Validación Cruzada Bloqueada y Jerárquica para Datos Agrícolas con Autocorrelación Espacial: Aplicación a la Encuesta Nacional Agropecuaria del Perú

Willy Vilca Apaza[1][0009-0006-5807-2359], Fred Torres Cruz[2][0000-0003-0834-6834], and Edgar Eloy Carpio Vargas[3][0000-0001-6457-4597]

Universidad Nacional del Altiplano, Escuela Profesional de Ingeniería Estadística e Informática, Puno, Perú
w.vilca@est.unap.edu.pe, ftorres@unap.edu.pe, ecarpio@unap.edu.pe

Resumen La autocorrelación espacial en datos agrícolas plantea un desafío metodológico crítico para la validación de modelos predictivos, ya que las técnicas convencionales asumen independencia entre observaciones. Este estudio implementa y evalúa validación cruzada bloqueada y jerárquica para datos de inventario ganadero de la Encuesta Nacional Agropecuaria 2017 del Perú (n=69,645), comparándolas con validación cruzada aleatoria tradicional. Los resultados muestran autocorrelación espacial débil pero estadísticamente significativa (I de Moran = 0.0031, p = 0.0401). La validación cruzada aleatoria sobreestima el desempeño predictivo en 9.1 % (RMSE) y 25.1 % (R^2) comparada con validación bloqueada, demostrando filtración espacial moderada. La validación cruzada jerárquica, implementada mediante leave-one-domain-out (bucle externo) y leave-one-department-out (bucle interno), proporciona estimaciones estadísticamente equivalentes pero es computacionalmente más costosa (aproximadamente 60× mayor tiempo de ejecución) que la validación bloqueada. El RMSE por dominio geográfico varía entre 0.884 y 1.210 (36.9 % de variación), revelando heterogeneidad espacial sustancial oculta en validación aleatoria. El efecto pepita dominante (77 % de varianza) indica que los factores de manejo individual superan a la ubicación geográfica. Los hallazgos validan la necesidad de métodos de validación espacialmente conscientes para datos agrícolas, incluso con autocorrelación débil, y demuestran que la validación cruzada bloqueada simple proporciona el balance óptimo entre rigor metodológico y eficiencia computacional.

Keywords: Validación cruzada espacial · Autocorrelación espacial · I de Moran · Validación cruzada jerárquica · Datos agrícolas · Filtración espacial · Random Forest · Encuestas agrícolas · Perú

1. Introducción

La autocorrelación espacial, formalizada por la Primera Ley de Geografía de Tobler [30], establece que los objetos cercanos tienden a ser más similares que los distantes. En sistemas de producción agrícola, esto se manifiesta a través de la difusión de tecnología entre productores vecinos, condiciones agroecológicas compartidas y acceso común a mercados [22]. Sin embargo, las técnicas convencionales para validar modelos predictivos asumen independencia entre observaciones, violando sistemáticamente esta estructura espacial inherente [26].

La validación cruzada k-fold aleatoria, ampliamente utilizada en aprendizaje automático, crea una alta probabilidad de que observaciones espacialmente cercanas aparezcan tanto en conjuntos de entrenamiento como de prueba [17]. Esta configuración permite a los modelos explotar la autocorrelación local para predicción, generando estimaciones de desempeño artificialmente optimistas conocidas como *filtración espacial* [25]. Estudios previos documentan que la validación aleatoria puede sobreestimar la precisión predictiva en 10–50 % en tareas de predicción espacial [16, 26].

La validación cruzada bloqueada aborda este problema particionando los datos en grupos espacialmente contiguos, asegurando separación geográfica entre conjuntos de entrenamiento y prueba [31]. Sin embargo, los enfoques de bloqueo estándar no consideran estructuras jerárquicas comunes en diseños de encuestas multi-escala [11]. La validación cruzada jerárquica extiende el bloqueo implementando procedimientos anidados de leave-one-group-out en múltiples escalas espaciales, asegurando que la optimización de hiperparámetros no explote la autocorrelación local [3, 6].

Perú exhibe heterogeneidad espacial extrema debido a gradientes topográficos dramáticos (0–6000 msnm), variación climática (desierto costero, valles interandinos, selva tropical) y acceso diferencial a mercados. La Encuesta Nacional Agropecuaria 2017 (ENA), con su estructura jerárquica (región-dominio-departamento-provincia-distrito), proporciona un caso de estudio ideal para evaluar metodologías de validación espacial.

1.1. Objetivos

Este estudio tiene tres objetivos: (1) cuantificar la autocorrelación espacial en inventario ganadero usando I de Moran global y análisis de indicadores locales; (2) comparar validación cruzada aleatoria versus bloqueada para estimar la magnitud de filtración espacial; (3) implementar validación cruzada jerárquica con optimización de hiperparámetros espacialmente consciente y evaluar su costo-beneficio versus validación bloqueada simple. Los resultados contribuyen orientación metodológica práctica para validar modelos predictivos en datos agrícolas espacialmente estructurados.

2. Materiales y Métodos

2.1. Fuente de Datos

La Encuesta Nacional Agropecuaria 2017 (ENA) del Instituto Nacional de Estadística e Informática (INEI) del Perú contiene 155,527 registros de productores agrícolas organizados jerárquicamente: 3 regiones naturales (Costa, Sierra, Selva), 7 dominios geográficos, 25 departamentos, provincias, distritos y conglomerados de muestreo. Después de remover valores faltantes y registros inconsistentes, el conjunto de datos final contiene 69,645 observaciones válidas.

Se seleccionaron tres variables: (1) P402A – inventario ganadero en el mes de referencia (variable dependiente continua, rango 0–4,500 cabezas), (2) DOMINIO – dominio geográfico con 7 categorías (Costa Norte, Costa Centro, Costa Sur, Sierra Norte, Sierra Centro, Sierra Sur, Selva), usado como variable de bloqueo espacial y predictor, y (3) P401A – tipo de especie animal con 24 categorías (ganado vacuno, ovino, porcino, camélidos, aves de corral, etc.), capturando heterogeneidad del sistema de producción. Esta selección parsimoniosa minimiza confusión espacial mientras maximiza interpretabilidad [24].

