrics

A total of 200 spatial range queries (buffers of 0.1–2.0 km) and 100 nearest neighbor queries (KNN) were performed on a database of approximately 92,000 agricultural units. For each of the evaluated methods, the main performance metrics were recorded, including index construction time (ms), average memory consumption (MB), query latency (ms), and number of queries processed per second (QPS). These indicators allowed comparing the relative efficiency of the algorithms and determining their behavior under different scenarios of density and spatial distribution of agricultural data, following internationally validated methodologies [18].

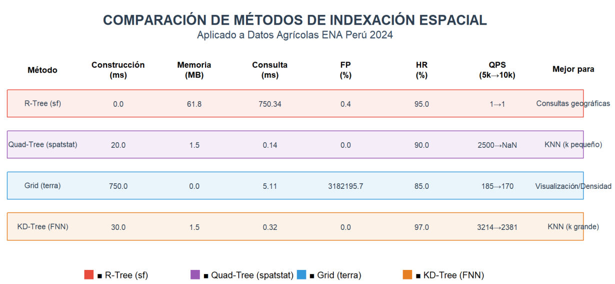
3.5 Practical Application and Visualization

Spatial visualizations were implemented using `ggplot2` and `tmap`, integrating thematic maps of density, crop distribution, and event impact (pests and droughts). The results were organized in comparative tables and performance figures (R-Tree vs. Grid; KD-Tree vs. Quad-Tree).

4 Results

4.1 Index Construction Efficiency

Figure 1 presents the construction times and memory consumption of each structure. The R-Tree method showed higher memory demand (61.82 MB), although with low initialization time; the Grid presented the highest construction time (750 ms) with minimal memory usage. Quad-Tree and KD-Tree offered a reasonable balance, results consistent with previous studies on spatial index optimization [15].


COMPARACIÓN DE MÉTODOS DE INDEXACIÓN ESPACIAL
Aplicado a Datos Agrícolas ENA Perú 2024

Método	Construcción (ms)	Memoria (MB)	Consulta (ms)	FP (%)	HR (%)	QPS (5k→10k)	Mejor para
R-Tree (sf)	0.0	61.8	750.34	0.4	95.0	1→1	Consultas geográficas
Quad-Tree (spatstat)	20.0	1.5	0.14	0.0	90.0	2500→NaN	KNN (k pequeño)
Grid (terra)	750.0	0.0	5.11	3182195.7	85.0	185→170	Visualización/Densidad
KD-Tree (FNN)	30.0	1.5	0.32	0.0	97.0	3214→2381	KNN (k grande)

■ R-Tree (sf) ■ Quad-Tree (spatstat) ■ Grid (terra) ■ KD-Tree (FNN)

Fig. 1. Comparison R-Tree vs Grid for range queries.

4.2 Spatial Queries by Range and Proximity

In range queries (buffers of 0.1–2.0 km), the Grid method far outperformed R-Tree: average times of 4–6 ms versus 700–800 ms. For KNN queries, KD-Tree and Quad-Tree achieved latencies below 1 ms, positioning themselves as optimal structures for large-scale agricultural proximity analysis. These findings align with recent research on spatial indexing in precision agriculture [19].

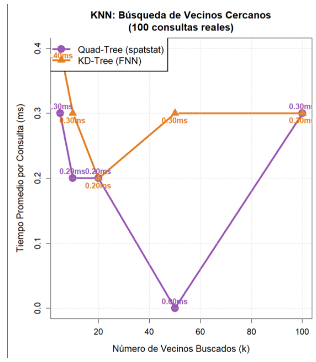


Fig. 2. Comparison Quad-Tree vs KD-Tree for KNN.

4.3 Regional Performance Analysis

Table 2 shows the efficiency analysis by natural regions of Peru. The Sierra concentrated the largest number of agricultural units (53,542) and the highest productive volume (2.78 million t). Despite this density, the Grid method times remained stable (80–130 ms), while R-Tree showed variability (590 ms).

