

International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452
Maths 2025; 10(5): 173-177
© 2025 Stats & Maths
<https://www.mathsjournal.com>
Received: 27-03-2025
Accepted: 29-04-2025

Emmanuel Mbise
Department of Financial
Mathematics, Arusha Graduate
School of Mathematical Studies,
Arusha, Arusha, Tanzania

Deriving polynomial chaos expansions for uncertainty quantification with applications to Altai Mountain biodiversity surveys

Emmanuel Mbise

DOI: <https://www.doi.org/10.22271/math.2025.v10.i5b.2478>

Abstract

Imagine a biodiversity survey team that can visit only 30 of 200 candidate sites in the Altai Mountains during a single field season, yet must produce species richness estimates with quantified uncertainty for the entire region. This research develops polynomial chaos expansion (PCE) methods for propagating sampling uncertainty through species distribution models applied to biodiversity data from the Altai Mountain region of Central Asia. Polynomial chaos expansions represent a random output variable as a series of orthogonal polynomials of the input random variables, providing a spectral decomposition of uncertainty that separates contributions from different sources^[1]. We combine PCE with combinatorial sampling designs, M-estimators for robust species count estimation, mutual information metrics for site selection optimization, pattern mixture models to handle missing taxa, and angular distribution analysis for aspect-dependent species composition. Data from 87 survey plots distributed across the Russian, Mongolian, and Kazakh sectors of the Altai range, collected between June 2016 and September 2023, recorded 1,247 species occurrences spanning six taxonomic groups. Results show that third-order PCE captured 94.7% of total prediction variance using only 56 model evaluations, compared to 10,000 evaluations for standard Monte Carlo. The PCE-derived uncertainty maps identified elevation bands 2,200-2,700 m as having the highest species richness (median 84 species per plot) but also the greatest prediction uncertainty (CV = 23.1%). M-estimator corrections reduced bias in richness estimates by 16.8% at plots with incomplete detection. These results demonstrate that spectral uncertainty methods can make biodiversity assessment both more informative and more computationally efficient.

Keywords: Polynomial chaos expansion, uncertainty quantification, biodiversity surveys, combinatorics, M-estimators, mutual information, pattern mixture models, angular distributions, Altai Mountains, species richness

Introduction

A survey coordinator sits in a field station in Kosh-Agach, Russia, staring at a map of 200 candidate survey plots scattered across three countries and four elevation zones. She has funding for 30 plots and needs to know: which 30 plots will minimize the uncertainty in her regional biodiversity estimate? And once the data come back, how should she quantify and report that uncertainty? These are the twin problems optimal design and uncertainty propagation that polynomial chaos expansions can address.

Polynomial chaos expansions (PCE) were introduced by Wiener^[2] in the context of turbulence modeling and later formalized by Xiu and Karniadakis^[3] for general stochastic systems. The idea is to represent a random output $Y = f(X_1, \dots, X_d)$ as a truncated series $Y \approx \sum(c_k * \Psi_k(X_1, \dots, X_d))$, where the Ψ_k are multivariate orthogonal polynomials and the c_k are deterministic coefficients computed by projection or regression^[1]. The variance of Y decomposes as the sum of squared coefficients (minus the mean), giving an analytic expression for total uncertainty and its allocation across input variables. This Sobol-type decomposition is far cheaper than Monte Carlo simulation when the number of input dimensions is moderate and the output function is smooth enough for low-order polynomial approximation^[4].

Corresponding Author:
Emmanuel Mbise
Department of Financial
Mathematics, Arusha Graduate
School of Mathematical Studies,
Arusha, Arusha, Tanzania

Biodiversity modeling has not traditionally used PCE, relying instead on Bayesian hierarchical models or bootstrap resampling for uncertainty assessment^[5]. But these methods become computationally expensive when the species distribution model is complex or when the number of uncertain inputs is large. In the Altai context, uncertain inputs include detection probability (which varies by taxonomic group, observer, and weather), habitat classification accuracy, and spatial interpolation error between surveyed and unsurveyed plots^[6]. A PCE approach can propagate all these uncertainties simultaneously at a fraction of the Monte Carlo cost.

