Research Statement

Spatial Analysis of Agricultural Production and Irrigation Practices in Peru

Using Multigaussian and Plurigaussian Geostatistical Simulation

Rosmery Luna Turpo
Universidad Nacional del Altiplano, Puno, Peru
Graduate Course: Spatial Statistics – Academic Year 2024-2025

1 Introduction

Peru's agricultural landscape exhibits remarkable spatial heterogeneity shaped by dramatic topographic gradients spanning from coastal deserts through Andean highlands to Amazonian rainforests [1]. Traditional statistical approaches often fail to capture the complex spatial dependencies inherent in agricultural systems, particularly when quantifying uncertainty in production patterns across such diverse environments [2]. This research applies advanced geostatistical simulation methods to microdata from the 2024 National Agricultural Survey, explicitly modeling spatial correlation structures while generating multiple equiprobable scenarios that honor observed data and quantify prediction uncertainty [3].

The study leverages two complementary simulation frameworks: Sequential Gaussian Simulation for continuous variables like yield and harvested area [4], and Plurigaussian simulation for categorical attributes such as irrigation types and crop species [5]. By integrating these methods with survey expansion factors and high-resolution environmental covariates, we aim to produce spatially explicit agricultural intelligence for evidence-based policy formulation [6]. Recent advances in computational geostatistics have made such large-scale analyses feasible through efficient algorithms and parallel processing capabilities [7].

2 Research Objectives

2.1 Main Objective

Develop a methodologically rigorous geostatistical framework combining multigaussian simulation for continuous agricultural variables with plurigaussian simulation for categorical management practices, preserving spatial interdependencies and providing robust uncertainty quantification across Peru's three natural regions.

2.2 Specific Objectives

- Establish reproducible preprocessing pipelines integrating survey microdata with segment-level georeferences and environmental covariates
- Estimate and model spatial correlation structures through variogram analysis, incorporating survey sampling weights
- Generate multiple spatial realizations through sequential simulation algorithms
- Produce spatially explicit uncertainty maps for agricultural productivity and irrigation infrastructure
- Deliver policy-relevant outputs aggregated to district and provincial scales

3 Data and Study Area

The 2024 National Agricultural Survey employs stratified multistage probability sampling covering 8,850 geographic segments across all 26 departments of Peru [8]. After rigorous quality control procedures that removed incomplete interviews and invalid coordinates, our analytical dataset comprises 81,095 producer-level observations with validated georeferences [9]. The survey integrates two modules: the household roster provides geographic coordinates with administrative codes and natural region classifications [10], while the agricultural production module documents detailed variables including harvested area, total output, crop species, and irrigation practices [11].

Our analysis focuses on key continuous variables including harvested area measured in hectares and production volume standardized to kilograms, allowing derivation of yield as the ratio of production to area [12]. These metrics reflect fundamental spatial drivers including soil quality, water availability, technology adoption, and market access that vary dramatically across

Peru's geography [13]. Categorical variables capture irrigation water sources spanning river diversions, groundwater extraction, reservoir storage, and rainfed systems, alongside irrigation technologies from traditional flood methods to modern drip and sprinkler systems [14]. The diversity of crop species exceeds 150 varieties ranging from traditional Andean staples to commercial exports, reflecting Peru's status as a global agricultural biodiversity hotspot [15].

To enhance predictions, we incorporate environmental covariates from multiple global datasets. SRTM provides 90-meter elevation data capturing Peru's extraordinary range from sea level to over 6,000 meters above sea level [16]. HydroSHEDS stream networks enable distance calculations to permanent water sources, critical for predicting irrigation feasibility in arid coastal valleys [17]. WorldClim 2.1 climate surfaces at 1-kilometer resolution characterize temperature and precipitation gradients spanning hyperarid coastal deserts receiving less than 50mm annual rainfall to humid Amazon forests with over 4,000mm [18]. Additional soil property maps from SoilGrids provide information on texture, organic carbon, and pH at 250-meter resolution [19].

