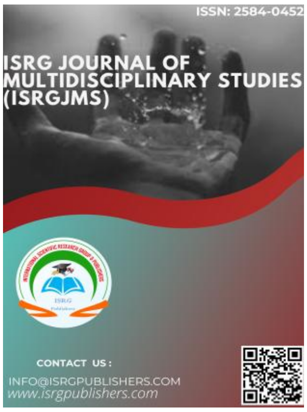
ISRG Journal of Multidisciplinary Studies (ISRGJMS)

ISRG PUBLISHERS
Abbreviated Key Title: isrg j. multidiscip. Stud.
ISSN: 2584-0452 (Online)
Journal homepage: https://isrgpublishers.com/isrgjms/
Volume – III, Issue - II (February) 2025
Frequency: Monthly



ISSN: 2584-0452
ISRG JOURNAL OF MULTIDISCIPLINARY STUDIES (ISRGJMS)
CONTACT US :
INFO@ISRGPUBLISHERS.COM
www.isrgpublishers.com

Performance of Beta Ridge Regression Estimator in Addressing Multicollinearity within Beta Distribution

Ratna Arum Sari[1], Netti Herawati[2*], Misgiyati[3], Khoirin Nisa[4]

[1, 2, 3, 4] Department of Mathematics, University of Lampung, Lampung, Indonesia

| **Received:** 27.01.2025 | **Accepted:** 03.02.2025 | **Published:** 04.02.2025

*__Corresponding author:__ Netti Herawati

Department of Mathematics, University of Lampung, Lampung, Indonesia

Abstract

Beta Ridge Regression (BRR) is a ridge method applied in the beta regression model used to overcome the problem of multicollinearity, which is a condition in which the independent variables in the regression model have a high correlation. This problem can cause parameter estimates to be unstable and less accurate. This study aims to determine the performance of BRR estimator in overcoming multicollinearity in simulated data with small sample size. The analysis is done by comparing the estimation results based on the Mean Squared Error (MSE) and Mean Absolute Error (MAE) values. The results show that the proposed BRR estimator has superior performance compared to the Maximum Likelihood Estimation (MLE) method, by producing lower MSE and MAE values than MLE.

Keywords: *Multicollinearity, Beta Ridge Regression, Beta Distribution, Simulated Data, Mean Squared Error, Mean Absolute Error*

1. INTRODUCTION

The beta regression model is a model that provides accurate and efficient parameter estimators compared to the ordinary least squares method, when the response variable is not symmetrically distributed, or when heteroscedasticity problems occur [1]. Beta regression models are used to model data that is limited to the interval (0,1) generally such as ratios or percentages. However, multicollinearity problems often arise in data analysis involving highly correlated independent variables. The occurrence of multicollinearity causes the least squares estimator to have a large variance [2].

DOI: 10.5281/zenodo.14800646

In an attempt to overcome multicollinearity, beta ridge regression (BRR) has been proposed as an alternative method. BRR incorporates the ridge regression technique into the beta regression model by adding a ridge parameter (k) to reduce the variance of the estimate and the impact of high correlation between independent variables. Previous research by Abonazel and Taha [3] and Qasim, et al. [4] showed that BRR performed better than the MLE method in overcoming multicollinearity.

This study aims to evaluate the performance of BRR in overcoming multicollinearity in simulated data with small sample sizes. In addition, this study also compares several ridge parameters $(k_1, k_2, k_3, k_4, k_5)$ to determine the parameter value that provides the best results based on MSE and MAE criteria.

2. LITERATURE REVIEW

2.1 Beta Distribution

The beta distribution is one of the continuous probability distributions that is often used in various statistical applications, especially as a model for random variables limited to the interval 0-1 and is determined by two positive parameters, namely a and b. These two parameters act as exponents of the random variable, which affect the shape of the beta distribution [5]. The probability density function of the beta distribution is expressed as follows:

$$f(y; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\, y^{a-1}(1-y)^{b-1},$$

where $0 < y < 1$, $\Gamma(.)$ is the gamma function, a and b are the beta distribution parameters.

In the beta distribution, the mean and variance with parameters a and b are as follows:

$$E(Y) = \frac{a}{a+b} \;\&\; Var(Y) = \frac{ab}{(a+b)^2(a+b+1)}.$$

2.2 Beta Regression Model

The beta regression model is a statistical method designed to model the relationship between an independent variable and a dependent variable, where the dependent variable follows a beta distribution.