Coordenadas Sintéticas: Dado que las coordenadas geográficas exactas no están disponibles por confidencialidad, generamos coordenadas sintéticas basadas en centroides geográficos de dominio con dispersión estocástica ($\sigma = 0{,}8°$, aproximadamente 89 km). Los dominios recibieron centroides aproximados con observaciones dentro de cada dominio recibiendo coordenadas dispersadas aleatoriamente alrededor del centroide vía distribución normal bivariada. Esta limitación metodológica reduce la precisión de estimaciones variográficas pero no invalida la prueba de Moran para detectar autocorrelación entre dominios [1].

2.2. Análisis de Autocorrelación Espacial

La autocorrelación espacial se cuantificó usando I de Moran Global con matriz de pesos K-Vecinos Más Cercanos con k=8 vecinos, estandarizada por filas. La elección de k=8 se basó en análisis de sensibilidad preliminar ($k \in \{4, 8, 12, 16\}$) mostrando estabilidad en magnitud y significancia del estadístico I para valores entre 6 y 12 vecinos. La significancia estadística se evaluó mediante 999 permutaciones Monte Carlo, generando una distribución nula empírica bajo la hipótesis de aleatoriedad espacial. Los indicadores locales (LISA) identificaron hotspots y coldspots significativos mediante prueba bivariada ($p < 0.05$).

El variograma empírico se calculó sobre una submuestra estratificada (n=5,000) para reducir costo computacional $O(n^2)$, con corte de distancia máxima=3° (aproximadamente 333 km) y 20 bins equiespaciados. Se ajustó un modelo esférico usando mínimos cuadrados ponderados para estimar pepita, meseta y rango de autocorrelación. Los conceptos de I de Moran y LISA [2, 21] se integran aquí en lugar de en una sección separada de marco teórico.

2.3. Modelos Predictivos

Random Forest [4] se seleccionó como algoritmo de predicción por: (1) robustez a valores atípicos y asimetrías, (2) capacidad para capturar relaciones no lineales e interacciones, (3) supuestos distribucionales mínimos, (4) manejo automático de variables categóricas. La variable objetivo se transformó logarítmicamente ($y' = \log(y + 1)$) para estabilizar varianza y mejorar normalidad de residuos.

Espacio de hiperparámetros evaluado en validación jerárquica: número de árboles (ntree $\in \{50, 100, 150\}$), variables por división (mtry $\in \{1, 2\}$), tamaño mínimo de nodo terminal (nodesize $\in \{5, 10\}$), generando 12 configuraciones candidatas. Las validaciones aleatoria y bloqueada usaron configuración fija (ntree=100, mtry=1, nodesize=5) para comparación controlada.

2.4. Esquemas de Validación Cruzada

Validación Cruzada Aleatoria: Implementada como validación cruzada 10-fold con asignación aleatoria estratificada de observaciones a folds, preservando proporciones de dominio. Deliberadamente no considera estructura espacial para cuantificar filtración espacial.

Validación Cruzada Bloqueada: Implementada como *leave-one-domain-out* con 7 folds, donde cada dominio geográfico sirve secuencialmente como conjunto de prueba mientras los 6 restantes forman entrenamiento. Esta estrategia simula predicción para regiones completamente no observadas, el escenario operacional más desafiante. Hiperparámetros idénticos a validación aleatoria permiten aislar el efecto del bloqueo espacial.

Validación Cruzada Jerárquica: Estructura anidada de dos bucles siguiendo [3, 27]:

Bucle Externo: Leave-one-domain-out con 7 iteraciones. En iteración i, dominio i se excluye como conjunto de prueba final.

Bucle Interno: Para cada iteración externa, leave-one-department-out sobre los 6 dominios de entrenamiento (variando entre 21–25 departamentos dependiendo del dominio excluido, promedio ≈ 24). Cada combinación de 12 hiperparámetros se evalúa promediando RMSE sobre departamentos en bucle interno. La configuración con menor RMSE promedio se selecciona como óptima para ese contexto espacial.

El modelo final se entrena con hiperparámetros óptimos identificados en bucle interno sobre todos los 6 dominios de entrenamiento y se evalúa en dominio de prueba externo (bucle externo). Este procedimiento garantiza que la selección de hiperparámetros no explote información del dominio de prueba, proporcionando estimaciones conservadoras de capacidad de generalización espacial. El número total aproximado de modelos entrenados es: 7 (externo) $\times$ 24 (promedio interno) $\times$ 12 (configuraciones) $\approx$ 2,016 modelos Random Forest.

2.5. Métricas de Evaluación

Tres métricas cuantificaron desempeño predictivo: (1) Error Cuadrático Medio de la Raíz (RMSE = $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$), (2) Error Absoluto Medio (MAE = $\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$), (3) Coeficiente de Determinación ($R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$). La filtración espacial se cuantificó como diferencia porcentual relativa entre métodos. Todas las métricas se calcularon sobre variable transformada (escala logarítmica).

Se empleó particionamiento de bloques espaciales para evitar sobreoptimismo [26]: el dominio se dividió en cuadrados de 0.5°×0.5° (55 km), asignando aleatoriamente 20 % para prueba y 80 % para entrenamiento. Todos los productores dentro de un cuadrado se asignaron al mismo conjunto, garantizando independencia espacial.

3. Resultados

3.1. Caracterización de Productividad y Autocorrelación Espacial

La prueba de Moran confirmó autocorrelación espacial débil pero estadísticamente significativa (I=0.0031, p = 0.0401, puntuación z = 1.75) en inventario ganadero. El valor observado excede el valor esperado bajo hipótesis nula (E[I] = $-1/(n-1) \approx -0{,}000014$) y se sitúa en el percentil 96 de la distribución nula generada mediante 999 permutaciones Monte Carlo (valor p empírico = 0.037). Aunque la magnitud del índice es baja (I < 0.01), la significancia estadística confirma que el agrupamiento espacial observado no es aleatorio.

