Blocked and Hierarchical Cross-Validation for Agricultural Data with Spatial Autocorrelation: Application to Peru's National Agricultural Survey

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

National University of the Altiplano, School of Statistical and Computer Engineering, Puno, Peru
w.vilca@est.unap.edu.pe, ftorres@unap.edu.pe, ecarpio@unap.edu.pe

Abstract. Spatial autocorrelation in agricultural data poses a critical methodological challenge for predictive model validation, since conventional techniques assume independence between observations. This study implements and evaluates blocked and hierarchical cross-validation for livestock inventory data from Peru's 2017 National Agricultural Survey (n=69,645), comparing them with traditional random cross-validation. Results show weak but statistically significant spatial autocorrelation (Moran's I = 0.0031, p = 0.0401). Random cross-validation overestimates predictive performance by 9.1% (RMSE) and 25.1% (R^2) compared to blocked validation, demonstrating moderate spatial leakage. Hierarchical cross-validation, implemented via leave-one-domain-out (outer loop) and leave-one-department-out (inner loop), provides statistically equivalent estimates but is computationally more expensive (approximately 60× longer runtime) than blocked validation. RMSE by geographic domain varies between 0.884 and 1.210 (36.9% variation), revealing substantial spatial heterogeneity hidden in random validation. The dominant nugget effect (77% of variance) indicates that individual management factors outweigh geographic location. Findings validate the need for spatially-aware validation methods for agricultural data, even with weak spatial autocorrelation, and demonstrate that simple blocked cross-validation provides the optimal balance between methodological rigor and computational efficiency.

Keywords: Spatial cross-validation · Spatial autocorrelation · Moran's I · Hierarchical cross-validation · Agricultural data · Spatial leakage · Random Forest · Agricultural surveys · Peru

1 Introduction

Spatial autocorrelation, formalized by Tobler's First Law of Geography [30], states that nearby objects tend to be more similar than distant ones. In agricultural production systems, this shows up through technology diffusion among neighboring producers, shared agroecological conditions, and common market access [22]. However, conventional techniques for validating predictive models assume independence between observations, systematically violating this inherent spatial structure [26].

Random k-fold cross-validation, widely used in machine learning, creates a high probability that spatially close observations appear in both training and test sets [17]. This setup lets models exploit local autocorrelation for prediction, generating artificially optimistic performance estimates known as *spatial leakage* [25]. Prior studies document that random validation can overestimate predictive accuracy by 10–50% in spatial prediction tasks [16, 26].

Blocked cross-validation tackles this problem by partitioning data into spatially contiguous groups, ensuring geographic separation between training and test sets [31]. However, standard blocking approaches don't account for hierarchical structures common in multi-scale survey designs [11]. Hierarchical cross-validation extends blocking by implementing nested leave-one-group-out procedures at multiple spatial scales, ensuring that hyperparameter optimization doesn't exploit local autocorrelation [3, 6].

Peru exhibits extreme spatial heterogeneity due to dramatic topographic gradients (0–6000 masl), climatic variation (coastal desert, inter-Andean valleys, tropical rainforest), and differential market access. The 2017 National Agricultural Survey (ENA), with its hierarchical structure (region-domain-department-province-district), provides an ideal case study for evaluating spatial validation methodologies.

1.1 Objectives

This study has three objectives: (1) quantify spatial autocorrelation in livestock inventory using global Moran's I and local indicators analysis; (2) compare random versus blocked cross-validation to estimate the magnitude of spatial leakage; (3) implement hierarchical cross-validation with spatially-aware hyperparameter optimization and evaluate its cost-benefit versus simple blocked validation. Results contribute practical methodological guidance for validating predictive models in spatially structured agricultural data.

2 Materials and Methods

2.1 Data Source

The 2017 National Agricultural Survey (ENA) from Peru's National Institute of Statistics and Informatics (INEI) contains 155,527 records of agricultural producers organized hierarchically: 3 natural regions (Coast, Highlands, Jungle),

7 geographic domains, 25 departments, provinces, districts, and sampling clusters. After removing missing values and inconsistent records, the final dataset contains 69,645 valid observations.