4 Methodology

4.1 Multigaussian Simulation

The multigaussian approach treats continuous agricultural variables as realizations of spatially correlated Gaussian random fields, generating multiple equiprobable scenarios that honor observed data while quantifying uncertainty [2]. Raw agricultural data typically exhibit strongly skewed distributions incompatible with Gaussian modeling assumptions [20]. We therefore apply normal score transformation to convert skewed data to standard normal distribution, incorporating survey expansion factors to maintain population representativeness [21]. This transformation preserves rank order while ensuring compatibility with kriging algorithms [22].

Spatial correlation is characterized through variogram analysis quantifying how dissimilarity increases with geographic distance [9]. The experimental variogram is computed from all available point pairs with survey weights [23]:

$$\hat{\gamma}(h) = \frac{1}{2|N(h)|} \sum_{N(h)} w_i w_j [Y(s_i) - Y(s_j)]^2 \tag{1}$$

where w_i are survey weights ensuring population inference. We fit parametric Matérn covariance functions to the empirical variogram, estimating key parameters including partial sill representing total spatial variance, range indicating the distance at which spatial correlation effectively disappears, and nugget effect capturing measurement error plus microscale variability [24]. Directional variograms reveal potential anisotropy where spatial correlation strength varies with direction—a common pattern in agricultural landscapes where valleys exhibit stronger correlation along-axis than across topographic gradients [25].

For multivariate analysis where agricultural variables are correlated, we employ the Linear Model of Coregionalization that preserves cross-correlations between variables such as yield and irrigation type [26]. This ensures that simulated realizations maintain realistic relationships between different agricultural attributes [27]. Sequential Gaussian Simulation visits unsampled locations in random order, using simple kriging to derive local conditional distributions from which values are drawn and immediately added to conditioning data [4]. Repeating this process 100 times produces an ensemble of realizations whose mean provides expected value estimates and variance quantifies spatial uncertainty [28]. This stochastic approach provides more realistic uncertainty quantification compared to deterministic interpolation methods [29].

4.2 Plurigaussian Simulation

Categorical variables require specialized methods that preserve realistic spatial transitions between discrete classes while maintaining spatial continuity [5]. The plurigaussian framework operates by simulating underlying continuous Gaussian random fields and applying truncation rules that partition continuous space into discrete facies zones [30]. This approach was originally developed for geological modeling but has proven valuable for categorical agricultural attributes [31]. Target proportions for each category are estimated from survey data with proper weighting, and these proportions can vary spatially to reflect regional agricultural specialization patterns [32].

For variables with ordered classes, horizontal truncation of a single Gaussian field suffices to generate realistic spatial patterns [33]. More complex variables with non-hierarchical relationships require two independent Gaussian fields with oblique truncation boundaries that accommodate gradual spatial transitions characteristic of real agricultural landscapes [34]. Spatial continuity of each facies is characterized through indicator variography, which measures how spatial correlation of presence-absence indicators varies with distance [35]. Short-range indicator variograms suggest patchy, fragmented

distributions typical of smallholder agriculture, while long-range structures indicate clustered spatial patterns found in commercial farming regions [36].

Conditional simulation ensures that simulated Gaussian field values at observed data locations fall within the appropriate truncation zones corresponding to observed categories [37]. This conditioning process maintains fidelity to observed spatial patterns while allowing variation in unsampled areas. After truncation, we compute facies probability maps by tallying how frequently each category appears at each location across the 100 realizations, providing spatially explicit uncertainty quantification for categorical predictions [38]. These probability maps are particularly valuable for identifying transition zones where categorical assignments are most uncertain and additional data collection would be most beneficial [39].

4.3 Integration and Aggregation

Multigaussian and plurigaussian realizations are analyzed jointly to address conditional queries relevant for agricultural planning, such as estimating expected yield specifically within drip-irrigated areas [40]. This joint analysis requires careful bookkeeping to ensure corresponding realizations from continuous and categorical simulations are paired correctly [41]. Spatial aggregation to administrative units employs area-weighted averaging, with calculations repeated across all 100 realizations producing predictive distributions and Monte Carlo confidence intervals suitable for policy decision support [42]. This approach properly propagates spatial uncertainty into aggregate statistics, avoiding the false precision of deterministic methods [43].

