Prior selection for Maize yield prediction

Guang Yang, Jerry Liang, Yuanzhuo Xu
Department of Applied Mathematics and Statistics
Johns Hopkins University
Baltimore, MD 21218

December 20, 2023

Abstract

This report evaluates models constructed by different selections of prior from different countries. It compares the model with classical statistics, non-informative prior, and Informative prior. The data from eight European countries(Austria, Belgium, France, Germany, Italy, Spain, Switzerland, and Poland) will be used.

1 Introduction

1.1 Purpose of study

As a central difference with the method of classical statistics, prior plays a significant role in Bayesian statistics. Given a specific distribution (e.g., Poisson distribution) on the likelihood function, the posterior distribution will be the gathered distribution of prior and the data. For example, suppose the likelihood function follows Poisson distribution, $P(Y|\theta) = \theta^{\sum y_i} e^{-n\theta}$. The conjugate prior is $Gamma(a, b)$

$$\{\theta|Y\} \sim gamma(a + \sum_{i=1}^{n} Y_i, b + n)$$

Hence, the posterior result from the Bayesian model is the combination of information from sampled data and prior. Then, prior selection could be essential as prior can be even more effective than the sampled data. This report will evaluate what prior should be chosen in practical study.

In the more general problem, informative prior and non-informative prior are two kinds of prior that can be used. Informative prior is the prior with specific distribution and gives some information to the result. Non-informative prior, such as flat prior or Jeffery prior, is the prior that does not provide information about the result. Then, which kind of prior should be better in practice? This report will compare and examine the impact of informative and non-informative prior on the model.

1.2 Data

This paper analyzes the effect of different variables on crop yield prediction. The effect of pesticides, temperature, rainfall, and year on Maize yield in Europe is

evaluated. The data comes from FAO[1], TradingEconomics [2] and WorldBank[3]
In general studies in Bayesian, results from previous studies are applied as prior. However, the actual model differs significantly from the model decades ago due to political or other factors. Hence, for the prediction model, finding sufficient data for prior in one country to make the prediction is problematic.
Instead of choosing data from one country, data from neighboring countries can be assumed to have a similar prediction model and used to train the model. However, in practice, the actual model in each country can be very different due to policy and climate. Thus, choosing countries as prior is a concern. The report will evaluate how the selection among a set of eight neighboring countries in Europe(Austria, Belgium, France, Germany, Italy, Spain, Switzerland, and Poland) as prior and prediction would provide better results.

1.3 Model

In this report, the linear regression model will be applied.

$$Y = X\beta + \epsilon$$

In the Bayesian approach, β follows a multivariate normal distribution. ϵ follows the inverse-Gamma distribution. The posterior of this Bayesian model is

$$P(\beta, \sigma^2|X, Y) \propto L(\beta, \sigma^2|X, y) * p(\beta) * p(\sigma^2)$$

1.4 Accuracy indicator

This report will evaluate the model's prediction performance based on MSE and R^2 of the model and later years' data. MSE is the average sum of the squares of the difference between the predicted and actual values. Less MSE means a better model. R^2 is $1 - \frac{SSR_{res}}{SSR_{tot}}$. It is between $-\infty$ and 1. Higher R^2 means a better model.

2 Literature review

Multiple previous studies of the effect of informative prior and non-informative prior have been done. In Wioletta Grzenda's study it evaluates unemployment in individual districts in Poland. Applying previous years' results as the prior shows that the accuracy of the model with informative prior is higher. In Tahir Abbas Malik and Muhammad Aslam's study, the results are more appropriate for data from medical research using Informative Priors. Tahir Abbas Malik and Muhammad Aslam's study of medical data compared normal prior as informative prior and Jeffrey's prior and Haldane prior as non-informative prior. The results are more appropriate when using Informative Priors. Thus, the non-informative result should generally be better than the informative prior result.