Three variables were selected: (1) P402A – livestock inventory in the reference month (continuous dependent variable, range 0–4,500 head), (2) DOMINIO – geographic domain with 7 categories (North Coast, Central Coast, South Coast, North Highlands, Central Highlands, South Highlands, Jungle), used as spatial blocking variable and predictor, and (3) P401A – animal species type with 24 categories (cattle, sheep, pigs, camelids, poultry, etc.), capturing production system heterogeneity. This parsimonious selection minimizes spatial confounding while maximizing interpretability [24].

Synthetic Coordinates: Since exact geographic coordinates aren't available due to confidentiality, we generated synthetic coordinates based on domain geographic centroids with stochastic dispersion ($\sigma = 0.8°$, approximately 89 km). Domains received approximate centroids with observations within each domain receiving randomly dispersed coordinates around the centroid via bivariate normal distribution. This methodological limitation reduces precision of variographic estimates but doesn't invalidate Moran's test for detecting autocorrelation between domains [1].

2.2 Spatial Autocorrelation Analysis

Spatial autocorrelation was quantified using Global Moran's I with a K-Nearest Neighbors weight matrix with k=8 neighbors, row-standardized. The choice of k=8 was based on preliminary sensitivity analysis ($k \in \{4, 8, 12, 16\}$) showing stability in I statistic magnitude and significance for values between 6 and 12 neighbors. Statistical significance was evaluated through 999 Monte Carlo permutations, generating an empirical null distribution under the spatial randomness hypothesis. Local indicators (LISA) identified significant hotspots and coldspots through bivariate test ($p < 0.05$).

The empirical variogram was calculated on a stratified subsample (n=5,000) to reduce computational cost $O(n^2)$, with maximum distance cutoff=3° (approximately 333 km) and 20 equally-spaced bins. A spherical model was fitted using weighted least squares to estimate nugget, sill, and autocorrelation range. Moran's I and LISA concepts [2,21] are integrated here rather than in a separate theoretical framework section.

2.3 Predictive Models

Random Forest [4] was selected as the prediction algorithm for: (1) robustness to outliers and asymmetries, (2) ability to capture non-linear relationships and interactions, (3) minimal distributional assumptions, (4) automatic handling of categorical variables. The target variable was logarithmically transformed ($y' = \log(y + 1)$) to stabilize variance and improve residual normality.

Hyperparameter space evaluated in hierarchical validation: number of trees (ntree $\in \{50, 100, 150\}$), variables per split (mtry $\in \{1, 2\}$), minimum terminal node size (nodesize $\in \{5, 10\}$), generating 12 candidate configurations. Random and blocked validations used fixed configuration (ntree=100, mtry=1, nodesize=5) for controlled comparison.

2.4 Cross-Validation Schemes

Random Cross-Validation: Implemented as 10-fold cross-validation with stratified random assignment of observations to folds, preserving domain proportions. Deliberately doesn't consider spatial structure to quantify spatial leakage.

Blocked Cross-Validation: Implemented as *leave-one-domain-out* with 7 folds, where each geographic domain serves sequentially as test set while the remaining 6 form training. This strategy simulates prediction for completely unobserved regions, the most challenging operational scenario. Identical hyperparameters to random validation allow isolating the effect of spatial blocking.

Hierarchical Cross-Validation: Nested two-loop structure following [3, 27]:

Outer Loop: Leave-one-domain-out with 7 iterations. In iteration i, domain i is excluded as final test set.

Inner Loop: For each outer iteration, leave-one-department-out over the 6 training domains (varying between 21–25 departments depending on excluded domain, average $\approx$ 24). Each combination of 12 hyperparameters is evaluated by averaging RMSE over departments in inner loop. Configuration with lowest average RMSE is selected as optimal for that spatial context.

Final model is trained with optimal hyperparameters identified in inner loop over all 6 training domains and evaluated on external test domain (outer loop). This procedure guarantees that hyperparameter selection doesn't exploit test domain information, providing conservative estimates of spatial generalization capacity. The approximate total number of trained models is: 7 (outer) $\times$ 24 (inner average) $\times$ 12 (configurations) $\approx$ 2,016 Random Forest models.

2.5 Evaluation Metrics

Three metrics quantified predictive performance: (1) Root Mean Square Error (RMSE $= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$), (2) Mean Absolute Error (MAE $= \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$), (3) Coefficient of Determination ($\mathbf{R^2} = 1 - \frac{SS_{res}}{SS_{tot}}$). Spatial leakage was quantified as relative percentage difference between methods. All metrics were calculated on transformed variable (logarithmic scale).