5 Expected Outcomes

We will produce high-resolution prediction maps at 5-kilometer resolution covering yield, harvested area, and associated uncertainty metrics expressed as coefficient of variation [6]. Probability maps will characterize the spatial distribution of irrigation technologies and crop species, while uncertainty maps identify regions where predictions are most reliable versus areas requiring additional data collection [44]. All spatial products will be delivered as GeoTIFF rasters and ESRI shapefiles with comprehensive metadata documenting coordinate systems, processing workflows, and uncertainty characteristics following INSPIRE guidelines [45].

Model validation will employ leave-one-out cross-validation to assess predictive accuracy, supplemented by residual diagnostics examining spatial structure in prediction errors and sensitivity analyses testing robustness to modeling assumptions [46]. We will compare our geostatistical approach against simpler deterministic methods to quantify the value added by explicit uncertainty quantification [47]. The complete analytical pipeline will be documented in R Markdown, integrating data processing, variogram estimation, simulation algorithms, and visualization in a fully reproducible workflow archived with a permanent digital object identifier [48]. Policy synthesis reports will translate technical results into actionable recommendations identifying priority intervention zones, crop suitability assessments, and targeting strategies for agricultural extension services [1].

Our 16-week implementation timeline allocates four weeks each to: data acquisition, quality control, and exploratory spatial analysis; variogram modeling, anisotropy assessment, and fitting multivariate coregionalization structures; simulation implementation with careful convergence diagnostics and computational optimization; and joint analysis, map production, and completion of technical and policy-oriented reporting.

6 Computational Strategy and Challenges

Generating 100 realizations for 81,095 observations at 5-kilometer national resolution demands substantial computational resources, with memory requirements potentially exceeding available RAM on standard workstations [7]. We address this through parallel processing distributing realizations across multiple processor cores, adaptive neighborhood search limiting kriging to the 50 nearest neighbors to reduce computational burden, and sparse matrix approximations for large covariance structures [41]. Recent developments in scalable geostatistical algorithms make such large-scale analyses increasingly feasible [49].

Positional uncertainty arises because survey coordinates represent segment centroids rather than precise field boundaries, introducing spatial error that can propagate into predictions [50]. We will handle this through controlled spatial jittering within 250-meter radius and comprehensive sensitivity analysis to assess how positional error affects final results [23]. Survey design effects violate simple random sampling assumptions underlying classical geostatistics, requiring us to employ weighted variogram estimators that properly incorporate expansion factors to ensure valid population inference [8].

Missing data patterns include both systematic absences where certain crops are biologically incompatible with specific

regions, and sporadic non-response from survey participants [51]. We will address this through facies-stratified analysis that handles systematic patterns and model-based imputation leveraging spatial correlation to predict missing values in a statistically principled manner [52]. Finally, data privacy compliance with national statistical agency protocols necessitates spatial blurring at fine resolutions and cell suppression rules to prevent re-identification of individual respondents while maintaining analytical utility [53].

References

[1] Lobell, D.B., et al.: Climate trends and global crop production since 1980. Science 333, 616–620 (2011). https://doi.org/10.1126/science.1204531

[2] Goovaerts, P.: Geostatistics for Natural Resources Evaluation. Oxford University Press, New York (1997). https://doi.org/10.1093/oso/9780195115383.001.0001

[3] Chilès, J.P., Delfiner, P.: Geostatistics: Modeling Spatial Uncertainty, 2nd edn. Wiley, New York (2012). https://doi.org/10.1002/9781118136188

[4] Deutsch, C.V., Journel, A.G.: GSLIB: Geostatistical Software Library and User's Guide, 2nd edn. Oxford University Press, New York (1998)

[5] Armstrong, M., et al.: Plurigaussian Simulations in Geosciences, 2nd edn. Springer, Berlin (2011). https://doi.org/10.1007/978-3-642-19607-2

[6] Hengl, T.: A Practical Guide to Geostatistical Mapping. EUR 22904 EN, Luxembourg (2009). https://doi.org/10.2788/92788

[7] Datta, A., et al.: Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. J. Am. Stat. Assoc. 111, 800–812 (2016). https://doi.org/10.1080/01621459.2015.1044091

[8] Pebesma, E.: Simple features for R: Standardized support for spatial vector data. R J. 10, 439–446 (2018). https://doi.org/10.32614/RJ-2018-009

