Validación Cruzada Bloqueada y Jerárquica para Datos de Producción Pecuaria con Autocorrelación Espacial

Willy Vilca Apaza

3 de octubre de 2025

1. Introducción

La autocorrelación espacial representa un desafío fundamental en el modelamiento estadístico de datos agrícolas, donde observaciones geográficamente próximas exhiben características similares debido a condiciones compartidas [12]. Las técnicas convencionales de validación cruzada asumen independencia entre observaciones, suposición frecuentemente violada en conjuntos de datos con estructura geográfica [24]. Cuando se ignora la dependencia espacial, los procedimientos de validación producen estimaciones optimistas del rendimiento del modelo, conduciendo a predicciones pobres [23].

Los datos de la Encuesta Nacional Agropecuaria (ENA) 2017 del Perú exhiben fuerte agrupamiento espacial a través de dominios geográficos, donde los niveles de inventario ganadero muestran correlación significativa dentro de regiones que comparten características agroecológicas. Esta estructura espacial, organizada jerárquicamente desde regiones nacionales hasta distritos, requiere estrategias de validación especializadas [5].

La validación cruzada bloqueada particiona datos en grupos espacialmente contiguos, emergiendo como solución robusta para evaluar modelos predictivos con dependencia espacial [29]. La validación cruzada jerárquica extiende estas estrategias mediante procedimientos anidados de leave-one-group-out en múltiples escalas espaciales [14]. Este estudio propone aplicar una metodología integrada a datos de inventario ganadero del Perú, seleccionando tres variables: inventario ganadero (P402A) como dependiente, dominio geográfico (DOMINIO) como factor de bloqueo espacial, y tipo de especie animal (P401A) como variable de control productivo [22].

2. Marco Teórico

2.1. Autocorrelación Espacial

La autocorrelación espacial en sistemas ganaderos se manifiesta mediante difusión de tecnología entre productores vecinos, acceso compartido a servicios veterinarios y restricciones ambientales comunes [17]. Ignorarla conduce a estimaciones sesgadas, tasas infladas de error tipo I e incertidumbre no confiable [1]. El estadístico I de Moran y el análisis de variogramas diagnostican estructuras de dependencia espacial [11]. El análisis geoestadístico revela agrupamiento espacial desde vecindarios locales hasta zonas regionales [25].

2.2. Limitaciones de Validación Convencional

La validación cruzada k-fold estándar asigna aleatoriamente observaciones a pliegues, creando alta probabilidad de que observaciones próximas aparezcan en entrenamiento y prueba [19]. Esta filtración espacial permite explotar autocorrelación para predicción, inflando métricas de validación [24]. Estudios previos demuestran que validación aleatoria sobreestima exactitud entre 20-50 % comparado con métodos de bloqueo espacial [23].

2.3. Validación Bloqueada y Jerárquica

La validación bloqueada particiona datos en pliegues espacialmente contiguos basados en límites administrativos, asegurando separación espacial entre entrenamiento y prueba [29]. Para encuestas administrativas, unidades naturales de bloqueo incluyen provincias o zonas agroecológicas [8]. La validación jerárquica implementa procedimientos anidados leave-one-group-out en diferentes escalas organizacionales [14]. Ignorar estructura jerárquica conduce a pseudoreplicación donde observaciones dependientes son tratadas como independientes [18]. La integración de ambos enfoques crea un marco comprensivo para datos espaciales complejos con estructuras anidadas, implementando bloqueo espacial en nivel jerárquico alto mientras respeta agrupamiento de nivel inferior [21].

3. Metodología

3.1. Datos y Variables

El conjunto de datos ENA 2017 contiene 55 variables organizadas jerárquicamente: regiones (3 categorías), dominios geográficos (7 categorías: Costa Norte, Costa Centro, Costa Sur, Sierra Norte, Sierra Centro, Sierra Sur, Selva), departamentos, provincias, distritos y conglomerados. Se seleccionaron tres variables: **P402A** (inventario ganadero, variable dependiente continua susceptible a agrupamiento espacial), **DOMINIO** (variable espacial independiente con 7 bloques naturalmente contiguos), y **P401A** (tipo de especie animal, 24 categorías con estructura espacial inherente). Esta selección minimiza confusión espacial mientras maximiza interpretabilidad [22].

3.2. Diagnósticos Espaciales

Previo al modelamiento, la autocorrelación espacial en P402A será cuantificada usando I de Moran Global a nivel de distrito, con significancia evaluada mediante 999 permutaciones Monte Carlo [11]. Los indicadores LISA identificarán hotspots y coldspots de concentración ganadera [2]. El análisis de variogramas estimará el rango de correlación espacial, proporcionando orientación para selección de distancia de buffer [15].

