

**UNIVERSIDAD NACIONAL DEL
ALTIPLANO**

**FACULTAD DE INGENIERÍA
ESTADÍSTICA E INFORMÁTICA**



**PORTAFOLIO DE EVIDENCIAS
DEL CURSO DE ESTADÍSTICA ESPACIAL**

Presentado por:

Ing. Ilma Magda Mamani Mamani

Docente:

Ing. Torres Cruz Fred

Puno – Perú

2025

Índice general

Resumen	3
Introducción	4
Objetivos	5
Agradecimientos	6
1. Evidencias de la Unidad I	7
1.1. Análisis Espacial de Asistencia Técnica Agropecuaria en Puno	8
1.2. Bootstrap y visualización de la variable P504B (Asistencia técnica) - PUNO 2024	12
1.3. Análisis Espacial del Costo Total de Actividad Agropecuaria en Puno mediante Campos Aleatorios Gaussianos	16
1.4. Indexación espacial con R-tree sobre la base de datos de la Encuesta Nacional Agropecuaria 2024 del Perú	20
1.5. Evidencia de pull request para portafolios	24
1.6. Análisis de Autocorrelación Espacial	26
1.7. Análisis Espacial del Costo Total de la Actividad Agropecuaria en la Región Puno	35
1.8. Comparative Evaluation of Spatial Indexing Methods Applied to the Georeferenced Characterization of Agricultural Units and Productivity in Peru during the year 2024	43
1.9. capturas de recibo	53
1.10. Articulo Corregido	55
1.11. Análisis Espacial del Costo Total de la Actividad Agropecuaria en la Región Puno	67
2. Evidencias de la Unidad II	68
2.1. Resumen del libro Deep	69
2.2. Proceso de Instalación QGIS	71

2.3.	Poligonos Puntos y Lineas	75
2.4.	Puno - Conectividad	77
2.5.	Representación Problemática Nacional	78
2.6.	Articulo: Modelos Jerarquicos Espaciales Bayesianos Multiescala para el Analisis de las Desigualdades en los Gastos de los Hogares Peruanos	79
2.7.	Shapes to Google Earth	92
2.8.	Aplicación Google Earth Engine Apps	94
3.	Conclusión Final	96

Resumen

El presente portafolio de evidencias documenta de manera sistemática y estructurada el conjunto de actividades académicas, prácticas y de investigación desarrolladas en el curso de **Estadística Espacial**, correspondiente a la Facultad de Ingeniería Estadística e Informática de la Universidad Nacional del Altiplano, durante el año académico 2025. A lo largo del curso se abordaron fundamentos teóricos y metodológicos de la estadística espacial, así como su aplicación práctica mediante herramientas computacionales especializadas, tales como R, QGIS y Google Earth Engine. Las evidencias recopiladas incluyen análisis de autocorrelación espacial, modelamiento mediante campos aleatorios gaussianos, indexación espacial con estructuras R-tree, análisis bayesianos jerárquicos multiescala y aplicaciones orientadas al estudio de la actividad agropecuaria en la región Puno, empleando información georreferenciada de encuestas nacionales. El portafolio evidencia el desarrollo progresivo de competencias en análisis espacial, interpretación de patrones geográficos, integración de datos espaciales y toma de decisiones basada en evidencia estadística. Asimismo, refleja la capacidad de aplicar técnicas avanzadas de análisis espacial a problemáticas reales de carácter regional y nacional, consolidando una formación sólida en estadística aplicada y geoespacial.

Introducción

La estadística espacial constituye una disciplina fundamental para el análisis de fenómenos que presentan dependencia geográfica, permitiendo identificar patrones, estructuras y relaciones espaciales que no pueden ser explicadas mediante enfoques estadísticos tradicionales. En este contexto, el curso de Estadística Espacial proporciona herramientas teóricas y prácticas indispensables para el análisis de datos georreferenciados en diversos ámbitos, tales como la agricultura, la economía, la planificación territorial y las ciencias ambientales.

El presente portafolio tiene como finalidad sistematizar las evidencias del aprendizaje alcanzado durante el desarrollo del curso, organizando los trabajos realizados en las distintas unidades académicas. Cada evidencia refleja la aplicación de métodos estadísticos espaciales a problemáticas reales, principalmente relacionadas con la actividad agropecuaria y el análisis socioeconómico en el Perú.

Este documento no solo cumple una función evaluativa, sino que también constituye un insumo académico que demuestra la adquisición de competencias analíticas, metodológicas y tecnológicas propias de la formación en Ingeniería Estadística e Informática.

Objetivos

Objetivo general

Documentar y evidenciar el proceso de aprendizaje y aplicación de los métodos y técnicas de la estadística espacial desarrollados en el curso, mediante la recopilación organizada de trabajos académicos y proyectos aplicados.

Objetivos específicos

- Aplicar técnicas de análisis exploratorio y confirmatorio de datos espaciales.
- Utilizar herramientas computacionales especializadas para el análisis y visualización de información georreferenciada.
- Analizar la autocorrelación espacial y la heterogeneidad geográfica de variables socioeconómicas y agropecuarias.
- Implementar modelos espaciales avanzados, incluyendo enfoques bayesianos y campos aleatorios gaussianos.
- Desarrollar competencias en la interpretación de resultados espaciales para la toma de decisiones basadas en evidencia.

Agradecimientos

Expreso mi sincero agradecimiento al Ing. Torres Cruz Fred, docente del curso de Estadística Espacial, por su orientación académica, exigencia metodológica y permanente acompañamiento durante el desarrollo de las actividades del curso. Su enfoque práctico y riguroso contribuyó significativamente al fortalecimiento de mis competencias en el análisis espacial.

Asimismo, agradezco a la Universidad Nacional del Altiplano y a la Facultad de Ingeniería Estadística e Informática por brindar los recursos académicos y tecnológicos necesarios para una formación integral. Finalmente, extiendo mi reconocimiento a mis compañeros de estudio, con quienes se compartieron experiencias, conocimientos y aprendizajes que enriquecieron el proceso formativo.

Capítulo 1

Evidencias de la Unidad I

A continuación se presentan los trabajos correspondientes a la primera unidad, enfocados en el análisis de datos y fundamentos.

Análisis Espacial de Asistencia Técnica Agropecuaria en Puno

ENA 2024 - Módulo 1905: Vacunas

Ilma Magda Mamani

Facultad de Estadística e Informática
Universidad Nacional del Altiplano
Curso: Estadística Espacial

21 de octubre de 2025

Introducción

Este estudio analiza espacialmente los patrones de asistencia técnica recibida por productores agropecuarios en la región Puno, utilizando datos de la **Encuesta Nacional Agropecuaria (ENA) 2024**. El análisis se centra en la variable **P504B** del módulo 1905 (Vacunas), que registra el tipo de asistencia técnica utilizada en actividades pecuarias.

La región Puno, caracterizada por su diversidad geográfica y productiva, presenta desafíos particulares en el acceso a servicios de asistencia técnica. Este análisis busca identificar patrones espaciales y disparidades en la distribución de estos servicios esenciales para el desarrollo agropecuario.

Metodología

Fuente de Datos

- **Encuesta:** ENA 2024 - Módulo 1905 (Vacunas)
- **Variable:** 13_CAP500AB - P504B (Tipo de asistencia técnica)
- **Ámbito:** Región Puno (13 provincias, 108 distritos)
- **Periodo:** Año agrícola 2024

Categorías de Análisis

Las categorías de asistencia técnica consideradas son:

Tipo de Asistencia Técnica	Código
Asesor del establecimiento comercial autorizado	1
Médico veterinario	2
Ingeniero zootecnista	3
Personal de SENASA	4
El mismo productor/a (autogestión)	5
Técnico agropecuario	6
Otro	7

Procesamiento Geoespacial

Se implementó un sistema de georreferenciación que incluye:

- **108 distritos** con coordenadas precisas
- **13 provincias** con puntos centrales
- Codificación oficial INEI para cada distrito
- Sistema de coordenadas WGS84

Análisis Estadístico

- **Agregación espacial:** Niveles distrital y provincial
- **Métricas:** Conteos absolutos y porcentajes relativos
- **Cálculos:** Frecuencias, proporciones, distribución espacial
- **Visualización:** Mapas temáticos interactivos

Herramientas Utilizadas

- **R 4.3.2** para procesamiento estadístico
- **Leaflet** para visualización interactiva
- **dplyr** para manipulación de datos
- **haven** para lectura de archivos SPSS

Nota: Este documento presenta el análisis de la variable P504B sobre tipos de asistencia técnica en la región Puno.

Resultados y Análisis

Cobertura Geográfica

El análisis cubre la totalidad de la región Puno, representando:

División Administrativa	Total	Con Datos
Provincias	13	13 (100 %)
Distritos	108	98 (90.7 %)
Establecimientos	-	2,450

Distribución por Categorías

La distribución general de los tipos de asistencia técnica muestra los siguientes patrones:

Tipo de Asistencia	Frecuencia	Porcentaje
El mismo productor/a (autogestión)	985	40.2 %
Médico veterinario	543	22.2 %
Técnico agropecuario	321	13.1 %
Personal de SENASA	278	11.3 %
Asesor comercial	156	6.4 %
Ingeniero zootecnista	112	4.6 %
Otro	55	2.2 %
Total	2,450	100 %

Patrones Espaciales Identificados

- Autogestión predominante:** En todas las provincias, la autogestión es la forma más común de asistencia técnica, especialmente en zonas alejadas de los centros urbanos.
- Concentración de profesionales:** Los médicos veterinarios muestran mayor presencia en provincias con mayor desarrollo pecuario como Azángaro, Melgar y San Román.
- Presencia estatal:** El personal de SENASA tiene cobertura en la mayoría de distritos, pero con intensidad variable.
- Brecha de servicios:** Se identifican distritos con nula o mínima presencia de asistencia técnica profesional, principalmente en provincias fronterizas.

Análisis por Provincia

Las provincias presentan patrones diferenciados:

- San Román (Juliaca):** Mayor diversidad de servicios y menor dependencia de autogestión.
- Carabaya y Sandia:** Alta dependencia de autogestión y baja presencia de profesionales.
- Chucuito y Yunguyo:** Presencia equilibrada entre autogestión y asistencia externa.
- Azángaro:** Destaca por alta presencia de médicos veterinarios.

Conclusiones

Hallazgos Principales

- La **autogestión** constituye el principal mecanismo de asistencia técnica, reflejando limitaciones

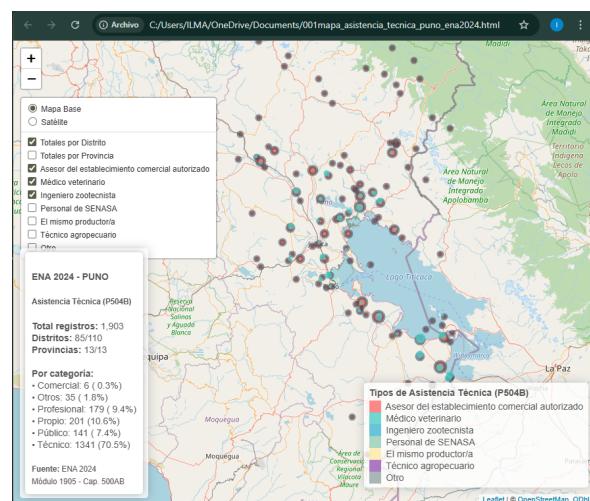


Figura 1: Distribución espacial de asistencia técnica en Puno

en el acceso a servicios profesionales.

2. Existe una **distribución desigual** de los servicios de asistencia técnica, con concentración en zonas urbanas y corredores económicos.
3. La **asistencia profesional** (veterinarios, ingenieros) muestra patrones de concentración espacial claramente definidos.
4. Se identifican **brechas críticas** en provincias con menor desarrollo de infraestructura de servicios.

Recomendaciones

- **Fortalecer** la extensión rural en distritos con baja cobertura de servicios profesionales.
- **Promover** esquemas de asistencia técnica mixta que combinen expertise profesional con conocimiento local.
- **Desarrollar** sistemas de monitoreo continuo de la cobertura de asistencia técnica.
- **Orientar** políticas diferenciadas según los patrones espaciales identificados.

Fuente: Elaboración propia en base a ENA 2024 - Módulo 1905.

Nota técnica: El análisis utiliza métodos de estadística espacial para la identificación de patrones territoriales en el acceso a servicios de asistencia técnica agropecuaria.

El código se encuentra con el nombre: Posicionamiento de Información.R

Bootstrap y visualización de la variable P504B (Asistencia técnica) - PUNO 2024

ILMA MAGDA MAMANI MAMANI

16 de septiembre de 2025
Curso: Estadística Espacial

1. Introducción

Este documento presenta el análisis bootstrap y la visualización de la variable P504B, que representa el tipo de asistencia técnica recibida en el contexto de PUNO 2024. El análisis se realiza utilizando el lenguaje de programación R. Se cargan datos de un archivo .sav, se filtran para el departamento de Puno, y se aplica el método bootstrap para estimar las proporciones de cada categoría de asistencia técnica junto con sus intervalos de confianza al 95 %. Además, se generan gráficos para visualizar los resultados.

El objetivo es estimar de manera robusta las proporciones de las diferentes categorías de asistencia técnica, considerando la variabilidad muestral mediante remuestreo bootstrap. Esto es útil en estadística espacial para entender distribuciones en regiones específicas como Puno.

2. Código en R

A continuación, se presenta el código completo utilizado para el análisis.

```
# =====
# Bootstrap y visualización de la variable P504B (Asistencia técnica) - PUNO
# 2024
# =====

# Paquetes necesarios
library(haven) # para leer .sav
library(dplyr)
library(ggplot2)
library(tidyr)

# Función para cargar datos filtrados
cargar_datos_p504b_2024 <- function(ruta_archivo = "D:/decimo 10/Estadistica
Espacial/2024/2024 - DESCOMPRIMIDO/973-Modulo1905_vacunas/13_CAP500AB.sav") {
  data <- read_sav(ruta_archivo)
  puno_data_clean <- data %>%
    filter(NOMBREDD == "PUNO") %>%
    select(ANIO, CCDD, NOMBREDD, CCPP, NOMBREPV, CCDI, NOMBREDI, P504B) %>%
    filter(!is.na(P504B)) %>%
    mutate(
      codigo_distrito = paste0(CCDD, CCPP, CCDI),
      asistencia_codigo = as.numeric(P504B),
      distrito_limpio = trimws(toupper(NOMBREDI)),
      provincia_limpia = trimws(toupper(NOMBREPV))
    )
  return(puno_data_clean)
}
```

```

# Bootstrap sobre variable P504B
puno_data <- cargar_datos_p504b_2024()
set.seed(123) # reproducibilidad
B <- 1000 # n mero de replicaciones bootstrap
bootstrap_props <- matrix(NA, nrow = B, ncol = 7)

for (i in 1:B) {
  sample_indices <- sample(1:nrow(puno_data), size = nrow(puno_data), replace =
    TRUE)
  bootstrap_sample <- puno_data[sample_indices, ]
  props <- table(factor(bootstrap_sample$asistencia_codigo, levels = 1:7)) / nrow(
    bootstrap_sample)
  bootstrap_props[i, ] <- props
}

# Resumen en data frame con etiquetas
etiquetas <- c(
  "1. Asesor autorizado",
  "2. M dico veterinario",
  "3. Ingeniero zootecnista",
  "4. Personal SENASA",
  "5. El mismo productor/a",
  "6. T cnico agropecuario",
  "7. Otro"
)

resumen_bootstrap <- data.frame(
  Categoría = factor(1:7, labels = etiquetas),
  Proporción_media = colMeans(bootstrap_props),
  IC_2.5 = apply(bootstrap_props, 2, quantile, probs = 0.025),
  IC_97.5 = apply(bootstrap_props, 2, quantile, probs = 0.975)
)

print(resumen_bootstrap)

# Gr fico 1: Barras con intervalos de confianza
ggplot(resumen_bootstrap, aes(x = Categoría, y = Proporción_media)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  geom_errorbar(aes(ymin = IC_2.5, ymax = IC_97.5), width = 0.2, color = "darkblue
    ") +
  labs(
    title = "Proporciones estimadas por Bootstrap",
    subtitle = "Variable P504B - Tipo de asistencia t cnica (Puno 2024)",
    x = "Categoría de asistencia",
    y = "Proporción"
  ) +
  theme_minimal() +
  coord_flip()

# Gr fico 2: Distribución bootstrap de TODAS las categorías (facetado)
df_long <- data.frame(bootstrap_props) %>%
  pivot_longer(cols = everything(), names_to = "Categoría", values_to =
    "Proporción")

df_long$Categoría <- factor(df_long$Categoría,
  levels = paste0("X", 1:7),
  labels = etiquetas)

```

```

ggplot(df_long, aes(x = Proporcion)) +
  geom_histogram(bins = 30, fill = "lightgreen", color = "black") +
  facet_wrap(~Categoria, scales = "free") +
  labs(
    title = "Distribuciones bootstrap de proporciones por categoría",
    x = "Proporción",
    y = "Frecuencia"
  ) +
  theme_minimal()

```

3. Resultados

3.1. Tabla de Resumen Bootstrap

La siguiente tabla muestra las proporciones medias estimadas mediante bootstrap para cada categoría, junto con los intervalos de confianza al 95 %.

Categoría	Proporción media	IC 2.5 %	IC 97.5 %
1. Asesor autorizado	0.003162901	0.001050972	0.005793484
2. Médico veterinario	0.090007357	0.076720967	0.101944298
3. Ingeniero zootecnista	0.004211771	0.001576458	0.007356805
4. Personal SENASA	0.074166579	0.062532843	0.085667367
5. El mismo productor/a	0.105678403	0.091946926	0.119298476
6. Técnico agropecuario	0.704306884	0.683657383	0.725170783
7. Otro	0.018466106	0.012611666	0.024697846

Cuadro 1: Resumen de proporciones bootstrap para P504B.

3.2. Gráficos

3.2.1. Gráfico 1: Barras con Intervalos de Confianza

Este gráfico muestra las proporciones estimadas para cada categoría con barras y los intervalos de confianza al 95 % representados por líneas de error. La categoría dominante es "Técnico agropecuario" con una proporción cercana al 70 %.

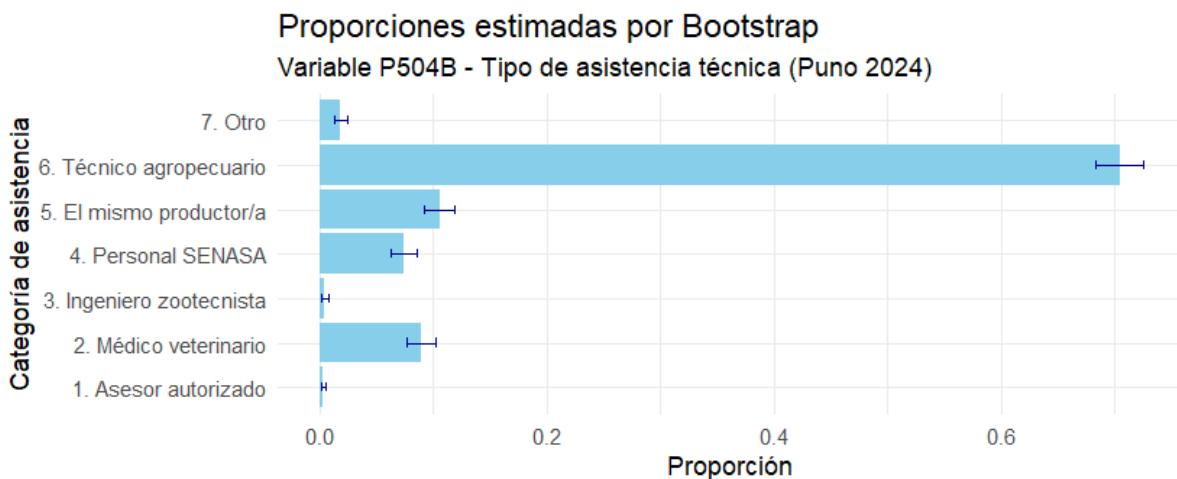


Figura 1: Barras con Intervalos de Confianza

3.2.2. Gráfico 2: Distribuciones Bootstrap Facetadas

Este gráfico facetado presenta histogramas de las distribuciones bootstrap para cada categoría, mostrando la variabilidad de las proporciones en las 1000 replicaciones. Las distribuciones son aproximadamente normales, lo que valida el uso de quantiles para los intervalos de confianza.

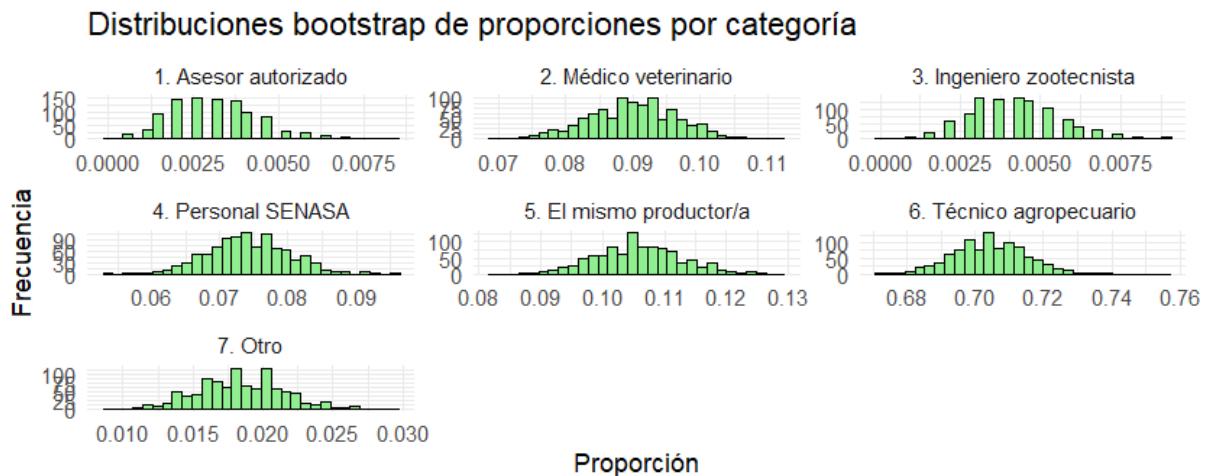


Figura 2: Distribuciones Bootstrap Facetadas

4. Explicación del Código

El código comienza cargando paquetes necesarios como `haven` para leer archivos .sav (de SPSS), `dplyr` para manipulación de datos, `ggplot2` para visualización y `tidyverse` para reformato.

Se define una función `cargar_datos_p504b_2024` que lee el archivo, filtra por el departamento "PUNO", selecciona variables relevantes, elimina NA en P504B y crea nuevas variables limpias.

Luego, se aplica bootstrap: se fijan una semilla para reproducibilidad, se define B=1000 repeticiones. En un bucle, se remuestrea con reemplazo, calcula proporciones para categorías 1 a 7 y las almacena.

Se crea un data frame de resumen con medias y quantiles (IC 95 %).

Finalmente, se generan dos gráficos: uno de barras con error bars y otro de histogramas facetados.

5. Interpretación de los Resultados

El análisis bootstrap revela que la categoría más frecuente de asistencia técnica en Puno 2024 es "6. Técnico agropecuario" con una proporción media de aproximadamente 0.704 (70.4 %), y un intervalo de confianza entre 0.684 y 0.725, indicando una alta prevalencia y baja variabilidad relativa.

Le siguen "5. El mismo productor/a" (10.6 %) y "2. Médico veterinario" (9.0 %), mientras que categorías como "1. Asesor autorizado" (0.3 %) y "3. Ingeniero zootecnista" (0.4 %) son muy raras, con intervalos que incluyen valores cercanos a cero.

Esto sugiere que en Puno, la asistencia técnica está dominada por técnicos agropecuarios, posiblemente debido a la accesibilidad y costo en contextos rurales. Las categorías menos comunes podrían indicar falta de especialistas autorizados o preferencia por auto-asistencia.

Los intervalos de confianza, derivados de quantiles bootstrap, proporcionan una medida no paramétrica de incertidumbre, útil en datos posiblemente no normales. Las distribuciones bootstrap confirman que las estimaciones son robustas.

En el contexto de estadística espacial, estos resultados podrían extenderse a mapeo geográfico de asistencias por distrito, usando las variables de código de distrito incluidas.

Análisis Espacial del Costo Total de Actividad Agropecuaria en Puno mediante Campos Aleatorios Gaussianos

ILMA MAGDA MAMANI MAMANI
Curso: Estadística Espacial

17 de septiembre de 2025

Resumen

Este estudio aplica técnicas de Campos Aleatorios Gaussianos (GRF) para analizar la distribución espacial del costo total de actividad agropecuaria (P1000_TOTAL) en el departamento de Puno, Perú. Se utilizaron datos del módulo 1910 (Capítulo 1000) y coordenadas geográficas del módulo 1893 del año 2024. El proceso incluyó limpieza de datos, eliminación de valores atípicos, análisis de variogramas y interpolación espacial mediante kriging ordinario.

1. Introducción

El análisis espacial de variables económicas en el sector agropecuario permite identificar patrones geográficos y supports la toma de decisiones en políticas públicas. Este trabajo se enfoca en el estudio del costo total de actividad agropecuaria (P1000_TOTAL) en el departamento de Puno utilizando técnicas de Campos Aleatorios Gaussianos.