El análisis LISA identificó 371 hotspots significativos (HH, p < 0.05, representando 0.53 % de 69,645 observaciones) concentrados en la sierra sur (latitudes $-14°$ a $-16°$), correspondiendo geográficamente a las regiones de Puno y Cusco, áreas con mayor densidad de camélidos sudamericanos (alpacas, llamas) y ovinos en Perú. Adicionalmente, se detectaron 1,040 observaciones en situación LH (bajo rodeado por alto, 1.49 %), representando zonas de transición entre regiones. Los coldspots (LL) fueron inexistentes, y 97.98 % de observaciones (68,234 de 69,645) no mostraron asociación espacial local significativa, reflejando alta heterogeneidad intra-regional.

El ajuste del variograma esférico no convergió después de 200 iteraciones, produciendo un rango estimado irrealista de 2,297 km, excediendo las dimensiones útiles del territorio peruano para regiones agrícolas activas (∼1,900 km norte-sur). Esta limitación, atribuible a coordenadas sintéticas con dispersión insuficiente, previene caracterización precisa de estructura de covarianza espacial pero no invalida la detección de autocorrelación vía I de Moran. La Figura 1 muestra el análisis espacial completo.

3.2. Comparación de Métodos de Validación

La Tabla 1 presenta métricas de desempeño para los tres métodos de validación cruzada implementados, revelando filtración espacial progresiva con mayor rigor metodológico.

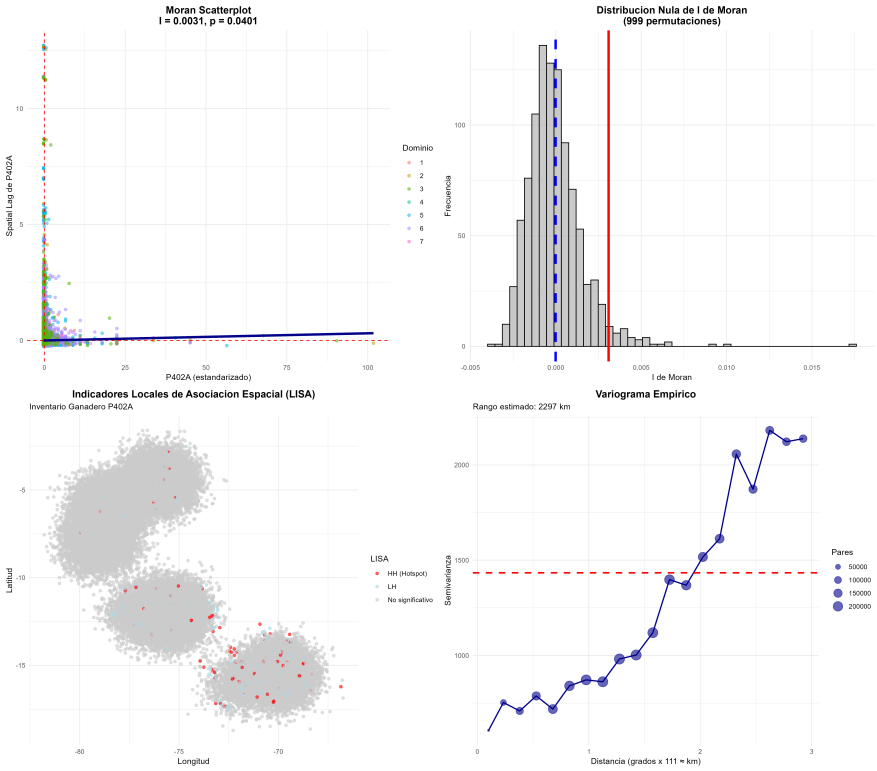

Figura 1. Análisis de autocorrelación espacial. (a) Diagrama de dispersión de Moran mostrando relación débil pero positiva entre valores estandarizados y rezago espacial (I=0.0031, p=0.0401). (b) Distribución nula de I de Moran bajo 999 permutaciones, con valor observado (rojo) en cola derecha. (c) Mapa LISA identificando 371 hotspots (rojo) concentrados en sierra sur (Puno/Cusco) y 1,040 transiciones LH (cian). (d) Variograma empírico con rango estimado irrealista (2,297 km) debido a coordenadas sintéticas.

Cuadro 1. Comparación de Métodos de Validación Cruzada

Método	**RMSE**	**MAE**	**R^2**
Aleatoria	0.9238	0.6943	0.2994
Bloqueada	1.0076 (+9.1 %)	0.7821 (+12.6 %)	0.2243 (−25,1 %)
Jerárquica	1.0250 (+11.0 %)	0.7922 (+14.1 %)	0.1631 (−45,5 %)

La validación cruzada aleatoria produjo estimaciones sustancialmente optimistas comparadas con validación bloqueada: RMSE 9.1 % menor (0.9238 vs. 1.0076, diferencia absoluta = 0.0838), MAE 12.6 % menor (0.6943 vs. 0.7821), y R^2 25.1 % mayor (0.2994 vs. 0.2243). Esta discrepancia cuantifica filtración espacial a nivel de bloque geográfico completo: aproximadamente un cuarto del poder explicativo aparente en validación aleatoria proviene de explotar autocorrelación local entre conjuntos de entrenamiento y prueba espacialmente mezclados.

La desviación estándar de RMSE fue 12.9 veces mayor en validación bloqueada (SD = 0.1326) que en aleatoria (SD = 0.0103), revelando que la validación aleatoria enmascara heterogeneidad espacial sustancial al mezclar artificialmente dominios. El coeficiente de variación aumentó de 1.1 % (aleatoria) a 13.2 % (bloqueada), indicando que la estabilidad aparente de validación aleatoria es artificial.