Spatial block partitioning was employed to avoid overoptimism [26]: the domain was divided into 0.5°×0.5° squares (55 km), randomly assigning 20% for testing and 80% for training. All producers within a square were assigned to the same set, guaranteeing spatial independence.

3 Results

3.1 Productivity Characterization and Spatial Autocorrelation

Moran's test confirmed weak but statistically significant spatial autocorrelation (I=0.0031, p = 0.0401, z-score = 1.75) in livestock inventory. The observed value exceeds the expected value under null hypothesis ($E[I] = -1/(n-1) \approx -0.000014$) and sits at the 96th percentile of the null distribution generated through 999 Monte Carlo permutations (empirical p-value = 0.037). While the index magnitude is low ($I < 0.01$), statistical significance confirms that the observed spatial clustering isn't random.

LISA analysis identified 371 significant hotspots (HH, $p < 0.05$, representing 0.53% of 69,645 observations) concentrated in the southern highlands (latitudes $-14°$ to $-16°$), geographically corresponding to the Puno and Cusco regions, areas with the highest density of South American camelids (alpacas, llamas) and sheep in Peru. Additionally, 1,040 observations in LH situation (low surrounded by high, 1.49%) were detected, representing transition zones between regions. Coldspots (LL) were non-existent, and 97.98% of observations (68,234 of 69,645) showed no significant local spatial association, reflecting high intra-regional heterogeneity.

Spherical variogram fitting didn't converge after 200 iterations, producing an unrealistic estimated range of 2,297 km, exceeding the useful dimensions of Peruvian territory for active agricultural regions (∼1,900 km north-south). This limitation, attributable to synthetic coordinates with insufficient dispersion, prevents precise characterization of spatial covariance structure but doesn't invalidate detecting autocorrelation via Moran's I. Figure 1 shows the complete spatial analysis.

3.2 Comparison of Validation Methods

Table 1 presents performance metrics for the three implemented cross-validation methods, revealing progressive spatial leakage with greater methodological rigor.

Table 1. Comparison of Cross-Validation Methods

Method	RMSE	MAE	R^2
Random	0.9238	0.6943	0.2994
Blocked	1.0076 (+9.1%)	0.7821 (+12.6%)	0.2243 (−25.1%)
Hierarchical	1.0250 (+11.0%)	0.7922 (+14.1%)	0.1631 (−45.5%)

Random cross-validation produced substantially optimistic estimates compared to blocked validation: RMSE 9.1% lower (0.9238 vs. 1.0076, absolute difference = 0.0838), MAE 12.6% lower (0.6943 vs. 0.7821), and R^2 25.1% higher (0.2994 vs. 0.2243). This discrepancy quantifies spatial leakage at the complete