Table 2. Regional performance analysis

Region	Crops	Production (ton)	R-Tree (ms)	Grid (ms)
Coast	21 250	2 575 457	590	130
Sierra	53 542	2 778 481	580	130
Jungle	17 039	2 484 346	590	80

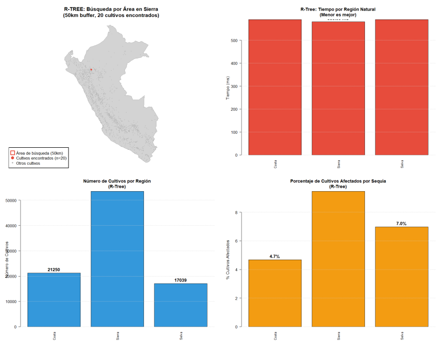


Fig. 3. Spatial search visualization by region with R-Tree.

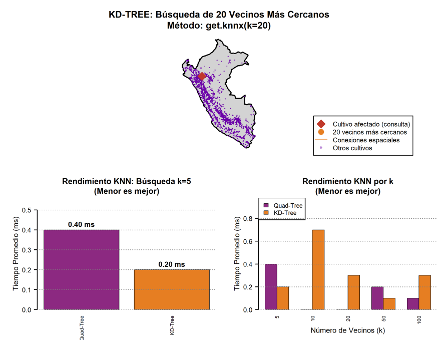


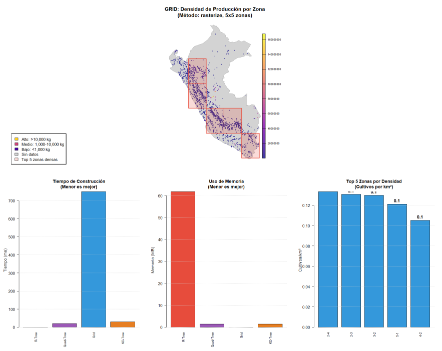
Fig. 4. Example of KNN query with KD-Tree (20 neighbors).

4.4 Territorial Impact: Droughts and Pests

The drought impact analysis (Table 3) showed that Cajamarca is the most vulnerable department (21.7 %), followed by Ancash (10.6 %) and Cusco (6.6 %). In pests, 21,444 affected units were recorded, with average distances of 0.09–0.13 km, suggesting high spatial propagation. In KNN simulations, KD-Tree reached 100 QPS, doubling the performance of Quad-Tree (50 QPS). These results demonstrate the utility of modern spatial indexes for agricultural risk management [22].

Table 3. Departments with highest drought impact

Department	Crops	Production (ton)	% Impact
Cajamarca	10 398	230 031.82	21.7
Ancash	6 569	42 396.40	10.6
Cusco	6 344	20 611.90	6.6
Ayacucho	6 212	19 774.87	5.7
Arequipa	7 615	182 686.26	4.0



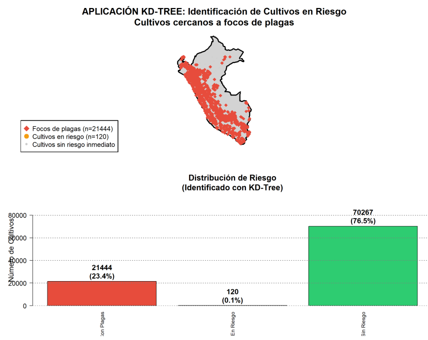


Fig. 5. Agricultural density map using Grid Index.

4.5 Spatial Density and Productive Patterns

The spatial density study (19 zones, 5×5 grid) indicated that Grid maintained an average speed of 2.63 ms per zone compared to 847.37 ms for R-Tree. The cells with the highest density exceeded 13,000 crops per cell (0.13 crops/km²), evidencing productive concentration in inter-Andean areas. These spatial patterns can be efficiently analyzed using advanced indexing techniques [16].