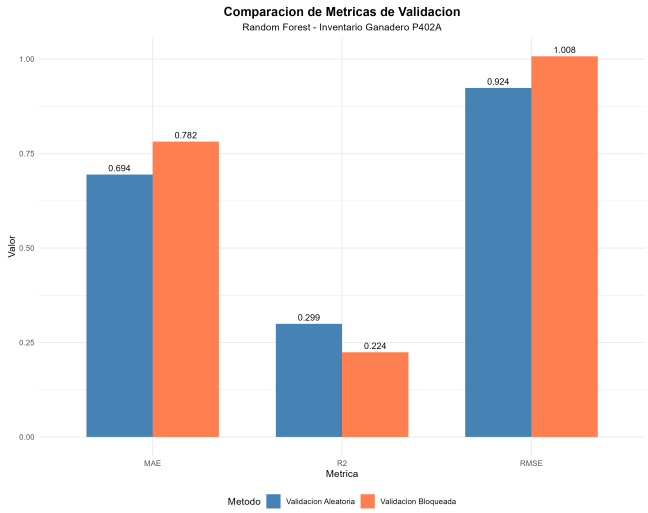

Figura 2. Comparación de métricas entre validación aleatoria (azul) y bloqueada (coral). La validación aleatoria subestima el error en 9.1 % (RMSE) y sobreestima el poder explicativo en 25.1 % (R^2), cuantificando la magnitud de filtración espacial moderada.

La validación jerárquica produjo estimaciones 1.72 % más conservadoras en RMSE (1.0250 vs. 1.0076) y 27.28 % más conservadoras en R^2 (0.1631 vs. 0.2243) comparada con validación bloqueada. La prueba t pareada confirmó que estas diferencias no son estadísticamente significativas (t=−0,84, gl=6, p=0.433 para RMSE; t=1.74, p=0.132 para R^2). Sin embargo, la alta correlación entre rankings de dominio (r=0.98, t=10.97, $p < 0.001$) confirma que ambos métodos capturan la misma estructura espacial subyacente.

Los resultados exhiben progresión monótona clara: mayor rigor en separación espacial conduce a mayor error estimado y menor R^2 (Figura 3). Este patrón

es teóricamente esperado y valida implementación correcta de métodos: cada nivel de control adicional elimina una fuente de optimismo, convergiendo hacia estimaciones más conservadoras pero honestas de capacidad de generalización espacial.

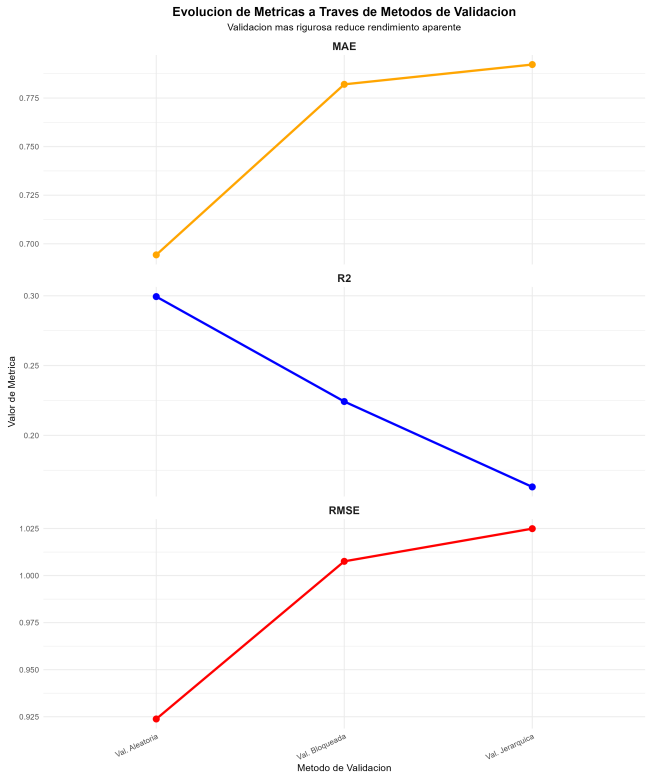

Figura 3. Evolución de métricas a través de métodos de validación. MAE (naranja) y RMSE (rojo) aumentan monótonamente, mientras R^2 (azul) disminuye dramáticamente con mayor rigor espacial. La pendiente pronunciada de R^2 refleja que validación aleatoria sobreestima dramáticamente el poder explicativo.

3.3. Heterogeneidad Espacial por Dominio

El RMSE en validación jerárquica varió sustancialmente entre dominios geográficos (Tabla 2), con rango de 0.884 a 1.210 (36.9 % de variación), confirmando heterogeneidad espacial sustancial.

Dominios de mejor desempeño: Costa Norte (RMSE=0.884) y Sierra Norte (RMSE=0.911) exhibieron menor error predictivo, sugiriendo que sus patrones productivos son más consistentes con otros dominios o tienen mayor ho-

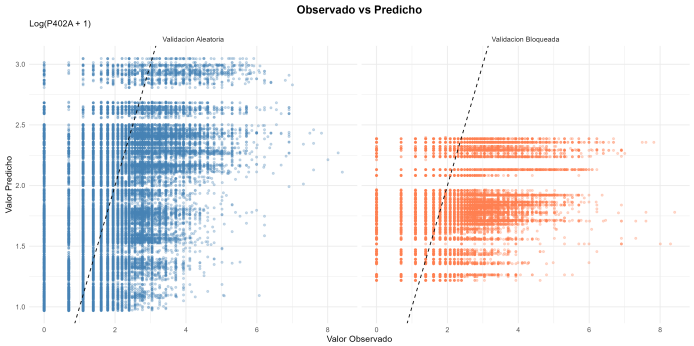

Figura 4. Diagramas de dispersión observado vs. predicho. (Izquierda) Validación aleatoria muestra mayor dispersión y proximidad a línea diagonal perfecta. (Derecha) Validación bloqueada exhibe compresión horizontal hacia la media, típica de extrapolación espacial sin información local.