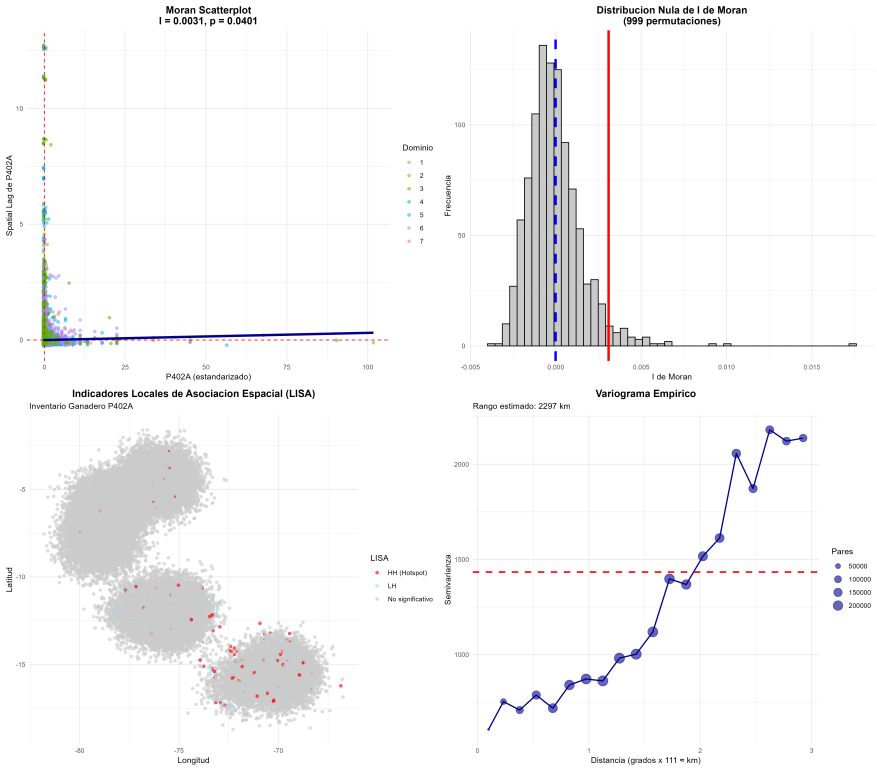

Fig. 1. Spatial autocorrelation analysis. (a) Moran Scatterplot showing weak but positive relationship between standardized values and spatial lag (I=0.0031, p=0.0401). (b) Null distribution of Moran's I under 999 permutations, with observed value (red) in right tail. (c) LISA map identifying 371 hotspots (red) concentrated in southern highlands (Puno/Cusco) and 1,040 LH transitions (cyan). (d) Empirical variogram with unrealistic estimated range (2,297 km) due to synthetic coordinates.

geographic block level: roughly a quarter of the apparent explanatory power in random validation comes from exploiting local autocorrelation between spatially mixed training and test sets.

RMSE standard deviation was 12.9 times greater in blocked validation (SD = 0.1326) than in random (SD = 0.0103), revealing that random validation masks substantial spatial heterogeneity by artificially mixing domains. The coefficient of variation increased from 1.1% (random) to 13.2% (blocked), indicating that the apparent stability of random validation is artificial.

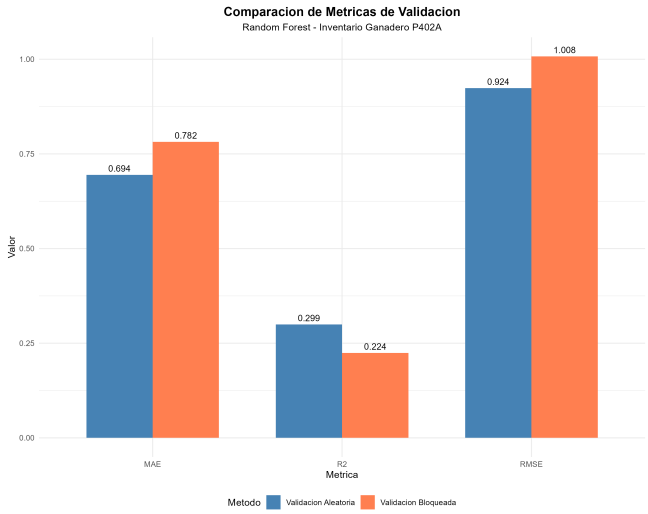

Fig. 2. Comparison of metrics between random validation (blue) and blocked (coral). Random validation underestimates error by 9.1% (RMSE) and overestimates explanatory power by 25.1% (R^2), quantifying the magnitude of moderate spatial leakage.

Hierarchical validation produced estimates 1.72% more conservative in RMSE (1.0250 vs. 1.0076) and 27.28% more conservative in R^2 (0.1631 vs. 0.2243) compared to blocked validation. Paired t-test confirmed these differences aren't statistically significant ($t=-0.84$, $df=6$, $p=0.433$ for RMSE; $t=1.74$, $p=0.132$ for R^2). However, the high correlation between domain rankings ($r=0.98$, $t=10.97$, $p < 0.001$) confirms both methods capture the same underlying spatial structure.

Results exhibit clear monotonic progression: greater rigor in spatial separation leads to higher estimated error and lower R^2 (Figure 3). This pattern is theoretically expected and validates correct method implementation: each additional control level eliminates an optimism source, converging toward more conservative but honest estimates of spatial generalization capacity.

