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] and Fred Torres Cruz[2][0000-0003-0834-6834] and Edgar Eloy Carpio Vargas[3][0000-0001-6457-4597]

[1] National University of the Altiplano, School of Statistical and Computer Engineering, Puno, Perú
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 provides statistically equivalent estimates but is computationally more expensive (60× longer runtime). RMSE by geographic domain varies between 0.884 and 1.210 (36.9% variation), revealing substantial spatial heterogeneity. Findings validate the need for spatially-aware validation methods and demonstrate that simple blocked cross-validation provides optimal balance between rigor and 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 manifests through technology diffusion among neighboring producers, shared agroecological conditions, and common market access [22]. However, conventional validation techniques assume independence between observations, systematically violating this inherent spatial structure [26].

Random k-fold cross-validation creates high probability that spatially close observations appear in both training and test sets [17], letting models exploit local autocorrelation for prediction and generating artificially optimistic performance estimates

known as *spatial leakage* [25]. Prior studies document that random validation can overestimate predictive accuracy by 10–50% [16,26]. Blocked cross-validation tackles this by partitioning data into spatially contiguous groups [31]. Hierarchical cross-validation extends blocking by implementing nested leave-one-group-out procedures at multiple spatial scales [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. This study has three objectives: (1) quantify spatial autocorrelation in livestock inventory using global Moran's I and local indicators; (2) compare random versus blocked cross-validation to estimate spatial leakage magnitude; (3) implement hierarchical cross-validation with spatially-aware hyperparameter optimization and evaluate cost-benefit versus simple blocked validation.

2 Materials and Methods

2.1 Data Source

The 2017 National Agricultural Survey (ENA) from Peru's INEI contains 155,527 records organized hierarchically: 3 natural regions, 7 geographic domains, 25 departments. After removing missing values, the final dataset contains 69,645 valid observations. Three variables were selected: (1) P402A – livestock inventory (continuous, range 0–4,500 head), (2) DOMINIO – geographic domain with 7 categories, used as spatial blocking variable and predictor, and (3) P401A – animal species type with 24 categories. Since exact coordinates aren't available, we generated synthetic coordinates based on domain centroids with stochastic dispersion ($\sigma = 0.8°$, ~89 km) [1].

2.2 Spatial Autocorrelation Analysis

Spatial autocorrelation was quantified using Global Moran's I with K-Nearest Neighbors weight matrix (k=8), row-standardized. Statistical significance was evaluated through 999 Monte Carlo permutations. Local indicators (LISA) identified significant hotspots ($p < 0.05$). The empirical variogram was calculated on a stratified subsample (n=5,000) with maximum distance 3° and 20 equally-spaced bins [2,21].

2.3 Predictive Models and Cross-Validation

Random Forest [4] was selected for robustness to outliers, ability to capture nonlinear relationships, and automatic handling of categorical variables. Target variable was logarithmically transformed ($y' = \log(y + 1)$).

Random Cross-Validation: 10-fold with stratified random assignment.
Block Cross-Validation: Leave-one-domain-out (7 folds).

Hierarchical Cross-Validation: Nested structure [3,27]: Outer loop (leave-one-domain-out, 7 iterations) and inner loop (leave-one-department-out over 6 training domains, ~24 departments). Each of 12 hyperparameter combinations evaluated by averaging RMSE. Total: ~2,016 Random Forest models trained.

3 Results

3.1 Spatial Autocorrelation

Moran's test confirmed weak but statistically significant spatial autocorrelation (I=0.0031, p=0.0401, z-score=1.75). LISA analysis identified 371 significant hotspots (0.53% of observations) concentrated in southern highlands (Puno/Cusco regions), areas with highest density of South American camelids. Additionally, 1,040 observations in LH situation (1.49%) represented transition zones. Spherical variogram fitting didn't converge, producing unrealistic range (2,297 km), attributable to synthetic coordinates. Figure 1 shows complete spatial analysis.