[9] Bivand, R.S., et al.: Applied Spatial Data Analysis with R, 2nd edn. Springer, New York (2013). https://doi.org/10.1007/978-1-4614-7618-4

[10] Haining, R.: Spatial Data Analysis: Theory and Practice. Cambridge University Press, Cambridge (2003). https://doi.org/10.1017/CBO9780511754944

[11] Banerjee, S., et al.: Hierarchical Modeling and Analysis for Spatial Data, 2nd edn. Chapman & Hall, Boca Raton (2015). https://doi.org/10.1201/b17115

[12] Schabenberger, O., Gotway, C.A.: Statistical Methods for Spatial Data Analysis. Chapman & Hall, Boca Raton (2005)

[13] Kerry, R., Oliver, M.A.: Determining the effect of asymmetric data on the variogram. Comput. Geosci. 36, 916–925 (2010). https://doi.org/10.1016/j.cageo.2009.11.009

[14] Cressie, N.: Statistics for Spatial Data, Revised edn. Wiley, New York (1993). https://doi.org/10.1002/9781119115151

[15] Zimmerer, K.S., et al.: The biodiversity of food and agriculture (agrobiodiversity) in the anthropocene. Anthropocene 12, 4–15 (2015). https://doi.org/10.1016/j.ancene.2015.05.001

[16] Jarvis, A., et al.: Hole-filled SRTM for the globe Version 4. CGIAR-CSI SRTM 90m Database (2008). http://srtm.csi.cgiar.org

[17] Lehner, B., et al.: New global hydrography derived from spaceborne elevation data. Eos 89, 93–94 (2008). https://doi.org/10.1029/2008EO100001

[18] Fick, S.E., Hijmans, R.J.: WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. Int. J. Climatol. 37, 4302–4315 (2017). https://doi.org/10.1002/joc.5086

[19] Hengl, T., et al.: SoilGrids250m: Global gridded soil information based on machine learning. PLoS ONE 12, e0169748 (2017). https://doi.org/10.1371/journal.pone.0169748

[20] Lark, R.M., et al.: Analyzing spatially referenced data on soil properties. Geoderma 133, 1–21 (2006). https://doi.org/10.1016/j.geoderma.2005.07.003

[21] Pebesma, E.J.: Multivariable geostatistics in S: the gstat package. Comput. Geosci. 30, 683–691 (2004). https://doi.org/10.1016/j.cageo.2004.03.012

[22] Isaaks, E.H., Srivastava, R.M.: An Introduction to Applied Geostatistics. Oxford University Press, New York (1989)

[23] Cressie, N., Wikle, C.K.: Statistics for Spatio-Temporal Data. Wiley, Hoboken (2011)

[24] Minasny, B., McBratney, A.B.: The Matérn function as a general model for soil variograms. Geoderma 128, 192–207 (2005). https://doi.org/10.1016/j.geoderma.2005.04.003

[25] Wackernagel, H.: Multivariate Geostatistics: An Introduction with Applications, 3rd edn. Springer, Berlin (2003). https://doi.org/10.1007/978-3-662-05294-5

[26] Goulard, M., Voltz, M.: Linear coregionalization model: Tools for estimation and choice of cross-variogram matrix. Math. Geol. 24, 269–286 (1992). https://doi.org/10.1007/BF00893750

[27] Chilès, J.P., Delfiner, P.: Geostatistics: Modeling Spatial Uncertainty, 1st edn. Wiley, New York (1999). https://doi.org/10.1002/9780470316993

[28] Journel, A.G., Huijbregts, C.J.: Mining Geostatistics. Academic Press, London (1978)

[29] Heuvelink, G.B.M.: Error Propagation in Environmental Modelling with GIS. Taylor & Francis, London (1998). https://doi.org/10.4324/9780203016114

[30] Emery, X.: Simulation of geological domains using the plurigaussian model: New developments and computer programs. Comput. Geosci. 33, 1189–1201 (2007). https://doi.org/10.1016/j.cageo.2007.01.006

[31] Le Loc'h, G., et al.: Truncated plurigaussian simulations to characterize aquifer heterogeneity. Ground Water 49, 13–24 (2011). https://doi.org/10.1111/j.1745-6584.2010.00719.x