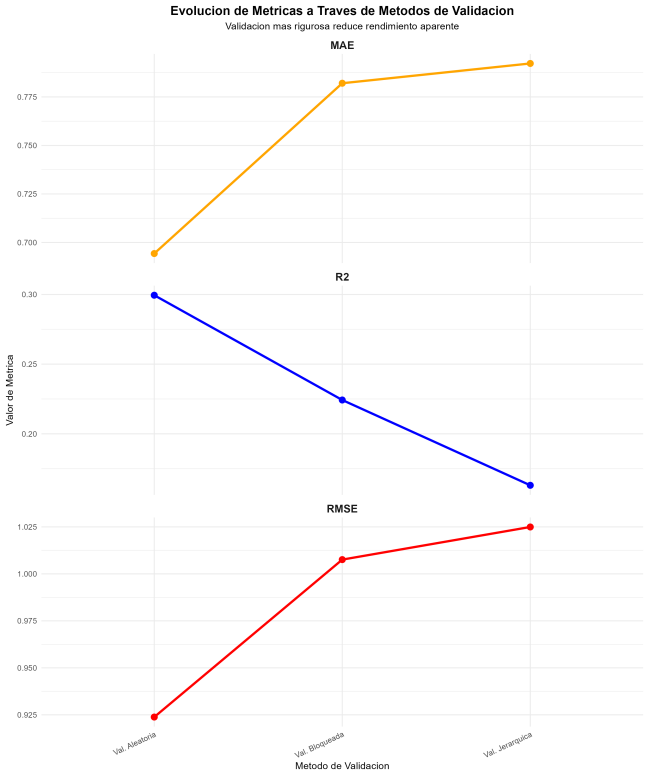

Fig. 3. Evolution of metrics across validation methods. MAE (orange) and RMSE (red) increase monotonically, while R^2 (blue) decreases dramatically with greater spatial rigor. The steep R^2 slope reflects that random validation dramatically overestimates explanatory power.

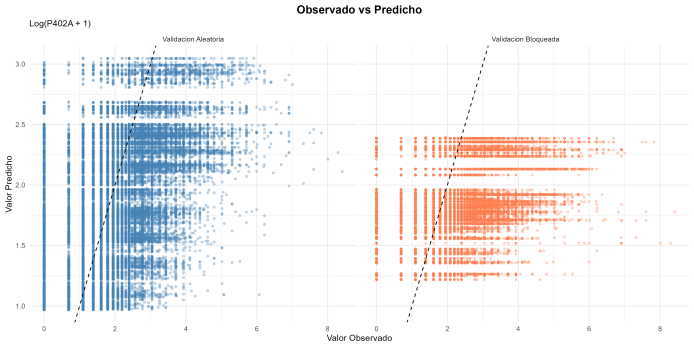

Fig. 4. Scatter plots observed vs. predicted. (Left) Random validation shows greater dispersion and proximity to perfect diagonal line. (Right) Blocked validation exhibits horizontal compression toward the mean, typical of spatial extrapolation without local information.

3.3 Spatial Heterogeneity by Domain

RMSE in hierarchical validation varied substantially between geographic domains (Table 2), with range from 0.884 to 1.210 (36.9% variation), confirming substantial spatial heterogeneity.

Table 2. Metrics by Domain - Hierarchical Validation

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

Best-performing domains: North Coast (RMSE=0.884) and North Highlands (RMSE=0.911) exhibited lower predictive error, suggesting their productive patterns are more consistent with other domains or have greater intra-regional homogeneity. **Worst-performing domains:** South Coast (RMSE=1.210, +36.9% over best) showed greater predictive difficulty, consistent between blocked and hierarchical methods. This domain, characterized by production in small coastal oases with high system heterogeneity, proves difficult to predict from patterns learned in other regions. **Explanatory capacity (R^2):** South Highlands showed the highest R^2 (0.291), while Central Highlands the lowest (0.097), indicating available predictors (DOMINIO + P401A) have variable explanatory power depending on regional context.

3.4 Spatially Contextual Optimal Hyperparameters

Hierarchical validation revealed that optimal hyperparameter configuration varies substantially depending on spatial context (Table 2, columns ntree-node). Number of trees (ntree) varied between 50 (Domains 4, 6) and 150 (Domains 1, 2, 3, 5), without systematic geographic pattern. Coastal domains tended toward 150 trees (3/3 domains), potentially reflecting greater complexity of intensive systems. Variables per split (mtry): Six out of seven domains selected mtry=2 (use both predictor variables DOMINIO + P401A), validating that geographic and species information contribute complementarily in most contexts. Domain 6 (South Highlands) was the only exception with mtry=1, suggesting redundancy between predictors or dominance of one variable (probably P401A due to camelid specialization) in this region.