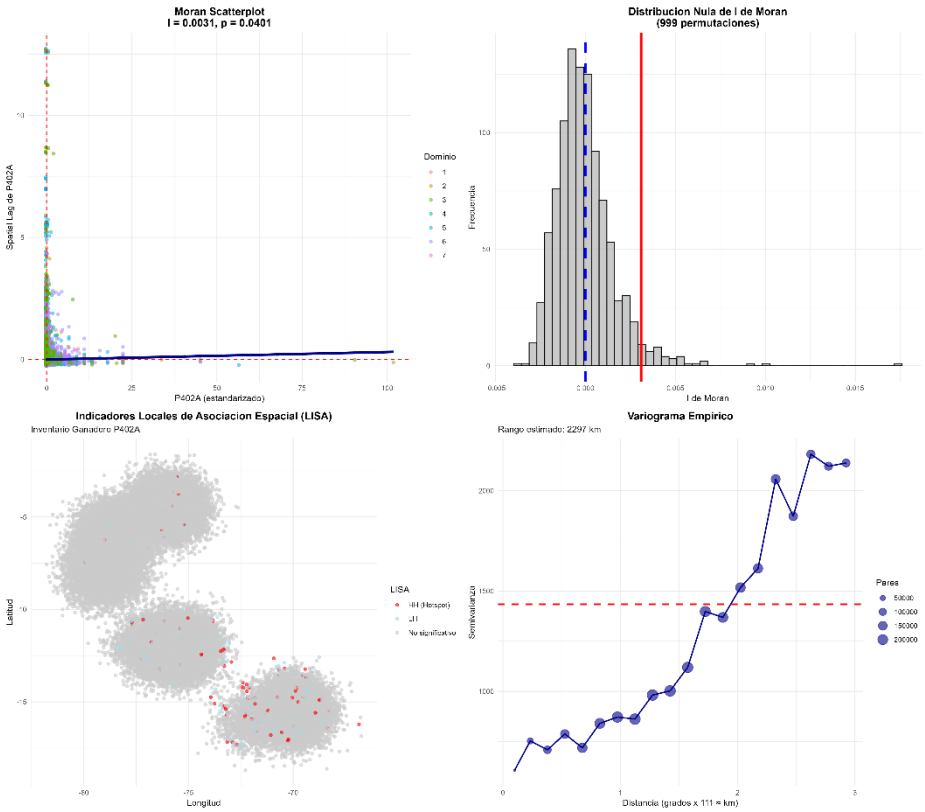

Fig. 1. Spatial autocorrelation analysis. (a) Moran Scatterplot (I=0.0031, p=0.0401). (b) Null distribution under 999 permutations. (c) LISA map: 371 hotspots (red) in

southern highlands and 1,040 LH transitions (cyan). (d) Empirical variogram with unrealistic range due to synthetic coordinates.

3.2 Comparison of Validation Methods

Table 1 presents performance metrics for the three cross-validation methods.

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: RMSE 9.1% lower, MAE 12.6% lower, and R^2 25.1% higher compared to blocked validation. This quantifies spatial leakage: roughly a quarter of apparent explanatory power comes from exploiting local autocorrelation. Hierarchical validation produced estimates 1.72% more conservative in RMSE (1.0250 vs. 1.0076) and 27.28% in R^2 (0.1631 vs. 0.2243) compared to blocked. Paired t-test confirmed these differences aren't statistically significant ($t=-0.84$, $p=0.433$ for RMSE). However, high correlation between domain rankings ($r=0.98$, $p<0.001$) confirms both methods capture the same spatial structure. Figures 2-4 illustrate these findings.