The beta regression model is a statistical approach used to analyze the relationship between independent variables and dependent variables, where the dependent variable follows a beta distribution.

This model is often used in the analysis of ratio or percentage data, such as success rates, proportions, and probabilities, because of its flexibility in handling various forms of data distribution, both symmetrical and asymmetrical. Beta regression is commonly applied in various fields, including economic, social, and other fields that involve analyzing data bounded on the interval (0,1). The beta regression model was first introduced by Ferrari and Cribari-Neto. Ferrari and Cribari-Neto [6] defined a parameterization to develop a beta distributed response regression model based on the initial equation of the beta regression density function. By supposing that $\mu = \frac{a}{a+b}$ and $\phi = a + b$ so that $a = \mu\phi$ and $b = \phi - \phi\mu = \phi(1-\mu)$. After the reparameterization, the function for the random variable y that follows the beta distribution is as follows:

$$f(y; \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma(\phi-\phi\mu)}\, y^{\mu\phi-1}(1-y)^{\phi-\phi\mu-1},$$

Where $0 < y < 1;\ \ 0 < \mu < 1;\ \ \phi > 0;\ \Gamma(.)$ is a gamma function, and ϕ is a parameter. ϕ is a parameter written by Bayer and Cribari-Neto [7] which is defined as follows:

$$\phi = \frac{1-\sigma^2}{\sigma^2}.$$

Therefore, the mean and variance of Y are expressed in the new parameterization as follows:

$$E(Y) = \mu \quad \& \quad Var(Y) = \frac{\mu(1-\mu)}{1+\phi}.$$

In beta regression, the relationship between the mean (μ) and the covariate X_i is expressed through a logit link function. The logit link function is as follows:

$$g(\mu) = \log\left(\frac{\mu}{1-\mu}\right).$$

Therefore, the beta regression model with a logit link function can be written as:

$$g(\mu_i) = X_i^T\beta \rightarrow \mu_i = \frac{\exp(X_i^T\beta)}{1+\exp(X_i^T\beta)}.$$

In estimating parameters, namely by using the MLE method by maximizing the likelihood function. MLE is a technique used to estimate the parameters of a population distribution by choosing parameter values that maximize the likelihood function of the observed data [8]. The log-likelihood function in the beta regression model is:

$$\ell(\mu, \phi) = \sum_{i=1}^{n}[\,\log\Gamma(\phi) \,-\log\,\Gamma\big(\mu_i(\phi)\big) - (\mu_i(\phi) - 1)\log(y_i) + \big((1-\mu_i)(\phi) - 1\big)\log(1-y_i)].$$

Furthermore, to obtain the value of the maximum likelihood estimator $\hat{\beta}$ where the coefficient vector uses the maximum likelihood method which is estimated by performing the first derivative of the log-likelihood function on β or called the score function. Where μ_i depends on β through the log-likelihood function. Equation The score function is as follows:

$$\mathcal{S}(\beta) = \frac{\partial \log \ell(\mu,\phi)}{\partial\beta}$$

$$= \sum_{i=1}^{n}\big[\big(-\psi(\mu_i(\phi)\big) - \psi\big(1 - \mu_i(\phi)\big) + \log(y_i) - \log(1 - y_i)\big].$$

Because the above equation is nonlinear and cannot be solved analytically. The solution of $\mathcal{S}(\beta)$ can be found through the iterative reweighted least square (IRLS) algorithm derived from the application of the Newton Raphson or Fisher Scoring method. The general Newton-Raphson equation for finding the roots of the log-likelihood is:

$$\beta_{m+1} = \beta_m + \{H(\beta_m)\}^{-1}\,\mathcal{S}(\beta_m),$$

then obtained as follows:

$$\beta_{m+1} = \beta_m + (X_m^T W_m X_m)^{-1}\, X_m^T W_m (z_m - X_m\beta_m).$$

After obtaining the simplification, the form of the logarithmic estimate is then associated with IRLS as follows:

$$\beta_m = (X_m^T W_m X_m)^{-1} X_m^T W_m z_m,$$

where $W_m = diag\left(\frac{1+\phi}{\mu_i(1-\mu_i)}\right)$, and $z_m = \log(\mu_i) + \frac{y_i+\mu_i}{\mu_i(1-\mu_i)}$.