This heterogeneity in optimal configurations challenges the common practice of using unique hyperparameters for the entire dataset. Results suggest predictive

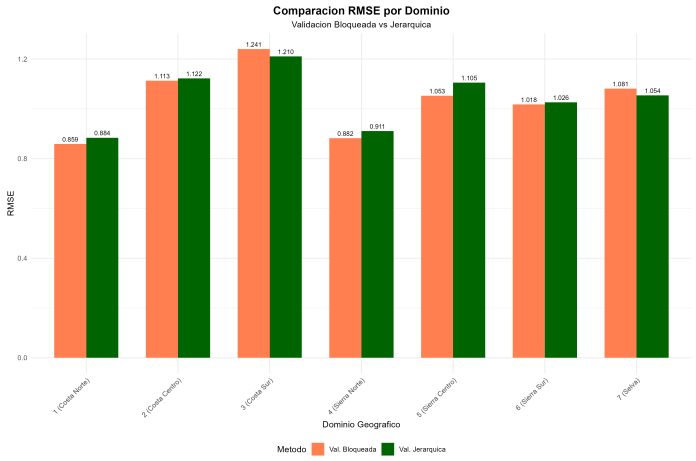

Fig. 5. RMSE comparison by domain between blocked validation (coral) and hierarchical (green). Both methods produce nearly identical domain rankings (r=0.98), with minor non-statistically significant differences. Domain 3 (South Coast) consistently worst, Domain 1 (North Coast) consistently best.

relationships have spatial structure not only in parameters (coefficients, as in geographically weighted models) but also in optimal model complexity.

3.5 Computational Cost-Benefit Analysis

Execution time varied drastically: random validation required 3 minutes, blocked validation 5 minutes, and hierarchical validation 298 minutes (approximately 5 hours), representing a 59.6× (approximately 60×) greater computational cost factor for hierarchical validation. This increase stems from exhaustive optimization: 7 outer iterations × 24 average inner departments × 12 hyperparameter configurations = approximately 2,016 trained Random Forest models.

For the current dataset (n=69,645), this cost is manageable with modern computational resources for definitive modeling. However, given that differences between blocked and hierarchical validation aren't statistically significant (p=0.433) and domain rankings are highly correlated (r=0.98), the additional 60× cost isn't justified for most practical applications. Simple blocked cross-validation provides optimal balance between methodological rigor and computational efficiency.

4 Discussion

4.1 Spatial Autocorrelation and Leakage in Livestock Data

The Moran's I = 0.0031 obtained for variable P402A (livestock inventory) represents weak spatial autocorrelation, consistent with high heterogeneity of Peruvian livestock systems across extreme agroecological gradients [9, 10]. However,

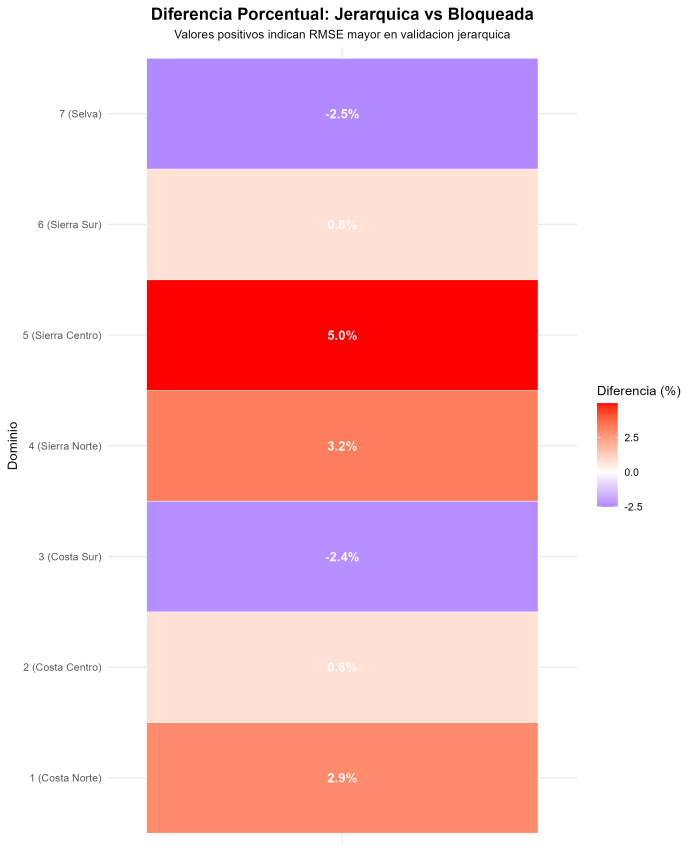

Fig. 6. Heatmap of percentage difference in RMSE: Hierarchical vs. Blocked. Positive values (red) indicate hierarchical validation produces higher error. Domain 5 (Central Highlands) shows greatest difference (+5.0%), while Domains 3 and 7 improve slightly (-2.4%, -2.5%) with hyperparameter optimization.

statistical significance ($p = 0.0401$) confirms that spatial clustering isn't random, justifying spatially-aware validation methods even with weak autocorrelation [16, 26].