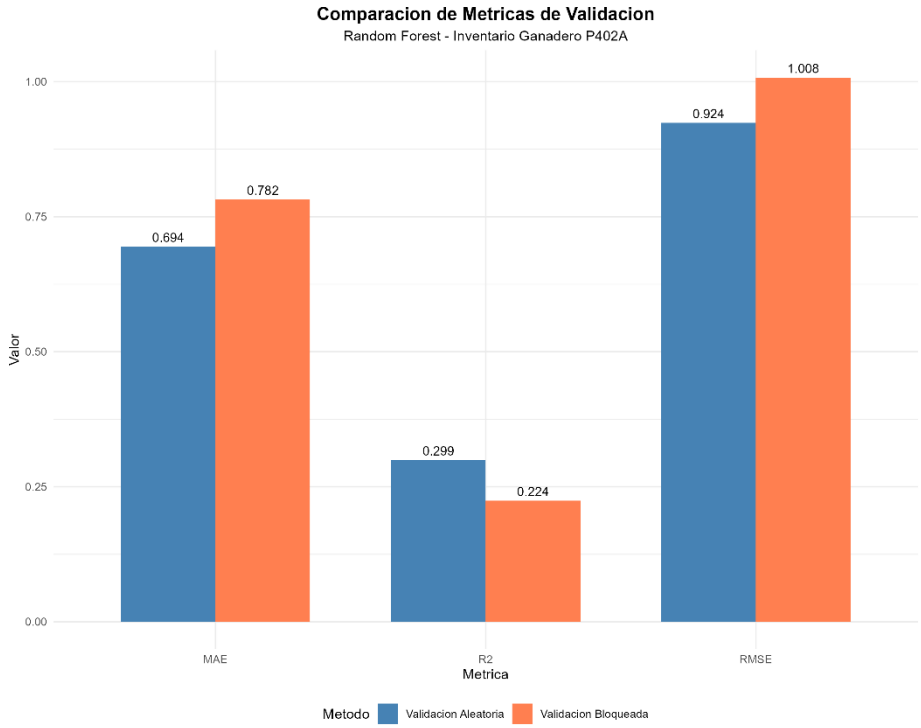

Fig. 2. Comparison of metrics between random (blue) and blocked (coral) validation. Random validation underestimates error by 9.1% (RMSE) and overestimates explanatory power by 25.1% (R^2).