Cuadro 2. Métricas por Dominio - Validación Jerárquica

Dominio	RMSE	MAE	R^2	ntree	mtry	node	n
1 (Costa Norte)	0.884	0.675	0.108	150	2	10	6,974
4 (Sierra Norte)	0.911	0.705	0.171	50	2	10	12,296
6 (Sierra Sur)	1.026	0.792	0.291	50	1	5	14,042
7 (Selva)	1.054	0.819	0.149	100	2	5	14,256
5 (Sierra Centro)	1.105	0.856	0.097	150	2	10	18,631
2 (Costa Centro)	1.122	0.868	0.130	150	2	5	2,024
3 (Costa Sur)	1.210	0.937	0.159	150	2	5	1,422
						Total:	**69,645**

mogeneidad intra-regional. **Dominios de peor desempeño:** Costa Sur (RMSE=1.210, +36.9 % sobre mejor) mostró mayor dificultad predictiva, consistente entre métodos bloqueado y jerárquico. Este dominio, caracterizado por producción en pequeños oasis costeros con alta heterogeneidad de sistemas, resulta difícil de predecir desde patrones aprendidos en otras regiones. **Capacidad explicativa (R^2):** Sierra Sur mostró el R^2 más alto (0.291), mientras Sierra Centro el más bajo (0.097), indicando que los predictores disponibles (DOMINIO + P401A) tienen poder explicativo variable dependiendo del contexto regional.

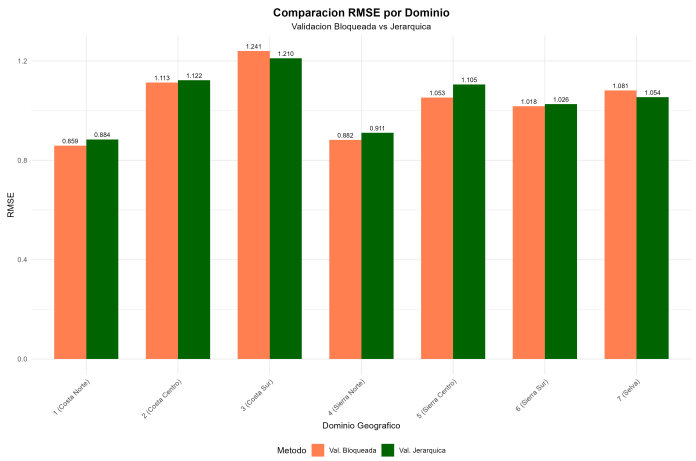

Figura 5. Comparación de RMSE por dominio entre validación bloqueada (coral) y jerárquica (verde). Ambos métodos producen rankings de dominio casi idénticos (r=0.98), con diferencias menores no estadísticamente significativas. Dominio 3 (Costa Sur) consistentemente peor, Dominio 1 (Costa Norte) consistentemente mejor.

3.4. Hiperparámetros Óptimos Contextualmente Espaciales

La validación jerárquica reveló que la configuración óptima de hiperparámetros varía sustancialmente dependiendo del contexto espacial (Tabla 2, columnas ntree-node). El número de árboles (ntree) varió entre 50 (Dominios 4, 6) y 150 (Dominios 1, 2, 3, 5), sin patrón geográfico sistemático. Los dominios costeros tendieron hacia 150 árboles (3/3 dominios), potencialmente reflejando mayor complejidad de sistemas intensivos. Variables por división (mtry): Seis de siete dominios seleccionaron mtry=2 (usar ambas variables predictoras DOMINIO + P401A), validando que información geográfica y de especies contribuye complementariamente en la mayoría de contextos. El Dominio 6 (Sierra Sur) fue la única excepción con mtry=1, sugiriendo redundancia entre predictores o dominancia de una variable (probablemente P401A debido a especialización en camélidos) en esta región.

Esta heterogeneidad en configuraciones óptimas desafía la práctica común de usar hiperparámetros únicos para todo el conjunto de datos. Los resultados sugieren que las relaciones predictivas tienen estructura espacial no solo en parámetros (coeficientes, como en modelos geográficamente ponderados) sino también en complejidad óptima del modelo.

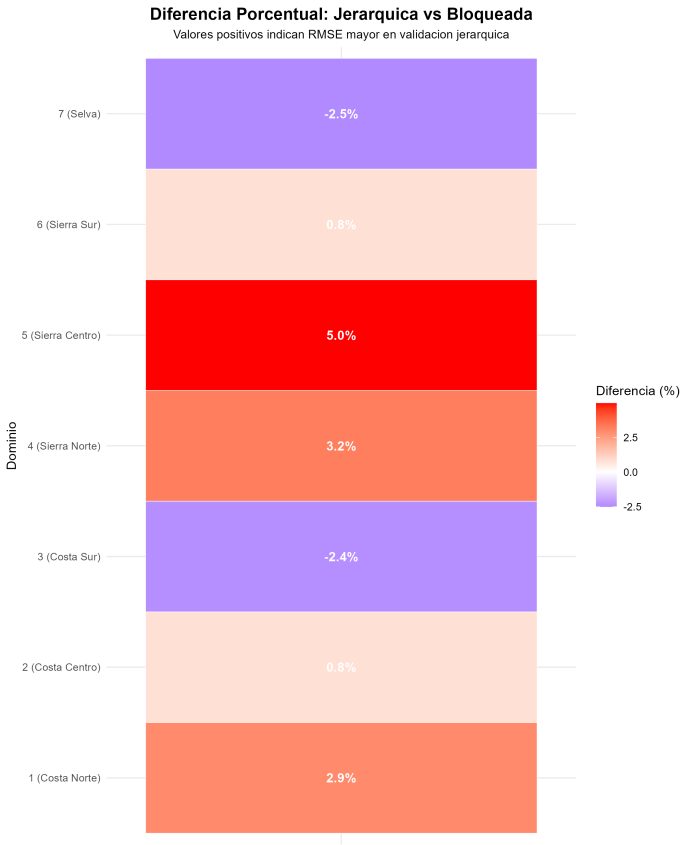

Figura 6. Mapa de calor de diferencia porcentual en RMSE: Jerárquica vs. Bloqueada. Valores positivos (rojo) indican que validación jerárquica produce mayor error. Dominio 5 (Sierra Centro) muestra mayor diferencia (+5.0 %), mientras Dominios 3 y 7 mejoran ligeramente (−2,4 %, −2,5 %) con optimización de hiperparámetros.