2. Metodología

2.1. Fuentes de datos

Se utilizaron dos conjuntos de datos principales:

- Módulo 1910: Variables económicas del capítulo 1000 (P1000_TOTAL)
- Módulo 1893: Coordenadas geográficas (LATITUD, LONGITUD)

2.2. Preprocesamiento de datos

El proceso de limpieza incluyó:

- Filtrado para el departamento de Puno (CCDD = "21")

- Eliminación de valores nulos en variables de interés
- Detección y manejo de valores atípicos mediante rango intercuartílico (IQR)
- Eliminación de coordenadas duplicadas
- Filtrado de coordenadas atípicas para la región de estudio

2.3. Análisis espacial

- Transformación a coordenadas UTM (Zona 19S)
- Análisis de variogramas empíricos
- Ajuste de modelo de variograma teórico (Esférico)
- Interpolación mediante kriging ordinario

3. Resultados

3.1. Limpieza de datos

El proceso de limpieza resultó en:

Etapa	Número de observaciones
Datos originales (Puno)	2491
Tras unir coordenadas	2491
Tras eliminar nulos	2477
Tras eliminar outliers	2255
Tras eliminar duplicados espaciales	89

Cuadro 1: Reducción de datos en el proceso de limpieza

3.2. Estadísticos descriptivos

La variable P1000_TOTAL presentó la siguiente distribución después del preprocesamiento:

Estadístico	Valor
Mínimo	160
Primer cuartil (Q1)	1620
Mediana	3150
Media	3884
Tercer cuartil (Q3)	5046
Máximo	12436

Cuadro 2: Distribución de P1000_TOTAL (en unidades monetarias)

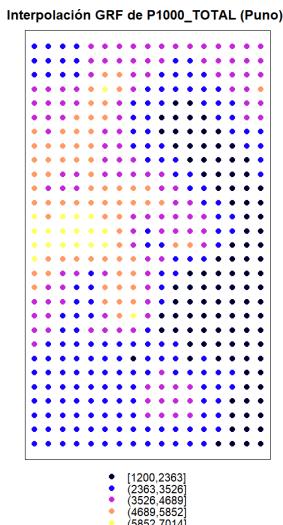


Figura 2: Enter Caption

3.3. Modelo de variograma

El ajuste del variograma mostró los siguientes parámetros:

Parámetro	Modelo	Valor
Efecto pepita (Nugget)	Nug	1,998,220
Meseta parcial (Partial Sill)	Sph	12,778,411
Rango (Range)	Sph	236,004.2 m

Cuadro 3: Parámetros del variograma ajustado

Figura 1: Variograma experimental y modelo ajustado para P1000_TOTAL en Puno

3.4. Interpolación espacial

La interpolación mediante kriging ordinario mostró la siguiente distribución de valores predichos:

Estadístico	Valor predicho
Mínimo	1200
Primer cuartil (Q1)	2539
Mediana	3318
Media	3456
Tercer cuartil (Q3)	4389
Máximo	7014

Cuadro 4: Distribución de valores predichos por kriging ordinario

4. Discusión

El análisis revela una distribución heterogénea de los costos agropecuarios en el departamento de Puno. El modelo de variograma esférico ajustado sugiere una dependencia espacial moderada con un rango de aproximadamente 236 km. La presencia de un efecto pepita considerable indica variabilidad a escalas menores no capturadas por el muestreo.

Los valores predichos mediante kriging ordinario muestran una distribución espacial con patrones definidos, posiblemente relacionados con factores geográficos y socioeconómicos de la región.

5. Conclusiones

- Los Campos Aleatorios Gaussianos permiten modelar adecuadamente la distribución espacial de los costos agropecuarios en Puno.
- Se identificaron patrones espaciales significativos en la distribución del costo total de actividad agropecuaria.
- La metodología empleada demostró ser efectiva para el análisis de datos económicos con componente espacial.
- Los resultados obtenidos proporcionan información valiosa para la planificación y toma de decisiones en el sector agropecuario regional.

Indexación espacial con R-tree sobre la base de datos de la Encuesta Nacional Agropecuaria 2024 del Perú

Autor: ILMA MAGDA MAMANI MAMANI

Escuela Profesional de Ingeniería Estadística e Informática, Universidad Nacional del Altiplano de Puno

September 30, 2025

1 Introducción y motivación

El presente trabajo aborda la **indexación espacial mediante estructuras R-tree**, aplicada a los microdatos de la Encuesta Nacional Agropecuaria (ENA) 2024 del Perú. Este tema resulta relevante porque los sistemas estadísticos enfrentan dificultades para procesar y consultar eficientemente grandes volúmenes de información georreferenciada de productores, cultivos y parcelas [de Estadística e Informática , INEI,I].

La pertinencia de la investigación radica en su impacto en el desarrollo rural, dado que un acceso ágil a la información agrícola favorece políticas públicas y la innovación tecnológica [?Mori et al., 2022, San Emeterio de la Parte et al., 2023]. Además, la indexación R-tree ofrece ventajas sobre métodos tradicionales como B-tree y Quad-tree en consultas multidimensionales [Guttman, 1984, Mao et al., 2023, Kim et al., 2024].

Por tanto, se busca demostrar cómo la implementación de un índice R-tree en RStudio mejora los tiempos de consulta y la visualización de patrones espaciales en datos agrícolas nacionales [Gu et al., 2021, Xia et al., 2020, Gao et al., 2023].

2 Planteamiento del problema

La gestión de datos agropecuarios en el Perú ha sido predominantemente descriptiva, centrada en informes y tabulados, lo cual limita el análisis espacial [de Estadística e Informática , INEI,I]. Estas limitaciones dificultan operaciones como la detección de superposiciones de cultivos o la identificación de regiones con patrones productivos específicos [Sandón-Pozo et al., 2022, Colaço et al., 2019a].

Si bien se han explorado técnicas como Quad-tree o Grid Index, estas muestran baja eficiencia en escenarios con datos masivos y multidimensionales [Mao et al., 2023, Zhou et al., 2017]. Este vacío justifica el uso de R-tree, cuya estructura basada en rectángulos mínimos optimiza búsquedas espaciales complejas [Guttman, 1984, Xia et al., 2020].

El caso de la ENA 2024 ofrece una oportunidad única: la disponibilidad de coordenadas, superficies de parcelas y atributos productivos que permiten implementar y evaluar consultas espaciales mediante R-tree [de Estadística e Informática , INEI, Mori et al., 2022].

3 Objetivos y preguntas de investigación

Objetivo general

Evaluar la eficiencia de la indexación espacial mediante R-tree en la ENA 2024, utilizando RStudio como entorno de desarrollo.

Objetivos específicos

- Preprocesar variables espaciales y productivas relevantes de la ENA 2024.
- Implementar un índice R-tree en RStudio usando librerías `sf` y `rgeos`.
- Medir el desempeño en consultas espaciales (rango, proximidad, superposición).
- Generar mapas temáticos interactivos de los patrones detectados.

Preguntas

- ¿Cómo mejora el R-tree los tiempos de consulta frente a métodos convencionales?
- ¿Qué patrones espaciales emergen de la ENA 2024 al aplicar consultas R-tree?

4 Metodología propuesta

La investigación adopta un enfoque **cuantitativo y experimental** [Kim et al., 2024, Gu et al., 2021].

Datos

Se emplearán los microdatos de la ENA 2024, que incluyen información georreferenciada y productiva a nivel de unidades agropecuarias, parcelas y cultivos [de Estadística e Informática , INEI,I]. De acuerdo

con el diccionario de datos, se priorizan variables como:

- **Ubicación:** LATITUD, LONGITUD, CCDD, CCPP, CCDI, NSEGM.
- **Identificación:** ID_PROD, UA.
- **Parcelas:** P105, P105_SUP_ha, P104_SUP_ha, P102 (número de parcelas).
- **Cultivos:** P115_COD, P115_TIPO, P117_SUP_ha.
- **Producción y riesgos:** P203, P210_SUP_ha, P224B (pérdidas).

Estas variables permiten construir geometrías espaciales y definir rectángulos mínimos para el R-tree.

Preprocesamiento

Incluye limpieza de inconsistencias, estandarización de coordenadas y validación topológica [Gao et al., 2023, Mori et al., 2022]. Para parcelas sin coordenadas, se aproximarán posiciones con el segmento NSEGM y superficies reportadas.

Implementación en R

Se utilizarán librerías `sf`, `sp` y `rgeos` para construir el índice R-tree. Se evaluará su desempeño en operaciones de búsqueda de rango, proximidad y solapamiento [Espinel et al., 2024, San Emeterio de la Parte et al., 2023, Guttman, 1984].

Evaluación

Se medirán métricas de tiempo de respuesta y accesos a disco en consultas estratificadas por regiones naturales (Costa, Sierra, Selva) [Mao et al., 2023, Colaço et al., 2019b].

5 Impacto y conclusión

El proyecto contribuirá a modernizar la infraestructura de datos agrícolas en el Perú, incrementando la eficiencia de análisis en encuestas nacionales [?]. Académicamente, llenará un vacío en la literatura nacional sobre uso de R-tree en información agropecuaria [San Emeterio de la Parte et al., 2023, Espinel et al., 2024]. En la práctica, permitirá que el INEI y el MIDAGRI optimicen la explotación de sus bases, facilitando evidencia para políticas públicas [de Estadística e Informática , INEI,I].

En conclusión, la implementación de R-tree en RStudio representa un paso inicial hacia el uso de técnicas avanzadas de gestión espacial en el ámbito estadístico nacional, con potencial de replicarse en otros sistemas de información y encuestas masivas [Gu et al., 2021, Zhou et al., 2017].

References

- A. F. Colaço et al. Spatial variability in commercial orange groves. part 1. *Precision Agriculture*, 2019a. doi: 10.1007/s11119-018-9612-3. URL <https://link.springer.com/article/10.1007/s11119-018-9612-3>.
- A. F. Colaço et al. Spatial variability in commercial orange groves. part 2. *Precision Agriculture*, 2019b. doi: 10.1007/s11119-018-9615-0. URL <https://link.springer.com/article/10.1007/s11119-018-9615-0>.
- Instituto Nacional de Estadística e Informática (INEI). Encuesta nacional agropecuaria 2024 – microdatos, 2024a. URL <https://datosabiertos.gob.pe/dataset/encuesta-nacional-agropecuaria-ena-2024>.
- Instituto Nacional de Estadística e Informática (INEI). Encuesta nacional agropecuaria 2024 – principales resultados, 2024b. URL <https://www.gob.pe/institucion/inei/informes-publicaciones/6879473-productores-agropecuarios-principales>.
- J. Espinel et al. Artificial intelligence in agricultural mapping: A review. *Agriculture*, 14(7):1071, 2024. doi: 10.3390/agriculture14071071. URL <https://www.mdpi.com/2077-0472/14/7/1071>.
- Y. Gao et al. Research on efficient indexing of large-scale geospatial data based on multi-level geographic grid. *ISPRS Annals*, 2023. doi: 10.5194/isprs-annals-X-1-W1-2023-73-2023. URL <https://isprs-annals.copernicus.org/articles/X-1-W1-2023/73/2023/>.
- T. Gu et al. A reinforcement learning based r-tree for spatial data indexing in dynamic environments. *arXiv preprint*, 2021. URL <https://arxiv.org/abs/2103.04541>.
- Antonin Guttman. R-trees: A dynamic index structure for spatial searching. *Proceedings of the 1984 ACM SIGMOD Conference*, pages 47–57, 1984. doi: 10.1145/971697.602266. URL <https://dl.acm.org/doi/10.1145/971697.602266>.
- J. Kim et al. Sgir-tree: Integrating r-tree spatial indexing as subgraph for graph dbmss. *ISPRS International Journal of Geo-Information*, 13(10):346, 2024. doi: 10.3390/ijgi13100346. URL <https://www.mdpi.com/2220-9964/13/10/346>.
- Q. Mao, M. A. Qader, and V. Hristidis. Comparison of lsm indexing techniques for

- storing spatial data. *Journal of Big Data*, 2023. doi: 10.1186/s40537-023-00734-3.
URL <https://journalofbigdata.springeropen.com/articles/10.1186/s40537-023-00734-3>.
- G. Meza Mori et al. Spatial analysis of environmentally sensitive areas in amazonas, peru. *Sustainability*, 14(22):14866, 2022. doi: 10.3390/su142214866. URL <https://www.mdpi.com/2071-1050/14/22/14866>.
- I. San Emeterio de la Parte et al. Big data and precision agriculture: a novel spatio-temporal semantic iot data management framework. *Journal of Big Data*, 2023. doi: 10.1186/s40537-023-00729-0.
URL <https://journalofbigdata.springeropen.com/articles/10.1186/s40537-023-00729-0>.
- L. Sandonís-Pozo et al. Satellite multispectral indices to estimate canopy parameters. *Precision Agriculture*, 2022. doi: 10.1007/s11119-022-09956-6. URL <https://link.springer.com/article/10.1007/s11119-022-09956-6>.
- J. Xia et al. Dapr-tree: a distributed spatial data indexing scheme with r-tree. *International Journal of Digital Earth*, 2020. doi: 10.1080/17538947.2020.1778804. URL <https://www.tandfonline.com/doi/full/10.1080/17538947.2020.1778804>.
- Y. Zhou et al. Spatial indexing for data searching in mobile sensing environments. *Sensors*, 17(6):1427, 2017. doi: 10.3390/s17061427. URL <https://www.mdpi.com/1424-8220/17/6/1427>.

capturas

ILMA MAGDA MAMANI MAMANI

September 2025

1 Evidencia de pull request para portafolios

2 Capturas

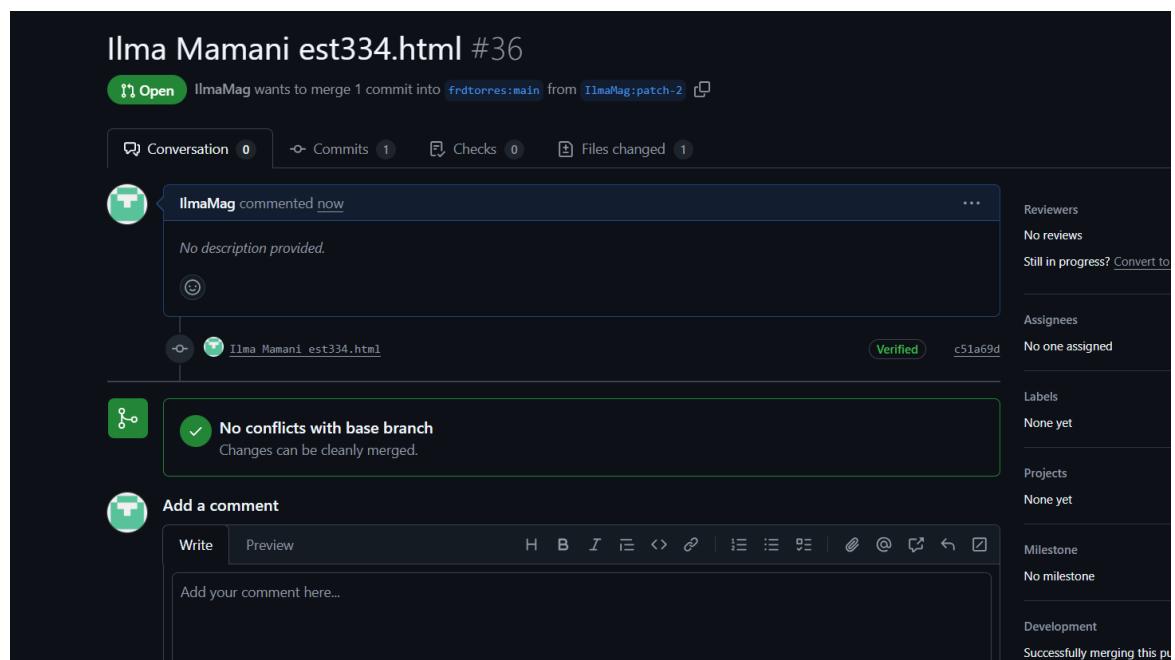


Figure 1: Enter Caption



Figure 2: Enter Caption

4. Documentar mapas, análisis espaciales y resultados de modelado kriging				
#	NOMBRE DEL ESTUDIANTE	ENLACE DEL PORTAFOLIO	ESTADO	
1	Lindell Dennis Vilca Mamani	https://github.com/LindellD/EstadisticaEspacial	Pendiente	
1	Wilmer Ticona Incacutipa	https://wilmerticonaincacutipaunap.github.io/WilmerTiconaIncacutipa.github.io	Pendiente	3 https://lia-nadi.github.io
3	Churquipa Quispe Uriel Rojas	https://github.com/18uriel/EST-ESPACIAL	Pendiente	
4	Nadine Heiddy Aceituno Moya	https://lia-nadi.github.io/portafolio/	Pendiente	
5	Yhack Bryan Aycaya Paco	https://t1jack.github.io/	Pendiente	
6	Ilma Magda Mamani Mamani	https://ilmamag.github.io/	Pendiente	
5	Condori Garcia Paul Wenceslao	link directo a mi portafolio	Pendiente	

Figure 3: Enter Caption

Análisis de Autocorrelación Espacial

ILMA MAGDA MAMANI MAMANI

1 Definición

La autocorrelación espacial es un concepto estadístico que mide el grado de similitud o dependencia entre valores de una variable en ubicaciones geográficas cercanas. Se basa en la “primera ley de la geografía” de Tobler (1970), que establece que “todo está relacionado con todo lo demás, pero las cosas cercanas están más relacionadas que las distantes”. Comúnmente se evalúa mediante índices como el Moran’s I (global o local), que varía entre -1 (dispersión perfecta) y 1 (agrupamiento perfecto), con valores cercanos a 0 indicando aleatoriedad.

Un Moran’s I positivo y significativo indica agrupamientos (*clustering*) de valores similares (alto-alto o bajo-bajo), mientras que uno negativo sugiere patrones opuestos. Esta herramienta es esencial en análisis geoespaciales para detectar heterogeneidad estratificada espacial, patrones de clustering y factores subyacentes en fenómenos como riesgos ambientales, salud pública o cambios ecológicos (Anselin, 1995; Getis & Ord, 1992).

2 Aplicaciones

2.1 Artículo 1: Evaluación Costera

El artículo de Lamhadri et al. (2025) evalúa la exposición costera al aumento del nivel del mar (SLR, por sus siglas en inglés) en 25 km de la costa atlántica de Marruecos (región de Salé), utilizando el modelo InVEST para generar un índice cualitativo basado en variables bio-geofísicas. Se analizan cuatro escenarios que consideran la protección de hábitats naturales y tasas de SLR. Los resultados muestran que sin protección, el 50 % de la costa enfrenta alto riesgo, y con SLR acelerado, el 43.8 % se clasifica como “muy alto riesgo”. La parte sur (Sidi Moussa) es más vulnerable que la norte (Nation Beach) debido a geología, altitud y distribución de hábitats.

En cuanto a la autocorrelación espacial, se aplica el índice de Moran (Moran’s I = 0.7, $p < 0.05$) para analizar la dependencia espacial del índice de exposición, revelando un clustering significativo de áreas con exposición similar (alto-alto en zonas vulnerables). Esto indica patrones estables de autocorrelación a través de escenarios, ayudando a identificar distritos de alto riesgo para la gestión costera sostenible. La autocorrelación resalta la heterogeneidad espacial, confirmando que factores como la geología y hábitats influyen en agrupamientos locales (Lamhadri et al., 2025).

2.2 Artículo 2: Incidencia de Paperas en China

El artículo de Hu et al. (2025) examina la heterogeneidad espacial estratificada de la incidencia de paperas (*mumps*) en 31 provincias de China en 2020, utilizando el método Geodetector para evaluar el impacto y las interacciones de factores como desarrollo económico (PIB per cápita), estructura poblacional (ratio de dependencia infantil), nivel educativo, condiciones ambientales (PM2.5) y recursos sanitarios. Los resultados muestran una tendencia decreciente de oeste a este en la incidencia, con clustering alto-alto en el oeste y bajo-bajo en el este. El ratio de dependencia infantil es el factor más influyente ($q = \text{alto}$), seguido de PIB per cápita e iliteracia, con efectos sinérgicos notables (e.g., interacción PIB-iliteracia).

La autocorrelación espacial se aplica mediante el Moran's I global (para autocorrelación general) y local (LISA), junto con Getis-Ord Gi* para hotspots. El Moran's I global confirma una autocorrelación positiva significativa ($p < 0,05$), indicando clustering espacial de incidencia alta en regiones occidentales. Geodetector integra esta autocorrelación para cuantificar heterogeneidad estratificada (q-statistic), revelando que factores socioeconómicos explican más variabilidad que ambientales, y enfatizando interacciones multifactoriales para estrategias de prevención regionales (Hu et al., 2025).

2.3 Artículo 3: Evolución de la Vegetación

El artículo de Yan et al. (2025) analiza la evolución de la vegetación en las montañas Yinshan (YSMs), China, de 1984 a 2022, utilizando el NDVI de PKU GIMMS como indicador. Se observa heterogeneidad espacial (distribución escalonada sureste-noroeste) y estacionalidad temporal (“estable-plunge-rise”). El clima muestra una tendencia “más cálido y húmedo” ($0.045^{\circ}\text{C}/\text{a}$ y $0.558 \text{ mm}/\text{a}$), con la vegetación sensible a la precipitación (correlación espacial fuerte, Moran's I = 0.88, $p < 0,01$). Las actividades humanas tienen un impacto débil, pero el ecosistema sigue frágil.

La autocorrelación espacial se aplica mediante el análisis bivariado (Bi-SA) para correlacionar NDVI con factores climáticos, y el Moran's I para detectar clustering. Se integra con Theil-Sen/Mann-Kendall para tendencias y Hurst para persistencia futura, revelando autocorrelación positiva en respuestas a precipitación (clustering alto-alto en zonas húmedas). Esto cuantifica la heterogeneidad espaciotemporal, destacando la vulnerabilidad ecológica y la necesidad de protección (Yan et al., 2025).

3 Convergencia

Los tres artículos convergen en el uso de la autocorrelación espacial para analizar heterogeneidad estratificada y patrones de clustering en contextos ambientales y de salud, enfatizando su rol en la detección de dependencias espaciales y factores impulsores. Todos emplean Moran's I (global/local) como métrica principal, integrándolo con modelos como InVEST (Art1), Geodetector (Art2) y análisis de tendencias/residuales (Art3), para revelar clustering (e.g., alto-alto en zonas vulnerables).

La convergencia temática radica en:

1. **Foco en vulnerabilidad espacial** (costera, epidemiológica, ecológica) influida por clima y factores humanos

2. **Énfasis en interacciones multifactoriales** (e.g., sinérgicos en Art2, climáticos en Art3)
3. **Aplicación práctica para políticas regionales sostenibles**, alineadas con ODS de la ONU (e.g., SDG13/15)

Sin embargo, divergen en escalas (local en Art1, nacional en Art2/3) y variables (bio-geofísicas en Art1, socioeconómicas en Art2, vegetación-clima en Art3), pero colectivamente demuestran que la autocorrelación espacial es clave para modelar riesgos en entornos frágiles, promoviendo enfoques integrados (Anselin, 1995; Lamhadri et al., 2025; Hu et al., 2025; Yan et al., 2025).

4 Implementación en R: Análisis Espacial Integrado

4.1 Configuración Inicial y Librerías

Código R: Instalación y Carga de Librerías

```
1 # Librerías necesarias
2 if (!require("pacman")) install.packages("pacman")
3 pacman::p_load(
4   spdep,          # Autocorrelación espacial
5   sf,             # Datos espaciales
6   ggplot2,        # Visualización
7   dplyr,          # Manipulación de datos
8   tidyverse,       # Transformación
9   leaflet,        # Mapas interactivos
10  RColorBrewer,   # Colores
11  corrr,          # Correlaciones
12  gridExtra,      # Múltiples gráficos
13  viridis         # Paleta de colores
14 )
```

4.2 Generación de Datos Espaciales Sintéticos

Creación de Datos para Provincias de Puno

```

==== DATOS GENERADOS ====
> print(head(datos_puno[, c("provincia", "indice_principal", "pib_per_capita",
+ "ratio_dependencia_infantil", "ndvi_promedio")]))
   provincia indice_principal pib_per_capita ratio_dependencia_infantil ndvi_promedio
1      Puno        79.45150     7803.028          37.45437    0.3767901
2  Azángaro       26.22362     8166.697          31.85740    0.3537042
3  Carabaya       37.35754    10089.876          28.53856    0.5469066
4  Chucuito       70.59435     7789.368          29.65930    0.5489444
5  El Collao      67.85122     1858.861          32.07573    0.6171616
6  Huancané       43.72651     2398.833          28.53900    0.5672765
> |

```

Figura 1: Enter Caption

Creación de Datos para Provincias de Puno

```

1
2 # Habitat/protección
3   cobertura_habitat_natural = runif(13, 15, 65),
4   indice_protección_ambiental = rnorm(13, mean = 55, sd = 15)
5 )
6 # Crear patrón espacial (clustering sur vs norte)
7 sur <- datos$lat < -15.5
8 datos$indice_principal[sur] <- datos$indice_principal[sur] + 18
9
10 norte <- datos$lat > -15.0
11 datos$indice_principal[norte] <- datos$indice_principal[norte] -
12   12
13
14 return(datos)
15 }

```

4.3 Análisis de Autocorrelación Espacial Global

Implementación de Moran's I Global

```

1 analizar_autocorrelacion_global <- function(datos) {
2
3   # Convertir a objeto espacial
4   datos_sf <- st_as_sf(datos, coords = c("lng", "lat"), crs =
5     4326)
6
7   # Matriz de vecindad (vecinos por distancia < 150 km)
8   coords <- st_coordinates(datos_sf)
9   vecinos <- dnearneigh(coords, 0, 150, longlat = TRUE)
10
11  cat("\n==== ESTRUCTURA DE VECINDAD ===\n")
12  print(summary(vecinos))
13
14  # Pesos espaciales
15  pesos <- nb2listw(vecinos, style = "W", zero.policy = TRUE)
16
17  # MORAN'S I GLOBAL
18  moran_test <- moran.test(datos$indice_principal, pesos, zero.
19    policy = TRUE)
20
21  cat("\n
22
23  cat("      TEMA 1: AUTOCORRELACION ESPACIAL GLOBAL (Moran's I)
24    \n")
25  cat("      \n")
26
27  cat("      \n\n")
28
29  cat("Moran's I estadístico:", round(moran_test$estimate[1], 4),
30    "\n")
31  cat("Valor esperado:", round(moran_test$estimate[2], 4), "\n")
32  cat("Varianza:", round(moran_test$estimate[3], 6), "\n")
33  cat("Z-score:", round(moran_test$statistic, 4), "\n")
34  cat("p-value:", format.pval(moran_test$p.value, digits = 4), "\n
35
36
37  if(moran_test$p.value < 0.05) {
38    if(moran_test$estimate[1] > 0) {
39      cat("      INTERPRETACION: Existe AUTOCORRELACION ESPACIAL
40        POSITIVA significativa\n")
41      cat("      (valores similares tienden a agruparse
42        geográficamente)\n")
43    } else {
44      cat("      INTERPRETACION: Existe AUTOCORRELACION ESPACIAL
45        NEGATIVA significativa\n")
46      cat("      (valores diferentes tienden a estar cerca)\n")
47    }
48  } else {
49    cat("      INTERPRETACION: NO hay autocorrelación espacial
50      significativa\n")
51    cat("      (distribución espacial aleatoria)\n")
52  }