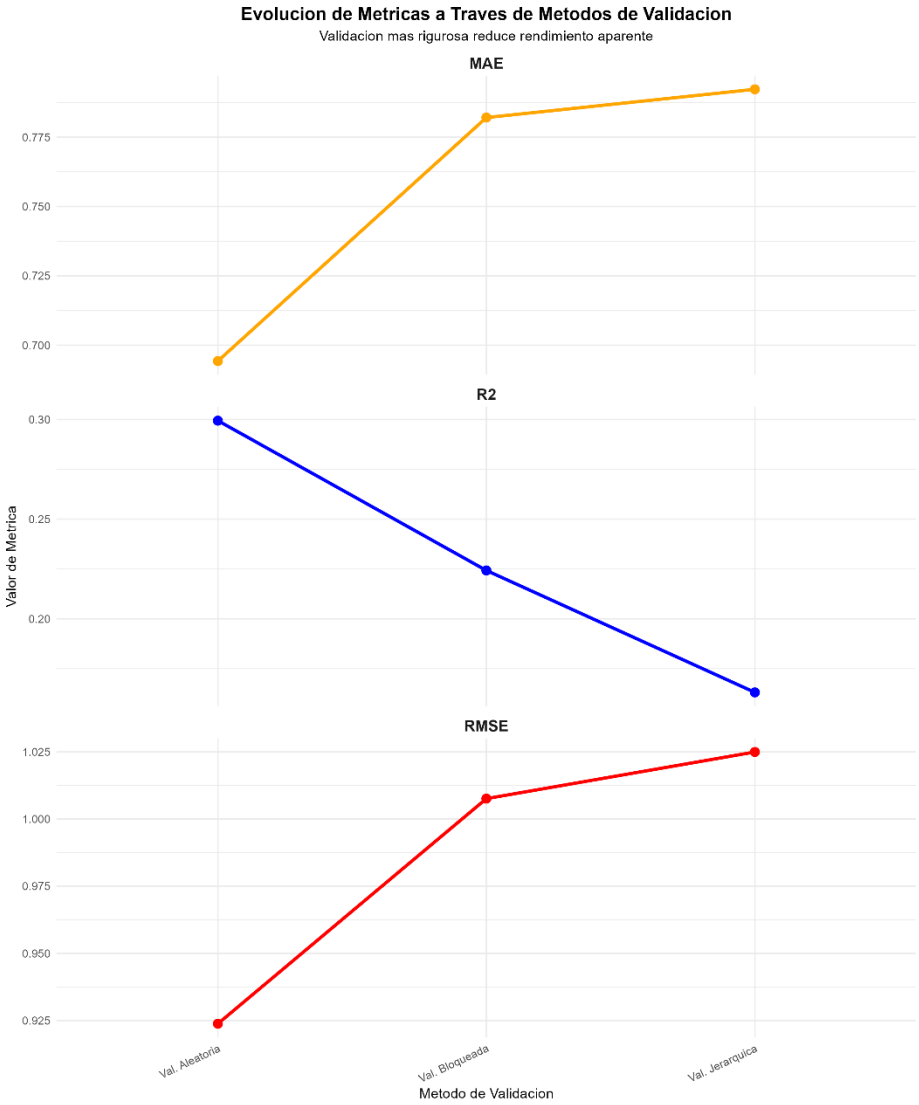

Fig. 3. Evolution of metrics across validation methods. MAE (orange) and RMSE (red) increase monotonically, while R^2 (blue) decreases dramatically with greater spatial rigor.

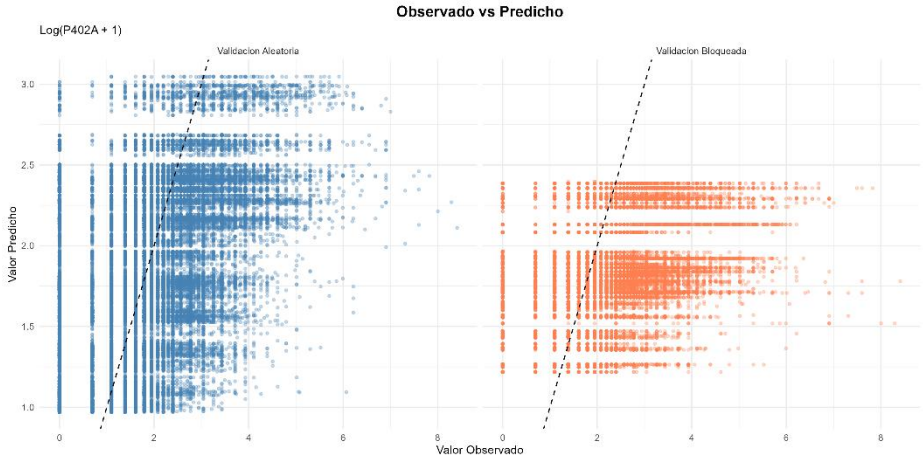

Fig. 4. Scatter plots observed vs. predicted. (Left) Random validation shows greater dispersion. (Right) Blocked validation exhibits horizontal compression toward mean, typical of spatial extrapolation.

3.3 Spatial Heterogeneity by Domain

RMSE in hierarchical validation varied substantially between domains (Table 2), with range from 0.884 to 1.210 (36.9% variation), confirming substantial spatial heterogeneity. North Coast (0.884) and North Highlands (0.911) exhibited lower error, while South Coast (1.210, +36.9%) showed greatest difficulty. Optimal hyperparameters varied by context: coastal domains preferred 150 trees, while six of seven domains selected mtry=2. Execution time varied drastically: random (3 min), blocked (5 min), hierarchical (298 min, 60× factor). Given non-significant differences (p=0.433) and high correlation (r=0.98), the 60× cost isn't justified. Figures 5-6 illustrate domain-level patterns.