[1]https://www.fao.org/
[2]https://tradingeconomics.com/
[3]https://climateknowledgeportal.worldbank.org/

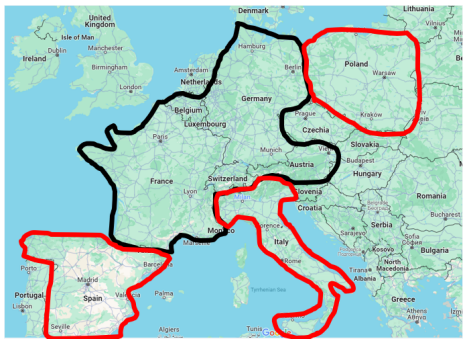

Figure 1: Training Countries and Test Countries

3 Methods

3.1 Variables selection

In predicting maize yield across eight different countries, it is crucial to consider factors such as annual temperature, rainfall, and pesticide usage, as these elements significantly impact crop growth and productivity. Firstly, annual temperature plays a pivotal role in maize cultivation; optimal temperature ranges are essential for seed germination, plant growth, and the development of maize ears. Temperatures too high or too low can hinder these processes, reducing yields. Rainfall is also essential since proper growth of maize depends on having enough water during its growing season. Drought stress can result from insufficient rainfall, and yield loss can be caused by excessive rainfall that causes flooding. Finally, pesticide usage is crucial in managing pests and diseases that can devastate maize crops. Effective pest control ensures healthier plants and can significantly boost yields. However, excessive or inappropriate use of pesticides can lead to environmental harm and may even negatively impact the crop. By analyzing these factors using linear regression and Bayesian methods, we can gain insights into how they collectively influence maize yield, enabling more accurate predictions and informed agricultural practices. After doing the correlation matrix using data from several countries, we found a significant correlation between the year and the pesticide usage amount. Thus, we need to delete one variable between them to avoid multicollinearity. Therefore, we set two regressions, the first using year and the second using pesticide usage, and compare them. The group with year has lower R^2, so choose year instead of pesticide.

3.2 Compared three regressions method

Then, we split our data set into two groups: the first group is the training data, which is constructed by Austria, Belgium, France, Germany, and Switzerland(the black circle below). The informative prior is obtained by OLS from three other countries(Spain, Italy, and Poland, the red circles above).

The results of our comparative analysis suggest that Ordinary Least Squares (OLS) outperformed both Bayesian regression with non-informative priors and Bayesian regression using OLS estimates from three countries as priors. Addi-

Table 1: PRESS and R^2 Values

Methods	PRESS	R^2
OLS	51441220549	0.07305
Non-informative	65790286542974	«0
Prior from 3 countries	102222260841309	«0

Table 2: Parameter Values

Method	β_0	β_1(Year)	β_2(Temp)	β_3(Rainfall)
OLS	-2137288.92	1097.028	0.15	24.12
Non-informative	-2188240.74	1139.53	-0.38	-33.46
Prior from 3 countries	-2054951.91	1072.77	-0.31	-41.41

tionally, the Bayesian model with priors from the three other countries resulted in the most significant Prediction Error Sum of Squares (PRESS), indicating that this model had the poorest predictive performance. Differences in Agricultural Practices may cause this: The agricultural practices, policies, and environmental conditions may vary significantly across countries, making transferring prior information from one set of countries to another less applicable. Moreover, If the underlying model based on the three countries does not capture the true relationship between predictors and maize yield, using its OLS estimates as priors can lead to a misspecified model for the other five countries.

Here, we need to explain that this R^2 is not the same as we expected in normal regression; in normal regression, the range of R^2 will be between 0 and 1, that is because $SSR_{res}<SSR_{tot}$ always exists, but here, including all R^2 (including the next part) is calculated by predicted data. Thus, those R^2s may be less than 0 for the SSR_{res} and less than the SSR_{tot}.

Variability in Weather Patterns: Weather patterns and their impact on maize yield could be highly specific to each country. The predictors that worked for the three countries may not be as predictive for the others, especially if there are microclimatic variations. The graph shows that the β_3 and β_2 are very different with OLS methods. This means that those three tries (Spain, Italy, and Poland) have very different temperatures and rainfalls since the information is unsuitable for predicting other European countries.