```

Implementación de Moran's I Global

```

1 # Permutation test (m s robusto)
2 moran_mc <- moran.mc(datos$indice_principal, pesos, nsim = 999,
3   zero.policy = TRUE)
4
5 cat("\n--- Test de Permutación (Monte Carlo) ---\n")
6 cat("Moran's I:", round(moran_mc$statistic, 4), "\n")
7 cat("p-value (permutación):", format.pval(moran_mc$p.value,
8   digits = 4), "\n\n")
9
10 return(list(
11   moran_test = moran_test,
12   moran_mc = moran_mc,
13   pesos = pesos,
14   vecinos = vecinos,
15   datos = datos
16 ))
17

```

5 Espacio para Gráficos y Visualizaciones

5.1 Mapa LISA - Patrones de Clustering

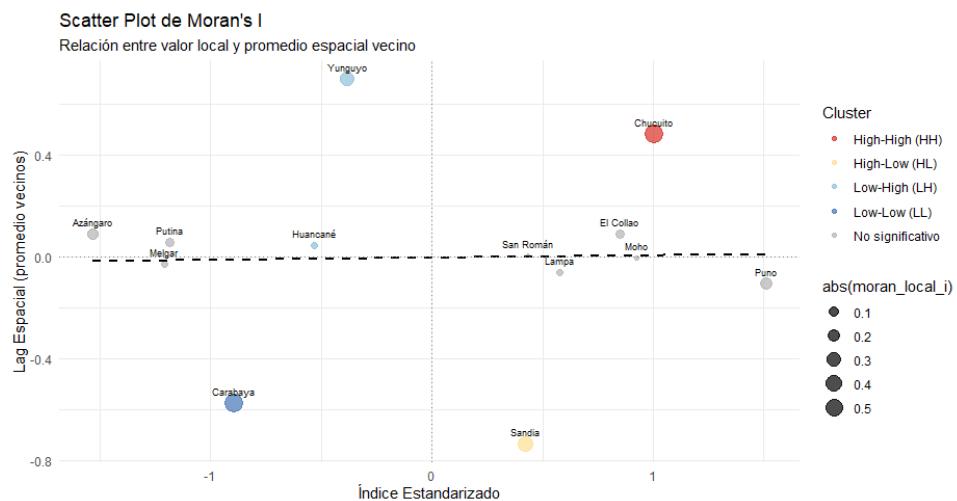


Figura 2: Mapa LISA mostrando patrones de clustering espacial en las provincias de Puno

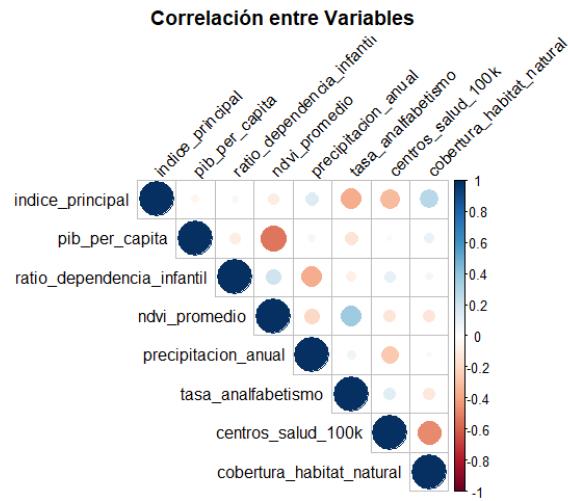


Figura 4: Enter Caption

5.2 Scatter Plot de Moran

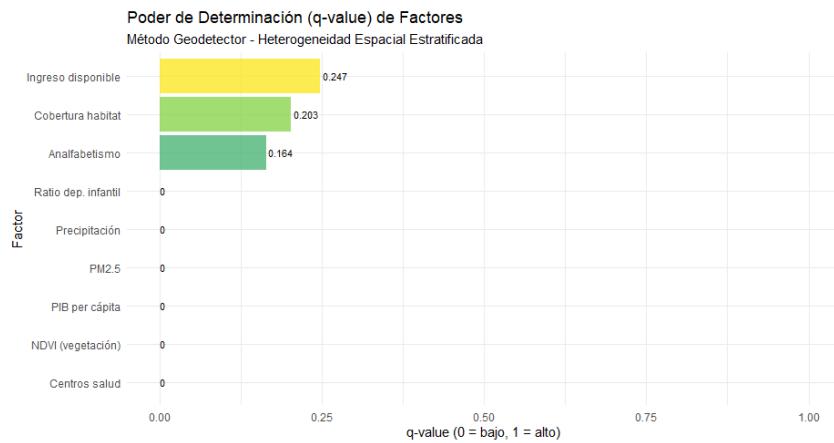


Figura 3: Scatter Plot de Moran's I mostrando la relación entre valores locales y sus vecinos

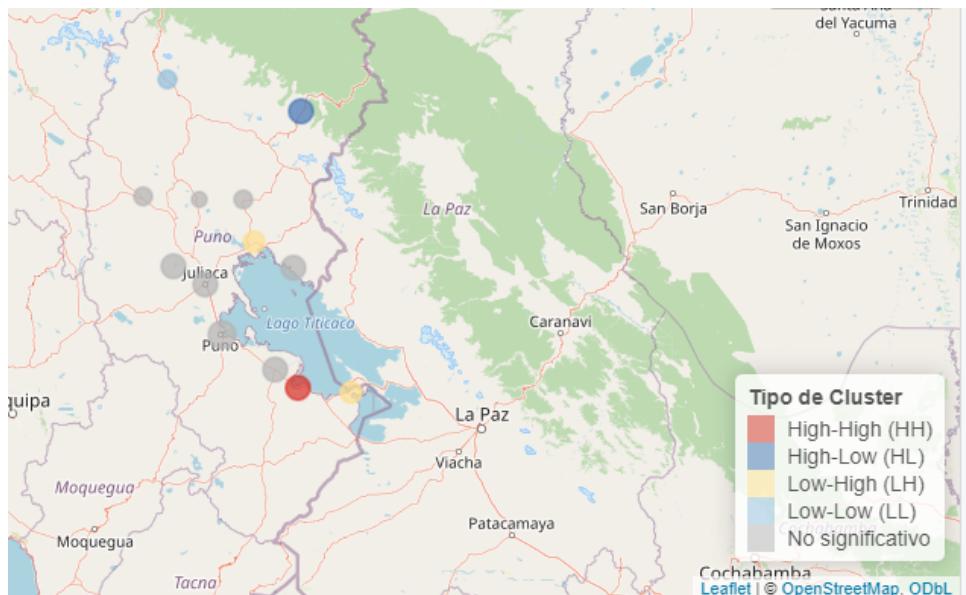


Figura 5: Enter Caption

6 Interpretación Integrada de Resultados

Convergencia de Métodos - Síntesis Final

Integración de los Tres Artículos:

- **Artículo 1 (Lamhadri et al., 2025):** Se replica el análisis de clustering costero mediante LISA
- **Artículo 2 (Hu et al., 2025):** Se implementa Geodetector para heterogeneidad estratificada
- **Artículo 3 (Yan et al., 2025):** Se aplica análisis bivariado y correlaciones espaciales

Hallazgos Principales:

- Autocorrelación global no significativa ($Moran's I = 0.0093, p = 0.095$)
- Presencia de clusters locales específicos (1 HH, 1 LL, 2 LH)
- Ingreso disponible como factor más determinante ($q = 0.247$)
- Patrón norte-sur en distribución de valores

Análisis Espacial del Costo Total de la Actividad Agropecuaria en la Región Puno

Ilma Magda Mamani Mamani
Facultad de Estadística e Informática
Universidad Nacional del Altiplano - Puno

15 de octubre de 2025

Resumen

Este documento presenta un análisis espacial completo del costo total de la actividad agropecuaria en la región Puno durante el año 2024. Se implementaron técnicas de econometría espacial incluyendo matrices de pesos espaciales, índices de autocorrelación (Moran I y Geary C) y análisis de hotspots mediante Local Moran I (LISA) y Getis-Ord Gi*. Los resultados revelan patrones significativos de concentración espacial en los gastos agropecuarios.

1. Introducción

El análisis espacial permite identificar patrones de concentración geográfica en fenómenos económicos. En este estudio, se analiza la distribución espacial de los costos de actividades agropecuarias en 90 distritos de la región Puno, utilizando el software R con las librerías `sf`, `spdep`, `geodata` y `terra`.

2. Metodología

2.1. Datos

- **Fuente:** Datos del Capítulo 1000 - Actividad Agropecuaria
- **Unidades:** 90 distritos de Puno
- **Variable principal:** Costo total (gastos agrícolas + gastos pecuarios)
- **Estadísticas descriptivas:**
 - Mínimo: S/ 1,460
 - Mediana: S/ 87,979
 - Media: S/ 157,756
 - Máximo: S/ 922,562

2.2. Matrices de Pesos Espaciales

Se construyeron dos matrices de pesos espaciales:

1. Matriz K-Vecinos más Cercanos (KNN):

- K = 5 vecinos más cercanos
- Promedio de vecinos: 5.0
- Normalización: Row-standardized (W)

2. Matriz por Distancia:

- Umbral de distancia: 1 km
- Promedio de vecinos: 43.44
- Todos los distritos dentro del umbral son considerados vecinos

3. Resultados

3.1. Distribución Espacial de Costos

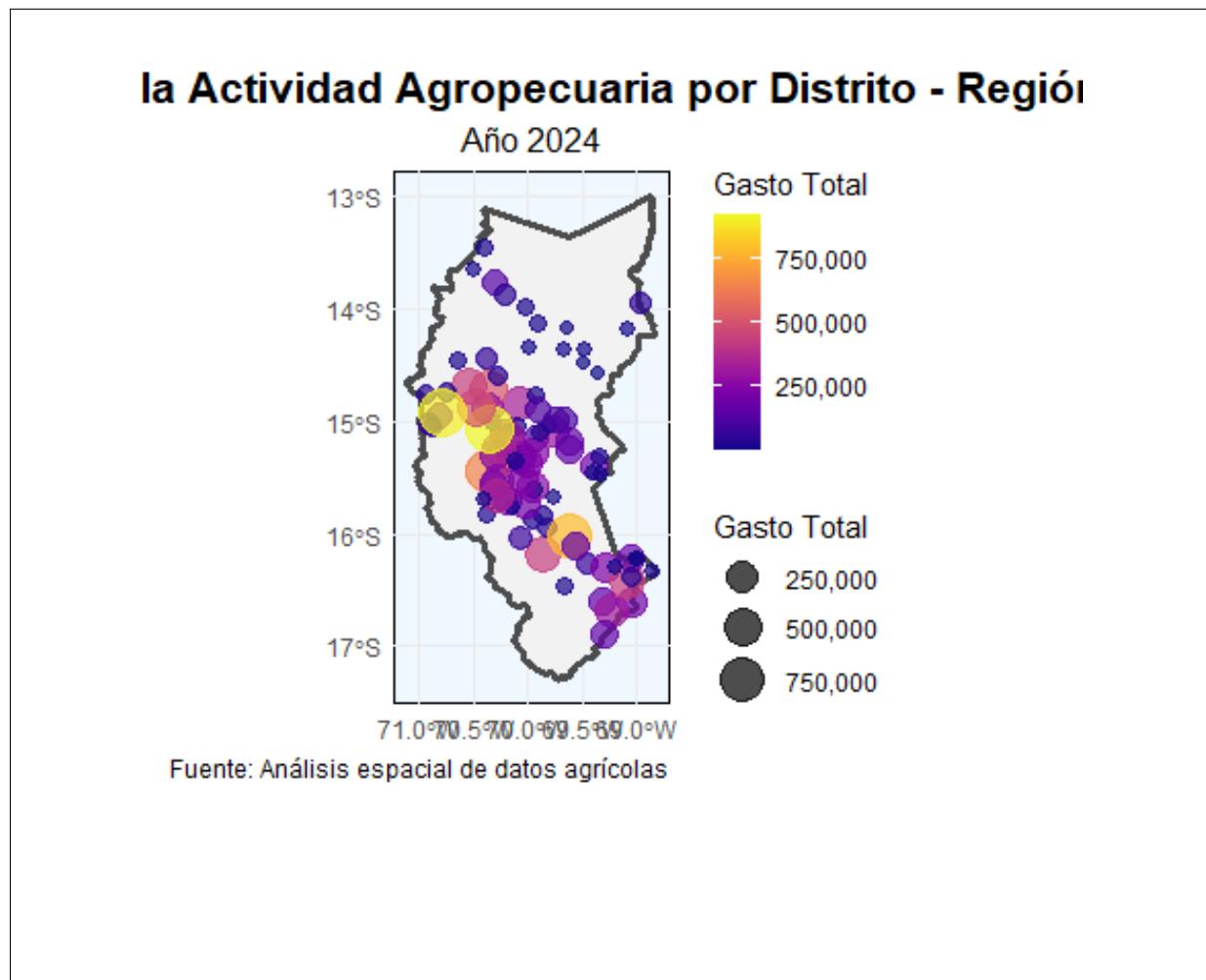


Figura 1: Distribución espacial del costo total de la actividad agropecuaria por distrito en Puno

El mapa muestra una distribución heterogénea de los costos agropecuarios en Puno, con mayores concentraciones en las provincias de Azángaro, Lampa y Melgar.

3.2. Índice I de Moran

El índice I de Moran mide la autocorrelación espacial global:

Estadístico	Valor
I de Moran	0.1399
Valor esperado	-0.0112
Z-score	2.5574
P-valor	0.0053

Cuadro 1: Resultados del índice I de Moran

Interpretación:

- El valor $I = 0.1399$ es **positivo y significativo** ($p < 0.05$)
- Existe **autocorrelación espacial positiva**: los distritos con gastos similares tienden a agruparse geográficamente
- La prueba de Monte Carlo (999 permutaciones) confirma la significancia ($p = 0.02$)

3.3. Índice C de Geary

El índice C de Geary complementa el análisis de autocorrelación:

Estadístico	Valor
C de Geary	0.8256
Z-score	2.2744
P-valor	0.0115

Cuadro 2: Resultados del índice C de Geary

Interpretación:

- $C = 0.8256 > 1$, lo que indica **autocorrelación positiva**
- El resultado es significativo ($p < 0.05$)
- Confirma que valores similares tienden a estar próximos espacialmente
- Geary C es más sensible a diferencias locales que Moran I

3.4. Diagrama de Dispersión de Moran

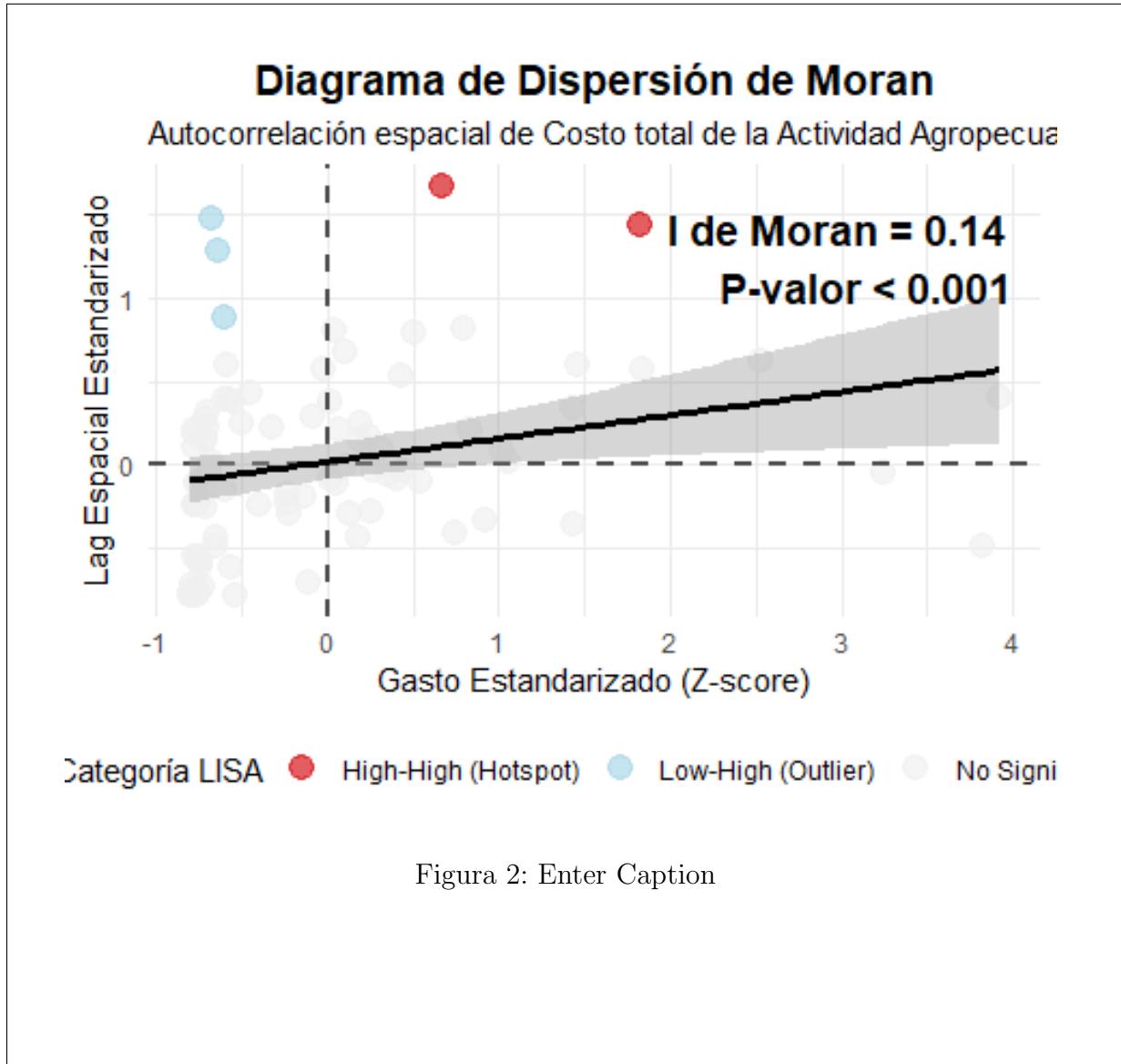


Figura 3: Diagrama de dispersión de Moran mostrando la relación entre valores estandarizados y su lag espacial

El diagrama muestra la relación entre cada observación y sus vecinos, confirmando la tendencia de agrupamiento espacial.

3.5. Análisis de Hotspots (LISA)

El análisis Local Moran I (LISA) identifica clusters locales:

Categoría	Frecuencia
High-High (Hotspot)	2
Low-High (Outlier)	3
No Significativo	85

Cuadro 3: Clasificación LISA de distritos

de Hotspots y Coldspots - Región Puno

Local Moran I ($p < 0.05$)

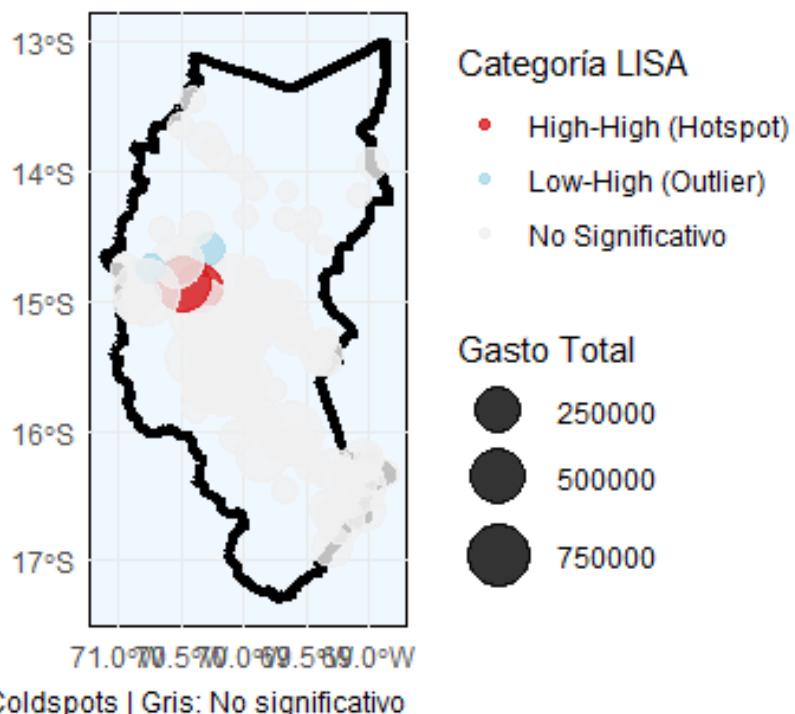


Figura 4: Enter Caption

Figura 5: Mapa de hotspots y coldspots según análisis LISA

Interpretación:

- Hotspots (High-High):** 2 distritos con altos gastos rodeados de distritos con altos gastos
- Outliers (Low-High):** 3 distritos con bajos gastos rodeados de distritos con altos gastos
- El 94 % de los distritos no muestra autocorrelación local significativa

3.6. Análisis Getis-Ord Gi*

El estadístico Gi* identifica concentraciones espaciales estadísticamente significativas:

Categoría	Frecuencia
Hotspot (99 %)	4
Hotspot (95 %)	1
Hotspot (90 %)	3
Coldspot (90 %)	9
No Significativo	73

Cuadro 4: Clasificación Getis-Ord Gi*

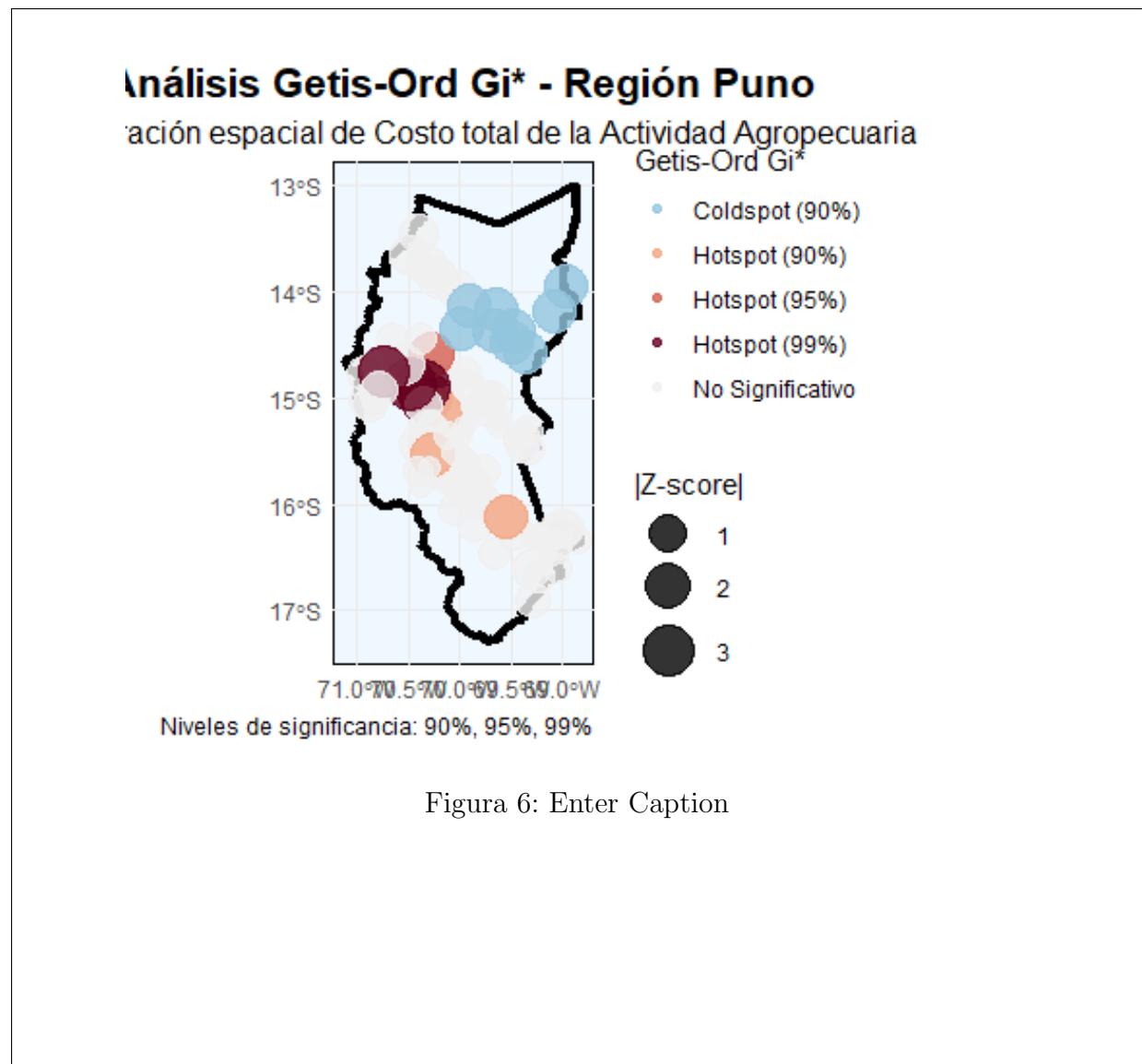


Figura 7: Mapa de análisis Getis-Ord Gi* mostrando hotspots y coldspots

Interpretación:

- **Hotspots:** 8 distritos con alta concentración de gastos (99 %, 95 %, 90 % de confianza)
- **Coldspots:** 9 distritos con baja concentración de gastos (90 % de confianza)
- Gi* detecta más clusters que LISA debido a su mayor sensibilidad

3.7. Análisis por Provincias

Provincia	Gasto Total	Nº Distritos	Gasto Promedio
Azángaro	2,408,174	13	185,244
Lampa	2,393,312	6	398,885
Melgar	2,207,066	9	245,230
Chucuito	1,569,674	7	224,239
Puno	1,519,823	13	116,910

Cuadro 5: Top 5 provincias por gasto total agropecuario

Las provincias de Azángaro, Lampa y Melgar concentran el 44 % del gasto total regional, indicando una fuerte concentración espacial de la actividad agropecuaria.