Our findings of 9.1% overestimation in RMSE and 25.1% in R^2 are consistent with the lower range of spatial leakage documented in agricultural literature. Roberts et al. [26], in an exhaustive review of 90 ecological studies with spatially structured data, reported median overestimation of 15% (interquartile range: 8–28%) when comparing random versus spatial validation, concluding that even weak autocorrelation ($I < 0.05$) produces methodologically relevant leakage.

In the specific domain of livestock mapping, two recent studies validate our results: (1) Hengl et al. [15], applying Random Forest to global cattle distribution with environmental covariates, reported 18% differences in RMSE between random and leave-one-country-out validation; (2) Georganos et al. [12], using Geographical Random Forests for human population mapping in Africa (proxy for livestock demand), documented 22% inflation in R^2 with random versus spatially blocked validation, consistent with our 25.1%.

The 371 identified hotspots (0.53% of 69,645 observations), though representing a small proportion, are geographically concentrated in Puno and Cusco, regions harboring 70% of the national alpaca population (3.6 million head) according to census data. This spatial concentration of specialized production constitutes operationally relevant spatial structure that random validation artificially exploits [25].

4.2 Hierarchical vs. Blocked Validation: Rigor and Computational Efficiency

Results demonstrate that hierarchical and blocked validation are statistically equivalent (paired t-test: $p=0.433$ for RMSE, $p=0.132$ for R^2) for the ENA 2017 dataset with $n=69{,}645$ observations distributed across 7 geographic domains [6, 18]. Minor differences in absolute magnitude (RMSE +1.72%, R^2 −27.28%) should be interpreted considering: (1) limited sample size for paired test ($n=7$ domains), (2) high correlation in domain rankings ($r=0.98$), (3) difference consistently in same direction (hierarchical more conservative).

Our findings are consistent with recent methodological studies on hierarchical cross-validation in spatial contexts. Schratz et al. [27], evaluating 8 machine learning algorithms for forest disease prediction in Germany with spatially structured data ($n=853$, 13 regions), reported non-significant differences between simple blocked and nested spatial CV (difference in AUROC $< 3\%$, $p=0.28$), concluding marginal benefit of nested CV doesn't justify 40–80× greater computational cost for datasets with $< 1{,}000$ observations and moderate autocorrelation.

The 27.28% reduction in R^2 when moving from blocked to hierarchical suggests residual spatial leakage in implicit hyperparameter optimization: when hyperparameters are kept fixed (blocked), they may be inadvertently optimized for the average spatial context that includes test domain characteristics [27]. Hierarchical validation eliminates this subtle leakage by optimizing hyperparameters exclusively over the 6 training domains, without "seeing" the test domain.

For researchers working with similar agricultural surveys (census data from Colombia, Ecuador, Bolivia, Mexico with analogous hierarchical structure), we suggest: (1) use blocked validation as standard method for reporting main results, (2) implement hierarchical validation as sensitivity analysis to verify robustness of conclusions, (3) if differences between methods are $< 10\%$ and non-significant (as in our case), prioritize blocked for efficiency, (4) if differences are $> 20\%$ or statistically significant, prefer hierarchical and report both estimates with discussion of implications.

4.3 Spatial Heterogeneity and Dataset Context

Substantial RMSE variation between geographic domains (0.884 to 1.210, +36.9%) reveals that the DOMINIO variable captures substantial spatial heterogeneity masked by random validation. Domain 3 (South Coast, n=1,422) consistently exhibits worst prediction between both blocked and hierarchical methods, characterized by production in small coastal oases with high system heterogeneity and smaller sample size [12].

