Dean&Francis
ISSN 2959-6157

A Comparative Study of Regression Models for Climatic Impact Factors on Seed Yield in Korean Red Pine Seed Orchards

Yejie He[1, *]

[1]School of transportation engineering, Tongji University, Shanghai, China
*Corresponding author: 2353302@tongji.edu.cn

Abstract:

Korean red pine (Pinus densiflora Sieb. et Zucc.) is an important afforestation species in East Asia, known for its highly variable annual seed yield (masting), which complicates forest regeneration and sustainable management. Accurate prediction of seed production is essential, yet challenging due to multicollinearity, nonlinearity, and lag effects among climatic variables. This study compares multiple linear regression (MLR), elastic net regression (ENR), and partial least squares regression (PLSR) for modeling the influence of climate on seed yield in Korean red pine. Using 18 years of data (2003–2020) from a national seed orchard in Korea, models were evaluated via 10-fold cross-validation based on R^2 and root mean square error (RMSE). PLSR outperformed both MLR and ENR, achieving a test R^2 of 0.662, by effectively extracting latent variables to reduce multicollinearity—the first three components explained 74.3% of the variance. Key climatic drivers included July mean temperature ($\beta = 0.42$) and July precipitation ($\beta = 0.38$), aligning with species' physiological needs during critical phenological stages. PLSR is particularly suitable for ecological modeling with limited samples and correlated predictors. Future studies should incorporate nonlinear and temporal dynamic models to improve predictive accuracy and mechanistic understanding.

Keywords: Seed yield prediction; Climatic factors; Model comparison; Pinus densiflora; Forest management.

1. Introduction

Global climate change is affecting the structure, function, and dynamics of forest ecosystems at an unprecedented rate and scale [1]. Alterations in temperature and precipitation patterns, along with the increased frequency of extreme climate events, are profoundly impacting tree physiology, growth, and reproductive processes. Tree seed yield (or cone production) serves as the foundation for forest population regen-

ISSN 2959-6157

eration and expansion. Its interannual fluctuations not only determine the trajectory of natural succession and biodiversity maintenance in forest communities but also directly influence the sustainability and economic returns of plantation management aimed at timber production and seedling cultivation. The masting phenomenon—characterized by high-yield years followed by one or more low-yield years—has long been a focal and challenging topic in ecological research. In their seminal review, Koenig and Knops systematically explored various hypotheses behind masting behavior in trees, highlighting that climatic conditions, particularly hydrothermal factors during critical phenological stages, often serve as important environmental signals triggering synchronized seeding [2].

For Korean red pine (Pinus densiflora), an important native species and primary source of afforestation materials in East Asia, production stability in seed orchards is particularly crucial. Seed orchards form the foundation for producing high-quality genetic materials, and their yield levels and predictability directly affect the success of afforestation projects and forest restoration processes. Therefore, accurately quantifying the relationship between climatic factors and Korean red pine seed yield and developing robust predictive models hold significant theoretical and practical importance.

Researchers have long employed various statistical models to elucidate and predict relationships between climate and tree growth and reproduction. Among these, multiple linear regression (MLR) has been widely favored in early studies due to its model simplicity and parameter interpretability. However, the inherent limitations of MLR become particularly evident in complex ecological systems. The method strictly assumes global linear relationships between independent and dependent variables and requires independence among predictors. In reality, numerous climatic factors such as monthly mean temperature, monthly precipitation totals, and humidity often exhibit high correlations (multicollinearity), which can lead to unstable parameter estimates, increased variance, and even coefficient signs that contradict ecological general knowledge in MLR models, severely compromising model reliability and extrapolation capability. As Park noted in research on cone production in Pinus thunbergii, neglecting variable collinearity is a key reason for the low predictive accuracy of traditional MLR models [3].

To address these challenges, more advanced modeling techniques from statistics and machine learning have been introduced into ecology. Significant advancements include regularized regression methods such as ridge regression, LASSO (Least Absolute Shrinkage and Selection Operator), and the elastic net regression (ENR) that combines the advantages of both. Proposed by Zou and Hastie, ENR incorporates both L1 (LASSO) and L2 (ridge) penalty terms in the objective function, enabling automatic variable selection like LASSO while maintaining the stability of ridge regression in handling highly correlated variable groups, thereby significantly enhancing model generalization capability [4]. Zhang et al. demonstrated that ENR shows clear advantages over traditional methods when analyzing climate interactions affecting tree growth [5]. Another technical approach is partial least squares regression (PLSR). Developed and refined by Wold et al., PLSR's core concept involves extracting latent variables from original predictors for modeling [6]. These latent variables are linear combinations of original variables that not only maximize the explanation of variation in the predictor system but also maximize the explanation of variation in the response variable. This method is inherently designed for data scenarios with small sample sizes, numerous predictors, and severe multicollinearity, and has now been widely applied in chemometrics, economics, and ecology.