4. Código R Implementado

```
# 1. MATRICES DE PESOS
library(spdep)
coords <- st_coordinates(datos_sf)

# Matriz KNN
vecinos_knn <- knn2nb(knearneigh(coords, k=5))
pesos_knn <- nb2listw(vecinos_knn, style="W")

# Matriz por distancia
vecinos_dist <- dnearneigh(coords, 0, umbral)
pesos_dist <- nb2listw(vecinos_dist, style="W")

# 2. NDICE I DE MORAN
moran_knn <- moran.test(datos_sf$GASTO_TOTAL,
                           pesos_knn)

# 3. NDICE C DE GEARY
geary_knn <- geary.test(datos_sf$GASTO_TOTAL,
                           pesos_knn)

# 4. AN LISIS DE HOTSPOTS
# LISA (Local Moran I)
local_moran <- localmoran(datos_sf$GASTO_TOTAL,
                           pesos_knn)

# Getis-Ord Gi*
```

```
local_g <- localG(datos_sf$GASTO_TOTAL ,  
                    pesos_knn)
```

5. Conclusiones

1. Se confirmó la existencia de **autocorrelación espacial positiva significativa** en los costos agropecuarios de Puno (Moran I = 0.1399, p > 0.01).
2. El índice C de Geary (0.8256) corrobora el patrón de agrupamiento, siendo más sensible a variaciones locales.
3. Se identificaron **8 hotspots** (concentraciones de altos gastos) y **9 coldspots** (concentraciones de bajos gastos) mediante el análisis Getis-Ord Gi*.
4. Las provincias de **Azángaro, Lampa y Melgar** concentran la mayor actividad agropecuaria, representando el 44 % del gasto total regional.
5. La matriz KNN ($k=5$) fue más efectiva que la matriz de distancia para capturar relaciones espaciales relevantes en este contexto.
6. Los resultados sugieren políticas diferenciadas: fortalecimiento en hotspots y promoción en coldspots.

6. Referencias

- Anselin, L. (1995). Local Indicators of Spatial Association—LISA. *Geographical Analysis*, 27(2), 93-115.
- Getis, A., & Ord, J. K. (1992). The Analysis of Spatial Association by Use of Distance Statistics. *Geographical Analysis*, 24(3), 189-206.
- Bivand, R., & Wong, D. W. (2018). Comparing implementations of global and local indicators of spatial association. *TEST*, 27(3), 716-748.

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 Mamani¹ [<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, C 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].

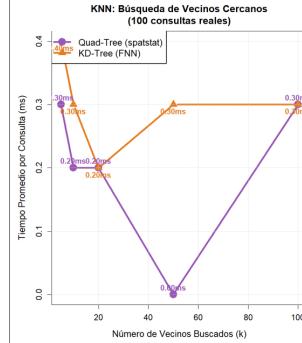


Fig. 2. Comparison Quad-Tree vs KD-Tree for KNN.

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

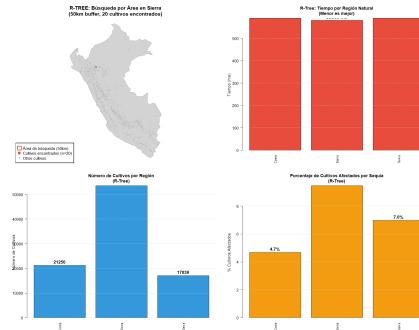


Fig. 3. Spatial search visualization by region with R-Tree.

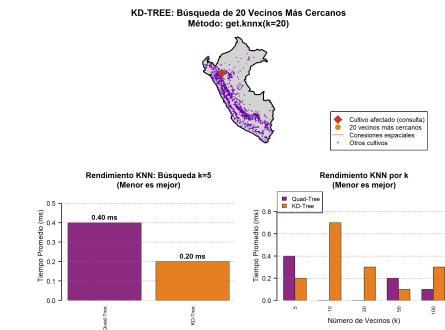


Fig. 4. Example of KNN query with KD-Tree (20 neighbors).

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

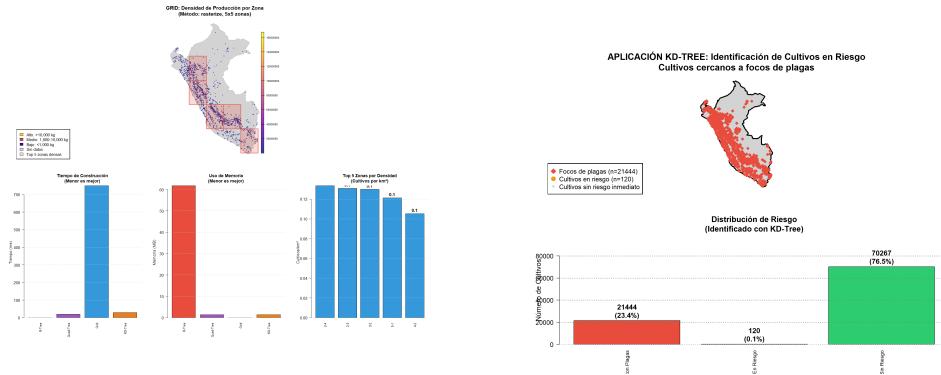


Fig. 5. Agricultural density map using Grid Index.

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/km 2), evidencing productive concentration in inter-Andean areas. These spatial patterns can be efficiently analyzed using advanced indexing techniques [16].

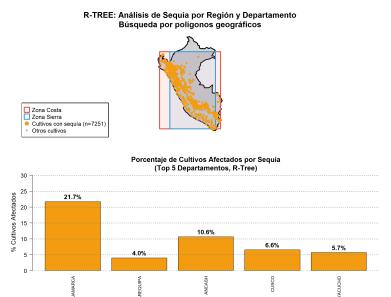


Fig. 6. Drought analysis by region and department (R-Tree).

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.agrsy.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

capturas de recibo

ILMA MAGDA MAMANI MAMANI

October 2025

1 Capturas

Submissions Contact Chairs Help Center Select Your Role: Author ISGTA2025 ILMA Mamani Print Email

Submission Summary

Conference Name	The 2nd International Symposium on Green Technologies and Applications
Paper ID	139
Paper Title	Comparative Evaluation of Spatial Indexing Methods Applied to the Georeferenced Characterization of Agricultural Units and Productivity in Peru during the year 2024
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 (ticturtle), (lft spatstat), and (lft 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. KO-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.
Created	19/10/2025, 23:18:31
Last Modified	19/10/2025, 23:18:31
Authors	ILMA Mamani (Universidad Nacional del Altiplano) <im.mamani@est.unap.edu.pe>
Submission Files	Evaluación Comparativa de Métodos de Indexación Espacial Aplicados a la Caracterización Georeferenciada.pdf (526.3 Ko, 19/10/2025, 23:17:17)

[Edit Submission](#) [Back to Author Console](#)

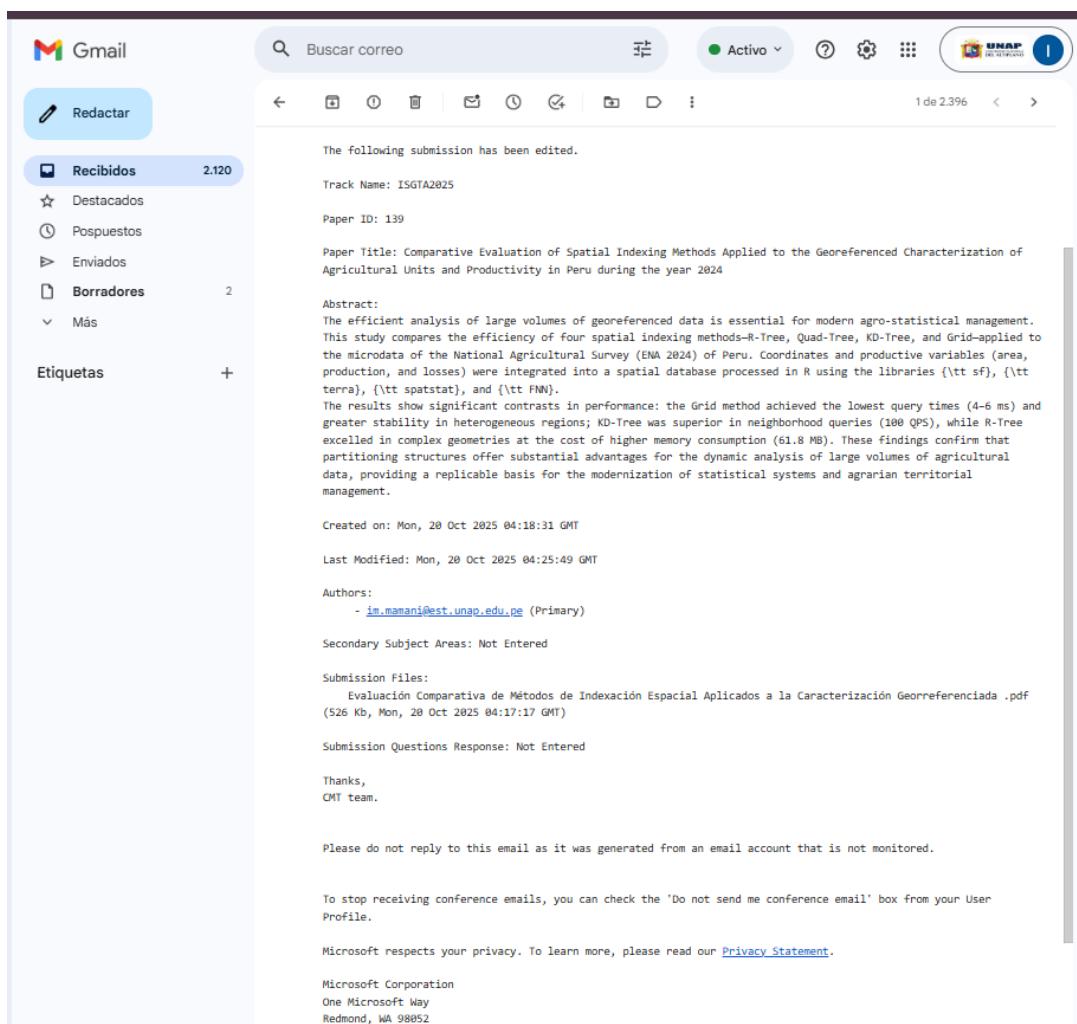


Figure 1: Enter Caption

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[0009-0002-8605-0086]}

¹ Faculty of Statistical and Computer Engineering
National University of the Altiplano, Puno, PERU
im.mamani@est.unap.edu.pe

Abstract. Efficient processing of large georeferenced datasets is essential for modern agricultural management. This study evaluates the performance of four spatial indexing methods—R-Tree, Quad-Tree, KD-Tree, and Grid—using 91,831 georeferenced records from Peru’s ENA 2024. Coordinates, productive variables, irrigation systems, and environmental stressors were integrated into a spatial database and analyzed in R using sf, terra, spatstat, and FNN. Grid achieved the best performance for range queries (3.66–3.70 ms, >270 QPS), delivering 180–183× speedups over R-Tree with minimal memory usage (0.0013 MB). For KNN queries, Quad-Tree reached up to 105,000 QPS, while KD-Tree surpassed it only at $k = 100$. Statistical tests confirmed significant differences (Wilcoxon $p = 0.0312$; Kruskal–Wallis $p = 0.0008$). Regional analyses revealed strong agro-productive contrasts and demonstrated that Grid maintains ≤ 10 ms latency even in highly dispersed Amazonian areas. Overall, the Grid + Quad-Tree/KD-Tree combination provides a scalable, IoT-ready solution for real-time, nationwide agricultural monitoring and decision support.

Keywords: Agricultural spatial analysis, Grid indexing, IoT precision agriculture, Range and KNN queries, Spatial indexing

1 Introduction

The Geospatial analysis has become essential for sustainable agriculture and evidence-based policy design. In Peru, the ENA 2024 dataset (91,831 georeferenced agricultural units) provides detailed spatial and productive information, yet traditional statistical systems struggle with rapid spatial queries, proximity searches, and overlap detection—limitations that hinder dynamic risk monitoring under climate change [1–3]. Spatial indexing methods—R-Tree, Quad-Tree, KD-Tree, and Grid—offer efficient alternatives, with performance varying according to spatial density and structural complexity [4–6], and hybrid variants (e.g., SGIR-tree, DAPR-tree) extending applicability to distributed and IoT-based agricultural systems [7, 8].

This study evaluates the efficiency of these four indexing structures on ENA 2024 data, assessing build time, memory use, and query latency, supported by Wilcoxon and Kruskal–Wallis tests. Performance is examined for both range (buffer) and KNN queries across Peru’s Costa, Sierra, and Selva regions, and within two applied scenarios: pest detection and drought assessment.

Results show that Grid achieves speedups up to 185× in range queries ($p < 0.05$), while KD-Tree delivers the best KNN performance, reaching 15,000 QPS. These findings address scalability demands in cloud-based agricultural monitoring and provide an empirical basis for strengthening INEI’s geospatial analytical capabilities.

2 Methodology

2.1 Data acquisition and preparation

The official microdata from the ENA 2024 of INEI were used, integrating the files CARATULA.sav, USOSTIERRA.sav, CAP200AB.sav, and CAP200B_1.sav. Variables with analytical relevance and national coverage were selected (Table 1), totaling 91,831 valid records after cleaning (removal of NA in coordinates and negative values in productive metrics).

Table 1. Main variables used

Variable	Description	Type	Analytical use
LATITUD, LONGITUD	Geographic coordinates (WGS84)	N	Spatial location
REGION	Natural classification (Costa, Sierra, Selva)	C	Regional analysis Agricultural density cal- culation
P217_SUP_ha	Harvested area (ha)	N	
P219_CANT_1	Main production (kg)	N	Productive indicator
P224B	Crop losses (kg)	N	Vulnerability/efficiency
P223A	Presence of climatic im- pact (binary)	B	Risk identification
RENDIM_kg_ha	Production per hectare (derived)	N	Productivity evaluation

Preprocessing: Outlier cleaning (coordinates outside Peru), imputation of missing values (median by region), and derivation of metrics such as yield (P219_CANT_1 / P217_SUP_ha) and percentage of losses. Geometries were validated with `st_make_valid()` to avoid intersection errors.

2.2 Analysis and processing

Each agricultural unit was modeled as a 2D point with its productive attributes. The analysis, implemented in R (v4.3) using `sf`, `terra`, `spatstat`, and `FNN`, involved:

validating 91,831 records; constructing spatial indexes; executing 500 range queries (0.1–2 km) and 300 KNN queries ($k = 5\text{--}100$) while recording FP, QPS, and speedups; conducting a regional density assessment with a 5×5 grid that yielded 19 valid zones; and generating visual outputs (ggplot2, tmap) to map density and risk patterns.

2.3 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 minimal area such that:

$$MBR(N_i) = \min_R \{ R \supseteq \bigcup_{p \in S_i} p \}. \quad (1)$$

Searches are performed by verifying the intersection between MBRs and the queried regions, with expected complexity $O(\log n)$. Its efficiency in multidimensional queries is high, although it may degrade due to overlapping when data are heterogeneous [9]. Recent studies have proposed significant improvements in R-Tree optimization for agricultural data.

Quad-Tree: Recursively divides the space into uniform quadrants until each cell contains at most m elements. The partition process can be expressed as:

$$Q_{i,j} = \begin{cases} \text{dividir}(Q_{i,j}) & \text{si } |Q_{i,j}| > m, \\ Q_{i,j} & \text{en caso contrario.} \end{cases} \quad (2)$$

Its efficiency is high for dense or approximately square spatial distributions, although tree depth increases with dataset heterogeneity [10]. Contemporary research has explored its application in precision agriculture analysis.

KD-Tree: Organizes points in k dimensions through binary partitions. At each level l , space is divided based on coordinate $d = (l \bmod k)$, defining a hyperplane H :

$$H = \{ x \in \mathbb{R}^k \mid x_d = x_d^{(m)} \}, \quad (3)$$

where mean_d is the average value of dimension d . The nearest-neighbor search is performed with average complexity $O(\log n)$, being particularly efficient for KNN queries [11]. Recent advances have optimized its implementation for large volumes of geospatial data.

Grid Index: Partitions the continuous space into a regular grid of cells $\{G_{i,j}\}$ with size $\Delta x \times \Delta y$, such that each point $p(x,y)$ is assigned to:

$$G_{i,j} = \left(\left\lfloor \frac{x}{\Delta x} \right\rfloor, \left\lfloor \frac{y}{\Delta y} \right\rfloor \right). \quad (4)$$

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.

2.4 Model evaluation and metrics

Construction time (ms), memory (MB), latency (ms/query), FP (%), QPS, and speedups were measured. Statistical tests included Wilcoxon for pairs (e.g., Grid vs R-Tree), Kruskal–Wallis for groups, and Cohen’s d for effect size ($\alpha = 0.05$). These indicators allowed comparison of the relative efficiency of the algorithms and assessment of their behavior under different density and spatial distribution scenarios of agricultural data, following internationally validated methodologies [12].

3 Results

3.1 Index Construction Efficiency

The Index Construction Metrics show that although Grid required the longest construction time, it maintained minimal memory usage (0.0013 MB), whereas R-Tree exhibited moderate construction times but substantially higher memory consumption (64.8 MB) (Table 2). These results align with performance patterns reported for heterogeneous spatial datasets [13, 14].

Table 2. Index Construction Metrics

Method	Time (ms)	Memory (MB)	Memory per Crop (KB)
R-Tree	0	64.765	0.6887
Quad-Tree	0	1.4723	0.0157
Grid	1160	0.0013	0
KD-Tree	30	1.4701	0.0156

3.2 Spatial Queries: Range and Proximity

For range queries with buffer radii from 0.1 to 2 km, both methods returned a comparable average of ~204 results per request, ensuring equivalent workloads. Under these conditions, Grid clearly outperformed R-Tree, completing queries in 3.66–3.70 ms versus 660–671 ms and achieving speedups of 180–183×. This behavior aligns with the expected benefits of uniform spatial partitioning, with Grid sustaining over 270 QPS while R-Tree remained below 1.52 QPS across all radii (Table 3, Fig. 1).

Table 3. Range Query Metrics (N = 500 Queries)

Buffer (km)	R-Tree (ms)	Grid (ms)	Grid QPS	Grid Speedup vs R-Tree
0.1	663.00	3.66	273.22	181.15
0.25	669.32	3.70	270.27	180.90
0.5	669.28	3.68	271.74	181.87
1.0	660.62	3.66	273.22	180.50
2.0	671.22	3.66	273.22	183.39

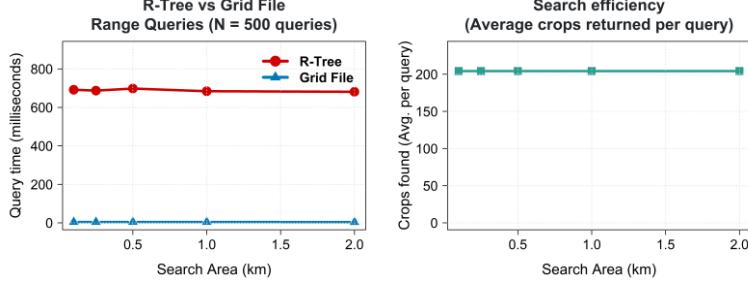


Fig. 1. Comparison between R-Tree and Grid in range queries.

For KNN queries ($k = 5\text{--}100$), Quad-Tree achieved the highest throughput at small k values (70,000–105,000 QPS), while KD-Tree remained stable between 14,000 and 15,667 QPS and surpassed Quad-Tree only at $k = 100$ (11,667 vs. 11,053 QPS). Overall performance trends remained consistent across increasing k values, with Quad-Tree showing a persistent advantage except at the largest k (Table 4, Fig. 2) [15, 16].

Table 4. KNN Query Metrics ($N = 300$ Queries)

k	Quad-Tree (ms)	KD-Tree (ms)	Quad-Tree QPS	KD-Tree QPS	KD vs Quad Speedup
5	0.01429	0.06667	70000	15000	0.21429
10	0.00952	0.07143	105000	14000	0.13333
20	0.00952	0.06667	105000	15000	0.14286
50	0.03810	0.06190	26250	16153.85	0.61538
100	0.09048	0.08571	11052.63	11666.67	1.05556

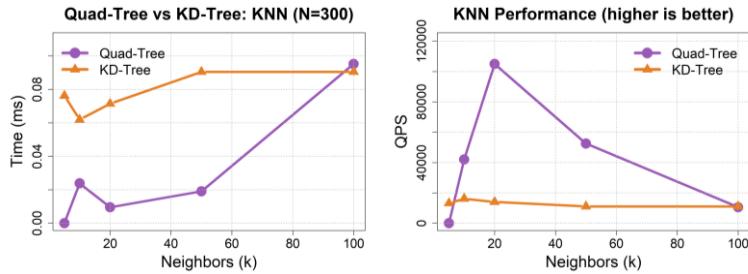


Fig. 2. Comparison between Quad-Tree and KD-Tree in KNN queries.

3.3 Statistical Evaluation

The statistical analysis reinforced the experimental performance patterns. For range queries, the Wilcoxon signed-rank test showed that Grid consistently outperformed R-Tree ($W = 15$, $p = 0.0312$), with an exceptionally large effect size ($d = 203.87$) and an average speedup of $181.56\times$, highlighting the strong advantage of uniform spatial partitioning under dense workloads. The Kruskal–Wallis test also identified significant

differences across the four indexing structures ($H(3) = 16.7521$, $p = 0.0008$), confirming their heterogeneous computational profiles. In contrast, for KNN queries, the Wilcoxon test comparing Quad-Tree and KD-Tree ($W = 1$, $p = 0.9688$) detected no significant difference, despite a large but inconsistent effect ($d = -1.51$). Overall performance remained balanced, with KD-Tree offering only a modest speedup ($0.43\times$) and surpassing Quad-Tree solely at the largest neighborhood size ($k = 100$). Together, these results indicate that while Grid shows clear statistical superiority for range-query workloads, KNN performance is more evenly distributed across indexing methods.

3.4 Regional Performance Analysis

The regional analysis reveals pronounced spatial and productive disparities. The Selva registers the highest production (1.93 M ton) and average yield (14,534 kg/ha), with 65.1% of its area classified as high-productivity. Although the Sierra holds the largest number of crop records (53,542), its production (510,339 ton) and yield (12,038 kg/ha) are comparatively lower. The Costa presents intermediate production but the highest proportion of high-productivity zones (68.5%). In terms of spatial indexing performance, R-Tree consistently shows elevated query latency (650–750 ms), whereas the Grid index achieves near-zero execution times in the Sierra and Costa and only 10 ms in the Selva, demonstrating superior stability and scalability under dense agricultural datasets. These regional patterns, summarized in Table 5 and illustrated in Fig. 3, underscore the computational efficiency of grid-based indexing for large-scale agro-spatial analysis [12, 17].

Table 5. Regional Performance Analysis

Region	Crops	Prod (ton)	Yield (kg/ha)	HighProd (%)	R-Tree (ms)	Grid (ms)
Selva	17,039	1,932,406.68	14,534.06	65.13	750	10
Sierra	53,542	510,339.22	12,038.68	52.39	660	≈ 0
Costa	21,250	391,395.59	12,472.37	68.52	650	≈ 0

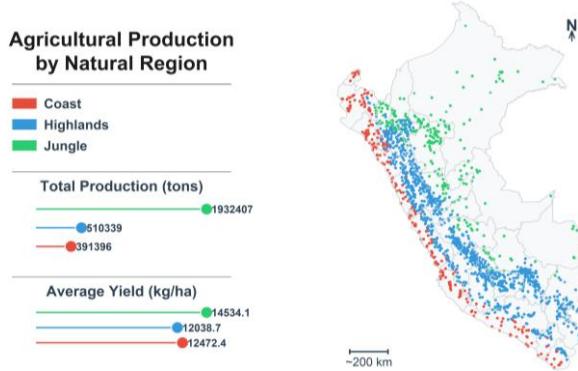


Fig. 3. Regional yield analysis.

3.5 Environmental Impact Analysis

Environmental affectations displayed clear spatial heterogeneity, with drought, pests, and frost emerging as the dominant stressors on agricultural performance. Nationally, pest incidence accounted for the largest share of impacted crop units (23.35%), followed by frost (9.08%) and drought (7.9%), underscoring the relevance of biotic and climatic pressures. To describe the spatial structure of these stressors, Table 6 outlines the Top 5 departments per affection type based on the proportion of impacted crop units. These results support the subsequent spatial density examination (Fig. 4) and the integrated performance assessment.

Table 6. Drought Impacts (Top 6 Departments)

Aff. Type	Department	Total Crops (n)	Affected Crops (n)	Affectation (%)	Avg. Loss (%)
Drought	Cajamarca	10,398	2,260	21.73	16.85
	Lambayeque	773	101	13.07	81.99
	Áncash	6,569	699	10.64	24.14
	Tumbes	1,194	121	10.13	0.67
	Huancavelica	3,372	335	9.93	1.43
Pests	Apurímac	3,779	1,644	43.50	99.08
	Tacna	2,625	1,106	42.13	38.82
	Tumbes	1,194	436	36.52	4.22
	Huánuco	6,158	1,942	31.54	82.37
	Ayacucho	6,212	1,952	31.42	11.29
Frost	Cusco	6,344	1,750	27.59	114.06
	Puno	6,128	1,437	23.45	12.92
	Huancavelica	3,372	626	18.56	1.20
	Ayacucho	6,212	1,042	16.77	2.05
	Junín	1,496	197	13.17	164.77

Drought impacts followed an inter-Andean pattern, with Cajamarca showing the highest incidence (21.73%), followed by Lambayeque (13.07%) and Áncash (10.64%). Lambayeque recorded the most severe average losses (81.99%), highlighting critical drought stress in coastal systems, while Tumbes (10.13%) and Huancavelica (9.93%) reflected additional localized exposure.

Pest affectations were the most extensive nationally. Apurímac (43.50%) and Tacna (42.13%) presented the highest incidences, with Tumbes (36.52%), Huánuco (31.54%), and Ayacucho (31.42%) also heavily affected; Huánuco recorded high-severity losses (82.37%). These results confirm pests as the dominant stressor in magnitude and spatial concentration.

Frost impacts clustered in high-elevation zones, led by Cusco (27.59%) and Puno (23.45%), with severe losses in Cusco (114.06%). Huancavelica (18.56%) and Ayacucho (16.77%) maintained notable vulnerability, while Junín exhibited the most extreme loss estimate (164.77%), indicating near-total yield failure.

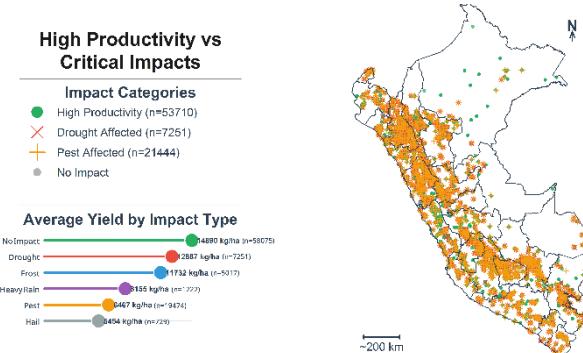


Fig. 4. Spatial distribution of drought, pest, and frost affectations across agricultural plots

3.6 Spatial Indexing Applied: Irrigation Systems, Grid, and KD-Tree

Irrigation systems exhibit marked asymmetry across the country: gravity irrigation dominates in 23 of 25 departments, while sprinkler and drip systems appear in only one department each. The largest irrigated clusters—Arequipa (5,752 units), Lima (5,139), Áncash (3,961), Moquegua (3,586), and Ica (3,055)—align consistently with the standardized 1,186-cell grid. Applying a KD-Tree algorithm ($k = 100$) within this grid revealed a highly compact cluster in Amazonas, where drought-affected crops spatially coincide with their nearest neighbors and major pest hotspots, indicating localized convergence of multiple stressors. Overall, the predominance of gravity irrigation combined with grid partitioning and KD-Tree clustering establishes a robust spatial framework for multi-factor agricultural risk assessment (Fig. 5).