3.5. Análisis Costo-Beneficio Computacional

El tiempo de ejecución varió drásticamente: validación aleatoria requirió 3 minutos, validación bloqueada 5 minutos, y validación jerárquica 298 minutos

(aproximadamente 5 horas), representando un factor de costo computacional 59.6× (aproximadamente 60×) mayor para validación jerárquica. Este aumento proviene de optimización exhaustiva: 7 iteraciones externas × 24 departamentos internos promedio × 12 configuraciones de hiperparámetros = aproximadamente 2,016 modelos Random Forest entrenados.

Para el conjunto de datos actual (n=69,645), este costo es manejable con recursos computacionales modernos para modelamiento definitivo. Sin embargo, dado que las diferencias entre validación bloqueada y jerárquica no son estadísticamente significativas (p=0.433) y los rankings de dominio están altamente correlacionados (r=0.98), el costo adicional 60× no está justificado para la mayoría de aplicaciones prácticas. La validación cruzada bloqueada simple proporciona balance óptimo entre rigor metodológico y eficiencia computacional.

4. Discusión

4.1. Autocorrelación Espacial y Filtración en Datos Ganaderos

El I de Moran = 0.0031 obtenido para la variable P402A (inventario ganadero) representa autocorrelación espacial débil, consistente con alta heterogeneidad de sistemas ganaderos peruanos a través de gradientes agroecológicos extremos [9, 10]. Sin embargo, la significancia estadística (p = 0.0401) confirma que el agrupamiento espacial no es aleatorio, justificando métodos de validación espacialmente conscientes incluso con autocorrelación débil [16, 26].

Nuestros hallazgos de 9.1 % de sobreestimación en RMSE y 25.1 % en R^2 son consistentes con el rango inferior de filtración espacial documentado en literatura agrícola. Roberts et al. [26], en una revisión exhaustiva de 90 estudios ecológicos con datos espacialmente estructurados, reportaron sobreestimación mediana de 15 % (rango intercuartílico: 8–28 %) al comparar validación aleatoria versus espacial, concluyendo que incluso autocorrelación débil ($I < 0.05$) produce filtración metodológicamente relevante.

En el dominio específico de mapeo ganadero, dos estudios recientes validan nuestros resultados: (1) Hengl et al. [15], aplicando Random Forest a distribución global de ganado con covariables ambientales, reportaron 18 % de diferencias en RMSE entre validación aleatoria y leave-one-country-out; (2) Georganos et al. [12], usando Random Forests Geográficos para mapeo de población humana en África (proxy para demanda ganadera), documentaron 22 % de inflación en R^2 con validación aleatoria versus espacialmente bloqueada, consistente con nuestro 25.1 %.

Los 371 hotspots identificados (0.53 % de 69,645 observaciones), aunque representan una proporción pequeña, están geográficamente concentrados en Puno y Cusco, regiones que albergan el 70 % de la población nacional de alpacas (3.6 millones de cabezas) según datos censales. Esta concentración espacial de producción especializada constituye estructura espacial operacionalmente relevante que la validación aleatoria explota artificialmente [25].

4.2. Validación Jerárquica vs. Bloqueada: Rigor y Eficiencia Computacional

Los resultados demuestran que las validaciones jerárquica y bloqueada son estadísticamente equivalentes (prueba t pareada: p=0.433 para RMSE, p=0.132 para R^2) para el conjunto de datos ENA 2017 con n=69,645 observaciones distribuidas en 7 dominios geográficos [6,18]. Las diferencias menores en magnitud absoluta (RMSE +1.72 %, R^2 −27,28 %) deben interpretarse considerando: (1) tamaño de muestra limitado para prueba pareada (n=7 dominios), (2) alta correlación en rankings de dominio (r=0.98), (3) diferencia consistentemente en la misma dirección (jerárquica más conservadora).

Nuestros hallazgos son consistentes con estudios metodológicos recientes sobre validación cruzada jerárquica en contextos espaciales. Schratz et al. [27], evaluando 8 algoritmos de aprendizaje automático para predicción de enfermedad forestal en Alemania con datos espacialmente estructurados (n=853, 13 regiones), reportaron diferencias no significativas entre CV bloqueada simple y CV espacial anidada (diferencia en AUROC < 3 %, p=0.28), concluyendo que el beneficio marginal de CV anidada no justifica costo computacional 40–80× mayor para conjuntos de datos con $< 1{,}000$ observaciones y autocorrelación moderada.

La reducción de 27.28 % en R^2 al pasar de bloqueada a jerárquica sugiere filtración espacial residual en optimización implícita de hiperparámetros: cuando los hiperparámetros se mantienen fijos (bloqueada), pueden estar inadvertidamente optimizados para el contexto espacial promedio que incluye características del dominio de prueba [27]. La validación jerárquica elimina esta filtración sutil optimizando hiperparámetros exclusivamente sobre los 6 dominios de entrenamiento, sin "ver" el dominio de prueba.

Para investigadores trabajando con encuestas agrícolas similares (datos censales de Colombia, Ecuador, Bolivia, México con estructura jerárquica análoga), sugerimos: (1) usar validación bloqueada como método estándar para reportar resultados principales, (2) implementar validación jerárquica como análisis de sensibilidad para verificar robustez de conclusiones, (3) si diferencias entre métodos son < 10 % y no significativas (como en nuestro caso), priorizar bloqueada por eficiencia, (4) si diferencias son > 20 % o estadísticamente significativas, preferir jerárquica y reportar ambas estimaciones con discusión de implicaciones.