3.3. Validación Cruzada Bloqueada

La validación bloqueada utilizará DOMINIO como variable de bloqueo primaria, creando siete pliegues espacialmente contiguos [29]. En cada iteración, un dominio sirve como conjunto de prueba mientras los seis restantes comprenden entrenamiento. Se implementará zona de buffer excluyendo distritos dentro de 50 km del límite del dominio de prueba [8]. El rendimiento será evaluado usando RMSE, MAE y R^2 calculados separadamente

para cada pliegue [10]. Esta estrategia leave-one-domain-out simula escenarios realistas donde modelos deben extrapolar a dominios no observados.

3.4. Validación Cruzada Jerárquica

El componente jerárquico implementará validación anidada con dos bucles [4]. El **bucle externo** realiza evaluación usando leave-one-domain-out (7 iteraciones). El **bucle interno** ejecuta optimización de hiperparámetros usando leave-one-department-out dentro de cada conjunto de entrenamiento. El bucle interno asegura que ajuste de hiperparámetros no explote estructura espacial dentro de dominios [28]. Hiperparámetros con mejor rendimiento promedio en bucle interno son seleccionados para entrenar modelo final en conjunto de entrenamiento completo, evaluado luego en dominio de prueba externo [14].

3.5. Comparación y Evaluación

El rendimiento será comparado contra tres enfoques: (1) validación aleatoria 10-fold sin consideración espacial, (2) validación bloqueada sin estructura jerárquica, (3) validación jerárquica sin zonas de buffer [26]. Diferencias de rendimiento serán cuantificadas usando pruebas t pareadas en RMSE a nivel de pliegue, con corrección de Bonferroni. La magnitud de filtración espacial será estimada como diferencia relativa en R^2 entre método aleatorio y bloqueado jerárquico, anticipada en rango 15-40 % [23]. Autocorrelación espacial residual será re-evaluada usando I de Moran para verificar eliminación de estructura espacial en errores.

4. Discusión

4.1. Ventajas Metodológicas

La integración de validación bloqueada y jerárquica aborda limitaciones fundamentales de enfoques convencionales en datos espacialmente estructurados. El bloqueo espacial a nivel de dominio previene sobreajuste a patrones locales y proporciona estimaciones realistas de transferibilidad del modelo [21]. El componente jerárquico asegura que ajuste de hiperparámetros no explote agrupamiento de nivel inferior, mejorando robustez del modelo. Las zonas de buffer eliminan filtración espacial residual cerca de límites de pliegues [31]. El enfoque genera estimaciones honestas de incertidumbre que reflejan error de predicción fuera de muestra, facilitando inferencia confiable para política agrícola. Para el sector ganadero del Perú, esta metodología permite planificación basada en evidencia identificando si modelos generalizan a través de zonas agroecológicas diversas o requieren parametrización específica por región [30].

4.2. Limitaciones

El método requiere suficientes observaciones por bloque; dominios con muestras pequeñas producen estimaciones inestables [27]. La elección de unidades de bloqueo involucra compensaciones entre independencia espacial y tamaño de muestra [3]. El costo computacional de validación anidada crece cuadráticamente con niveles jerárquicos [7]. Los límites administrativos pueden no alinearse con gradientes ecológicos graduales, creando

discontinuidades artificiales [16]. El método asume estacionariedad de relaciones espaciales; procesos no estacionarios pueden requerir enfoques alternativos como regresión geográficamente ponderada [13].

4.3. Implicaciones para Política Agrícola

Esta investigación contribuye orientación metodológica para validar modelos espaciales en contextos de países en desarrollo. El sector agrícola del Perú exhibe heterogeneidad espacial extrema debido a gradientes topográficos y climáticos dramáticos, haciendo validación espacial rigurosa esencial [9]. El marco proporciona plantilla para futuros análisis de censos agropecuarios en América Latina. Al demostrar brecha de rendimiento entre validación convencional y espacialmente consciente, el estudio resalta riesgos de predicciones excesivamente confiadas en planificación agrícola, asignación de créditos subsidiados y focalización de programas de extensión [20]. La metodología se alinea con énfasis en investigación reproducible en estadística agrícola [6]. La implementación mediante software de código abierto en R (blockCV, sperrorest) o Python (scikit-learn, PySAL) facilita adopción por estadísticos gubernamentales e investigadores, promoviendo mejores prácticas en análisis de datos agropecuarios espacialmente estructurados.