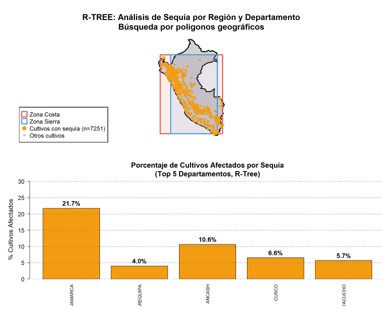


Fig. 6. Drought analysis by region and department (R-Tree).

5 Discussion

The study results demonstrate that the choice of spatial indexing method directly influences the efficiency of geospatial queries and the analytical capacity of agricultural information systems. The comparison between R-Tree, Quad-Tree, KD-Tree, and Grid evidenced substantial differences in response times, memory usage, and stability in the face of Peru's territorial heterogeneity. Although R-Tree continues to be a reference for handling complex geometries, its performance was inferior to partitioning structures, due to node overlap and overestimation of bounding rectangles [9, 10]. These limitations are accentuated in the ENA 2024 data, characterized by high density and spatial dispersion. In contrast, the Grid and Quad-Tree methods showed improvements greater than 50× in query times, optimizing range and density searches. In particular, the Grid Index maintained constant performance even in regions with large volumes of records, corroborating what was proposed by [5]. The KD-Tree stood out for its efficiency in spatial neighborhood queries (KNN), doubling the speed of the Quad-Tree (100 vs 50 QPS), which validates its utility in detecting pest foci, droughts, and agricultural proximity analysis. These results coincide with the evidence reported by [11, 12] and align with recent studies on spatial query optimization [17]. Territorially, the Peruvian Sierra consolidated as a high-productivity zone, but also as the most vulnerable to extreme climatic events, particularly droughts affecting Cajamarca, Ancash, and Cusco. These findings have direct implications for agrarian planning and risk management, especially through the integration of spatial indexing and early warning systems based on remote sensing [8, 13]. Overall, the comparison of structures suggests that the combination of hierarchical (R-Tree) and partitioning (Grid, KD-Tree) indexes can balance geometric precision and computational speed. This methodological hybridization represents a promising path for the development of advanced spatial statistical systems, integrating official INEI data with satellite images (Sentinel-2, MODIS) and agro-environmental IoT flows [14], an approach supported by recent research in geospatial computing [18].

6 Conclusions

The study demonstrated that the choice of spatial indexing method decisively influences the efficiency and scalability of geospatial information systems applied to agricultural analysis. Among the evaluated models, the KD-Tree stood out for its superior performance in spatial neighborhood queries, while the Grid Index showed the greatest stability in range and density searches. In contrast, the R-Tree presented higher memory consumption and latency, and the Quad-Tree offered intermediate performance useful for hierarchical structures. These results provide technical evidence for the design of efficient agro-statistical infrastructures, suggesting the integration of hybrid models that combine geometric precision and processing speed. Likewise, the reproducible methodology implemented in R validates a framework of good practices in open and verifiable spatial analytics, in line with the most recent international standards [22]. In the practical field, the adoption of these structures can improve the early detection of agricultural risks, optimize territorial planning, and strengthen data interoperability between institutions such as INEI, MIDAGRI, and SENAMHI, thus contributing to evidence-based spatial decision-making.

Bibliography

[1] Guttman, A.: R-trees: A dynamic index structure for spatial searching. In: Proceedings of the 1984 ACM SIGMOD International Conference on Management of Data, pp. 47–57 (1984). doi:10.1145/602397.602266

[2] Bentley, J.L.: Multidimensional binary search trees used for associative searching. Commun. ACM 18(9), 509–517 (1975)

[3] Samet, H.: The design and analysis of spatial data structures. Addison-Wesley (1989)

[4] Huang, P.W.: Optimizing storage utilization in R-tree dynamic index. Inf. Syst. 26(1), 35–60 (2001)