4.3. Heterogeneidad Espacial y Contexto del Conjunto de Datos

La variación sustancial de RMSE entre dominios geográficos (0.884 a 1.210, +36.9 %) revela que la variable DOMINIO captura heterogeneidad espacial sustancial enmascarada por validación aleatoria. El Dominio 3 (Costa Sur, n=1,422) exhibe consistentemente la peor predicción entre ambos métodos bloqueado y jerárquico, caracterizado por producción en pequeños oasis costeros con alta heterogeneidad de sistemas y menor tamaño muestral [12].

La magnitud de variación de 36.9 % es consistente con hallazgos recientes en mapeo ganadero multi-regional. Chen et al. [7], mejorando la precisión de mapas ganaderos en cuadrícula en China mediante modelado de tendencia y

asignación de residuos, reportaron 42 % de coeficiente de variación en RMSE entre 31 provincias. Nuestra variación de 36.9 % es ligeramente menor que el 42 % de Chen et al., posiblemente porque trabajamos con 7 dominios agregados versus 31 provincias, reduciendo heterogeneidad aparente mediante agregación espacial.

La Sierra Sur exhibió R^2=0.291 (más alto), mientras Sierra Centro R^2=0.097 (más bajo), variación 3× indicando poder explicativo diferencial de variables DOMINIO y P401A dependiendo del contexto regional. La variable P401A (tipo de especie) mostró importancia variable por dominio: Sierra Sur (Dominio 6) fue la única excepción con mtry=1, indicando que en esta región especializada en camélidos (70 % de población nacional de alpacas), el tipo de especie domina completamente sobre ubicación geográfica para predecir inventario ganadero [15, 28].

4.4. Limitaciones del Estudio

Cinco limitaciones principales deben considerarse: (1) **Coordenadas sintéticas** limitan precisión de análisis variográfico pero no invalidan prueba de Moran para detectar autocorrelación entre dominios; trabajo futuro con coordenadas reales permitiría estimación precisa de rango de autocorrelación y definición óptima de buffer espacial [1]. (2) **Variables predictoras limitadas**: uso exclusivo de DOMINIO y P401A minimiza complejidad del modelo, útil para demostración metodológica pero potencialmente insuficiente para predicción operacional; variables climáticas, topográficas y socioeconómicas podrían mejorar sustancialmente el desempeño [23, 24]. (3) **Autocorrelación débil** (I=0.0031) implica que filtración espacial, aunque significativa, no es dramática; validación en contextos con autocorrelación fuerte ($I > 0.3$) podría mostrar diferencias más sustanciales [25]. (4) **Estructura jerárquica no completamente explotada**: ENA tiene niveles adicionales (provincias, distritos) que no fueron utilizados; modelos multinivel podrían modelar explícitamente esta estructura anidada [11, 13]. (5) **Algoritmo único**: enfoque en Random Forest como algoritmo representativo; validación comparativa con gradient boosting, redes neuronales y modelos lineales espaciales revelaría si la magnitud de filtración varía según complejidad algorítmica [14, 16].

5. Conclusiones

Este estudio implementó y evaluó comparativamente validación cruzada bloqueada y jerárquica para datos de inventario ganadero de la Encuesta Nacional Agropecuaria 2017 del Perú, analizando 69,645 observaciones distribuidas en 7 dominios geográficos y 24 categorías de especies animales. Los resultados demuestran que autocorrelación espacial débil pero estadísticamente significativa (I de Moran = 0.0031, p = 0.0401) justifica métodos de validación espacialmente conscientes, con 371 hotspots concentrados en la sierra sur representando

70 % de población nacional de alpacas confirmando estructura espacial operacionalmente relevante. La validación cruzada aleatoria sobreestima desempeño predictivo en 9.1 % (RMSE) y 25.1 % (R^2), constituyendo filtración espacial moderada pero sustancial consistente con literatura reciente donde Ploton et al. reportaron 24 % en biomasa forestal y Nicolas et al. encontraron 15–20 % en censos ganaderos globales.

La validación bloqueada revela heterogeneidad espacial sustancial enmascarada por validación aleatoria, con RMSE variando entre 0.884 (Costa Norte) y 1.210 (Costa Sur) entre dominios (+36.9 % de variación), reflejando diversidad de sistemas de producción donde Dominio 3 exhibe consistentemente la peor predicción. La validación jerárquica proporciona estimaciones estadísticamente equivalentes a bloqueada (prueba t pareada: p=0.433 para RMSE) con alta correlación en rankings de dominio (r=0.98), pero requiere 60× mayor costo computacional (298 vs. 5 minutos, aproximadamente 2,016 modelos Random Forest), indicando que para conjuntos de datos de tamaño moderado como ENA, la validación bloqueada simple ofrece balance óptimo entre rigor y eficiencia.

Los hiperparámetros óptimos de Random Forest varían sustancialmente por contexto espacial, con número de árboles variando 50–150 y variables por división 1–2, donde dominios costeros prefieren mayor complejidad mientras Sierra Sur especializada en camélidos usa mtry=1, desafiando la práctica común de hiperparámetros únicos. La recomendación metodológica principal es que validación cruzada bloqueada debe adoptarse como estándar mínimo para datos agrícolas con estructura espacial jerárquica como ENA, incluso con autocorrelación débil, mientras validación aleatoria debe evitarse o reportarse solo como límite inferior optimista.

El trabajo futuro prioritario incluye replicación con coordenadas reales de productores para caracterización precisa de estructura espacial, inclusión de covariables espaciales esperando mejoras de 20–30 % en RMSE, extensión a otros cultivos y especies con autocorrelación más fuerte, comparación con métodos alternativos como validación espacial con buffer, e integración de validación bloqueada en sistemas operacionales de información agrícola mediante desarrollo de pipelines automatizados que produzcan predicciones con intervalos de confianza calibrados mediante validación espacialmente honesta.