Referencias

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

[2] L. Anselin, "Local indicators of spatial association—LISA," *Geographical Analysis*, vol. 27, no. 2, pp. 93–115, 1995.

[3] G. Arbia et al., "A spatial econometric approach to empirical probability mapping," *Statistical Methods and Applications*, vol. 22, no. 1, pp. 3–27, 2013.

[4] S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection," *Statistics Surveys*, vol. 4, pp. 40–79, 2010.

[5] D. Bates et al., "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[6] C. G. Begley and J. P. Ioannidis, "Reproducibility in science," *Circulation Research*, vol. 116, no. 1, pp. 116–126, 2015.

[7] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer, 2006.

[8] A. Brenning, "Spatial cross-validation and bootstrap," in *IEEE Int. Geoscience and Remote Sensing Symp.*, Munich, 2012, pp. 5372–5375.

[9] K. G. Cassman et al., "A global perspective on sustainable intensification," *Nature Sustainability*, vol. 3, no. 4, pp. 262–268, 2020.

[10] T. Chai and R. R. Draxler, "Root mean square error or mean absolute error?," *Geoscientific Model Development*, vol. 7, no. 3, pp. 1247–1250, 2014.

[11] A. D. Cliff and J. K. Ord, *Spatial Processes: Models and Applications*. London: Pion, 1981.

[12] C. F. Dormann et al., "Methods to account for spatial autocorrelation," *Ecography*, vol. 30, no. 5, pp. 609–628, 2007.

[13] A. S. Fotheringham, C. Brunsdon, and M. Charlton, *Geographically Weighted Regression*. Chichester: Wiley, 2002.

[14] A. Gelman and J. Hill, *Data Analysis Using Regression and Multilevel Models*. Cambridge: Cambridge University Press, 2007.

[15] M. G. Genton, "Variogram fitting by generalized least squares," *Mathematical Geology*, vol. 30, no. 4, pp. 323–345, 1998.

[16] M. F. Goodchild, "The validity and usefulness of laws in GIS," *Annals of the AAG*, vol. 94, no. 2, pp. 300–303, 2004.

[17] M. Herrero et al., "Biomass use and greenhouse gas emissions from livestock," *PNAS*, vol. 110, no. 52, pp. 20888–20893, 2013.

[18] S. J. Hurlbert, "Pseudoreplication and the design of experiments," *Ecological Monographs*, vol. 54, no. 2, pp. 187–211, 1984.

[19] R. Kohavi, "A study of cross-validation and bootstrap," in *Proc. IJCAI*, Montreal, 1995, pp. 1137–1143.

[20] D. B. Lobell et al., "The critical role of extreme heat for maize," *Nature Climate Change*, vol. 3, no. 5, pp. 497–501, 2013.

[21] H. Meyer et al., "Predicting into unknown space?," *Methods in Ecology and Evolution*, vol. 12, no. 9, pp. 1620–1633, 2021.

[22] C. Paciorek, "The importance of scale for spatial-confounding bias," *Statistical Science*, vol. 25, no. 1, pp. 107–125, 2010.

[23] H. Ploton et al., "Spatial validation reveals poor predictive performance," *Nature Communications*, vol. 11, no. 1, pp. 1–11, 2020.

[24] D. R. Roberts et al., "Cross-validation strategies for data with structure," *Ecography*, vol. 40, no. 8, pp. 913–929, 2017.

[25] T. P. Robinson et al., "Mapping the global distribution of livestock," *PLoS ONE*, vol. 9, no. 5, p. e96084, 2014.

[26] P. Schratz et al., "Hyperparameter tuning using spatial data," *Ecological Modelling*, vol. 406, pp. 109–120, 2019.

[27] S. Stehman, "Selecting measures of thematic classification accuracy," *Remote Sensing of Environment*, vol. 62, no. 1, pp. 77–89, 1997.

[28] C. J. Stone, "Cross-validatory choice and assessment," *JRSS: Series B*, vol. 36, no. 2, pp. 111–133, 1974.

[29] R. Valavi et al., "blockCV: An R package for spatial cross-validation," *Methods in Ecology and Evolution*, vol. 10, no. 2, pp. 225–232, 2019.

[30] M. K. van Ittersum et al., "Yield gap analysis with global relevance," *Field Crops Research*, vol. 143, pp. 4–17, 2013.

[31] T. Wiegand et al., "Species associations in a heterogeneous forest," *The American Naturalist*, vol. 170, no. 4, pp. E77–E95, 2007.