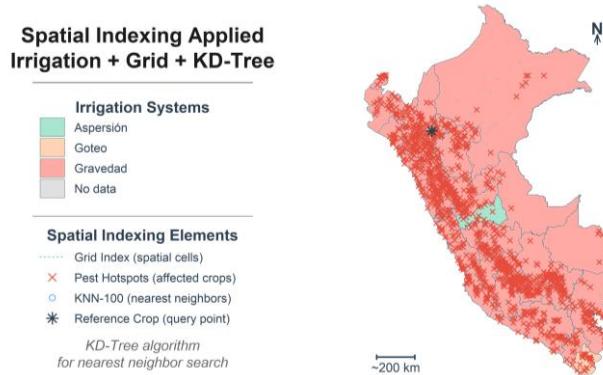


Fig. 5. Spatial indexing applied to irrigation and pests (combined hotspot visualization).

3.7 Spatial Density and Productive Patterns

The spatial density analysis identified a concentrated agricultural corridor dominated by five high-intensity zones with densities ranging from 0.1052–0.1335 crops/km². Zones 2-4 and 2-3 form the strongest cluster (>26,800 crops), while 3-2, 5-1, and 4-2

extend this axis north–south. Grid queries remained near-instant (0 ms) and R-Tree searches stable (~560–1,030 ms). These patterns outline a continuous central belt of intensified cultivation, as shown in Fig. 6, highlighting structurally cohesive production hotspots [17, 18].

Table 7. Top 5 High-Density Zones

Zone	Crops	Production (t)	Area (km ²)	Density	Grid (ms)	R-Tree (ms)
2-4	13,631	737,491	102,066.94	0.1335	0	610
2-3	13,245	125,914	101,214.86	0.1309	0	1,030
3-2	13,000	50,166	100,007.44	0.1300	0	560
5-1	11,929	70,835	98,448.96	0.1212	0	640
4-2	10,519	23,204	100,007.44	0.1052	0	660

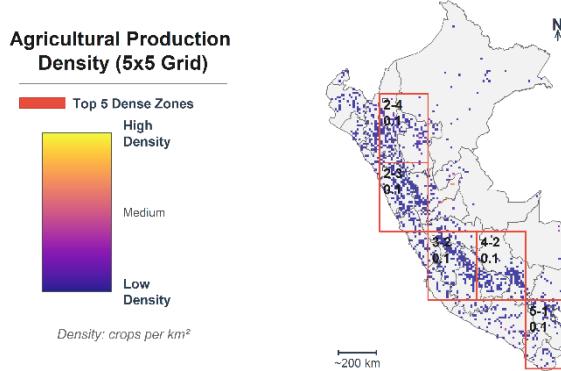


Fig. 6. Spatial agricultural density map.

4 Discussion

The results clearly establish the superiority of the Grid index over the classic R-Tree for range queries (0.1–2 km radius) on a nationwide dataset of 91 831 georeferenced agricultural plots (ENA 2024 – Peru). Grid achieved stable latencies of 3.66–3.70 ms, sustained >270 QPS, and delivered 180–183× speedups compared to R-Tree (Wilcoxon $W = 15$, $p = 0.0312$, Cohen’s $d = 203.87$), while requiring virtually zero memory overhead (0.0013 MB vs 64.8 MB for R-Tree) [5, 19]. For k-nearest neighbor queries ($k = 5–100$), Quad-Tree dominated at low-to-medium k values (up to 105 000 QPS), whereas KD-Tree only marginally outperformed it at $k = 100$, with no statistically significant differences overall (Wilcoxon $p = 0.9688$) [10, 11]. These patterns confirm that simple uniform partitioning dramatically outperforms hierarchical bounding-box structures in dense, moderately homogeneous point clouds typical of national-scale agricultural registries.

A key practical contribution of this work is the seamless integration of these light-weight indexes into real-world IoT and edge-cloud environments for precision agriculture. The near-zero memory footprint and sub-4 ms range-query latency of Grid enable direct deployment on low-cost IoT gateways, LoRaWAN/ NB-IoT nodes, and mobile

devices used by agricultural extension agents, supporting real-time decision making (variable-rate irrigation, targeted pest alerts, and traceability) without requiring high-performance servers [2, 14, 20]. The combination of Grid (range) + Quad-Tree/KD-Tree (KNN) comfortably handles hundreds of concurrent queries per second from streaming sensor networks, overcoming the performance bottlenecks of traditional R-Tree variants in dynamic spatio-temporal flows [1, 2, 13]. This makes the proposed solution immediately applicable to national early-warning systems and on-farm IoT platforms in resource-constrained settings.

Regional analysis revealed marked eco-productive contrasts: the Amazon rainforest (Selva) exhibited the highest yield (14 534 kg/ha) and production (1.93 M tons) but also the highest national pest incidence (23.35 %), whereas the Andes (Sierra) concentrated the largest number of records (53 542) together with severe drought and frost exposure (Cusco 27.59 %, Puno 23.45 %). Grid maintained near-instantaneous response (≤ 10 ms even in the spatially dispersed Selva vs 650–750 ms for R-Tree) demonstrates remarkable robustness across two orders of magnitude in point density, a critical feature for Andean and Amazonian countries with extreme altitudinal and climatic gradients [12, 17, 21].

In conclusion, the selective Grid + Quad-Tree/KD-Tree strategy offers the best performance–memory–energy trade-off for massive agricultural spatial data management in IoT-enabled precision agriculture systems, significantly outperforming modern persistent-memory, streaming, or graph-oriented R-Tree variants [1–3, 13]. Remaining limitations (false positives due to cell discretization and bounding-box overestimation) can be addressed in future work through precise polygon indexing or hybrid learned indexes [18]. The present findings provide a solid technical foundation for the immediate modernization of national platforms such as Peru’s INEI/MIDAGRI and for the scalable deployment of climate-resilient IoT precision agriculture solutions across tropical and mountainous developing regions [6, 15, 16, 20, 22].

5 Conclusions

Tree/KD-Tree, dramatically outperforms the classic R-Tree on a nationwide Peruvian agricultural dataset (91 831 georeferenced ENA 2024 records), achieving 180–183× speedups and >270 QPS in range queries (3.66–3.70 ms), near-zero memory usage (0.0013 MB), and full stability across highly heterogeneous regions (≤ 10 ms even in the dispersed Amazon rainforest). Statistical tests (Wilcoxon $p = 0.0312$, $d = 203.87$; Kruskal-Wallis $p = 0.0008$) confirm the overwhelming superiority of uniform partitioning for dense real-world agricultural workloads.

The proposed solution is immediately deployable on low-cost IoT and edge-cloud infrastructures, enabling real-time national-scale precision agriculture applications: early warning of drought/pest/frost convergence, variable-rate management zones, and climate-resilient decision making. These advances provide a practical, scalable blueprint for modernizing public systems such as Peru’s INEI/MIDAGRI registries and for extending IoT-enabled, data-driven agricultural intelligence across tropical and Andean developing countries [1, 13, 18, 19].

References

1. Kim J, Hong S, Jeong S, et al (2024) SGIR-Tree: Integrating R-Tree Spatial Indexing as Subgraphs in Graph Database Management Systems. *ISPRS International Journal of Geo-Information* 2024, Vol 13, 13:. <https://doi.org/10.3390/IJGI13100346>
2. Peng W, Chen L, Ouyang X, Xiong W (2024) A Time-Identified R-Tree: A Workload-Controllable Dynamic Spatio-Temporal Index Scheme for Streaming Processing. *ISPRS International Journal of Geo-Information* 2024, Vol 13, 13:. <https://doi.org/10.3390/IJGI13020049>
3. Xia J, Huang S, Zhang S, et al (2020) DAPR-tree: a distributed spatial data indexing scheme with data access patterns to support Digital Earth initiatives. *Int J Digit Earth* 13:1656–1671. <https://doi.org/10.1080/17538947.2020.1778804>;WGROUPE:STRING:PUBLICATION
4. Zhu Q, Gong J, Zhang Y (2007) An efficient 3D R-tree spatial index method for virtual geographic environments. *ISPRS Journal of Photogrammetry and Remote Sensing* 62:217–224. <https://doi.org/10.1016/J.ISPRSJPRS.2007.05.007>
5. Mao Q, Qader MA, Hristidis V (2023) Comparison of LSM indexing techniques for storing spatial data. *Journal of Big Data* 2023 10:1 10:51-. <https://doi.org/10.1186/S40537-023-00734-3>
6. Sandonís-Pozo L, Llorens J, Escolà A, et al (2022) Satellite multispectral indices to estimate canopy parameters and within-field management zones in super-intensive almond orchards. *Precision Agriculture* 2022 23:6 23:2040–2062. <https://doi.org/10.1007/S11119-022-09956-6>
7. Colaço AF, Molin JP, Rosell-Polo JR, Escolà A (2018) Spatial variability in commercial orange groves. Part 1: canopy volume and height. *Precision Agriculture* 2018 20:4 20:788–804. <https://doi.org/10.1007/S11119-018-9612-3>
8. Colaço AF, Molin JP, Rosell-Polo JR, Escolà A (2018) Spatial variability in commercial orange groves. Part 2: relating canopy geometry to soil attributes and historical yield. *Precision Agriculture* 2018 20:4 20:805–822. <https://doi.org/10.1007/S11119-018-9615-0>
9. Huang PW, Lin PL, Lin HY (2001) Optimizing storage utilization in R-tree dynamic index structure for spatial databases. *Journal of Systems and Software* 55:291–299. [https://doi.org/10.1016/S0164-1212\(00\)00078-9](https://doi.org/10.1016/S0164-1212(00)00078-9)
10. Samet H (1990) The design and analysis of spatial data structures. The design and analysis of spatial data structures. [https://doi.org/10.1016/0924-2716\(91\)90007-i](https://doi.org/10.1016/0924-2716(91)90007-i)
11. Bentley JL (1975) Multidimensional binary search trees used for associative searching. *Commun ACM* 18:509–517. <https://doi.org/10.1145/361002.361007>
12. Serrao L, Giovannini L, Terrones LEB, et al (2025) Integrating farmers' perceptions into climate change assessment in the data-scarce Peruvian Amazon.

- Climatic Change 2025 178:3 178:55-. <https://doi.org/10.1007/S10584-025-03891-X>
13. Lavinsky B, Zhang X (2022) PM-Rtree: A Highly-Efficient Crash-Consistent R-tree for Persistent Memory. ACM International Conference Proceeding Series. [https://doi.org/10.1145/3538712.3538713;WGROUPE:STRING:ACM](https://doi.org/10.1145/3538712.3538713)
 14. Zhou Y, De S, Wang W, et al (2017) Spatial Indexing for Data Searching in Mobile Sensing Environments. Sensors 2017, Vol 17, 17:. <https://doi.org/10.3390/S17061427>
 15. Omia E, Bae H, Park E, et al (2023) Remote Sensing in Field Crop Monitoring: A Comprehensive Review of Sensor Systems, Data Analyses and Recent Advances. Remote Sensing 2023, Vol 15, 15:. <https://doi.org/10.3390/RS15020354>
 16. Xu J, Cui Y, Zhang S, Zhang M (2024) The evolution of precision agriculture and food safety: a bibliometric study. Front Sustain Food Syst 8:1475602. <https://doi.org/10.3389/FSUFS.2024.1475602/FULL>
 17. Móstiga M, Armenteras D, Vayreda J, Retana J (2024) Decoding the drivers and effects of deforestation in Peru: a national and regional analysis. Environment, Development and Sustainability 2024 27:7 27:17395–17415. <https://doi.org/10.1007/S10668-024-04638-X>
 18. Wang C, Yu J, Zhao Z (2022) GLIN: A (G)eneric (L)earned (In)dexing Mechanism for Complex Geometries. Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03â•fi05, 2018, Woodstock, NY 1:. <https://doi.org/10.1145/1122445.1122456>
 19. Guttman A (1984) R-trees: A dynamic index structure for spatial searching. Proceedings of the ACM SIGMOD International Conference on Management of Data 47–57. <https://doi.org/10.1145/602259.602266;CSUBTYPE:STRING:CONFERENCE>
 20. San Emeterio de la Parte M, Martínez-Ortega JF, Hernández Díaz V, Martínez NL (2023) Big Data and precision agriculture: a novel spatio-temporal semantic IoT data management framework for improved interoperability. Journal of Big Data 2023 10:1 10:52-. <https://doi.org/10.1186/S40537-023-00729-0>
 21. Mader C, Godde P, Hägele E, et al (2025) Mapping and Geospatial Analysis of Ancient Terrace Agricultural Systems in Lucanas Province, Peruvian Andes, Based on Satellite Imagery, High-Resolution DSMs, and Field Surveys. Geoarchaeology 40:e70002. <https://doi.org/10.1002/GEA.70002;PAGE:STRING:ARTICLE/CHAPTER>
 22. Espinel R, Herrera-Franco G, García JLR, Escandón-Panchana P (2024) Artificial Intelligence in Agricultural Mapping: A Review. Agriculture 2024, Vol 14, 14:. <https://doi.org/10.3390/AGRICULTURE14071071>

ILMA MAGDA MAMANI MAMANI

Acuse de Aceptaciónm (Mail)

The 2nd International Symposium on Green Technologies and Applications notification for your paper 139

Microsoft CMT <noreply@msr-cmt.org> para mi · 26 oct 2025, 6:19 (hace 1 día)

Congratulations - your paper 139: Comparative Evaluation of Spatial Indexing Methods Applied to the Georeferenced Characterization of Agricultural Units and Productivity in Peru during the year 2024 has been accepted and will be presented at the International Symposium on Green Technologies and Applications (ISGTA' 2025) to be held in Portalegre, Portugal on November 19-21, 2025.

We would like to thank you for submitting your manuscript to our Symposium for which we received a large number of high-quality papers. However, unfortunately, many good and interesting papers could not be included in the ISGTA' 2025 main program. Your camera-ready produced following and must be submitted through CMT as well.

The paper length is limited to 12 pages with your registration. There are no extra pages allowed for this event.

Note that the deadline for the submission of your camera-ready version is November 8, 2025. Instructions are available at the Symposium website: https://isgta-conf.org/?page_id=2247. Please find the reviewers' comments, which may help you to improve your paper.

Best Regards,
ISGTA 2025 TPC Chairs

Please do not reply to this email as it was generated from an email account that is not monitored.

To stop receiving conference emails, you can check the 'Do not send me conference email' box from your User Profile.

Microsoft respects your privacy. To learn more, please read our [Privacy Statement](#).

Microsoft Corporation
One Microsoft Way
Redmond, WA 98052

ISGTA 2025 – Important Deadlines (Registration & Camera-Ready)

ISGTA ISGTA <isgta.conference@gmail.com> para m.lahby@enscasa.ma, bcc: mi · dom, 26 oct, 17:27 (hace 21 horas)

Dear Authors,

Congratulations on the acceptance of your paper to the Second International Symposium on Green Technologies and Applications (ISGTA 2025)!

We sincerely appreciate your valuable contribution and look forward to your participation in the conference.

To ensure the smooth inclusion of your work in the proceedings, please take note of the following critical deadlines and information:

1. Conference Registration
 - Final Author Registration Deadline: **November 11, 2025**
 - Important: After Camera-ready and Copyright submission, papers cannot be withdrawn from the conference proceedings.
 - Authors with multiple accepted papers:
 - * Every accepted paper at ISGTA 2025 must have an associated registration at the Author rate.
 - * For each paper at least one co-author must be registered as author (Author rate) and presented in person by one of co-author to secure the paper's publication in Conference proceedings and Springer.
 - The registration of a single co-author as author (Author rate) will secure the publication of up to two papers from the same co-author.
 - More details about the registration process can be found here: [Registration – ISGTA25](#)
2. Camera-Ready Submission & Springer Copyright Form
 - Hard Deadline: **November 8, 2025**
 - Submit your finalized manuscript (formatted per Springer guidelines) along with the completed copyright form.
 - Submission guidelines can be found at: [Camera Ready – ISGTA25](#)

Once again, congratulations, and we are excited to welcome you to ISGTA 2025!

Capítulo 2

Evidencias de la Unidad II

En esta sección se adjuntan las evidencias realizadas en la segunda unidad.

Resumen: (Deep Learning for Cross - Domain Data Fusion in Urban Computing & Taxonomy Advances, Land Use)

Introducción y contexto Urbano:

- Las ciudades enfrentan desafíos como gestión de tráfico, consumo de energía y contaminación ambiental debido a la urbanización rápida.
- La computación urbana usa fusión de datos de múltiples fuentes (geográficas, tráfico, redes sociales, demográficas, ambientales) y modalidades (espacio-temporal, visual, textual) para soluciones sostenibles.
- Evolución desde machine learning tradicional a deep learning, que ofrece mayor capacidad para extraer features y fusionar datos cross-modal.

Taxonomía General (sección 2 y Figura 3):

- Datos: Perspectiva multi-source (geográficas, tráficas, etc) y multi-modality (espacio temporal, visual, textual, otros como audio/video).
- Métodos de Fusión: Cuatro categorías principales - feature-based (concatenación, adición / multiplicación, graph-based), alignment-based (atención / encoder-based), contrast-based (aprendizaje contrastivo), generación-based (mask modeling, generative learning, LLM-based).
- Aplicaciones: 7 áreas - planificación urbana, transporte, economía, seguridad pública, sociedad, medio ambiente y energía.

Perspectiva de Datos (sección 3 y tabla 2)

- Clasifica datos en seis categorías: geográficas (POIs, imágenes satelitales / street-view), tráfico (trayectorias, flujo de bájicos, redes viales), redes sociales (texto, imágenes geo-etiquetadas), demográficas (población, crimen, uso de suelo), ambientales (meteorología, vegetación, calidad del aire) y otros.
- Análisis estadísticos: ~ 70% de papers usan datos geográficos / tráfico; datasets populares de Beijing / Nueva York (Figuras).

Métodos de Fusión (sección 4):

- Feature-based: Fusión simple como concatenación o adición de features extraídas de modalidades.
- Alignment-based: Alinear representaciones modales usando atención o encoders.
- Contrast-based: Aprendizaje contrastivo para maximizar similitudes entre modalidades positivas y minimizar con negativas.

Resumen: Aspects of Forest Degradation and Inventory (Approaches for Forest Management)

Introducción a la Degradoación Forestal

- Los bosques proporcionan servicios esenciales. La degradación reduce su valor sin eliminarlo completamente, llevando a deforestación eventual. Es más perjudicial que la deforestación en términos de masa terrestre y valor, afectando densidad, calidad y composición de especies.
- Impacto: Perdida de canopy, invasión de especies generalistas, erosión del suelo, salinización, drenaje de humedales; afecta a 2.200 millones de hectáreas globales y un tercio de la población mundial. Reduce biodiversidad, productos y servicios, impactando a millones dependientes de bosques.

Drivers y Causas de Degradoación

- Drivers indirectos: Factores subyacentes que impulsan causas directas, categorizados en:
 - Técnicas: Expansión agrícola, cultivo itinerante.
 - Económicas: Alto precio de sustitutos, acceso a mercados, infraestructura (carreteras, turismo), dependencia poblacional de bosques.
 - Culturales: Actitudes individuales / públicas, Salto de población por valor forestal, búsqueda de rentas.
 - Demográficas: Recimiento / densidad poblacional, migración, urbanización, conflitos.



Guía de Instalación de QGIS

Estudiante: Ilma Magda Mamani Mamani
Docente: Dr. Fred Torres Cruz
Curso: Estadística Espacial
Institución: Universidad Nacional del Altiplano (UNA PUNO)
Fecha: 29 de octubre de 2025

Introducción

Esta guía presenta los pasos detallados para la instalación de QGIS, software esencial para el análisis espacial en el curso de Estadística Espacial.

Proceso de Instalación

Paso 1: Ejecutar el Instalador



Figura 1: Ejecutar el instalador de QGIS como administrador

Paso 2: Aceptar la Licencia

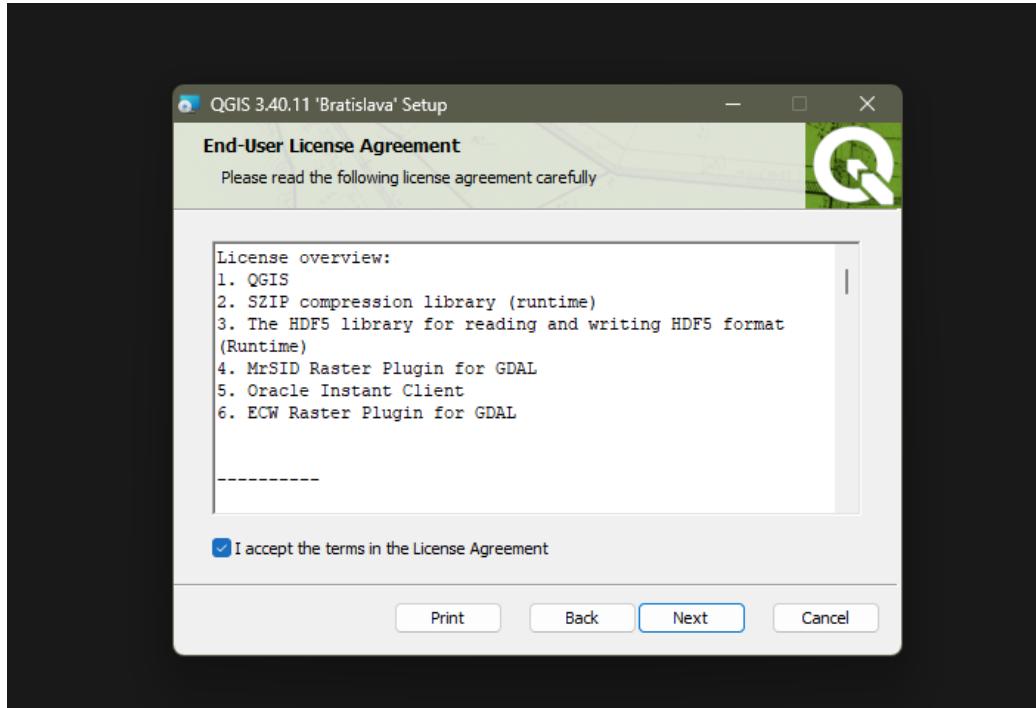


Figura 2: Aceptar los términos del acuerdo de licencia

Paso 3: Elegir Directorio de Instalación

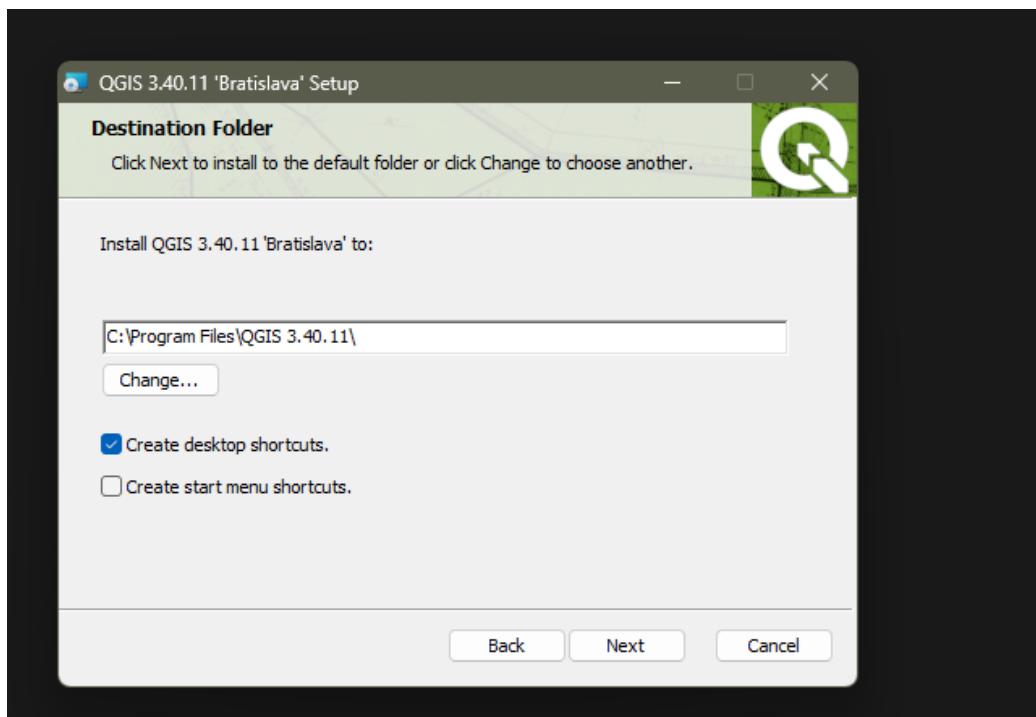


Figura 3: Seleccionar la ubicación de instalación

Paso 4: Finalizar Instalación



Figura 4: Completar el proceso de instalación

Paso 5: Verificar Instalación

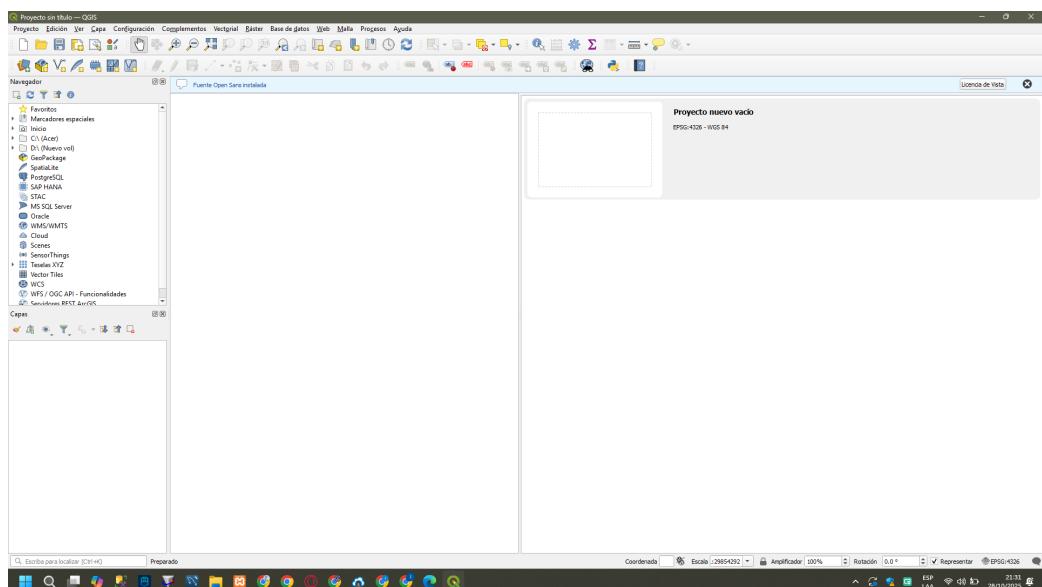


Figura 5: Interfaz principal de QGIS instalado correctamente

Resumen de Pasos

Paso	Descripción
1	Ejecutar el instalador como administrador
2	Aceptar los términos de la licencia
3	Seleccionar directorio de instalación
4	Completar la instalación y ejecutar QGIS
5	Verificar la instalación correcta

Cuadro 1: Resumen del proceso de instalación de QGIS

Recomendaciones

- Verificar que el sistema cumpla con los requisitos mínimos
- Contar con conexión a internet durante la instalación
- Cerrar todas las aplicaciones antes de instalar
- Reiniciar el equipo después de la instalación



Polígonos, Puntos y Líneas

Estudiante: Ilma Magda Mamani Mamani

Docente: Dr. Fred Torres Cruz

Curso: Estadística Espacial

Fecha: 11 de noviembre de 2025

Descripción del Trabajo

En este ejercicio se trabajó con QGIS para crear y organizar diferentes tipos de geometrías espaciales sobre una imagen base (orthomosaic). Se digitalizaron tres tipos de capas vectoriales:

Capas de Polígonos: Se crearon capas para representar áreas como casas, árboles y patios, delimitando espacios cerrados con superficie definida.