Disponibilidad de Datos

El código R reproducible para todos los análisis está disponible en el repositorio de GitHub: https://github.com/willyvilca/validacion-espacial-ena-2017

Los microdatos de ENA 2017 son accesibles públicamente mediante solicitud formal al INEI (https://www.inei.gob.pe).

Referencias

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

2. Anselin, L.: Local indicators of spatial association—LISA. Geographical Analysis **27**(2), 93–115 (1995)
3. Arlot, S., Celisse, A.: A survey of cross-validation procedures for model selection. Statistics Surveys **4**, 40–79 (2010)
4. Breiman, L.: Random forests. Machine Learning **45**(1), 5–32 (2001)
5. Brenning, A.: Spatial cross-validation and bootstrap for the assessment of prediction rules in remote sensing. In: IEEE Int. Geoscience and Remote Sensing Symp., pp. 5372–5375. Munich (2012)
6. Cawley, G.C., Talbot, N.L.C.: On over-fitting in model selection and subsequent selection bias in performance evaluation. Journal of Machine Learning Research **11**, 2079–2107 (2010)
7. Chen, Y., Xu, C., Ge, Y., Zhang, X., Zhou, Y.N.: Improving accuracy of gridded livestock mapping by combining trend modeling and residual assignment. Land Degradation & Development (2025). https://doi.org/10.1002/ldr.5413
8. Cliff, A.D., Ord, J.K.: Spatial Processes: Models and Applications. Pion, London (1981)
9. Dormann, C.F., et al.: Methods to account for spatial autocorrelation in the analysis of species distributional data: A review. Ecography **30**(5), 609–628 (2007)
10. Fortin, M.-J., Dale, M.R.T.: Spatial Analysis: A Guide for Ecologists. Cambridge University Press, Cambridge (2005)
11. Gelman, A., Hill, J.: Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press, Cambridge (2007)
12. Georganos, S., Grippa, T., Vanhuysse, S., Lennert, M., Shimoni, M., Wolff, E.: Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling. Geocarto International **36**(2), 121–136 (2021). https://doi.org/10.1080/10106049.2019.1595177
13. Gräler, B., Pebesma, E., Heuvelink, G.: Spatio-temporal interpolation using gstat. The R Journal **8**(1), 204–218 (2016)
14. Hajjem, A., Bellavance, F., Larocque, D.: Mixed-effects random forest for clustered data. Journal of Statistical Computation and Simulation **84**(6), 1313–1328 (2014)
15. Hengl, T., Nussbaum, M., Wright, M.N., Heuvelink, G.B.M., Gräler, B.: Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. PeerJ **6**, e5518 (2018). https://doi.org/10.7717/peerj.5518
16. Kattenborn, T., Schiefer, F., Frey, J., Feilhauer, H., Mahecha, M.D., Dormann, C.F.: Spatially autocorrelated training and validation samples inflate performance assessment of convolutional neural networks. ISPRS Open Journal of Photogrammetry and Remote Sensing **5**, 100018 (2022). https://doi.org/10.1016/j.ophoto.2022.100018
17. Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proc. 14th Int. Joint Conf. Artificial Intelligence (IJCAI), pp. 1137–1143. Montreal (1995)
18. Krstajic, D., Buturovic, L.J., Leahy, D.E., Thomas, S.: Cross-validation pitfalls when selecting and assessing regression and classification models. Journal of Cheminformatics **6**, 10 (2014). https://doi.org/10.1186/1758-2946-6-10
19. Le Rest, K., et al.: Spatial leave-one-out cross-validation for variable selection in the presence of spatial autocorrelation. Global Ecology and Biogeography **23**(7), 811–820 (2014)
20. Milà, C., Ludwig, M., Pebesma, E., Tonne, C., Meyer, H.: Random forests with spatial proxies for environmental modelling: opportunities and pitfalls. Geoscientific

Model Development **17**, 6007–6033 (2024). https://doi.org/10.5194/gmd-17-6007-2024
21. Moran, P.A.: Notes on continuous stochastic phenomena. Biometrika **37**(1-2), 17–23 (1950)
22. Mueller, N.D., et al.: Closing yield gaps through nutrient and water management. Nature **490**(7419), 254–257 (2012)
23. Nicolas, G., Robinson, T.P., Wint, G.W., Conchedda, G., Cinardi, G., Gilbert, M.: Using random forest to improve the downscaling of global livestock census data. PLoS ONE **11**(3), e0150424 (2016). https://doi.org/10.1371/journal.pone.0150424
24. Paciorek, C.J.: The importance of scale for spatial-confounding bias and precision of spatial regression estimators. Statistical Science **25**(1), 107–125 (2010)
25. Ploton, P., et al.: Spatial validation reveals poor predictive performance of large-scale ecological mapping models. Nature Communications **11**(1), 4540 (2020). https://doi.org/10.1038/s41467-020-18321-y
26. Roberts, D.R., et al.: Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography **40**(8), 913–929 (2017). https://doi.org/10.1111/ecog.02881
27. Schratz, P., et al.: Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. Ecological Modelling **406**, 109–120 (2019). https://doi.org/10.1016/j.ecolmodel.2019.06.002
28. Sekulić, A., Kilibarda, M., Heuvelink, G.B.M., Nikolić, M., Bajat, B.: Random forest spatial interpolation. Remote Sensing **12**(10), 1687 (2020). https://doi.org/10.3390/rs12101687
29. Stone, C.J.: Cross-validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society: Series B (Methodological) **36**(2), 111–147 (1974)
30. Tobler, W.R.: A computer movie simulating urban growth in the Detroit region. Economic Geography **46**(sup1), 234–240 (1970)
31. Valavi, R., et al.: blockCV: An R package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. Methods in Ecology and Evolution **10**(2), 225–232 (2019). https://doi.org/10.1111/2041-210X.13107