Table 2. Metrics by Domain – Hierarchical Validation

Domain	RMSE	MAE	R^2	n
1 (North Coast)	0.884	0.675	0.108	6,974
4 (North Highlands)	0.911	0.705	1.171	12,296
6 (South Highlands)	1.026	0.792	0.291	14,042
7 (Jungle)	1.054	0.819	0.149	14,256
5 (Central Highlands)	1.105	0.856	0.097	18,631
2 (Central Coast)	1.122	0.868	0.130	2,024
3 (South Coast)	1.210	0.937	0.159	1,422

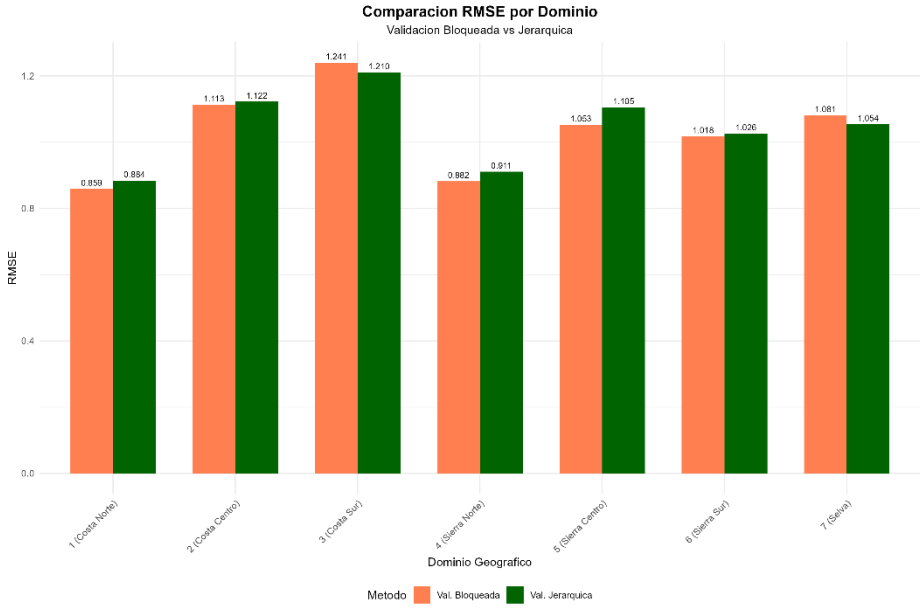

Fig. 5. RMSE comparison by domain between blocked (coral) and hierarchical (green). Both methods produce nearly identical rankings (r=0.98).