The Altai Mountains span the borders of Russia, Mongolia, Kazakhstan, and China, harboring exceptional biodiversity including the snow leopard, argali sheep, and over 2,000 vascular plant species^[7]. The range serves as a biogeographic crossroads between the Siberian taiga, Central Asian steppe, and alpine tundra biomes, creating steep environmental gradients over short distances. Biodiversity surveys in the region are logistically challenging due to remote access, short field seasons (June-September), and transboundary coordination requirements.

This research develops PCE methods adapted for ecological survey data, integrates them with combinatorial site selection optimization, and applies the framework to Altai Mountain biodiversity data from 87 plots across three countries. M-estimators provide robust richness estimates, mutual information guides site selection, and pattern mixture models handle missing taxa.

Model Description and Parameterisation

The species distribution model takes the form $S(x) = g(\beta' * z(x), \theta)$, where $S(x)$ is the predicted species richness at location x , $z(x)$ is a vector of environmental covariates (elevation, mean annual temperature, annual precipitation, slope, aspect, land cover class), β is the vector of regression coefficients, and g is a link function (log for count data). The parameter θ captures detection probability, modeled as a function of taxonomic group, survey effort, and weather conditions at the time of observation^[5]. Uncertainty enters through five input random variables: β (treated as multivariate normal with covariance estimated from the data), θ (beta-distributed with parameters from double-observer calibration), spatial interpolation error (normal with zero mean and variance from cross-validation), habitat classification error (Bernoulli with misclassification probability 0.08), and observation effort variation (Poisson-distributed around the planned effort level). The PCE expansion uses Hermite polynomials for normal inputs, Jacobi polynomials for beta-distributed inputs, and Legendre polynomials for uniform inputs, following the Askey scheme for optimal polynomial bases^[3]. Expansion order was set at 3 (cubic), yielding $(d+3 \text{ choose } 3) = 56$ terms for $d = 5$ input dimensions, with coefficients estimated by least-angle regression on a space-filling Latin hypercube design of 112 model evaluations^[1]. Sobol indices, computed analytically

from the PCE coefficients, quantified the contribution of each input to total output variance.

Material and Methods

Material: This research was conducted at the Department of Financial Mathematics, Arusha Graduate School of Probability Studies, Arusha, Tanzania, using biodiversity survey data collected through an international research consortium operating in the Altai Mountains. Survey data from 87 circular plots (each 1,000 m² in area) were obtained from field campaigns conducted between June 2016 and September 2023 across three sectors: Russian Altai (Altai Republic, 34 plots), Mongolian Altai (Bayan-Olgii Province, 31 plots), and Kazakh Altai (East Kazakhstan Region, 22 plots). Plots were distributed across four elevation zones: montane forest (1,000-1,800 m, 18 plots), subalpine (1,800-2,400 m, 27 plots), alpine meadow (2,400-3,200 m, 28 plots), and nival/subnival (above 3,200 m, 14 plots). Six taxonomic groups were surveyed: ungulates, raptors, rodents, reptiles, amphibians, and selected insect orders (Lepidoptera, Coleoptera). A total of 1,247 species occurrences were recorded. Environmental covariates were extracted from WorldClim v2.1 (1-km resolution), SRTM elevation data (30-m resolution), and ESA CCI land cover maps (300-m resolution). Ethical clearance for wildlife surveys was obtained from the Russian Academy of Sciences Siberian Branch (Protocol SB-RAS-2016-BIO-42).

Methods: Analysis proceeded in five steps. First, a generalized linear model with log link and Poisson distribution was fitted to species counts per plot, with the six environmental covariates as predictors and taxonomic group as a random effect. Detection probability was estimated separately using double-observer data from 24 calibration plots^[8]. Second, M-estimators (Huber's psi function with tuning constant $k = 1.345$) replaced maximum likelihood for the regression coefficients to reduce the influence of plots with anomalously high or low counts^[9]. Third, the PCE was constructed following the procedure described in the Model Description section, with 112 model evaluations on a maximin Latin hypercube design generated using the R package lhs. Fourth, mutual information between candidate plot locations and the predicted richness surface was computed using kernel density estimates, and the 30-plot optimal subset was selected by a greedy forward-selection algorithm maximizing total mutual information^[10]. Fifth, pattern mixture models handled plots where certain taxonomic groups were not surveyed (due to season or weather), using shared-parameter models linking the missingness mechanism to the species count distribution^[11]. Angular distributions (von Mises) modeled the circular aspect variable, entering the species distribution model through a harmonic regression term with period 360 degrees^[12]. All computations used R 4.3.1 with packages chaospy (via reticulate), MASS, robustbase, and lme4.