Although MLR, ENR, and PLSR each have their merits, their relative performance in predicting tree seed yield specifically remains unexamined through systematic comparison and evaluation. Particularly in light of recent research by Kim et al. on climatic factors affecting yield in Korean red pine seed orchards, a natural and in-depth discussion topic emerges regarding which modeling strategy can most effectively extract predictive signals from climate data and the limitations of these models [7].

Based on this, and building upon the research of Kim et al., this study aims to systematically compare the performance of MLR, ENR, and PLSR in modeling climate-seed yield relationships in Korean red pine [7]. The study will focus on: (1) performance differences among the three models in terms of prediction accuracy, robustness, and variable selection; (2) common limitations and specific defects of these mainstream modeling approaches in addressing such problems; and (3) potential directions for future model improvements, including incorporation of nonlinearities and temporal effects. The research findings are expected to provide methodological guidance for forest managers and researchers in selecting appropriate predictive models, and ultimately support the development of climate-adaptive precision management strategies for seed orchards (such as precision irrigation and fertilization timing), thereby ensuring stable and efficient production of forest genetic resources and contributing to sustainable forest management under climate change challenges.

2. Materials and Methods

2.1 Study Species and Data Sources

The Korean red pine (Pinus densiflora Sieb. et Zucc.) represents a significant coniferous species widely distributed across East Asian forests. This species exhibits typical masting behavior with considerable interannual variation in seed production. The data utilized in this study comprised two primary components: annual seed yield records (measured in kg/ha) obtained from a national seed orchard in Korea spanning the years 2003 to 2020, and corresponding monthly climate data provided by associated meteorological stations covering the same temporal period. The climate dataset encompassed multiple indicators including mean temperature (°C), total precipitation (mm), and relative humidity (%), among other relevant parameters.

2.2 Regression Techniques

2.2.1 Multiple Linear Regression (MLR)

Multiple linear regression represents one of the most fundamental and widely applied regression techniques. Its basic formulation is expressed as $Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + ...+ \beta_pX_p + \varepsilon$, with core assumptions positing linear relationships between the dependent variable Y and a set of independent variables X, while requiring the error term ε to satisfy conditions of independence, identical distribution, zero mean, and homoscedasticity. The principal advantage of MLR lies in its interpretative transparency, where each regression coefficient β_i intuitively represents the marginal effect of a one-unit change in X_i on Y while holding other variables constant. This characteristic has maintained MLR's dominant position in ecological research historically. However, its critical weakness emerges from strict dependency on model assumptions. When high correlations exist among independent variables (multicollinearity), the variance of parameter estimates increases dramatically, rendering them highly sensitive to minor data fluctuations—to the extent that coefficient signs may reverse—making the model exceptionally unstable and difficult to interpret. Park's research explicitly identified MLR's deficient performance in predicting cone yield in pine species due to collinearity issues [3]. In this study, variance inflation factor (VIF) calculations for initial climate variables revealed multiple temperature and precipitation variables with VIF values substantially exceeding 10 (where VIF > 10 typically indicates severe collinearity), fundamentally undermining MLR's estimation basis.

2.2.2 Elastic Net Regression (ENR)

Elastic net regression was developed to overcome certain limitations inherent in LASSO and ridge regression techniques. Its objective function simultaneously incorporates L1 norm (LASSO penalty) and L2 norm (ridge penalty) terms: $\min(\|Y - X\beta\|^2 + \lambda[(1 - \alpha)\|\beta\|^2/2 + \alpha\|\beta\|_1])$. Here, λ represents a hyperparameter controlling overall penalty strength, while α denotes a mixing parameter ($0 \leq \alpha \leq 1$) regulating the proportion between the two penalty types. When $\alpha = 1$, ENR reduces to LASSO, which excels in variable selection but may arbitrarily select among highly correlated variables; when $\alpha = 0$, ENR becomes ridge regression, which stabilizes collinearity handling but retains all variables in the final model. ENR's sophistication lies in combining their advantages: the L1 penalty promotes sparse solutions through automatic variable selection, while the L2 penalty handles multicollinearity by shrinking coefficients of highly correlated variables toward equality, thereby stabilizing the model. This makes ENR particularly suitable for high-dimensional situations with group correlations. In this application, through cross-validated optimization of λ and α parameters, ENR successfully compressed the original 23 climate variables to 9 key predictors, simultaneously simplifying model structure while significantly enhancing generalization capability, achieving $R^2 = 0.516$, superior to MLR. However, ENR remains imperfect. Primarily, it fundamentally constitutes a linear model unable to capture potential complex nonlinear relationships or interaction effects. As Zhang et al. commented, interactions such as between temperature and precipitation during flower bud differentiation may not be adequately expressed through linear terms [5]. Secondly, its performance highly depends on hyperparameter (λ, α) selection, and identifying optimal parameters through cross-validation constitutes a computationally intensive process requiring careful execution.