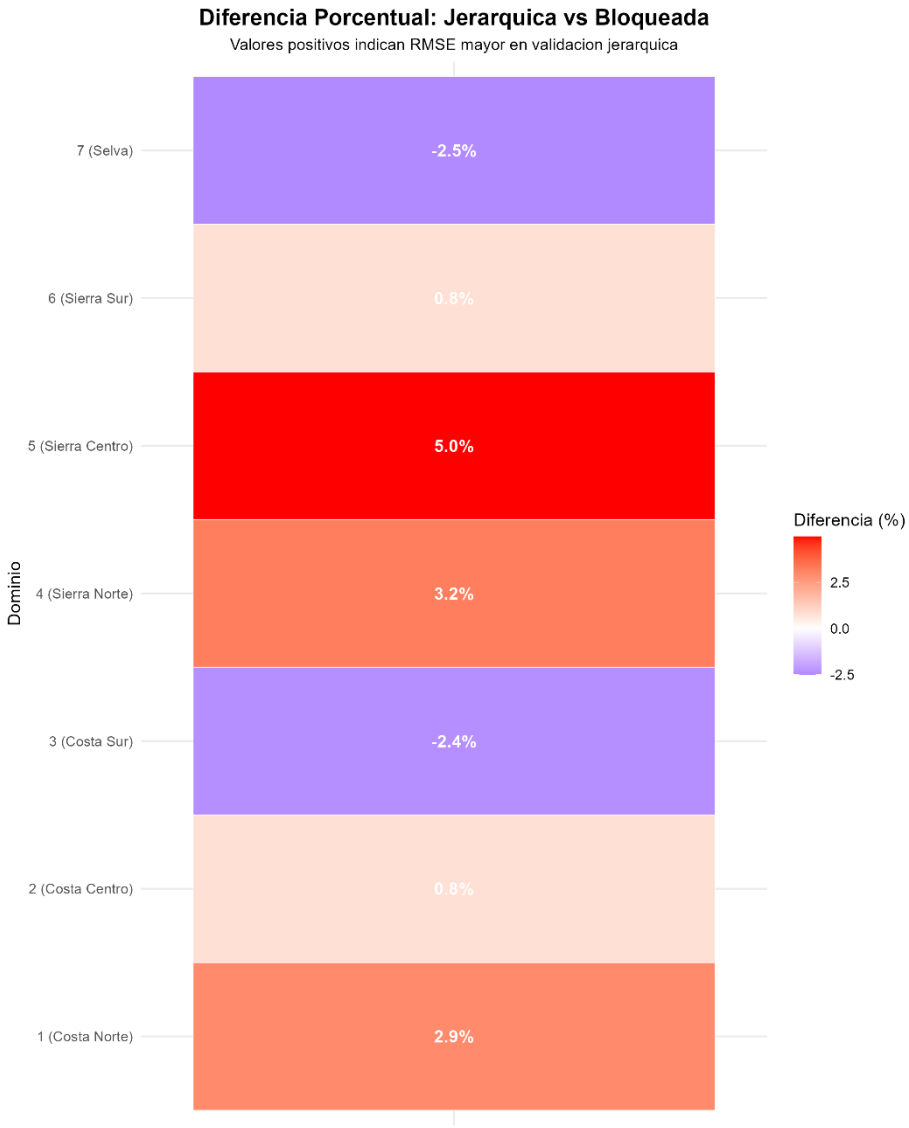

Fig. 6. Heatmap of percentage difference in RMSE: Hierarchical vs. Blocked. Positive values (red) indicate hierarchical produces higher error. Domain 5 shows greatest difference (+5.0%).

4 Discussion

Our Moran's I = 0.0031 represents weak spatial autocorrelation, consistent with heterogeneity of Peruvian livestock systems across extreme agroecological gradients [9,10]. However, statistical significance (p=0.0401) justifies spatially-aware valida-

tion even with weak autocorrelation [16,26]. Our 9.1% RMSE overestimation and 25.1% R^2 inflation are consistent with agricultural literature: Roberts et al. [26] reported median 15% overestimation (IQR: 8–28%), Hengl et al. [15] found 18% differences in global cattle mapping, and Georganos et al. [12] documented 22% R^2 inflation. The 371 hotspots (0.53%), though small proportion, are concentrated in Puno/Cusco (70% of national alpaca population), constituting operationally relevant structure that random validation exploits [25].

Hierarchical and blocked validation are statistically equivalent ($p=0.433$) with high correlation ($r=0.98$), consistent with Schratz et al. [27] who reported non-significant differences (AUROC<3%, $p=0.28$) and concluded marginal benefit doesn't justify 40–80× computational cost. The 36.9% RMSE variation between domains aligns with Chen et al. [7] who reported 42% coefficient of variation across Chinese provinces. For researchers with similar agricultural surveys (Colombia, Ecuador, Bolivia, Mexico), we suggest: (1) use blocked validation as standard, (2) implement hierarchical as sensitivity analysis, (3) if differences <10% and non-significant, prioritize blocked for efficiency.