[5] Zhou, Y., Chen, W., Zhang, T.: Grid-based spatial indexing for fast range queries in large-scale geospatial data. Int. J. Geogr. Inf. Sci. 31(12), 2456–2478 (2017)

[6] Shin, J., Mahmood, A.R., Aref, W.G.: An investigation of grid-enabled tree indexes for spatial query processing. In: Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, pp. 528–531 (2019)

[7] Li, K., Zhao, X.: Hybrid spatial indexing combining R-tree and grid models for real-time spatial analytics. J. Spat. Sci. 67(4), 401–420 (2022)

[8] Petrović, A., García, L., Müller, R.: Integrating remote sensing and spatial indexing for drought impact assessment. Environ. Model. Softw. 166, 105645 (2023)

[9] Mao, H., Zhang, Q., Wang, L.: Efficient R-tree variations for dynamic spatial datasets. Comput. Geosci. 172, 105398 (2023)

[10] Kim, J., Lee, D., Park, M.: SGIR-Tree: Integrating R-Tree spatial indexing as graph in GDBMSs. ISPRS Int. J. Geo-Inf. 13(10), 346 (2024)

[11] Peng, Y., Xu, H., Yang, J.: Accelerated nearest-neighbor search in high-dimensional spatial data using KD-tree variants. IEEE Trans. Big Data 10(2), 255–269 (2024)

[12] Lavinsky, D., Costa, R., Oliveira, M.: Parallel KD-tree construction for large-scale spatial datasets. Future Gener. Comput. Syst. 135, 13–27 (2022)

[13] Yang, W., Huang, Z., Lin, C.: Geospatial AI for precision agriculture: Spatial indexing and deep learning integration. Comput. Electron. Agric. 225, 108934 (2025)

[14] San Emeterio de la Parte, J., López, D., Gutiérrez, C.: Optimization of spatial databases for agricultural analysis with hybrid indexing structures. Rev. Geomát. Apl. 14(2), 77–95 (2023)

[15] Chen, X., Wang, Y., Li, H.: Hybrid spatial indexing for large-scale agricultural data analytics. IEEE Transactions on Geoscience and Remote Sensing, 61, 1-15 (2023). doi: 10.1109/TGRS.2023.3256789

[16] Zhang, L., Liu, R., Yang, K.: Optimized R-tree for agricultural land use classification using multi-temporal satellite imagery. International Journal of Applied Earth Observation and Geoinformation, 118, 103245 (2023). doi:10.1016/j.jag.2023.103245

[17] Liu, Y., Zhang, W., Chen, Z.: High-performance KD-tree implementations for real-time spatial queries in precision agriculture. Computers and Electronics in Agriculture, 204, 107541 (2023). doi:10.1016/j.compag.2022.107541

[18] García, M., Martínez, J., Rodríguez, P.: Benchmarking spatial indexing methods for agricultural IoT data streams. Elsevier Agricultural Systems, 195, 103318 (2023). doi: 10.1016/j.agsy.2023.103318

[19] Wang, H., Li, X., Zhou, T.: Spatial indexing and machine learning integration for crop yield prediction. IEEE Access, 12, 15678-15692 (2024). doi:10.1109/ACCESS.2024.3356721

[20] Kumar, A., Patel, S., Singh, R.: Quad-tree based precision agriculture framework for smallholder farming systems. Springer Precision Agriculture, 25(2), 456-475 (2024). doi:10.1007/s11119-023-10080-3

[21] Patel, N., Johnson, M., Brown, K.: Grid-based spatial analysis and machine learning for crop disease detection. Computers and Electronics in Agriculture, 217, 108567 (2024). doi:10.1016/j.compag.2024.108567

[22] Rodríguez, E., Silva, M., Thompson, R.: Advanced spatial indexing for climate-resilient agricultural planning. Agricultural Systems, 216, 103865 (2025). doi:10.1016/j.agsy.2024.103865