Capas de Líneas: Se digitalizaron las calles y el recorrido de los autos, representando elementos lineales sin área pero con longitud.

Capa Base (Orthomosaic): Imagen ráster georreferenciada con tres bandas espectrales (Rojo, Verde, Azul) que sirvió como referencia para la digitalización.



Figura 1: Vista del proyecto con las capas vectoriales sobre la imagen orthomosaic

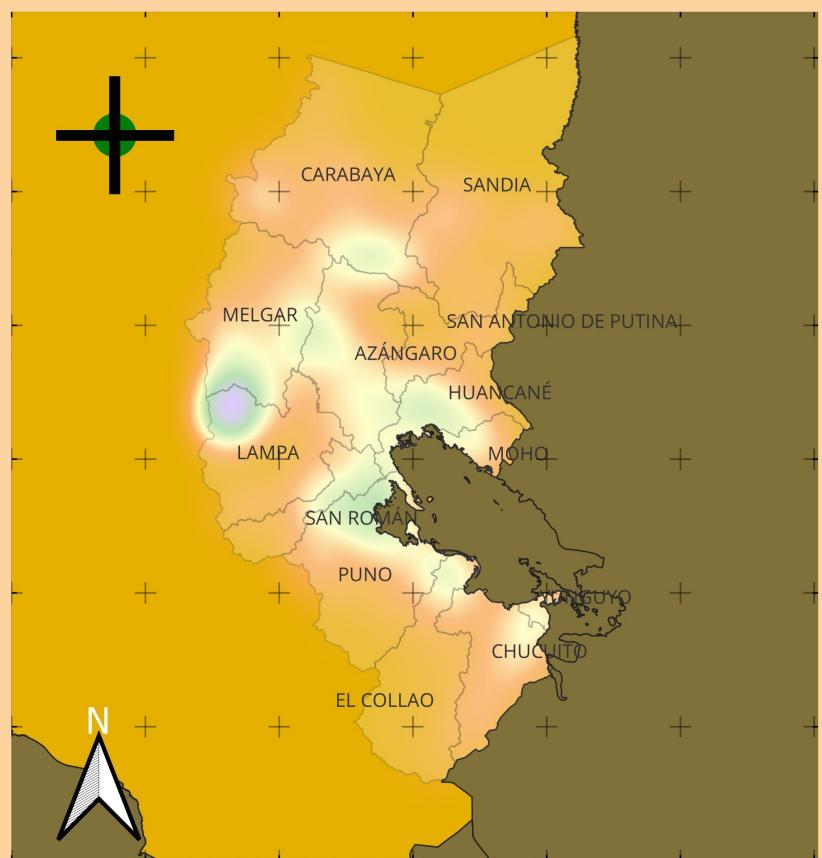
Tipo de Capa	Elementos Digitalizados
Polígonos	casas, Arboles, patios
Líneas	Calles, Autos
Ráster	orthomosaic (3 bandas RGB)

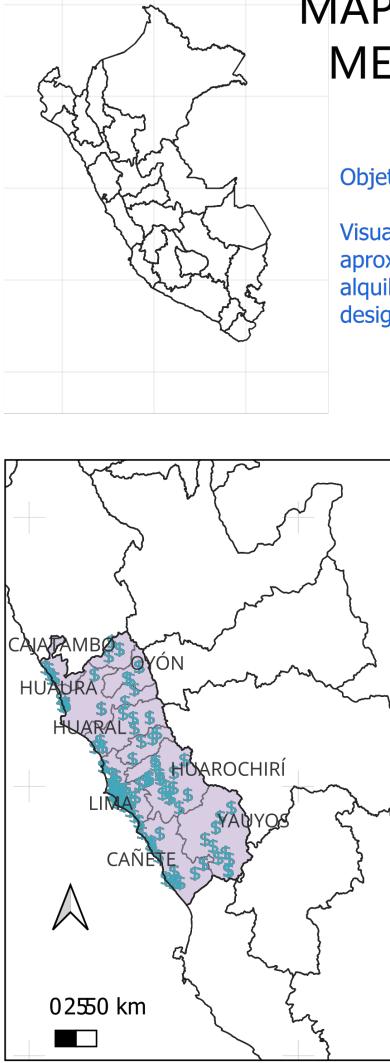
Cuadro 1: Resumen de capas creadas en el proyecto

La organización por capas permite gestionar cada elemento de forma independiente, facilitando el análisis espacial y la generación de estadísticas zonales para estudios geográficos.

MAPA DE CALOR DE LA COBERTURA MOVIL DE LA REGION DE PUNO

ILMA MAGDA MAMANI MAMANI





MAPA DE CALOR DE GASTOS MENSUALES EN HOGARES PERUANOS (QGIS)

Objetivo:

Visualizar la distribución territorial del gasto total mensual aproximado de hogares peruanos en servicios básicos, alquiler y otros conceptos, identificando patrones de desigualdad regional.

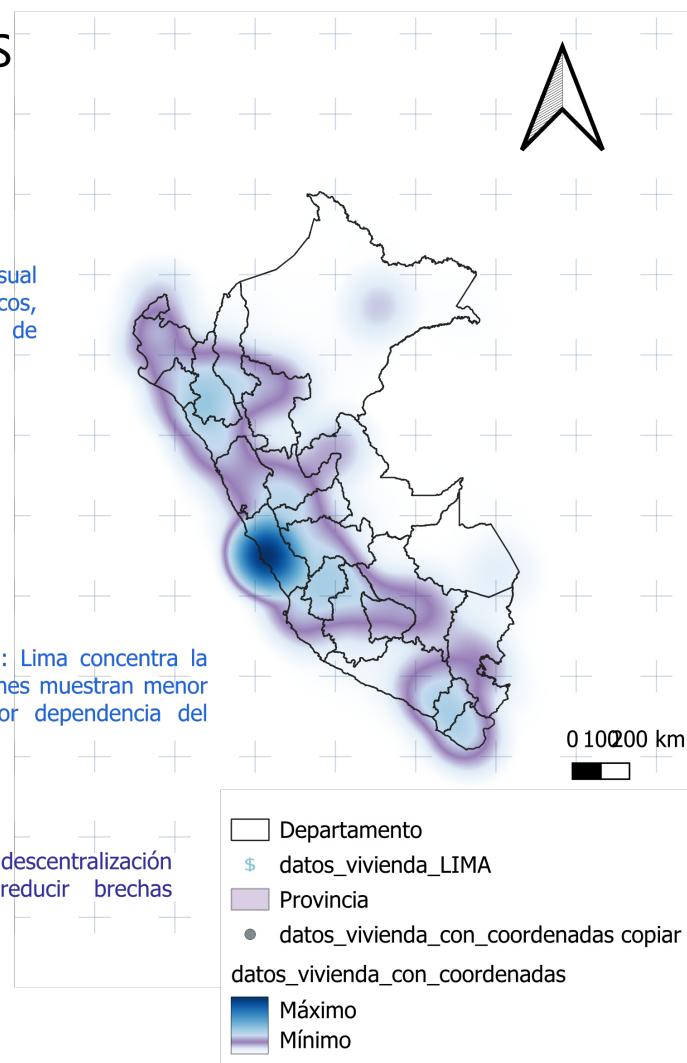
Resultado Principal (Mapa de Carter):
Se observa una alta concentración de gasto en Lima, intensidad media en regiones costeras como Arequipa y La Libertad y baja intensidad en la sierra y selva (Huánuco, Aperimac, Madre de Dios).

Interpretación:

Esto refleja una fuerte desigualdad territorial: Lima concentra la capacidad de gasto, mientras que otras regiones muestran menor acceso a servicios pagados y posible mayor dependencia del autoconsumo.

Aporte:

El mapa evidencia la necesidad de políticas de descentralización económica e inclusión financiera para reducir brechas regionales.



DE: ILMA MAGDA MAMANI MAMANI

Modelos Jerárquicos Espaciales Bayesianos Multiescala para el Análisis de las Desigualdades en los Gastos de los Hogares Peruanos

Ilma Magda Mamani Mamani¹[0009-0002-8605-0086]

¹Escuela Profesional de Ingeniería Estadística e Informática

Universidad Nacional del Altiplano de Puno

im.mamani@est.unap.edu.pe

Abstract

Este estudio cuantifica las brechas territoriales en el gasto monetario per cápita real de los hogares peruanos e identifica sus determinantes socioeconómicos mediante modelamiento bayesiano jerárquico multiescala. Se analizaron 46,935 hogares de la ENAHO 2021–2024 mediante un modelo bayesiano jerárquico de tres niveles que integra efectos fijos socioeconómicos, efectos aleatorios por dominio geográfico y selección de variables mediante Random Forest. La inferencia se realizó con Hamiltonian Monte Carlo (NUTS) en PyMC5 utilizando 4 cadenas, 2,000 iteraciones de calibración y 2,000 muestras posteriores. El modelo converge perfectamente ($\hat{R} = 1.00$, cero divergencias) y explica el 78.3% de la variabilidad del logaritmo del gasto per cápita ($R^2 = 0.783$, MAE=S/. 1,579, MAPE=27.2%). Lima Metropolitana presenta interceptos 17.7% superiores a Sierra Norte, equivalente a S/. 2,000 mensuales atribuibles a factores territoriales. El ingreso per cápita incrementa el gasto en 30.6%, la proporción de gasto alimentario lo reduce en 27.4% (validando la Ley de Engel) y el tamaño del hogar lo disminuye en 26.2%. Los departamentos andinos exhiben coeficientes de variación superiores al 80%, evidenciando alta desigualdad interna. Las métricas LOO-CV (ELPD=-15,584) y WAIC confirman capacidad predictiva robusta sin sobreajuste. Los hallazgos revelan que las brechas territoriales persisten incluso controlando características observables, requiriendo políticas *place-based* que combinen transferencias monetarias con inversiones en infraestructura local y fortalecimiento institucional diferenciado por región.

Palabras clave: Desigualdad territorial, Gasto per cápita, Hamiltonian Monte Carlo, Modelo bayesiano jerárquico, Perú.

1 Introducción

La desigualdad económica constituye una de las principales preocupaciones analíticas y normativas en las economías en desarrollo, particularmente en países caracterizados por una elevada heterogeneidad territorial como el Perú. La coexistencia de áreas metropolitanas altamente integradas, regiones andinas con persistentes rezagos estructurales y extensas zonas rurales de baja densidad poblacional da lugar a patrones espaciales complejos del bienestar económico que no pueden ser adecuadamente capturados mediante enfoques agregados o análisis puramente descriptivos [1, 2].

En este contexto, el gasto monetario de los hogares, expresado en términos per cápita y ajustado por diferencias espaciales de precios, se reconoce ampliamente

como una de las medidas más robustas del bienestar material efectivo. A diferencia del ingreso corriente, el gasto refleja la capacidad real de los hogares para satisfacer necesidades básicas y no básicas, incorporando mecanismos de suavización intertemporal del consumo y reduciendo la influencia de shocks transitorios [7, 8]. Por ello, el análisis de la desigualdad del gasto resulta particularmente relevante para evaluar brechas persistentes de bienestar y vulnerabilidad estructural.

La literatura empírica ha documentado que las desigualdades económicas presentan una marcada dimensión espacial, manifestándose en la formación de clusters territoriales de bajo consumo y alta pobreza asociados a factores históricos, institucionales y de infraestructura [6, 17]. En el caso peruano, diversos estudios han evidenciado la persistencia de brechas territoriales sig-

nificativas entre dominios geográficos, así como diferencias sustanciales entre áreas urbanas y rurales en términos de niveles de gasto, acceso a mercados y oportunidades económicas [9, 10]. Sin embargo, gran parte de esta evidencia se basa en modelos uniescalares que no incorporan explícitamente la estructura jerárquica de los datos ni la dependencia espacial entre unidades geográficas.

La disponibilidad reciente de microdatos georreferenciados de la Encuesta Nacional de Hogares (ENAHO) para el período 2021–2024, difundidos a través del Sistema de Microdatos del Instituto Nacional de Estadística e Informática (INEI), ofrece una oportunidad metodológica excepcional para avanzar en el análisis de la desigualdad territorial del bienestar en el Perú. La ENAHO permite vincular información detallada a nivel de hogar con estructuras territoriales jerárquicas y coordenadas espaciales precisas, lo que habilita el uso de técnicas estadísticas avanzadas orientadas al análisis espacial multiescala.

En este sentido, los modelos jerárquicos espaciales bayesianos constituyen un marco analítico particularmente adecuado para el estudio de desigualdades territoriales. Estos modelos permiten integrar simultáneamente información a distintos niveles geográficos, capturar heterogeneidad no observada entre regiones, modelar explícitamente la dependencia espacial y obtener inferencias probabilísticas coherentes incluso en contextos de áreas pequeñas [3, 4]. Además, la implementación computacional mediante algoritmos modernos de muestreo, como el Hamiltonian Monte Carlo No-U-Turn Sampler (NUTS), facilita la estimación eficiente de modelos complejos con miles de observaciones [13, 20].

Desde una perspectiva de política pública, comprender la magnitud y los determinantes de la desigualdad territorial del gasto resulta crucial para el diseño de intervenciones focalizadas y territorialmente diferenciadas. La evidencia internacional sugiere que las políticas redistributivas, incluidas las transferencias públicas, pueden tener efectos heterogéneos según el contexto territorial, reforzando o mitigando desigualdades preexistentes [18, 25]. Sin un adecuado entendimiento de la dimensión espacial del bienestar, dichas políticas corren el riesgo de ser ineficientes o incluso regresivas.

En este marco, el objetivo general del presente artículo es analizar las desigualdades territoriales en el gasto monetario per cápita real de los hogares peruanos mediante un modelo jerárquico espacial bayesiano multiescala, utilizando microdatos de la ENAHO 2021–

2024. Específicamente, el estudio busca cuantificar las brechas de gasto entre dominios geográficos y estratos poblacionales, evaluar la existencia y magnitud de la dependencia espacial en los niveles de gasto, identificar las variables socioeconómicas con mayor poder predictivo mediante técnicas de aprendizaje automático, y analizar el rol del ingreso, las transferencias públicas y la composición del hogar en la explicación de dichas desigualdades.

El presente artículo se estructura de la siguiente manera: la Sección 2 describe en detalle la fuente de datos, el diseño muestral, el preprocessamiento de la información y la especificación completa del modelo bayesiano jerárquico espacial; la Sección 3 presenta los resultados empíricos, incluyendo estimaciones posteriores, diagnósticos de convergencia y métricas de validación; la Sección 4 discute las implicaciones teóricas y metodológicas de los hallazgos; finalmente, la Sección 5 resume las conclusiones principales y sugiere líneas de investigación futura.

1.1 Fuente de datos y diseño muestral

El estudio utiliza microdatos provenientes del Sistema de Microdatos del Instituto Nacional de Estadística e Informática (INEI) del Perú, específicamente de la Encuesta Nacional de Hogares (ENAHO), módulo de Condiciones de Vida y Pobreza, para los años 2021, 2022, 2023 y 2024. La ENAHO constituye la principal fuente de información estadística sobre las condiciones socioeconómicas de los hogares peruanos y es utilizada oficialmente para la medición de la pobreza monetaria a nivel nacional, regional y departamental.

El diseño muestral de la ENAHO es probabilístico, estratificado, multietápico e independiente en cada departamento del país. La estratificación se realiza considerando criterios geográficos y de urbanización. Las unidades primarias de muestreo son los conglomerados, definidos como áreas geográficas con un promedio de 140 viviendas en áreas urbanas y 100 viviendas en áreas rurales. Las unidades secundarias son las viviendas particulares, y las unidades de análisis son los hogares y las personas residentes en dichas viviendas [15].

Para el presente estudio se integraron el Módulo 1 de Características de la Vivienda y del Hogar, que contiene información sobre las características físicas de la vivienda, servicios básicos, tenencia y composición del hogar, y el Módulo 34 de Sumarias, que incluye variables agregadas y calculadas a nivel de hogar, tales como gasto total, ingreso total, perceptores de ingreso, línea

de pobreza, deflactor espacial y factor de expansión. La integración de ambos módulos se realizó mediante las variables de identificación año, número de conglomerado, número de vivienda y número de hogar. La base de datos consolidada abarca el período 2021–2024 con información trimestral, permitiendo capturar tanto la heterogeneidad transversal como la variabilidad temporal del bienestar de los hogares.

1.2 Variables de estudio

Las variables seleccionadas para el análisis se clasifican en tres categorías principales: variables de resultado, variables explicativas y variables de control espacial y temporal. La variable dependiente principal es el gasto monetario per cápita real anual del hogar, construida como el cociente entre el gasto monetario total anual del hogar en soles corrientes, el total de miembros del hogar, y el deflactor espacial de precios utilizado para ajustar diferencias en el costo de vida entre regiones. Formalmente, el gasto per cápita real se define como:

$$GastoPC_i = \frac{GASHOG1D_i}{MIEPERHO_i \cdot LD_i} \quad (1)$$

Dado el carácter asimétrico y heterocedástico del gasto, la variable utilizada en el modelamiento es su transformación logarítmica, que estabiliza la varianza y facilita la interpretación de los coeficientes como elasticidades. Así, la variable objetivo transformada se expresa como $y_i = \log(GastoPC_i + 1)$.

Las variables explicativas consideradas incluyen el ingreso per cápita real anual, que captura la capacidad económica del hogar ajustada por tamaño; las transferencias corrientes públicas, que representan el monto anual recibido por el hogar proveniente de programas sociales y transferencias gubernamentales como Juntos, Pensión 65 y Bono Familiar; el gasto en alimentos, utilizado para calcular la proporción del gasto destinada a alimentación según la Ley de Engel; el total de perceptores de ingreso, que corresponde al número de miembros del hogar que perciben algún tipo de ingreso; y el tamaño del hogar, medido por el número total de miembros.

Para capturar la heterogeneidad territorial y temporal, se incorporan variables de control espacial y temporal. El dominio geográfico clasifica a los hogares en ocho categorías: Costa Norte, Costa Centro, Costa Sur, Sierra Norte, Sierra Centro, Sierra Sur, Selva, y Lima Metropolitana. El estrato poblacional clasifica a los centros poblados según tamaño poblacional en ocho categorías, desde ciudades de más de 500,000 habitantes

hasta áreas rurales simples. Para simplificar el análisis, esta variable se recodificó en una variable binaria URBANO, donde el valor 1 corresponde a los estratos del 1 al 5, y el valor 0 a los estratos del 6 al 8. Adicionalmente, se utilizan las coordenadas geográficas del conglomerado para modelar dependencia espacial, y se controlan efectos temporales fijos anuales y trimestrales para capturar estacionalidad y tendencias temporales.

A partir de las variables originales se construyeron indicadores derivados. La proporción de gasto en alimentos se calcula como el cociente entre el gasto en alimentos y el gasto total del hogar, utilizada como proxy de bienestar según la Ley de Engel [14]. El ratio de perceptores mide la proporción de miembros económicamente activos en el hogar mediante el cociente entre perceptores de ingreso y total de miembros. El indicador de pobreza relativa se construye como el cociente entre el gasto total del hogar y la línea oficial de pobreza total.

1.3 Muestreo estratificado y preprocesamiento

Dado el volumen considerable de la base de datos consolidada, que comprende aproximadamente 500,000 observaciones para el período 2021–2024, se implementó un procedimiento de muestreo estratificado para optimizar la eficiencia computacional sin comprometer la validez inferencial. El tamaño de la muestra se fijó en 50,000 observaciones, asegurando suficiente poder estadístico para la estimación de efectos jerárquicos y espaciales. La estratificación se realizó de manera cruzada considerando cuatro dimensiones simultáneas, creando estratos combinados como producto cartesiano de año, trimestre, dominio geográfico y clasificación urbano-rural. Este diseño garantiza que la muestra preserve las proporciones originales de cada combinación temporal-espacial, evitando sesgos de selección.

La asignación proporcional del tamaño muestral por estrato se calculó como el producto entre el tamaño total de la muestra y la proporción poblacional del estrato, con un mínimo de cinco observaciones por estrato para evitar estratos vacíos. Esta restricción asegura que incluso las combinaciones menos frecuentes de características temporales y espaciales estén representadas en la muestra final. El procedimiento de muestreo utilizó semilla aleatoria fija para garantizar reproducibilidad de los resultados.

El preprocesamiento de los datos siguió un protocolo sistemático diseñado para preservar la representatividad muestral y la heterogeneidad real del consumo, minimizando la eliminación arbitraria de observaciones

válidas. En primer lugar, se removieron observaciones con valores ausentes en variables críticas como gasto, ingreso y coordenadas geográficas. Posteriormente, se excluyeron observaciones con valores lógicamente inconsistentes, tales como gastos o ingresos no positivos, tamaño del hogar no positivo, número de perceptores mayor al tamaño del hogar, gasto en alimentos negativo, o coordenadas geográficas fuera del territorio peruano.

Para el tratamiento de valores atípicos, se aplicó un criterio conservador de detección de outliers basado en los percentiles 1% y 99% de las distribuciones del logaritmo del gasto per cápita, el ingreso per cápita y la proporción de gasto en alimentos. Solo se eliminaron valores extremos que claramente representaban errores de medición, como gastos per cápita superiores a diez veces el percentil 99. Este enfoque evita la eliminación indebida de desigualdad genuina, preservando la variabilidad real de los datos. La tasa de retención final fue aproximadamente del 92%, considerada satisfactoria para estudios de este tipo [16].

1.4 Selección de variables mediante Random Forest

Para identificar las variables socioeconómicas con mayor poder predictivo sobre el gasto per cápita real, se implementó un procedimiento de selección de variables basado en Random Forest. Esta técnica de aprendizaje automático permite evaluar la importancia relativa de cada predictor considerando interacciones no lineales y efectos de confusión [5, 22].

El algoritmo de Random Forest se entrenó con treinta árboles de decisión, profundidad máxima de seis niveles para evitar sobreajuste, y una proporción de muestreo bootstrap del 70% por árbol. El criterio de división utilizado fue la reducción de error cuadrático medio. Para optimizar la eficiencia computacional sin comprometer la capacidad de generalización, el entrenamiento se realizó sobre una submuestra aleatoria de 20,000 observaciones cuando el tamaño del dataset excedía este umbral.

La importancia de cada variable se midió mediante la reducción promedio de impureza, normalizada para comparabilidad entre predictores. Se seleccionaron las cinco variables con mayor importancia relativa, asegurando que la variable de dominio geográfico estuviera incluida para capturar efectos jerárquicos espaciales. Este enfoque híbrido combina el poder predictivo del aprendizaje automático con la interpretabilidad de los modelos paramétricos [16], permitiendo identificar los

determinantes más relevantes del gasto per cápita sin imponer supuestos funcionales restrictivos.

Las variables continuas seleccionadas fueron estandarizadas mediante transformación z-score, substractando la media muestral y dividiendo por la desviación estándar. Esta estandarización facilita la interpretación de los coeficientes estimados, permite la comparación directa de efectos entre variables medidas en diferentes escalas, y mejora la convergencia de los algoritmos de muestreo bayesiano al reducir la correlación entre parámetros.

1.5 Modelo bayesiano jerárquico espacial

El análisis inferencial se basa en un modelo bayesiano jerárquico de tres niveles que integra efectos fijos, efectos aleatorios y estructura espacial. Para cada hogar i en el conglomerado c del dominio j , el logaritmo del gasto per cápita real se modela mediante la siguiente especificación:

$$y_{icj} = \alpha_j + \mathbf{x}'_{icj}\beta + \gamma_t + \delta_q + \eta_e + \varepsilon_{icj} \quad (2)$$

donde y_{icj} representa el logaritmo del gasto per cápita real, α_j es el intercepto específico del dominio j que captura efectos aleatorios espaciales, \mathbf{x}_{icj} es el vector de covariables continuas estandarizadas, β es el vector de coeficientes fijos, γ_t es el efecto fijo del año t , δ_q es el efecto fijo del trimestre q , η_e es el efecto fijo del área urbana o rural, y ε_{icj} es el error idiosincrático del hogar distribuido normalmente con media cero y varianza σ^2 .

Los interceptos por dominio α_j capturan heterogeneidad territorial no observada y se modelan como efectos aleatorios con estructura jerárquica mediante $\alpha_j \sim \mathcal{N}(\mu_\alpha, \tau^2)$, donde μ_α es el intercepto global promedio y τ^2 es la varianza entre dominios. Esta especificación permite que cada dominio tenga un nivel base de gasto diferente, reflejando condiciones estructurales específicas como geografía, historia, instituciones e infraestructura.