It will be concluded that β_m converges to $\hat{\beta}_{MLE}$ when $m \rightarrow \infty$, so the final form of MLE is:

$$\hat{\beta}_{MLE} = \left(X^T\widehat{W}X\right)^{-1}X^T\widehat{W}z.$$

The variance of MLE is:

$$V\left(\hat{\beta}_{MLE}\right) = \phi(X^TWX)^{-1}.$$

The MSE of MLE is:

$$MSE\left(\hat{\beta}_{MLE}\right) = E\left(\hat{\beta}_{MLE} - \beta\right)^T\left(\hat{\beta}_{MLE} - \beta\right).$$

Since $\hat{\beta}_{MLE}$ is an unbiased estimator, meaning $E\left(\hat{\beta}_{MLE}\right) = \beta$, the MSE only depends on the variance (trace) which is as follows:

$$MSE\left(\hat{\beta}_{MLE}\right) \ = trace\{\phi(X^TWX)^{-1}\}$$

$$= \phi\sum_{j=1}^{p}\frac{1}{\lambda_j}$$

where λ_j is the jth eigenvalue in the X^TWX matrix.

2.3 Beta Ridge Regression Estimator

This section present the BRR method for beta regression model, which is a generalization of Hoerl and Kennard [9]. In the BRR method, the maximum likelihood principle is used to estimate the parameter estimates with the minimum *weight sum of square* (WSSE). Suppose we will choose an arbitrary estimator $\hat{B}$ other than $\hat{\beta}_{MLE}$, where $\hat{B}$ is a vector of β then the WSSE of this estimator can be defined as follows:

$$\theta \ = \left(y - \hat{B}\right)^T(y - \hat{B})$$

$$= \left(y - X\hat{\beta}_{MLE}\right)^T\left(y - X\hat{\beta}_{MLE}\right) + \quad \left(\hat{B} - \hat{\beta}_{MLE}\right)^T X^TWX(\hat{B} - \hat{\beta}_{MLE})$$

$$= \theta_{min} \ + \theta\left(\hat{B}\right).$$

where θ_{min} is the minimum value, and $\theta\left(\hat{B}\right) > 0$ is a fixed increment that increases the Weight Sum of Square when the $\hat{\beta}_{MLE}$ estimator is replaced by the $\hat{B}$ estimator. According to Hoerl and Kennard [9], to obtain parameter estimates in beta ridge regression, namely by minimizing the lagrange function (Q) using $\hat{B}^T\hat{B}$ with $\left(\hat{B} - \hat{\beta}_{MLE}\right)^TX^TWX(\hat{B} - \hat{\beta}_{MLE}) = \theta_0$ which will then be made in lagrangian form as:

$$Minimum\ (Q) = \left(y - \hat{B}\right)^T(y - \hat{B})$$

$$= \hat{B}^tB + \left(\frac{1}{k}\right)\left\{\left(\hat{B} - \ \hat{\beta}_{MLE}\right)^TX^TWX\left(\hat{B} - \quad \hat{\beta}_{MLE}\right) - \theta_0\right\},$$

where $\left(\frac{1}{k}\right)$ is the multiple lagrangian and θ_0 is the sum of squares error.

Then the lagrangian equation is derived against $\hat{B}$ and the result is equalized to zero, as follows:

$$\frac{\partial Q}{\partial\hat{B}} = 2\hat{B} + \frac{\left\{2X^TWX\left(\hat{B} - \hat{\beta}_{MLE}\right)\right\}}{k} = 0\,.$$

In the above equation $\hat{B}$ is considered as the beta ridge regression ($\hat{\beta}_{BRR}$), so the final equation of the beta ridge regression estimator can be defined as follows:

$$\hat{\beta}_{BRR} = (kI + X^TWX)^{-1}X^TWX\hat{\beta}_{MLE}\,,$$

where k is the ridge parameter and I is the identity matrix of order $p \times p$.