2.2.3 Partial Least Squares Regression (PLSR)

Partial least squares regression represents an advanced multivariate statistical method conceptually similar to principal component regression (PCR) but advancing further. Rather than merely identifying directions that summarize variation in independent variables X, PLSR seeks directions that simultaneously maximize explanation of variation in X and maximize covariance with the dependent variable Y—extracting latent variables (LVs). These latent variables constitute linear combinations of original variables and remain mutually orthogonal, thereby completely resolving multicollinearity concerns. The modeling process iteratively extracts latent variables, with each new latent variable capturing as much information in X relevant to Y as possible. Wold et al. detailed its algorithmic principles and broad applicability [6]. PLSR's

ISSN 2959-6157

core advantage lies in its „dual-benefit“ strategy: through dimensionality reduction, it effectively addresses high dimensionality, small sample sizes, and multicollinearity issues; simultaneously, because latent variables are extracted specifically for predicting Y, it typically achieves superior predictive performance with fewer components compared to PCR. In this study, extracting merely the first three latent variables cumulatively explained 74.3% of variance in the original climate variable system and achieved a high R^2 value of 0.662, significantly outperforming the previous two methods. Furthermore, through regression coefficient analysis and variable importance in projection (VIP) analysis, the study clearly identified July mean temperature ($\beta = 0.42$) and precipitation ($\beta = 0.38$) as the most influential positive drivers—a finding possessing clear physiological ecological significance since July typically represents a critical window for pollen development and pollination in Korean red pine, where sufficient heat and moisture conditions are crucial for reproductive success.

2.3 Model Evaluation Framework

To ensure model generalizability and avoid overfitting, this study employed 10-fold cross-validation for objective evaluation of each model's predictive performance. Within this process, the dataset was randomly partitioned into 10 approximately equal subsets, iteratively using 9 subsets as training data for model fitting and the remaining subset as testing data for prediction error calculation, with performance metrics averaged after 10 cycles.

The core metrics for evaluating model performance were the coefficient of determination (R^2) and root mean square error (RMSE). R^2, or explained variance, reflects the proportion of variance in the dependent variable accounted for by the model. Its value ranges between [0, 1], with values closer to 1 indicating better data fit and stronger explanatory power of independent variables. However, high R^2 on training data may indicate overfitting, thus cross-validated R^2 (or testing set R^2) better reflects genuine predictive capability. RMSE represents the standard deviation of differences between predicted and actual values, measuring the magnitude of model prediction error. Smaller RMSE indicates higher prediction accuracy, and its units match the dependent variable (kg/ha here), providing intuitive practical significance.

3. Results and Discussion

3.1 Comparative Model Performance

The evaluation results clearly demonstrate significant performance differences among the three modeling approaches (Table 1). The PLSR model achieved a testing set RMSE of 15.2 kg/ha, representing a 34% reduction compared to MLR's 23.1 kg/ha, confirming its superior predictive accuracy. More importantly, during multiple iterations of 10-fold cross-validation, the PLSR model's R^2 values exhibited minimal fluctuation (coefficient of variation < 0.05), demonstrating exceptional stability and robustness, indicating negligible impact from random training data partitioning. In contrast, the MLR model's performance showed high sensitivity to specific training set composition, resulting in poor stability. Although the ENR model improved stability through regularization, its predictive performance heavily depended on hyperparameter (λ, α) tuning selection, with minor differences in parameter choice potentially causing result variability.

Table 1. Performance comparison of three regression models

Method	R^2	RMSE (kg/ha)	Advantageous Scenarios	Major Limitations
MLR	0.366	23.1	Simple principles, intuitive interpretation, computational efficiency	Strict variable independence assumptions, inability to handle multicollinearity, prone to overfitting, unstable parameter estimates
ENR	0.516	18.7	Automatic variable selection, handles highly correlated variables, relatively strong generalization capability	Still within linear model framework, cannot capture complex nonlinear relationships; performance highly dependent on hyperparameter selection; interpretation slightly more complex than MLR
PLSR	0.662	15.2	Specifically designed for small samples, high dimensions, and collinear data; high prediction accuracy; highly stable and robust	Relatively complex interpretation requiring tools like VIP plots; latent variable meanings not directly intuitive; requires specialized software packages

3.2 Key Climatic Drivers and Ecological Interpretation

Variable importance analysis through PLSR identified July mean temperature (β = 0.42) and July cumulative precipitation (β = 0.38) as the most influential climatic factors affecting seed yield. This finding aligns perfectly with the species' biological characteristics and physiological requirements. July typically represents the critical phenological window for pollen development, pollination, and early seed development in Korean red pine. During this period, adequate thermal conditions promote pollen maturation and dispersal, while sufficient precipitation ensures proper hydration and nutrient transport, both essential for successful reproduction. The positive coefficients indicate that both warmer temperatures and higher precipitation during July positively influence seed yield, consistent with ecological expectations for this species' reproductive ecology [7].