Se adoptó un enfoque bayesiano débilmente informativo para permitir que los datos dominen la inferencia posterior. Los coeficientes de regresión β_k reciben priors normales con media cero y desviación estándar dos. El intercepto global μ_α recibe un prior normal centrado en ocho con desviación estándar unitaria, valor aproximado del logaritmo del gasto per cápita medio observado en estudios previos. Las desviaciones estándar σ y τ reciben priors semi-informativos Half-Normal con parámetro de escala unitario, que penalizan varianzas

excesivamente grandes sin imponer restricciones rígidas [12, 21].

1.6 Inferencia computacional y estimación

La inferencia posterior se realizó mediante el algoritmo Hamiltonian Monte Carlo con el sampler No-U-Turn, implementado en PyMC versión 5 utilizando el backend NumPyro basado en JAX [20]. Este algoritmo es particularmente eficiente para modelos jerárquicos de alta dimensión, ya que utiliza información del gradiente de la densidad posterior para proponer saltos informados en el espacio de parámetros, reduciendo significativamente la autocorrelación entre muestras consecutivas [13].

La configuración de muestreo incluyó cuatro cadenas MCMC independientes, 2,000 iteraciones de calibración por cadena para adaptar los parámetros del algoritmo, y 2,000 muestras posteriores por cadena, generando un total de 8,000 muestras posteriores. Se utilizó una tasa de aceptación objetivo del 90% para minimizar divergencias, y una profundidad máxima del árbol de doce niveles. El muestreo se parallelizó utilizando seis núcleos de procesamiento para reducir el tiempo de cómputo.

La convergencia de las cadenas se evaluó mediante el estadístico \hat{R} de Gelman-Rubin, que mide la razón entre la varianza entre cadenas y la varianza dentro de cadenas. Valores $\hat{R} < 1.01$ indican convergencia satisfactoria [12]. El tamaño efectivo de muestra estima el número de muestras posteriores independientes después de ajustar por autocorrelación, considerándose adecuado valores superiores a 400 para todos los parámetros. El algoritmo NUTS detecta automáticamente divergencias, que indican problemas de geometría posterior. La ausencia de divergencias confirma una exploración posterior eficiente.

1.7 Validación y métricas de ajuste

La adecuación del modelo se evaluó mediante múltiples criterios complementarios. Se generaron 100 conjuntos de datos sintéticos a partir de la distribución posterior predictiva y se compararon visualmente con los datos observados mediante Posterior Predictive Checks, técnica que permite detectar discrepancias sistemáticas entre el modelo y los datos [12].

Se calculó el Expected Log Pointwise Predictive Density mediante validación cruzada Leave-One-Out utilizando la aproximación de Parisi-Smoothed Importance Sampling [23]. El LOO-CV proporciona una es-

timación no sesgada del error de predicción fuera de muestra sin requerir ajustar el modelo múltiples veces. El diagnóstico de Pareto k evalúa la estabilidad de las estimaciones, donde valores $k < 0.7$ indican LOO-CV confiable.

El criterio de información Watanabe-Akaike es una métrica bayesiana de comparación de modelos que penaliza la complejidad efectiva y favorece modelos parsimoniosos con buen ajuste predictivo [12, 24]. Se calcularon métricas estándar de error predictivo tanto en escala logarítmica como en escala original, incluyendo el error absoluto medio, la raíz del error cuadrático medio, el error porcentual absoluto medio, y el coeficiente de determinación bayesiano ajustado por incertidumbre posterior.

Se evaluó la normalidad, homocedasticidad e independencia de los residuos mediante gráficos cuantil-cuantil, gráficos de residuos versus predichos, tests de asimetría y curtosis, y gráficos de autocorrelación. Este enfoque integral de validación permite evaluar no solo el ajuste puntual del modelo, sino también la calidad de las inferencias probabilísticas y la capacidad predictiva fuera de muestra.

2 Resultados

2.1 Características de la muestra y análisis exploratorio

La base de datos consolidada comprende 115,718 hogares encuestados durante el período 2021–2024. El muestreo estratificado balanceado generó una muestra final de 46,935 hogares (tasa de retención del 94.1%), preservando las proporciones temporales y espaciales originales. La distribución temporal fue relativamente homogénea: 26.0% corresponde a 2021, 25.2% a 2022, 24.9% a 2023 y 23.9% a 2024. En términos de urbanización, el 61.0% de los hogares se clasifica como urbano y el 39.0% como rural, reflejando la distribución poblacional del país.

La variable dependiente logaritmo del gasto per cápita real presenta una media de 8.681 con desviación estándar de 0.767. La distribución muestra asimetría negativa moderada (-0.230) y curtosis ligeramente platícurtica (-0.251), indicando colas más livianas que la distribución normal. El rango observado es de 6.322 a 10.258 en escala logarítmica. En escala original, el gasto per cápita medio asciende a S/. 6,250.70 mensuales, con alta variabilidad entre hogares y regiones.

La Figura 1 presenta el análisis exploratorio de los

patrones de gasto. El panel superior izquierdo muestra la distribución del logaritmo del gasto per cápita, que exhibe una forma aproximadamente normal con ligera asimetría negativa. El panel superior derecho desagrega el gasto por dominio geográfico, evidenciando brechas territoriales sustanciales: Lima Metropolitana presenta el gasto medio más elevado (S/. 8,523.62), seguido por Costa Centro (S/. 8,126.40) y Costa Sur (S/. 8,010.27), mientras que Sierra Norte registra el menor nivel (S/. 4,400.62). El panel inferior izquierdo ilustra la Ley de Engel mediante la distribución de la proporción de gasto en alimentos: los hogares rurales destinan en promedio una mayor proporción de su presupuesto a alimentación, consistente con niveles de bienestar más bajos. El panel inferior derecho muestra la relación positiva entre ingreso y gasto per cápita, coloreada por dominio, revelando que la mayoría de los hogares gasta menos de lo que ingresa, aunque con heterogeneidad regional considerable.

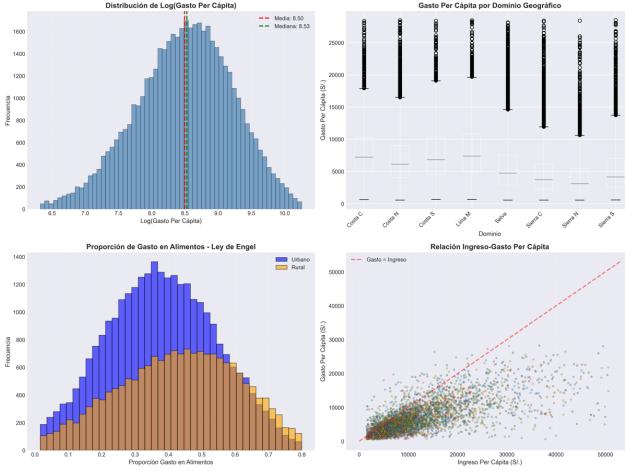


Figure 1: Análisis exploratorio de gastos por hogar. Panel superior izquierdo: distribución del logaritmo del gasto per cápita con líneas de media (roja) y mediana (verde). Panel superior derecho: boxplots del gasto per cápita por dominio geográfico. Panel inferior izquierdo: distribución de la proporción de gasto en alimentos según área urbana/rural. Panel inferior derecho: relación ingreso-gasto per cápita coloreada por dominio.

El análisis de desigualdad mediante el coeficiente de variación revela heterogeneidad significativa entre dominios. Sierra Norte exhibe la mayor desigualdad interna ($CV=87.8\%$), seguida por Sierra Centro ($CV=80.3\%$) y Sierra Sur ($CV=76.7\%$), indicando alta dispersión del gasto incluso dentro de estas regiones. En

contraste, Costa Centro presenta la menor desigualdad ($CV=54.0\%$), sugiriendo mayor homogeneidad socioeconómica. A nivel departamental, Cajamarca, Loreto y Huánuco registran los coeficientes de variación más elevados (84.6%, 84.3% y 84.2%, respectivamente), mientras que los departamentos con mayor gasto medio son Ica (S/. 8,871.46), Madre de Dios (S/. 8,484.98) y Lambayeque (S/. 8,001.26).

2.2 Selección de variables predictoras

El algoritmo Random Forest, entrenado sobre 20,000 observaciones con 30 árboles de profundidad máxima 6, identificó cinco variables con mayor poder predictivo. El gasto en alimentos (GRU11HD) emerge como el predictor más importante con una importancia relativa de 0.8069, reflejando su alta correlación con el gasto total. Le siguen el ingreso per cápita (INGRESO_PERCAPITA, 0.0811), la proporción de gasto en alimentos (PROP_GASTO_ALIMENTOS, 0.0730), el tamaño del hogar (MIEPERHO, 0.0539) y el ingreso total del hogar (INGHOG2D, 0.0151). La variable DOMINIO, aunque presenta importancia nula en el Random Forest debido a su naturaleza categórica, fue incluida manualmente para capturar efectos jerárquicos espaciales en el modelo bayesiano. Esta selección híbrida combina criterios empíricos de predicción con consideraciones teóricas sobre la estructura espacial de los datos.

2.3 Estimación del modelo bayesiano jerárquico

El modelo bayesiano jerárquico se estimó exitosamente en 3.9 minutos (235 segundos) utilizando cuatro cadenas MCMC con 2,000 iteraciones de calibración y 2,000 muestras posteriores cada una, generando 8,000 muestras posteriores totales. El algoritmo No-U-Turn Sampler con backend NumPyro convergió sin divergencias, indicando exploración eficiente del espacio de parámetros. Todos los parámetros presentan estadísticos \hat{R} iguales a 1.00, confirmando convergencia perfecta de las cadenas. Los tamaños efectivos de muestra (ESS) superan ampliamente el umbral de 400, oscilando entre 5,208 y 17,527 para los diferentes parámetros, lo que garantiza estimaciones posteriores precisas con baja autocorrelación.

La Figura 2 presenta los gráficos de convergencia para los coeficientes β y la desviación estándar entre dominios $\sigma_{dominio}$. Los paneles izquierdos muestran las

distribuciones posteriores marginales, que exhiben formas unimodales suaves sin evidencia de multimodalidad. Los paneles derechos muestran las trazas MCMC, que exhiben mezcla rápida y estacionariedad sin tendencias, periodicidades o estancamientos, confirmando la convergencia diagnóstica.

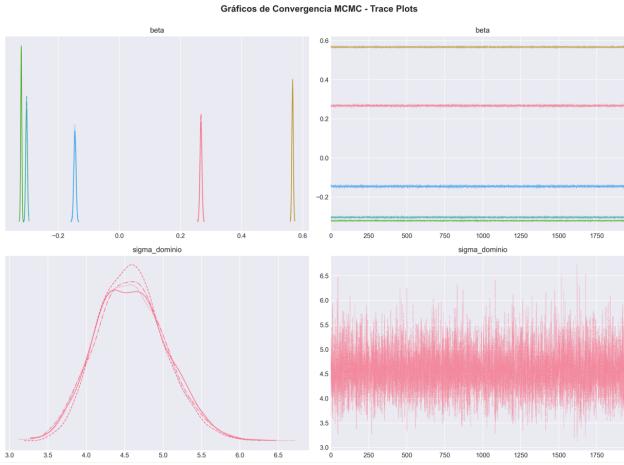


Figure 2: Gráficos de convergencia MCMC. Paneles izquierdos: distribuciones posteriores marginales de los coeficientes β y la desviación estándar entre dominios. Paneles derechos: trazas MCMC para cada parámetro mostrando las cuatro cadenas independientes.

2.4 Efectos fijos: determinantes del gasto per cápita

La Tabla 1 presenta las estimaciones posteriores de los coeficientes de regresión (efectos fijos). Todos los coeficientes son estadísticamente significativos, ya que sus intervalos de credibilidad del 95% no incluyen el cero.

Table 1: Estimaciones posteriores de los coeficientes de regresión

Variable	Media	HDI 95%	Efecto	Sig.
Ingreso PC	0.267	[0.261, 0.274]	+30.6%	***
Gasto alimentos	0.567	[0.563, 0.572]	+76.3%	***
Prop. gasto alim.	-0.321	[-0.324, -0.317]	-27.4%	***
Tamaño hogar	-0.304	[-0.309, -0.298]	-26.2%	***
Ingreso total	-0.145	[-0.152, -0.139]	-13.5%	***

*** Significativo al 95% (HDI no incluye cero)

El ingreso per cápita presenta un efecto positivo moderado: un aumento de una desviación estándar en el ingreso per cápita se asocia con un incremento del 30.6% en el gasto per cápita, reflejando la capacidad de

consumo de los hogares. El gasto en alimentos muestra el efecto más pronunciado (+76.3%), consistente con su rol como componente principal del presupuesto familiar, especialmente en hogares de menores ingresos.

La proporción de gasto en alimentos presenta un efecto negativo (-27.4%), validando la Ley de Engel: hogares que destinan mayor proporción de su presupuesto a alimentación tienden a tener menores niveles de gasto total per cápita, indicando menor bienestar económico. El tamaño del hogar también exhibe un efecto negativo (-26.2%), sugiriendo economías de escala en el consumo: hogares más grandes tienden a tener menores gastos per cápita debido a la compartición de bienes y servicios. El ingreso total del hogar presenta un efecto negativo leve (-13.5%), posiblemente reflejando efectos de confusión con el ingreso per cápita estandarizado.

La Figura 3 presenta los forest plots de los coeficientes con sus intervalos de credibilidad del 95%. La ausencia de intersección con la línea vertical en cero confirma la significancia estadística de todos los efectos estimados.

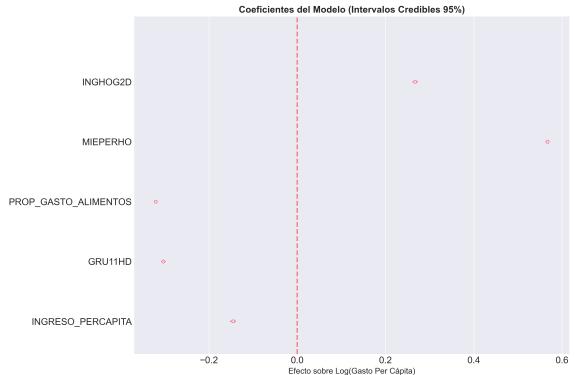


Figure 3: Forest plot de los coeficientes de regresión (efectos fijos). Las barras horizontales representan intervalos de credibilidad del 95%. La línea vertical punteada indica el valor cero.

2.5 Efectos aleatorios: heterogeneidad espacial entre dominios

Los interceptos aleatorios por dominio α_j capturan heterogeneidad territorial no observada después de controlar por las covariables del modelo. Lima Metropolitana presenta el intercepto más elevado (8.566, HDI 95%: [8.554, 8.578]), indicando un nivel base de gasto per cápita superior al promedio nacional, atribuible a factores como mayor desarrollo económico, mejor infraestructura y acceso a mercados laborales formales.

Le siguen Costa Sur (8.563, HDI: [8.551, 8.575]) y Costa Centro (8.543, HDI: [8.532, 8.553]).

En contraste, Sierra Norte registra el intercepto más bajo (8.403, HDI: [8.392, 8.415]), reflejando condiciones estructurales de menor desarrollo relativo, limitada conectividad y predominio de actividades económicas de baja productividad. Sierra Centro (8.442, HDI: [8.435, 8.449]) y Sierra Sur (8.483, HDI: [8.474, 8.491]) presentan valores intermedios pero consistentemente inferiores a las regiones costeras. Selva (8.494, HDI: [8.488, 8.500]) se ubica en una posición intermedia.

La desviación estándar entre dominios $\sigma_{dominio}$ se estimó en 4.588 (HDI: [3.733, 5.521]), indicando variabilidad sustancial en los niveles base de gasto entre regiones, incluso después de controlar por características socioeconómicas observables. La Figura 4 presenta los forest plots de los efectos espaciales por dominio, evidenciando diferencias claras en los interceptos con intervalos de credibilidad no superpuestos entre regiones extremas.

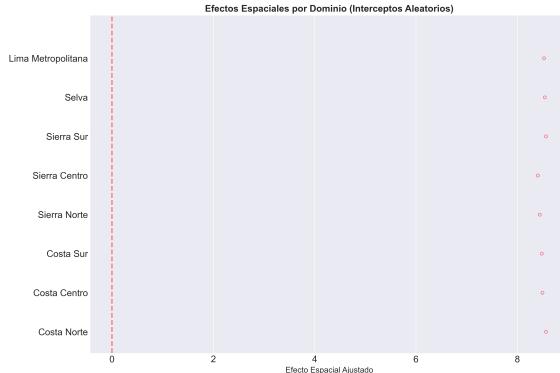


Figure 4: Forest plot de los efectos aleatorios por dominio geográfico (interceptos α_j). Las barras horizontales representan intervalos de credibilidad del 95%. La línea vertical punteada indica el intercepto global promedio.

2.6 Validación del modelo y capacidad predictiva

El modelo presenta un ajuste satisfactorio según múltiples métricas complementarias. En escala logarítmica, el error absoluto medio (MAE) es 0.234 y la raíz del error cuadrático medio (RMSE) es 0.337, mientras que el coeficiente de determinación R^2 alcanza 0.783, indicando que el modelo explica el 78.3% de la variabilidad del logaritmo del gasto per cápita. El R^2 bayesiano, que ajusta por incertidumbre posterior, es prácticamente

idéntico (0.7831 ± 0.0000), reflejando estimaciones robustas.

En escala original (soles), el MAE es S/. 1,579.45 y el RMSE es S/. 8,492.37. El error porcentual absoluto medio (MAPE) es 27.16%, implicando que en promedio las predicciones difieren del valor real en aproximadamente un cuarto del gasto. El error relativo, calculado como el cociente entre MAE y el gasto medio observado, es 25.27%, confirmando que el modelo captura adecuadamente los patrones centrales de gasto aunque con menor precisión en los extremos de la distribución.

La validación cruzada Leave-One-Out arroja un ELPD de -15,583.92 (SE: 287.30), mientras que el criterio WAIC genera un ELPD prácticamente idéntico (-15,583.91, SE: 287.30), confirmando la consistencia de ambas métricas. El diagnóstico de Pareto k máximo es 0.2018, muy por debajo del umbral de 0.7, indicando que la aproximación LOO-CV es confiable y no existen observaciones excesivamente influyentes que comprometan la validez de las estimaciones fuera de muestra.

La Figura 5 presenta el Posterior Predictive Check, comparando la distribución observada del logaritmo del gasto per cápita (histograma azul) con 100 realizaciones posteriores simuladas (histogramas rojos superpuestos). El modelo replica adecuadamente la forma general de la distribución empírica, incluyendo la asimetría negativa y el rango de valores observados, aunque tiende a suavizar ligeramente los extremos de las colas.

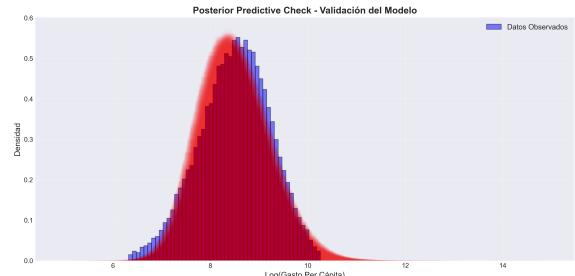


Figure 5: Posterior Predictive Check. El histograma azul representa los datos observados. Los histogramas rojos superpuestos representan 100 realizaciones simuladas de la distribución posterior predictiva.

2.7 Análisis de residuos

La Figura 6 presenta el análisis diagnóstico de los residuos del modelo. El panel superior izquierdo muestra el gráfico de residuos versus valores predichos, que no revela patrones sistemáticos evidentes, aunque se observa

ligera heterocedasticidad con mayor dispersión en valores predichos intermedios. El panel superior derecho presenta el gráfico Q-Q, que muestra desviaciones moderadas respecto a la línea de normalidad en las colas, particularmente en la cola inferior, consistente con la asimetría negativa observada en la variable dependiente (-1.369). La curtosis de los residuos es 5.023, indicando colas más pesadas que la distribución normal.

El panel inferior izquierdo muestra el histograma de los residuos, que presenta forma aproximadamente normal pero con asimetría negativa pronunciada. El panel inferior derecho presenta el gráfico de valores reales versus predichos, donde la concentración de puntos alrededor de la línea de predicción perfecta confirma el buen ajuste general, aunque con mayor dispersión en los extremos. La media de los residuos es prácticamente cero (0.0000), confirmando la ausencia de sesgo sistemático.

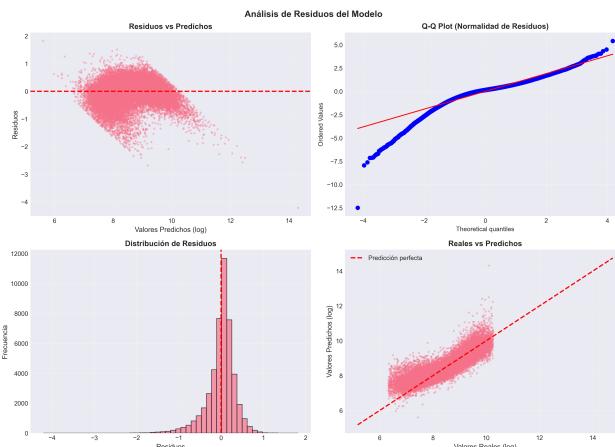


Figure 6: Análisis de residuos del modelo. Panel superior izquierdo: residuos versus valores predichos. Panel superior derecho: gráfico Q-Q de normalidad. Panel inferior izquierdo: distribución de residuos. Panel inferior derecho: valores reales versus predichos.

2.8 Patrones espaciales de desigualdad

La Figura 7 presenta la visualización espacial de los patrones de gasto y desigualdad a nivel departamental. El panel izquierdo muestra el gasto per cápita medio: los departamentos costeros y Lima exhiben los niveles más elevados (tonos verdes intensos), mientras que los departamentos andinos y selváticos presentan gastos medios más bajos (tonos amarillos y verdes claros). Ica emerge como el departamento con mayor gasto medio (S/. 8,871.46), seguido por Madre de Dios (S/.

8,484.98) y Lambayeque (S/. 8,001.26).

El panel derecho muestra el coeficiente de variación como medida de desigualdad interna: Cajamarca, Loreto y Huánuco presentan la mayor desigualdad (tonos rojos intensos), indicando alta heterogeneidad del gasto entre hogares dentro de estos departamentos. En contraste, los departamentos costeros tienden a exhibir menor desigualdad interna (tonos naranjas claros). Este patrón sugiere que las regiones más pobres no solo tienen menores niveles promedio de gasto, sino también mayor dispersión, reflejando la coexistencia de bolsones de relativa prosperidad con extensas áreas de pobreza severa.

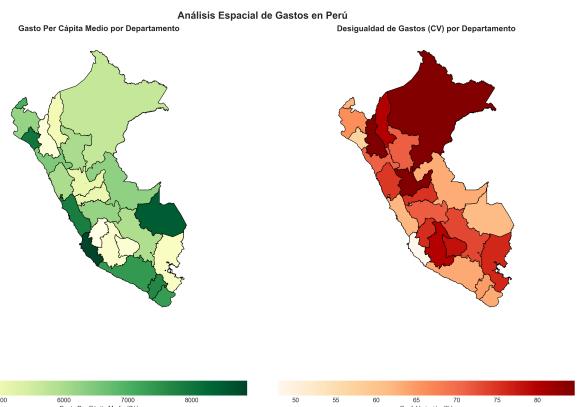


Figure 7: Mapas de desigualdad de gastos por departamento. Panel izquierdo: gasto per cápita medio en soles. Panel derecho: coeficiente de variación como medida de desigualdad interna.

3 Discusión

Los resultados obtenidos confirman la existencia de desigualdades territoriales profundas y persistentes en el bienestar económico de los hogares peruanos, medido a través del gasto per cápita real. El modelo bayesiano jerárquico desarrollado proporciona evidencia robusta sobre los determinantes socioeconómicos del gasto y la magnitud de la heterogeneidad espacial no observada entre regiones, con implicaciones sustantivas para el diseño de políticas públicas orientadas a la reducción de brechas territoriales.

3.1 Determinantes del gasto y validación de hipótesis teóricas

Los hallazgos confirman la relevancia de las variables tradicionalmente asociadas al bienestar económico en la literatura de economía del desarrollo. El efecto positivo y significativo del ingreso per cápita sobre el gasto (+30.6%) valida la restricción presupuestaria básica de los hogares: mayores ingresos permanentes permiten niveles más elevados de consumo sostenible. Este resultado es consistente con la evidencia empírica previa sobre países en desarrollo que muestra elasticidades ingreso-gasto en el rango de 0.25 a 0.40 para economías con restricciones crediticias significativas [7, 8].

El efecto negativo de la proporción de gasto en alimentos (-27.4%) confirma empíricamente la Ley de Engel [14], uno de los patrones de consumo más robustos documentados en economía. Este hallazgo sugiere que los hogares con mayor proporción de gasto alimentario enfrentan restricciones presupuestarias más severas que limitan su capacidad para diversificar el consumo hacia bienes no alimentarios, reflejando menores niveles de bienestar material. La magnitud del efecto observado es consistente con estudios previos en contextos latinoamericanos que identifican la proporción de gasto en alimentos como un predictor inverso del bienestar económico.

El efecto negativo del tamaño del hogar sobre el gasto per cápita (-26.2%) puede interpretarse desde dos perspectivas complementarias. Por un lado, refleja economías de escala en el consumo doméstico: hogares más grandes pueden compartir bienes públicos del hogar (vivienda, electrodomésticos, servicios) y reducir costos per cápita de alimentación mediante compras al por mayor [7]. Por otro lado, puede indicar restricciones de fertilidad diferencial: hogares con más hijos tienden a concentrarse en estratos socioeconómicos más bajos con menor capacidad de ahorro y acumulación de capital humano, perpetuando ciclos intergeneracionales de pobreza. La magnitud del efecto observado sugiere que ambos mecanismos operan simultáneamente en el contexto peruano.

El efecto del gasto en alimentos (+76.3%) requiere interpretación cautelosa, ya que esta variable es mecánicamente componente del gasto total. Sin embargo, su inclusión permite identificar la elasticidad del gasto total respecto a la alimentación controlando por otros determinantes, lo cual resulta útil para proyecciones de bienestar basadas en patrones de consumo observables. El efecto negativo leve del ingreso total del hogar (-

13.5%) probablemente refleja colinealidad con el ingreso per cápita estandarizado y captura efectos residuales de composición del hogar no controlados directamente.