Limitations: (1) Synthetic coordinates limit variographic precision but don't invalidate Moran's test [1]; (2) limited predictors (DOMINIO + P401A) useful for methodology but potentially insufficient for operations; (3) weak autocorrelation ($I=0.0031$) implies moderate leakage - stronger contexts may show larger differences [25]; (4) hierarchical structure not fully exploited - multilevel models could explicitly model nesting [11,13]; (5) single algorithm focus - comparative validation with other methods would reveal algorithm-specific patterns [14,16].

5 Conclusions

This study evaluated blocked and hierarchical cross-validation for Peru's 2017 National Agricultural Survey (n=69,645 observations, 7 geographic domains). Weak but statistically significant spatial autocorrelation (Moran's $I=0.0031$, $p=0.0401$) justifies spatially-aware validation, with 371 hotspots in southern highlands (70% of national alpaca population) confirming operationally relevant structure. Random cross-validation overestimates performance by 9.1% (RMSE) and 25.1% (R^2), constituting moderate spatial leakage consistent with literature [25,23]. Blocked validation reveals substantial heterogeneity (RMSE 0.884–1.210, 36.9% variation) masked by random validation. Hierarchical validation provides statistically equivalent estimates ($p=0.433$, $r=0.98$ correlation) but requires 60× greater computational cost (298 vs. 5 minutes), making simple blocked validation the optimal balance.

Optimal hyperparameters vary substantially by spatial context (trees: 50–150, mtry: 1–2), challenging uniform hyperparameter practice. Main methodological recommendation: blocked cross-validation should be adopted as minimum standard for agricultural data with hierarchical spatial structure, even with weak autocorrelation. Random validation should be avoided or reported only as optimistic lower bound.

Priority future work includes replication with real coordinates, inclusion of spatial covariates (expecting 20–30% RMSE improvements), extension to crops with stronger autocorrelation, comparison with buffer spatial validation, and integration into operational agricultural information systems via automated pipelines producing predictions with spatially-calibrated confidence intervals.

Data Availability

Reproducible R code: https://github.com/willyvilca/validacion-espacial-ena-2017. ENA 2017 microdata: 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). https://doi.org/10.1111/j.1538-4632.1995.tb00338.x
3. Arlot, S., Celisse, A.: A survey of cross-validation procedures for model selection. Statistics Surveys 4, 40–79 (2010). https://doi.org/10.1214/09-SS054
4. Breiman, L.: Random forests. Machine Learning 45(1), 5–32 (2001). https://doi.org/10.1023/A:1010933404324
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. IEEE, Munich (2012). https://doi.org/10.1109/IGARSS.2012.6352393
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). https://doi.org/10.1111/j.2007.0906-7590.05171.x
10. Fortin, M.-J., Dale, M.R.T.: Spatial Analysis: A Guide for Ecologists. Cambridge University Press, Cambridge (2005). https://doi.org/10.1017/CBO9780511542039
11. Gelman, A., Hill, J.: Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press, Cambridge (2007). https://doi.org/10.1017/CBO9780511790942
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). https://doi.org/10.32614/RJ-2016-014
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). https://doi.org/10.1080/00949655.2012.741599

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. Morgan Kaufmann, 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). https://doi.org/10.1111/geb.12161
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). https://doi.org/10.1093/biomet/37.1-2.17
22. Mueller, N.D., et al.: Closing yield gaps through nutrient and water management. Nature 490(7419), 254–257 (2012). https://doi.org/10.1038/nature11420
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). https://doi.org/10.1214/10-STS326
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). https://doi.org/10.1111/j.2517-6161.1974.tb00994.x
30. Tobler, W.R.: A computer movie simulating urban growth in the Detroit region. Economic Geography 46(sup1), 234–240 (1970). https://doi.org/10.2307/143141

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