2.4 Choosing the Ridge Parameter k

In beta ridge regression to handle the multicollinearity problem, a ridge parameter called ridge parameter k is required. Taken from the research of Alkhamisi, et al [10] and Kibria [11], we can modify the estimation of k based on Abonazel and Taha [3] for the beta regression model as follows:

1. $k_1 = max\left(\frac{1}{\phi\hat{\alpha}_j^2}\right)$
2. $k_2 = median\left(\frac{1}{\phi\hat{\alpha}_j^2}\right)$
3. $k_3 = mean\left(\frac{1}{\phi\hat{\alpha}_j^2}\right)$
4. $k_4 = \left(\frac{\lambda_{max}}{\phi\hat{\alpha}_{max}^2}\right)$
5. $k_5 = \left(\frac{\lambda_{min}}{\phi\hat{\alpha}_{min}^2}\right)$

With λ_{max} the maximum eigenvalues of the X^TWX matrix and λ_{min} is the minimum eigenvalues of the X^TWX matrix, α_j^2 is the value of $Q^T\hat{\beta}_{MLE}$, where $Q^T$is the eigenvector element of the X^TWX matrix.

3. METHODOLOGY

The data used in this study are simulated data containing multicollinearity generated using RStudio Software. The simulation data is data generated with 6 independent variables (p = 6) with a correlation level between independent variables of 0.6 and 0.99 ($\rho = 0.6, 0.99$) and the number of samples used is $n = 10,\ 20,\ 30,\ 50,$ and 75 with repetition $L = 1000$ times and the parameter ϕ used is ($\phi = 1{,}2{,}3$).

To obtain multicollinearity data on each data set X_p is generated using Monte Carlo simulation based on McDonald and Galarneau [12] with the following equation:

$$X_p = \sqrt{1 - \rho^2}z_{ij} + \rho z_{i(p+1)},$$

where $i = 1, 2, \dots, n;$

$j = 1, 2, \dots, p;\ z_{ij} \sim Normal\ (0{,}1).$

In this study, the model estimation performance is analyzed using the MSE and MAE criteria MSE and MAE values can be calculated as follows:

$$MSE = \frac{1}{L}\sum_{i=1}^{L}\left(\widehat{\beta}_{\imath} - \beta\right)^2$$

$$MSE = \frac{1}{L}\sum_{i=1}^{L}|\,\widehat{\beta}_{\imath} - \beta|$$

4. RESULT AND DISCUSSION

To see the performance of Beta Ridge Regression, the first step that must be done is to calculate the VIF value for the partial correlation between the independent variables and the full correlation. This correlation value is needed to see if there is a multicollinearity problem between the independent variables. The VIF results are shown in Table 1 and Table 2.

DOI: 10.5281/zenodo.14800646

Table 1. VIF Value $\rho = 0.6$

n	$\rho = 0.6$					
	X_1	X_2	X_3	X_4	X_5	X_6
10	5.24	123.1	8.78	82.24	1.96	58.29
20	2.40	52.46	2.78	1.70	54.39	64.72
30	1.86	3.28	43.79	75.54	37.58	1.74
50	49.40	1.38	27.36	1.75	38.57	1.67
75	32.82	42.81	38.41	1.73	1.43	1.52

Table 2. VIF Value $\rho = 0.99$

n	$\rho = 0.99$					
	X_1	X_2	X_3	X_4	X_5	X_6
10	212.6	163.3	222.3	266.9	291.8	101.2
20	30.35	28.02	55.74	30.47	29.50	42.30
30	51.02	18.93	26.73	34.79	36.32	43.20
50	50.61	44.76	67.28	41.73	42.26	43.53
75	36.41	36.76	41.17	33.47	29.23	35.56

Table 1 shows that there is a partial correlation between the independent variables indicated by the VIF value > 10 for $\rho = 0.6$ with $n = 10, 20, 30, 50, 75$. In addition the VIF result at $\rho = 0.99$ for $n = 10, 20, 30, 50, 75$ is displayed in Table 2. Table 2 shows that there is a full correlation between independent variables. Both of these indicate that if VIF >10 between independent variables, it can be believed that multicollinearity is presence.

Next is to calculate the MSE values for MLE and BRR at $n = 10, 20, 30, 50, 75$ and $\rho = 0.6 \,\&\, 0.99$ to find which one is better at handling the multicollinearity present in the model. The results of the analysis can be seen in Table 3 and Table 4.