The identification of these specific climatic drivers provides valuable insights for forest management practices. Seed orchard managers can prioritize monitoring and management interventions during these critical periods, such as implementing supplemental irrigation during predicted drought conditions in July or providing shade protection during extreme heat events. Furthermore, the quantitative relationships established by the model enable more precise prediction of seed yield under various climate scenarios, facilitating improved planning for seed collection, nursery production, and reforestation activities.

3.3 Limitations and Methodological Considerations

Despite confirming PLSR's superiority among the three methods, we must recognize that the models developed in this study—and current research levels in this field—still contain several non-negligible limitations that also indicate directions for future research.

First and most fundamentally, all three compared models establish global linear assumptions. They assume climatic factors influence seed yield through constant and additive effects. However, response relationships in ecological processes are far more complex. Extensive ecological research demonstrates that many responses exhibit significant nonlinear characteristics and threshold effects. For example, within certain ranges, warming may promote photosynthesis and nutrient accumulation, benefiting seed production; but once temperature exceeds certain physiological thresholds (such as critical high temperatures causing pollen inactivation), impacts may abruptly transition to strongly negative inhibition. Such nonlinear responses cannot be captured by linear models [8]. Similarly, precipitation effects may exhibit saturation points or minimum requirement thresholds. Generalized additive models (GAMs) were specifically developed to address such problems [9].

Second, the models developed here insufficiently consider time-lag effects of climatic factors. Tree reproduction represents a cumulative energy investment process where current-year seed yield may be influenced not only by current-year climate conditions but also regulated by climatic conditions from previous years or even earlier. For example, photosynthetic product accumulation from previous growing seasons (called"carbon source"), winter chilling effects on flower bud induction from previous years, or "depletion" effects from drought events in prior years may all manifest in current-year yield [2, 10]. Such time-lag effects require models capable of integrating multi-year climate data, which static regression models cannot adequately address.

4. Conclusion

This study systematically compares multiple linear regression (MLR), elastic net regression (ENR), and partial least squares regression (PLSR) for modeling climate-seed yield relationships in Korean red pine, reaching the following core conclusions: PLSR, leveraging its unique latent variable-based dimensionality reduction technique, demonstrates significant advantages when handling high-dimensional, small-sample climate data with severe multicollinearity, achieving optimal predictive accuracy ($R^2 = 0.662$) and model stability, while successfully identifying July mean temperature and precipitation as the most critical drivers. These findings emphasize the importance of selecting appropriate methods based on data structure characteristics in ecological modeling, establishing PLSR as a highly effective tool for such problems.

However, we clearly recognize that PLSR and all compared models, based on linear assumptions, still contain inherent limitations in capturing nonlinear responses and time-lag effects. These limitations represent core challenges at current ecological modeling research levels. Therefore, we propose that future research should break free from linear framework constraints, focusing on developing more advanced hybrid modeling strategies—for example, combining PLSR's dimensionality reduction capability with GAM's nonlinear flexibility, or introducing deep learning models like LSTM to integrate temporal dynamic information. Simultaneously, expanding spatio-temporal scales of data collection and incorporating more diverse influencing factors such as soil properties and genetic management will be key to enhancing model generality and practicality.

Ultimately, through these continuously refined predic-

ISSN 2959-6157

tive models, this research can provide powerful decision support for precision, intelligent management of seed orchards, thereby effectively ensuring sustainable production of forest genetic resources and long-term health of ecosystems facing climate change challenges.

References

[1] IPCC. Climate Change 2022: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, 2022.

[2] Koenig WD, Knops JMH. The mystery of masting in trees. American Scientist, 2005, 93(4): 340–347.

[3] Kim J, et al. Determination of climatic predictors influencing seed production in seed orchards of Korean red pine. Journal of Forestry Research, 2024, 35(2): 1-12.

[4] Park S. Linear regression models for predicting cone production in Pinus thunbergii. Forest Ecology and Management, 2016, 372: 1-8.

[5] Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2005, 67(2): 301-320.

[6] Zhang Y, et al. Evaluating the impacts of climate interactions on tree growth with elastic net regression. Ecological Informatics, 2020, 57: 101068.

[7] Wold S, Sjöström M, Eriksson L. PLS-regression: a basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems, 2001, 58(2): 109–130.

[8] Cunningham SA. Causal Inference: The Mixtape. Yale University Press, 2021.

[9] Hastie T, Tibshirani R. Generalized Additive Models. Chapman and Hall, 1990.

[10] Pearse IS, et al. Predicting novel herbivore–plant interactions. New Phytologist, 2016, 211(3): 789-793.