[32] Gneiting, T., Raftery, A.E.: Strictly proper scoring rules, prediction, and estimation. J. Am. Stat. Assoc. 102, 359–378 (2007). https://doi.org/10.1198/016214506000001437

[33] Beucher, H., et al.: Truncated Gaussian and derived methods. C. R. Geosci. 348, 510–519 (2016). https://doi.org/10.1016/j.crte.2015.10.004

[34] Armstrong, M., et al.: Bayesian updating of plurigaussian simulations. Math. Geol. 35, 969–989 (2003). https://doi.org/10.1023/B:MATG.0000011588.99715.a9

[35] Oliver, M.A., Webster, R.: Basic Steps in Geostatistics: The Variogram and Kriging. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-15865-5

[36] Galli, A., et al.: The pros and cons of the truncated Gaussian method. In: Armstrong, M., Dowd, P.A. (eds.) Geostatistical Simulations, pp. 217–233. Springer, Dordrecht (1994). https://doi.org/10.1007/978-94-015-8267-4_18

[37] Mariethoz, G., et al.: The Direct Sampling method to perform multiple-point geostatistical simulations. Water Resour. Res. 45, W11407 (2009). https://doi.org/10.1029/2008WR007621

[38] Allard, D., et al.: Probability aggregation methods in geoscience. Math. Geosci. 44, 545–581 (2012). https://doi.org/10.1007/s11004-012-9396-3

[39] Emery, X., Peláez, M.: Assessing the accuracy of sequential Gaussian simulation and cosimulation. Comput. Geosci. 15, 673–689 (2011). https://doi.org/10.1007/s10596-011-9235-5

[40] Banerjee, S., et al.: Gaussian predictive process models for large spatial data sets. J. R. Stat. Soc. B 70, 825–848 (2008). https://doi.org/10.1111/j.1467-9868.2008.00663.x

[41] Lindgren, F., et al.: An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. J. R. Stat. Soc. B 73, 423–498 (2011). https://doi.org/10.1111/j.1467-9868.2011.00777.x

[42] Rue, H., et al.: Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. J. R. Stat. Soc. B 71, 319–392 (2009). https://doi.org/10.1111/j.1467-9868.2008.00700.x

[43] Diggle, P.J., Ribeiro Jr, P.J.: Model-based Geostatistics. Springer, New York (2007). https://doi.org/10.1007/978-0-387-48536-2

[44] Webster, R., Oliver, M.A.: Geostatistics for Environmental Scientists, 2nd edn. Wiley, Chichester (2007). https://doi.org/10.1002/9780470517277

[45] Masser, I., et al.: Building European Spatial Data Infrastructures, 2nd edn. ESRI Press, Redlands (2010)

[46] Roberts, D.R., et al.: Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography 40, 913–929 (2017). https://doi.org/10.1111/ecog.02881

[47] Mueller, N.D., et al.: Closing yield gaps through nutrient and water management. Nature 490, 254–257 (2012). https://doi.org/10.1038/nature11420

[48] Pebesma, E., Bivand, R.: Spatial Data Science: With Applications in R. Chapman & Hall, Boca Raton (2023). https://doi.org/10.1201/9780429459016

[49] Heaton, M.J., et al.: A case study competition among methods for analyzing large spatial data. J. Agric. Biol. Environ. Stat. 24, 398–425 (2019). https://doi.org/10.1007/s13253-018-00348-w

[50] Magnussen, S., et al.: Prediction of tree diameter from tree height using a mixed-effects model. Can. J. For. Res. 29, 1497–1506 (1999). https://doi.org/10.1139/x99-103

[51] Little, R.J.A., Rubin, D.B.: Statistical Analysis with Missing Data, 3rd edn. Wiley, Hoboken (2019). https://doi.org/10.1002/9781119482260

[52] Ribeiro Jr, P.J., Diggle, P.J.: geoR: A package for geostatistical analysis. R News 1, 14–18 (2001)

[53] Monmonier, M.: How to Lie with Maps, 3rd edn. University of Chicago Press, Chicago (2018). https://doi.org/10.7208/chicago/9780226436081.001.0001