Table 3. MSE Value $\rho = 0.6$

n	$\rho = 0.6$					
	MLE	k_1	k_2	k_3	k_4	k_5
$\phi = 1$						
10	85.12	0.307	0.506	0.470	0.212	77.93
20	2.003	0.406	0.609	0.509	0.245	0.310
30	1.021	0.498	0.795	0.704	0.295	0.241
50	0.757	0.649	0.723	0.715	0.481	0.300
75	0.624	0.581	0.614	0.607	0.573	0.347
$\phi = 2$						
10	85.12	0.555	0.801	0.740	0.312	81.42
20	2.003	0.590	0.719	0.734	0.244	0.558
30	1.021	0.730	0.974	0.895	0.249	0.264
50	0.757	0.724	0.748	0.745	0.400	0.307
75	0.624	0.612	0.621	0.619	0.485	0.340
$\phi = 3$						
10	85.12	0.721	1.157	1.039	0.383	82.62
20	2.003	0.766	0.950	0.971	0.228	0.633
30	1.021	0.829	0.987	0.957	0.229	0.225
50	0.757	0.742	0.753	0.752	0.366	0.320
75	0.624	0.618	0.618	0.623	0.666	0.445

Table 4. MSE Value $\rho = 0.99$

n	$\rho = 0.99$					
	MLE	k_1	k_2	k_3	k_4	k_5
$\phi = 1$						
10	139.4	0.050	0.259	0.151	0.011	137.8
20	5.847	0.498	2.208	1.247	0.384	1.809
30	0.902	0.467	0.865	0.675	0.428	0.320
50	1.656	0.892	1.098	1.108	0.541	0.414
75	1.105	0.749	1.099	0.977	0.445	0.342
$\phi = 2$						
10	139.4	0.113	10.43	1.725	0.014	72.49
20	5.847	1.216	3.804	2.849	0.382	0.638
30	0.902	0.680	0.844	0.816	0.393	0.339
50	1.656	1.097	1.512	1.436	0.446	0.613
75	1.105	0.956	1.103	1.064	0.394	0.342
$\phi = 3$						
10	139.4	1.424	9.156	6.393	0.013	6.507
20	5.847	1.802	5.262	3.865	0.396	0.482
30	0.902	0.776	0.873	0.859	0.369	0.339
50	1.656	1.209	1.640	1.542	0.423	0.612
75	1.105	1.028	1.104	1.085	0.376	0.342

In Table 3 and Table 4, it can be seen that the MSE of the BRR is smaller than MLE for all ridge parameters used. If we look in more detail at Table 3 and Table 4, ridge parameters k_4 and k_5 have a smaller MSE than ridge parameters k_1, k_2, k_3.

In addition, MAE value is also evaluated to see the performance of the model. The MAE values of the simulation results are shown in Table 5 and Table 6.

Table 5. MAE Value $\rho = 0.6$

n	$\rho = 0.6$					
	MLE	k_1	k_2	k_3	k_4	k_5
$\phi = 1$						
10	6.725	0.511	0.604	0.577	0.443	6.412

20	1.047	0.525	0.538	0.536	0.486	0.489
30	0.762	0.502	0.668	0.624	0.539	0.439
50	0.772	0.714	0.754	0.750	0.692	0.533
75	0.713	0.683	0.706	0.702	0.606	0.586
$\phi = 2$						
10	6.725	0.647	0.845	0.802	0.513	6.564
20	1.047	0.538	0.554	0.562	0.485	0.538
30	0.762	0.637	0.744	0.712	0.492	0.509
50	0.772	0.755	0.767	0.766	0.631	0.522
75	0.713	0.704	0.711	0.710	0.695	0.572
$\phi = 3$						
10	6.725	0.787	1.027	0.975	0.547	6.617
20	1.047	0.573	0.667	0.677	0.465	0.539
30	0.762	0.683	0.749	0.737	0.466	0.448
50	0.772	0.764	0.770	0.769	0.604	0.518
75	0.713	0.709	0.712	0.711	0.666	0.568