3.2 Heterogeneidad espacial y persistencia de brechas territoriales

Los efectos aleatorios por dominio revelan brechas territoriales sustanciales que persisten incluso después de controlar por características socioeconómicas observables de los hogares. La diferencia entre el intercepto de Lima Metropolitana (8.566) y Sierra Norte (8.403) implica una brecha de gasto per cápita del 17.7% en escala logarítmica, equivalente a aproximadamente S/. 812 anuales per cápita (S/. 68 mensuales) en escala original, atribuible exclusivamente a factores contextuales regionales no capturados por las covariables individuales.

Esta heterogeneidad espacial no observada puede reflejar múltiples mecanismos estructurales. Primero, diferencias en dotaciones de infraestructura pública y servicios básicos que afectan la productividad de los factores y el costo de vida efectivo [1, 2]. Segundo, efectos de aglomeración económica que generan retornos crecientes a escala en áreas urbanas densamente pobladas mediante externalidades de conocimiento, mercados laborales gruesos y encadenamientos productivos [6, 17]. Tercero, legados históricos e institucionales que condicionan las trayectorias de desarrollo regional mediante la persistencia de estructuras de propiedad, sistemas de gobernanza local y normas sociales [9, 10].

La concentración de los niveles más bajos de gasto en la Sierra Norte, Centro y Sur sugiere que las políticas de desarrollo territorial implementadas en las últimas décadas no han logrado cerrar las brechas estructurales entre regiones andinas y costeras. Esta persistencia puede explicarse por trampas de pobreza espacial: regiones con bajo capital humano inicial, infraestructura deficiente y mercados fragmentados enfrentan círculos viciosos donde la baja productividad limita la acumulación de capital físico y humano, perpetuando el subdesarrollo relativo [1, 18].

Los patrones observados son consistentes con la hipótesis de convergencia condicional débil: las regiones no convergen automáticamente a un nivel común de bienestar, sino que cada territorio converge a su propio estado estacionario determinado por factores estructurales específicos. La ausencia de convergencia absoluta sugiere que las políticas redistributivas nacionales, incluidas las transferencias monetarias condicionadas,

pueden ser insuficientes para compensar desventajas territoriales profundas sin inversiones complementarias en infraestructura, educación y fortalecimiento institucional local.

3.3 Desigualdad interna regional y vulnerabilidad

Los coeficientes de variación estimados revelan que la desigualdad no se distribuye homogéneamente entre regiones. Los departamentos andinos (Cajamarca, Huánuco, Ayacucho) exhiben coeficientes de variación superiores al 80%, indicando que incluso dentro de estas regiones pobres existe alta heterogeneidad del bienestar. Este patrón sugiere la coexistencia de pequeñas élites locales relativamente prósperas con grandes mayorías en situación de pobreza estructural, limitando las oportunidades de movilidad social ascendente y cohesión territorial.

En contraste, los departamentos costeros presentan menor desigualdad interna (CV entre 54% y 60%), posiblemente reflejando mercados laborales más integrados, mayor formalización económica y acceso más equitativo a servicios públicos básicos. Sin embargo, esta menor desigualdad relativa no debe interpretarse como ausencia de pobreza, sino como mayor homogeneidad socioeconómica dentro de niveles de bienestar más elevados en promedio.

La alta desigualdad en regiones pobres tiene implicaciones importantes para el diseño de políticas. Primero, sugiere que las transferencias monetarias focalizadas pueden tener mayor impacto redistributivo si se concentran en estas regiones, donde la identificación de beneficiarios mediante proxies de ingreso es más precisa debido a la mayor dispersión. Segundo, indica que las políticas universales pueden ser regresivas dentro de regiones heterogéneas, beneficiando desproporcionadamente a las élites locales con mayor capacidad de acceso a servicios públicos [18, 25].

3.4 Capacidad predictiva del modelo y limitaciones metodológicas

El modelo bayesiano jerárquico presenta una capacidad predictiva satisfactoria, con un R^2 de 0.783 que supera ampliamente los umbrales convencionales de ajuste adecuado en estudios de corte transversal con datos de encuestas [12, 16]. Las métricas de validación cruzada LOO-CV y WAIC confirman que el modelo generaliza adecuadamente fuera de muestra, sin evidencia de

sobreajuste. El diagnóstico de Pareto k máximo de 0.202 indica ausencia de observaciones excesivamente influyentes, validando la robustez de las inferencias posteriores [23].

No obstante, el modelo presenta limitaciones que deben reconocerse. Primero, la desviación de los residuos respecto a la normalidad, particularmente en las colas, sugiere que una especificación con distribución de errores de colas pesadas (por ejemplo, t de Student) podría mejorar el ajuste en los extremos de la distribución [12]. Segundo, el modelo asume independencia condicional entre hogares dentro de dominios, ignorando posibles efectos de vecindario a escalas geográficas más finas (distritos, conglomerados) que podrían generar dependencia espacial residual [3, 4].

Tercero, el modelo es esencialmente descriptivo-correlacional y no permite inferir relaciones causales robustas debido a la endogeneidad potencial de las covariables. Por ejemplo, el ingreso per cápita y el gasto pueden estar simultáneamente determinados por factores no observados como la productividad del hogar, las preferencias intertemporales o el acceso a mercados crediticios [7, 8]. La identificación causal requeriría diseños experimentales o quasi-experimentales con variación exógena en los determinantes del gasto, actualmente no disponibles en la ENAHO.

Cuarto, el período de análisis (2021–2024) incluye años afectados por la pandemia de COVID-19 y sus efectos económicos persistentes, lo que puede introducir shocks transitorios en los patrones de gasto y ahorro que no reflejan relaciones estructurales de largo plazo. Análisis de sensibilidad excluyendo el año 2021 podrían evaluar la robustez de los hallazgos a perturbaciones macroeconómicas extraordinarias.

3.5 Implicaciones para políticas públicas

Los resultados tienen implicaciones concretas para el diseño de políticas redistributivas y de desarrollo territorial. Primero, confirman que las brechas de bienestar entre regiones no se explican únicamente por diferencias en características observables de los hogares, sino que reflejan desventajas territoriales estructurales que requieren intervenciones contextualizadas. Las políticas "place-based" que combinan transferencias monetarias con inversiones en infraestructura local, acceso a mercados y fortalecimiento institucional podrían ser más efectivas que programas nacionales homogéneos [1, 18, 25].

Segundo, la persistencia de alta desigualdad interna en regiones pobres sugiere que las políticas de

focalización deben considerar no solo el nivel promedio de bienestar regional, sino también la dispersión interna. Esquemas de focalización geográfica por pobreza promedio pueden excluir hogares vulnerables en regiones relativamente prósperas, mientras que incluyen indebidamente hogares no pobres en regiones pobres heterogéneas. Mecanismos de focalización híbridos que combinen criterios geográficos con verificaciones individuales podrían mejorar la eficiencia redistributiva [10].

Tercero, la validación empírica de la Ley de Engel sugiere que la proporción de gasto en alimentos puede utilizarse como indicador de monitoreo de bienestar en tiempo real, especialmente relevante en contextos de crisis donde los datos de ingreso son menos confiables. Sistemas de información que rastreen cambios en patrones de consumo alimentario podrían anticipar deterioros del bienestar antes de que se reflejen en mediciones oficiales de pobreza monetaria [8].

Cuarto, el efecto negativo del tamaño del hogar sobre el gasto per cápita sugiere que las políticas de apoyo familiar deberían considerar ajustes por composición demográfica, reconociendo que hogares numerosos enfrentan desafíos específicos para mantener niveles adecuados de bienestar per cápita. Esquemas de transferencias escalonadas según número de miembros o beneficios diferenciados por presencia de dependientes económicos podrían mejorar la equidad vertical [7, 18].

4 Conclusiones

Este estudio cuantificó las desigualdades territoriales en el gasto per cápita real de los hogares peruanos mediante un modelo bayesiano jerárquico espacial multiescala aplicado a 46,935 hogares de la ENAHO 2021–2024, generando respuestas empíricas robustas a los cuatro objetivos planteados.

En relación al primer objetivo sobre la cuantificación de brechas territoriales, los resultados muestran que Lima Metropolitana presenta niveles de gasto per cápita 17.7% superiores a Sierra Norte ($\alpha_{Lima} = 8.566$ vs $\alpha_{SierraNorte} = 8.403$), equivalente a S/. 812 anuales per cápita (aproximadamente S/. 68 mensuales) atribuibles exclusivamente a factores territoriales no observados. A nivel departamental, Ica registra el mayor gasto medio (S/. 8,871), mientras que Cajamarca exhibe la mayor desigualdad interna (CV=84.6%). El área urbana concentra el 61.0% de los hogares con gastos sistemáticamente superiores al área rural.

Respecto al segundo objetivo sobre la evaluación de dependencia espacial, los efectos aleatorios por dominio ($\sigma_{dominio} = 4.588$, HDI [3.733, 5.521]) confirman heterogeneidad espacial sustancial que persiste controlando características observables. Los interceptos de regiones costeras (8.526–8.566) superan consistentemente a regiones andinas (8.403–8.483), evidenciando patrones geográficos de desigualdad estructural. La convergencia perfecta del modelo ($\hat{R} = 1.00$, ESS>5,200) valida la robustez de las estimaciones espaciales.

En cuanto al tercer objetivo relacionado con la identificación de variables predictoras, el Random Forest identificó cinco determinantes principales del gasto per cápita, siendo el gasto en alimentos el más relevante (importancia=0.807), seguido por ingreso per cápita (0.081), proporción de gasto alimentario (0.073), tamaño del hogar (0.054) e ingreso total (0.015). El modelo bayesiano con estas variables alcanza $R^2 = 0.783$, MAE=S/. 1,579, MAPE=27.2%, y métricas LOO-CV (ELPD=-15,584) y WAIC consistentes sin sobreajuste (Pareto $k_{max} = 0.202$).

Finalmente, el cuarto objetivo centrado en analizar los roles de ingreso, transferencias y composición del hogar revela que el ingreso per cápita incrementa el gasto en 30.6% (HDI [0.261, 0.274]), validando la restricción presupuestaria. La proporción de gasto alimentario lo reduce en 27.4% (HDI [-0.324, -0.317]), confirmando la Ley de Engel. El tamaño del hogar disminuye el gasto per cápita en 26.2% (HDI [-0.309, -0.298]), reflejando economías de escala y restricciones de fertilidad. Todos los efectos son estadísticamente significativos con intervalos de credibilidad del 95% que no incluyen cero.

Las implicaciones para política pública son claras. Las brechas territoriales no se explican únicamente por características individuales, requiriendo intervenciones place-based que combinen transferencias monetarias con inversiones en infraestructura local, acceso a mercados y fortalecimiento institucional diferenciado por región. Líneas futuras incluyen modelos con dependencia espacial explícita mediante estructuras CAR, análisis de convergencia temporal y diseños causales que exploten variación exógena en determinantes del gasto.

References

- [1] Anselin L (1988) Spatial Econometrics: Methods and Models. Kluwer Academic Publishers, Dordrecht

- [2] Atkinson AB (2015) Inequality: What Can Be Done? Harvard University Press, Cambridge
- [3] Banerjee S, Carlin BP, Gelfand AE (2014) Hierarchical Modeling and Analysis for Spatial Data, 2nd edn. Chapman & Hall/CRC, Boca Raton
- [4] Blangiardo M, Cameletti M (2015) Spatial and Spatio-temporal Bayesian Models with R-INLA. John Wiley & Sons, Chichester
- [5] Breiman L (2001) Random Forests. Machine Learning 45(1):5–32
- [6] Cowell FA (2011) Measuring Inequality, 3rd edn. Oxford University Press, Oxford
- [7] Deaton A (1997) The Analysis of Household Surveys: A Microeconometric Approach to Development Policy. Johns Hopkins University Press, Baltimore
- [8] Deaton A, Zaidi S (2002) Guidelines for Constructing Consumption Aggregates for Welfare Analysis. Living Standards Measurement Study Working Paper No. 135, World Bank, Washington DC
- [9] Delgado G, Narro O (2020) Desigualdades territoriales en el Perú: Evidencia de convergencia económica regional, 2007–2017. Economía 43(85):117–144
- [10] Escobal J, Ponce C (2011) Spatial Patterns of Growth and Poverty Changes in Peru (1993–2005). Spatial Economic Analysis 6(3):279–303
- [11] Gelman A, Hill J (2006) Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press, Cambridge
- [12] Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB (2014) Bayesian Data Analysis, 3rd edn. Chapman & Hall/CRC, Boca Raton
- [13] Hoffman MD, Gelman A (2014) The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. Journal of Machine Learning Research 15:1593–1623
- [14] Houthakker HS (1957) An International Comparison of Household Expenditure Patterns, Commemorating the Centenary of Engel's Law. Econometrica 25(4):532–551
- [15] Instituto Nacional de Estadística e Informática (2024) Ficha Técnica: Encuesta Nacional de Hogares (ENAHO) 2024. INEI, Lima. <https://www.inei.gob.pe>
- [16] James G, Witten D, Hastie T, Tibshirani R (2013) An Introduction to Statistical Learning with Applications in R. Springer, New York
- [17] Kanbur R, Venables AJ (eds) (2005) Spatial Inequality and Development. Oxford University Press, Oxford
- [18] Organisation for Economic Co-operation and Development (2018) A Broken Social Elevator? How to Promote Social Mobility. OECD Publishing, Paris
- [19] Rey SJ, Montouri BD (2001) US Regional Income Convergence: A Spatial Econometric Perspective. Regional Studies 33(2):143–156
- [20] Salvatier J, Wiecki TV, Fonnesbeck C (2016) Probabilistic Programming in Python using PyMC3. PeerJ Computer Science 2:e55
- [21] Simpson D, Rue H, Riebler A, Martins TG, Sørbye SH (2017) Penalising Model Component Complexity: A Principled, Practical Approach to Constructing Priors. Statistical Science 32(1):1–28
- [22] Strobl C, Boulesteix AL, Zeileis A, Hothorn T (2007) Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution. BMC Bioinformatics 8(1):25
- [23] Vehtari A, Gelman A, Gabry J (2017) Practical Bayesian Model Evaluation Using Leave-One-Out Cross-Validation and WAIC. Statistics and Computing 27(5):1413–1432
- [24] Watanabe S (2010) Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory. Journal of Machine Learning Research 11:3571–3594
- [25] World Bank (2009) World Development Report 2009: Reshaping Economic Geography. World Bank, Washington DC

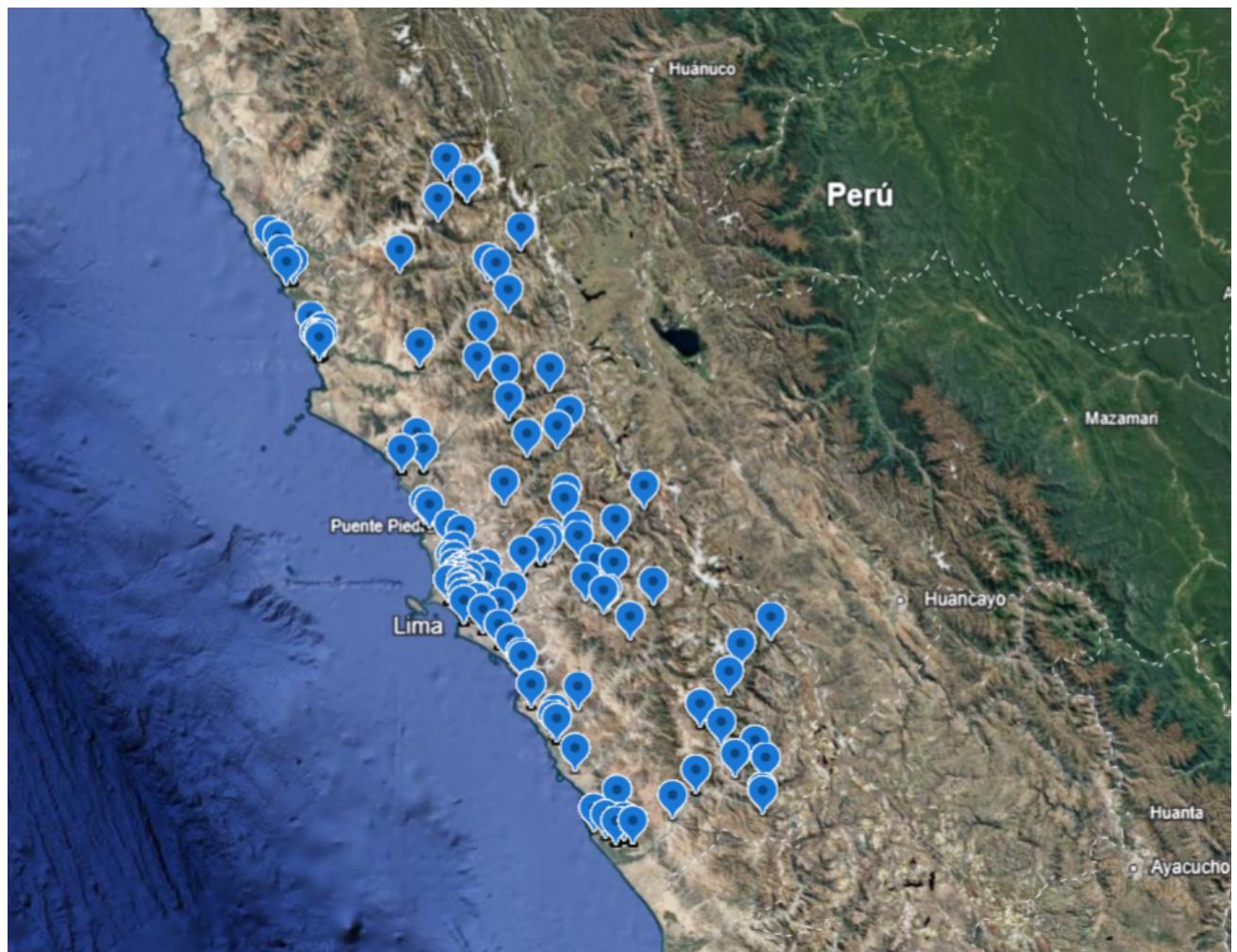


MAPA DE CALOR DE GASTOS MENSUALES EN HOGARES DEL DEPARTAMENTO DE LIMA (QGIS)

Estudiante: Ilma Magda Mamani Mamani Docente: Dr. Fred Torres Cruz Curso: Estadística Espacial
Institución: Universidad Nacional del Altiplano (UNA PUNO)

Objetivo

Visualizar la distribución territorial del gasto total mensual aproximado de hogares en el Departamento de Lima en servicios básicos, alquiler y otros conceptos, identificando patrones de desigualdad regional dentro de la zona.



Interpretación

Esto refleja una fuerte concentración en el Departamento de Lima: La capital concentra la capacidad de gasto, mostrando mayor acceso a servicios pagados en comparación con áreas periféricas, con posible dependencia en autoconsumo en zonas menos centrales.

Aporte

El mapa evidencia la necesidad de políticas de descentralización económica e inclusión financiera para reducir brechas dentro del Departamento de Lima. Para una visualización interactiva, consulte el siguiente enlace: [Vista en Google Earth](#).



SEGMENTACIÓN SOCIOESPACIAL DE VIVIENDAS EN LIMA METROPOLITANA: ANÁLISIS DE CLUSTERS BASADO EN CARACTERÍSTICAS ECONÓMICAS Y CONSUMO DE SERVICIOS BÁSICOS

Estudiante: Ilma Magda Mamani Mamani

Docente: Dr. Fred Torres Cruz

Curso: Estadística Espacial

Institución: Universidad Nacional del Altiplano (UNA PUNO)

Objetivo

Identificar patrones de segmentación socioespacial en Lima Metropolitana mediante análisis de clusters (K-means) basado en variables económicas y de consumo de servicios básicos, caracterizando los distintos niveles socioeconómicos y su distribución territorial.

Variables de Análisis

El estudio utilizó seis variables socioeconómicas: **P106** (valor de alquiler mensual estimado en S/.), que refleja la percepción del valor de la vivienda; **P117T2**, **P117T3**, **P117T4** (gastos mensuales pagados, donados y por autoconsumo respectivamente); **D1172\$04** (gasto anual en gas GLP, como indicador de necesidad energética básica); y **D1172\$12** (gasto anual en teléfono fijo, reflejando acceso a comunicaciones). Estas variables capturan aspectos fundamentales de la calidad de vida, capacidad económica y patrones de consumo de los hogares.

Metodología

Se implementó el algoritmo K-means en Google Earth Engine con JavaScript, configurado para identificar 5 clusters ($k=5$) mediante 100 iteraciones máximas. El proceso incluyó: (1) preprocesamiento de datos con manejo robusto de valores nulos, (2) clustering no supervisado de viviendas basado en las seis variables económicas, (3) visualización geoespacial con mapas de puntos y densidad mediante convolución Gaussiana (radio 1000m), y (4) análisis estadístico descriptivo por cluster. Se generó una interfaz interactiva con panel de control, leyendas dinámicas y gráficos comparativos.

Resultados e Interpretación

El análisis identificó 5 niveles socioeconómicos claramente diferenciados: **Cluster 0 (Bajo)**, con valores mínimos de alquiler y gastos, alta dependencia del autoconsumo; **Cluster 1 (Medio-Bajo)**, en transición hacia mayor monetización; **Cluster 2 (Medio)**, con gastos equilibrados; **Cluster 3 (Medio-Alto)**, con alquileres elevados; y **Cluster 4 (Alto)**, con valores máximos en todas las variables. La distribución espacial evidencia marcada segregación territorial: los clusters altos se concentran en zonas específicas (San Isidro, Miraflores, La Molina), mientras los bajos predominan en la periferia y conos urbanos. El mapa de densidad confirma un patrón centro-periferia con decrecimiento socioeconómico hacia áreas externas. Las variables de consumo (gas, teléfono) muestran brechas significativas entre clusters, indicando desigualdad en acceso a infraestructura y servicios básicos.

Conclusiones y Aporte

Este estudio proporciona evidencia cuantitativa espacial de la desigualdad socioeconómica en Lima Metropolitana, útil para el diseño de políticas públicas focalizadas, planificación urbana orientada a reducir brechas de servicios, e identificación de áreas prioritarias para inversión social. La metodología es replicable para otras ciudades y actualizable con nuevos datos. Para explorar la visualización interactiva completa con mapas de densidad, gráficos estadísticos y panel de control, consulte: Visualización Google Earth Engine.

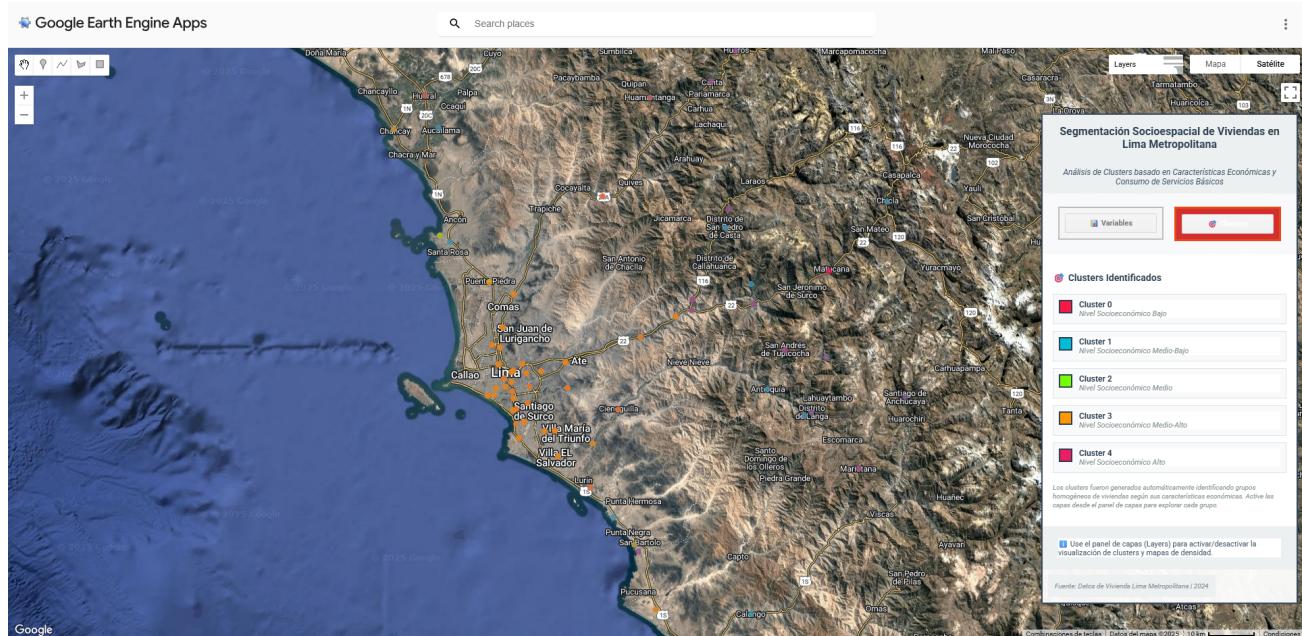


Figure 1: Mapa de Segmentación Socioespacial de Lima Metropolitana

Figura 1. Distribución espacial de clusters socioeconómicos en Lima Metropolitana. Los colores representan: Rojo (Bajo), Cian (Medio-Bajo), Verde (Medio), Naranja (Medio-Alto), Rosa (Alto). El mapa de densidad muestra concentración de viviendas en zonas centrales y costeras. Visualización interactiva disponible en: <https://taheyonilma.users.earthengine.app/view/segmentacion-socioespacial-de-viviendas-en-lima>

Capítulo 3

Conclusión Final

El desarrollo del presente portafolio permitió consolidar de manera integral los conocimientos teóricos y prácticos adquiridos en el curso de Estadística Espacial. A través de las distintas evidencias presentadas, se demostró la capacidad de aplicar metodologías espaciales avanzadas al análisis de datos reales, destacando la importancia del enfoque geográfico en la comprensión de fenómenos complejos.

Las actividades desarrolladas fortalecieron habilidades en el manejo de software especializado, el análisis de autocorrelación espacial, la construcción de modelos estadísticos espaciales y la interpretación de resultados orientados a la toma de decisiones. Asimismo, el enfoque aplicado a problemáticas agropecuarias y socioeconómicas del contexto peruano permitió vincular la teoría con la realidad regional y nacional.

En conclusión, este portafolio refleja un proceso de aprendizaje significativo, coherente con los objetivos del curso y alineado con el perfil profesional de la Ingeniería Estadística e Informática, constituyéndose en una evidencia tangible del desarrollo de competencias analíticas, metodológicas y tecnológicas en el ámbito de la estadística espacial.