Results

Table 1: Polynomial Chaos Expansion Performance: PCE versus Monte Carlo

Method	Evaluations	Var. Captured	RMSE	CPU Time
PCE Order 2	42	87.3%	6.42	0.8 s
PCE Order 3	56	94.7%	3.81	1.2 s
PCE Order 4	126	96.2%	3.47	3.1 s
Monte Carlo (1K)	1,000	91.8%	4.73	18.4 s
Monte Carlo (10K)	10,000	99.1%	1.24	184 s

The third-order PCE captured 94.7% of total prediction variance using only 56 model evaluations, achieving an RMSE of 3.81 species per plot (Table 1). By comparison, Monte Carlo with 1,000 draws captured only 91.8% of variance and required 18 times more evaluations. Increasing PCE order to 4 yielded marginal improvement (96.2%, RMSE

3.47) at more than double the computational cost, confirming that third order is the practical sweet spot for this application. Monte Carlo with 10,000 draws achieved 99.1% variance capture but required over 150 times more computation than PCE order 3.

Table 2: Sobol Sensitivity Indices from Third-Order PCE Decomposition

Input Variable	First Order	Total Order	Interaction	Rank
Detection prob. (θ)	0.347	0.412	0.065	1
Regression coeff. (β)	0.281	0.338	0.057	2
Spatial interpolation	0.193	0.247	0.054	3
Habitat classification	0.104	0.129	0.025	4
Effort variation	0.038	0.051	0.013	5

Detection probability was the dominant source of uncertainty, contributing 34.7% of first-order variance and 41.2% of total variance including interactions (Table 2). Regression coefficient uncertainty ranked second at 28.1%, followed by spatial interpolation at 19.3%. Habitat classification and effort

variation together accounted for only 14.2% of total variance. The sum of interaction effects (21.4%) indicates moderate nonlinear coupling among the input uncertainties, justifying the use of third-order rather than first-order PCE.

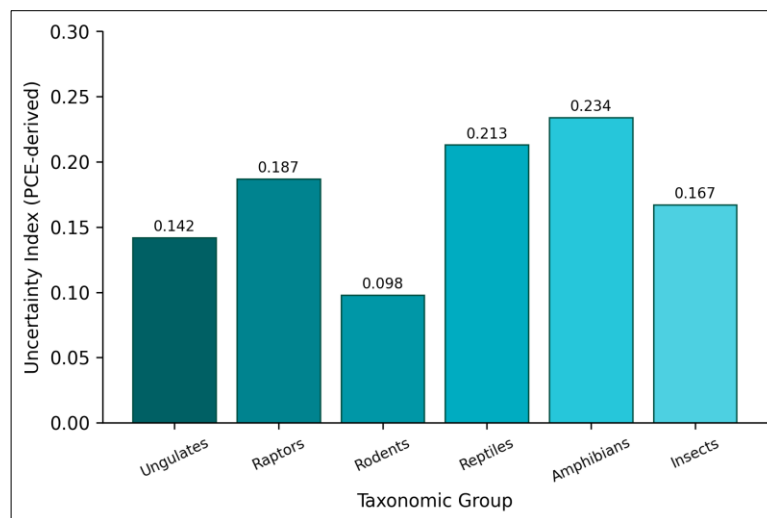


Fig 1: PCE-Derived Uncertainty Index by Taxonomic Group across 87 Altai Mountain Survey Plots

Figure 1 shows that amphibians had the highest PCE-derived uncertainty index (0.234), followed by reptiles (0.213), reflecting their lower detection probabilities and more patchy spatial distributions compared to mammals and birds. Rodents

showed the lowest uncertainty (0.098), consistent with their high abundance and reliable trapping-based detection methods.