Table 6. MAE Value $\rho = 0.99$

n	$\rho = 0.99$					
	MLE	k_1	k_2	k_3	k_4	k_5
$\phi = 1$						
10	9.946	0.177	0.365	0.282	0.101	9.855
20	2.140	0.624	1.364	0.990	0.616	1.225
30	0.772	0.625	0.760	0.713	0.654	0.562
50	1.062	0.850	0.919	0.922	0.735	0.636
75	0.775	0.593	0.772	0.701	0.667	0.584
$\phi = 2$						
10	9.946	0.247	2.308	0.942	0.103	6.169
20	2.140	0.971	1,784	1,550	0.610	0.725
30	0.772	0.715	0.756	0.750	0.626	0.582
50	1.062	0.920	1.029	1.011	0.667	0.718
75	0.775	0.688	0.774	0.752	0.628	0.584
$\phi = 3$						
10	9.946	0.856	2.162	1.808	0.103	1.824
20	2.140	1.211	2.053	1.796	0.607	0.615
30	0.772	0.740	0.762	0.759	0.607	0.582
50	1.062	0.953	1.058	1.037	0.650	0.718
75	0.775	0.731	0.775	0.764	0.613	0.584

From Table 5 and Table 6, in can be seen that the MAE value of BRR is smaller than MLE for all ridge parameters used. The ridge parameters k_4 and k_5 have smaller MAE values than the ridge parameters k_1, k_2, k_3.

These results indicate that the BRR estimator with ridge parameter k_1, k_2, k_3, k_4, k_5 has superior performance and is more reliable than the MLE in handling multicollinearity problems. Among the five proposed parameter ridge k, parameter ridge k_4 and k_5 provide the best estimation compared to ridge parameters $k_1, k_2,$ dan k_3 in handling multicollinearity, as they produce the smallest MSE and MAE values.

5. Conclusion

Multicollinearity is an issue in regression modeling that leads to a high variance in the least squares estimator. To address this problem in beta regression models, this study develops a beta ridge regression estimator. Several ridge parameters (k) are proposed, all of which outperform the MLE estimator in terms of MSE and MAE criteria. Among the proposed estimators, the parameters k_4 and k_5 yield more efficient and reliable results under the conditions examined in this study. Therefore, we recommend the beta ridge regression estimator for practitioners.

REFERENCES

1. Swearingen, C. J., Tilley, B. C., Adams, R. J., Rumboldt, Z., Nicholas, J. S., Bandyopadhyay, D. & Woolson, R. F. 2011. Application of beta regression to analyze ischemic stroke volume in NINDS rt-PA clinical trials. *Neuroepidemiology*. 37(2): 73-82.
2. Montgomery, D. C. & Peck, E. A. 1992. *Introduction to Linear Regression Analysis (2nd ed.).* Wiley, New York.
3. Abonazel, M. R. & Taha, I. M. 2021. Beta ridge regression estimators: simulation and application. *Communications in Statistics-Simulation and Computation*. 52(9): 4280-4292.
4. Qasim, M., Mansson, K. & Kibria, B. G. 2021. On some beta ridge regression estimators: method, simulation and application. *Journal of Statistical Computation and Simulation*. 91(9): 1699-1712.
5. Degroot, M. H. & Schervish, M. J. 2002. *Probability and Statistics.* Addison-Wesley, Boston.
6. Ferrari, S. & Cribari-Neto, F. 2004. Beta regression for modelling rates and proportions. *Journal of applied statistics*. 31(7): 799-815.
7. Bayer, F. M. & Cribari-Neto, F. 2017. Model Selection criteria in beta regression with varying dispersion. *Communications in Statistics-Simulation and Computation.* 46(1): 729-746.
8. Casella, G & Berger, R. L. 2002. *Statistical Inference (2nd ed.).* Duxbury Press, Pacific Grove.
9. Hoerl, A. E. & Kennard, R. W. 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*. 12 (1): 55-67.
10. Alkhamisi, M., Khalaf G. & Shukur, G. 2006. Some modifications for choosing ridge parameters. *Communications in Statistics-Theory and Methods*. 35(11): 2005-2020.
11. Kibria, B. G. 2003. Performance of some new ridge regression estimators. *Communications in Statistics-Simulation and Computation*. 32(2): 419-435.

12. McDonald, G. C., & Galarneau, D. I. 1975. A Monte Carlo evaluation of some ridge-type estimators. *Journal of the American Statistical Association* 70(350): 407-416.

DOI: 10.5281/zenodo.14800646