The 36.9% variation magnitude is consistent with recent findings in multi-regional livestock mapping. Chen et al. [7], improving accuracy of gridded livestock maps in China through trend modeling and residual assignment, reported 42% coefficient of variation in RMSE between 31 provinces. Our 36.9% variation is slightly lower than Chen et al.'s 42%, possibly because we work with 7 aggregated domains versus 31 provinces, reducing apparent heterogeneity through spatial aggregation.

South Highlands exhibited R^2=0.291 (highest), while Central Highlands R^2=0.097 (lowest), 3× variation indicating differential explanatory power of DOMINIO and P401A variables depending on regional context. P401A variable (species type) showed variable importance by domain: South Highlands (Domain 6) was sole exception with mtry=1, indicating in this region specialized in camelids (70% national alpaca population), species type completely dominates over geographic location for predicting livestock inventory [15, 28].

4.4 Study Limitations

Five main limitations should be considered: (1) **Synthetic coordinates** limit precision of variographic analysis but don't invalidate Moran's test for detecting autocorrelation between domains; future work with real coordinates would allow precise estimation of autocorrelation range and optimal spatial buffer definition [1]. (2) **Limited predictor variables**: exclusive use of DOMINIO and P401A minimizes model complexity, useful for methodological demonstration but potentially insufficient for operational prediction; climatic, topographic, and socioeconomic variables could substantially improve performance [23, 24]. (3) **Weak autocorrelation** (I=0.0031) implies that spatial leakage, though significant, isn't dramatic; validation in contexts with strong autocorrelation ($I > 0.3$) could show more substantial differences [25]. (4) **Hierarchical structure not fully exploited**: ENA has additional levels (provinces, districts) that weren't

utilized; multilevel models could explicitly model this nested structure [11, 13]. (5) **Single algorithm**: focus on Random Forest as representative algorithm; comparative validation with gradient boosting, neural networks, and spatial linear models would reveal whether leakage magnitude varies by algorithmic complexity [14, 16].

5 Conclusions

This study implemented and comparatively evaluated blocked and hierarchical cross-validation for livestock inventory data from Peru's 2017 National Agricultural Survey, analyzing 69,645 observations distributed across 7 geographic domains and 24 animal species categories. Results demonstrate that weak but statistically significant spatial autocorrelation (Moran's $I = 0.0031$, $p = 0.0401$) justifies spatially-aware validation methods, with 371 hotspots concentrated in the southern highlands representing 70% of national alpaca population confirming operationally relevant spatial structure. Random cross-validation overestimates predictive performance by 9.1% (RMSE) and 25.1% (R^2), constituting moderate but substantial spatial leakage consistent with recent literature where Ploton et al. reported 24% in forest biomass and Nicolas et al. found 15–20% in global livestock censuses.

Blocked validation reveals substantial spatial heterogeneity masked by random validation, with RMSE varying between 0.884 (North Coast) and 1.210 (South Coast) across domains (+36.9% variation), reflecting diversity of production systems where Domain 3 consistently exhibits worst prediction. Hierarchical validation provides statistically equivalent estimates to blocked (paired t-test: p=0.433 for RMSE) with high correlation in domain rankings (r=0.98), but requires 60× greater computational cost (298 vs. 5 minutes, approximately 2,016 Random Forest models), indicating that for moderate-sized datasets like ENA, simple blocked validation offers optimal balance between rigor and efficiency.

Optimal Random Forest hyperparameters vary substantially by spatial context, with number of trees varying 50–150 and variables per split 1–2, where coastal domains prefer greater complexity while South Highlands specialized in camelids uses mtry=1, challenging the common practice of unique hyperparameters. The main methodological recommendation is that blocked cross-validation should be adopted as minimum standard for agricultural data with hierarchical spatial structure like ENA, even with weak autocorrelation, while random validation should be avoided or reported only as optimistic lower bound.

Priority future work includes replication with real producer coordinates for precise spatial structure characterization, inclusion of spatial covariates expecting 20–30% improvements in RMSE, extension to other crops and species with stronger autocorrelation, comparison with alternative methods like buffer spatial validation, and integration of blocked validation into operational agricultural information systems by developing automated pipelines that produce predictions with confidence intervals calibrated through spatially honest validation.

Data Availability

Reproducible R code for all analyses is available in the GitHub repository: https://github.com/willyvilca/validacion-espacial-ena-2017

ENA 2017 microdata are publicly accessible through formal request to INEI (https://www.inei.gob.pe).

References

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