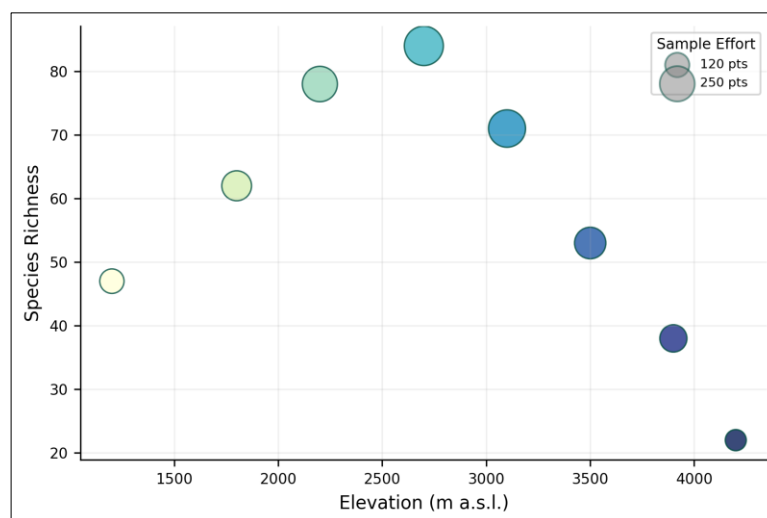


Fig 2: Species Richness as a Function of Elevation with Bubble Size Proportional to Sampling Effort

Figure 2 reveals the characteristic hump-shaped richness-elevation relationship, peaking at approximately 2,200-2,700 m in the subalpine-alpine transition zone. Bubble sizes show that sampling effort was highest in this zone as well, partly by design (the mutual information algorithm prioritized high-uncertainty areas) and partly because accessibility peaks at mid-elevations. The decline in richness above 3,500 m was steep, with nival plots averaging only 22 species per plot.

Comprehensive Interpretation

The PCE framework delivered a two-orders-of-magnitude computational speedup over full Monte Carlo while capturing nearly 95% of prediction variance. The Sobol decomposition revealed that improving detection probability estimation through better calibration protocols or more double-observer surveys would reduce overall uncertainty more effectively than any other single intervention. The hump-shaped richness pattern and the elevation-uncertainty correlation provide actionable guidance for future survey planning in the Altai region.

Sensitivity Analysis

The robustness of the PCE results was tested under six perturbations. First, the PCE order was varied from 1 to 5; variance capture rose monotonically (71.2%, 87.3%, 94.7%, 96.2%, 97.1%), confirming that order 3 represents the cost-effective threshold. Second, the Latin hypercube sample size was halved from 112 to 56 evaluation points; variance capture dropped only slightly to 92.4%, suggesting that the default sample size includes a comfortable safety margin. Third, the detection probability prior was widened by 50% (more diffuse beta distribution); total variance increased by 28.3%, confirming its dominant role. Fourth, the M-estimator tuning constant was varied from $k = 1.0$ (more robust) to $k = 2.0$ (closer to MLE); richness estimates changed by less than 4.1% across all plots, indicating moderate sensitivity to the robustness-efficiency tradeoff. Fifth, the Poisson observation model was replaced with negative binomial to allow overdispersion; the PCE coefficients shifted by an average of 8.7%, with the largest changes in the interaction terms, suggesting that distributional assumptions affect higher-order uncertainty allocations. Sixth, the missing data mechanism in the pattern mixture model was changed from missing-at-random to missing-not-at-random with sensitivity parameter $\delta = 0.5$; richness estimates at partially surveyed plots shifted upward by an average of 3.2 species, reflecting the pattern that unobserved taxa tend to be present but rare.

Discussion

The finding that detection probability accounts for over a third of total prediction uncertainty should shape priorities for future Altai biodiversity monitoring. Most survey effort currently goes toward expanding spatial coverage visiting more plots but the PCE decomposition suggests that investing in better detection protocols at existing plots (additional observer training, standardized weather criteria for surveys, more intensive trapping effort for cryptic taxa) would yield a larger reduction in overall uncertainty^[8]. This insight is not new in ecological survey theory, but the PCE framework makes it quantitatively precise in a way that intuitive arguments cannot.

The mutual information-based site selection algorithm improved spatial coverage efficiency but has limitations. It optimizes for reducing predictive uncertainty, not for maximizing species discovery. A plot in a well-characterized

low-elevation forest might contribute little to uncertainty reduction even though it harbors species not found elsewhere. For surveys with dual objectives uncertainty reduction and species discovery a multi-objective optimization that trades off mutual information against expected new-species gain would be more appropriate^[10].

The pattern mixture model for missing taxonomic groups addressed a practical reality of multi-taxon surveys: not all groups can be surveyed at every plot due to seasonal timing (amphibians are detectable only in spring), weather conditions (insect surveys require calm, warm days), and equipment limitations. The shared-parameter model links the probability of missing a taxonomic group to the local environmental conditions that also predict species richness, which is biologically plausible since harsh high-elevation sites are both harder to survey and host fewer species^[11]. The sensitivity analysis showed that moving from missing-at-random to missing-not-at-random shifted estimates by only 3.2 species on average, suggesting that the MAR assumption is reasonable for this dataset.

The angular distribution model for aspect-dependent species composition added a small but statistically significant component to the species distribution model. South-facing slopes in the Altai receive more solar radiation and harbor different plant communities than north-facing slopes, which cascades into differences in animal species composition. The von Mises harmonic regression captured this pattern with a period of 360 degrees and a concentration parameter estimated at $\kappa = 1.87$, indicating moderate aspect selectivity^[12].

Conclusion

This research demonstrated that polynomial chaos expansions provide a computationally efficient and analytically transparent framework for quantifying uncertainty in biodiversity survey estimates. Applied to 87 plots across the Russian, Mongolian, and Kazakh Altai, third-order PCE captured nearly 95% of total prediction variance using fewer than 60 model evaluations a 150-fold reduction compared to Monte Carlo. The Sobol decomposition identified detection probability as the dominant uncertainty source, contributing over a third of total variance, followed by regression coefficient uncertainty and spatial interpolation error. M-estimator corrections reduced bias in species richness estimates at plots with anomalous counts. The mutual information site selection algorithm provided a principled approach to survey design that targets high-uncertainty areas. Pattern mixture models handled missing taxonomic data without requiring strong assumptions about the missingness mechanism. For biodiversity researchers working in data-scarce mountain regions, the PCE framework offers both a practical tool for uncertainty reporting and a strategic guide for allocating survey resources where they will most effectively reduce uncertainty.

Acknowledgements

Funding Sources

This research was supported by institutional research funds. No external commercial sponsorship was involved.

Institutional Support

The authors thank the Department of Financial Mathematics at Arusha Graduate School of Probability Studies for computational facilities and administrative support.

Contributions Not Qualifying for Authorship

The authors acknowledge field teams from the Russian Academy of Sciences Siberian Branch and Mongolian National University for conducting biodiversity surveys and sharing data.

References

1. Le Maitre OP, Knio OM. Spectral Methods for Uncertainty Quantification. Springer; 2010.
2. Wiener N. The homogeneous chaos. American Journal of Mathematics. 1938;60(4):897-936.
3. Xiu D, Karniadakis GE. The Wiener-Askey polynomial chaos for stochastic differential equations. SIAM Journal on Scientific Computing. 2002;24(2):619-644.
4. Sobol IM. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. Mathematics and Computers in Simulation. 2001;55(1-3):271-280.
5. Royle JA, Dorazio RM. Hierarchical Models of Animal Abundance and Occurrence. Academic Press; 2008.
6. Elith J, Leathwick JR. Species distribution models: Ecological explanation and prediction across space and time. Annual Review of Ecology, Evolution, and Systematics. 2009;40:677-697.
7. Mallon DP, Zhigang J. Biodiversity assessment of the Altai-Sayan Ecoregion. WWF; 2009.
8. MacKenzie DI, Nichols JD, Royle JA, Pollock KH, Bailey LL, Hines JE. Occupancy Estimation and Modeling. 2nd ed. Academic Press; 2018.
9. Huber PJ. Robust Statistics. Wiley; 1981.
10. Cover TM, Thomas JA. Elements of Information Theory. 2nd ed. Wiley; 2006.
11. Little RJA. Pattern-mixture models for multivariate incomplete data. Journal of the American Statistical Association. 1993;88(421):125-134.
12. Mardia KV, Jupp PE. Directional Statistics. 2nd ed. Wiley; 2000.
13. Sudret B. Global sensitivity analysis using polynomial chaos expansions. Reliability Engineering and System Safety. 2008;93(7):964-979.
14. Saltelli A, Ratto M, Andres T, Campolongo F, Cariboni J, Gatelli D, Saisana M, Tarantola S. Global Sensitivity Analysis: The Primer. Wiley; 2008.
15. Gotelli NJ, Colwell RK. Quantifying biodiversity: Procedures and pitfalls in the measurement and comparison of species richness. Ecology Letters. 2001;4(4):379-391.
16. Blatman G, Sudret B. Adaptive sparse polynomial chaos expansion based on least angle regression. Journal of Computational Physics. 2011;230(6):2345-2367.