3.3 Choose the Best Prior Information

Country-Data Representation (For table rendering purposes, these are abbreviated to dX, where dX corresponds to dataX, where X is an integer between 1 and 8): Data1:Austria Data2:Belgium Data3:France Data4:Germany Data5:Italy Data6:Poland Data7:Spain Data8:Switzerland

In the specified research phase, we aim to determine the optimal prior information for Bayesian regression analysis by conducting an exploratory exercise using linear regression. In doing so, we will use eight data sets containing maize yield data for eight different countries. For clarity and ease of execution of the approach, the datasets have been labeled from data1 to data8, respectively.

The approach is methodical and structured into several steps:

Dataset Segregation: the eight datasets will be used to systematically part the data into two subsets with four datasets each. This implies that 70 possibilities

(C(8,4)) exist where both subsets will be used alternately as either a training or validation set.

Model Training: A linear regression model is constructed using the training set for each combination. In the model, maize yield is considered as a dependent variable. At the same time, Year, Temperature, and Rainfall are considered as independent variables having their effect on output as maize yield. The relationship can be formulated as yield Year + Temperature + Rainfall.

Coefficient Extraction: Four coefficients (which include the intercept) are extracted from the linear regression model during training. These coefficients encapsulate the learned relationship between predictors and maize yield for the training set.

Validation Process: these coefficients are used to predict maize yield against the corresponding validation set. This predictive model uses these coefficients to generate estimated yields for the four datasets that were not part of the training set.

Comparison and Evaluation: The predicted maize yields are compared against the actual recorded yields in the validation datasets. This comparison is quantified by calculating two statistical measures: the Mean Squared Error (MSE) and the coefficient of determination (R^2).

Table 3: Prior Selection Results

Index	Training data	Test data	MSE	R^2
70	d5, d6, d7, d8	d1, d2, d3, d4	2.11E+08	0.000141
42	d2, d3, d5, d8	d1, d4, d6, d7	3.63E+08	0.009751
34	d1, d5, d7, d8	d2, d3, d4, d6	4.43E+08	0.015191
6	d1, d2, d4, d5	d3, d6, d7, d8	3.23E+08	0.015216
62	d3, d5, d6, d7	d1, d2, d4, d8	1.83E+08	0.031926
36	d2, d3, d4, d5	d1, d6, d7, d8	3.43E+08	0.093091
32	d1, d5, d6, d7	d2, d3, d4, d8	1.73E+08	0.103407
2	d1, d2, d3, d5	d4, d6, d7, d8	2.95E+08	0.140489
57	d3, d4, d5, d7	d1, d2, d6, d8	4.28E+08	0.147623
68	d4, d5, d7, d8	d1, d2, d3, d6	3.99E+08	0.179816
27	d1, d4, d5, d7	d2, d3, d6, d8	3.77E+08	0.183410
64	d3, d5, d7, d8	d1, d2, d4, d6	3.93E+08	0.214371
21	d1, d3, d5, d7	d2, d4, d6, d8	3.67E+08	0.230000
28	d1, d4, d5, d8	d2, d3, d6, d7	3.73E+08	0.255780
22	d1, d3, d5, d8	d2, d4, d6, d7	3.61E+08	0.302362
58	d3, d4, d5, d8	d1, d2, d6, d7	3.8E+08	0.304349
16	d1, d3, d4, d5	d2, d6, d7, d8	3.41E+08	0.346530

It is imperative to recognize that the objective of this research is not merely to achieve high predictive accuracy through the utilization of extensive data. Instead, the aim is to ascertain the most effective prior information within constrained data availability. To this end, a methodological decision was made to delimit the scope of the study to a selection of only four countries from the available eight. This decision was driven by the need to simulate a more realistic scenario where data are unavailable, aligning with the research goal of identifying the best priors in a situation that is reflective of limited information. The selection process for these four countries was underpinned by rigorous statistical criteria,

ensuring that the priors chosen would provide a robust basis for subsequent predictive analyses within the Bayesian paradigm. The outcomes of this process are of considerable significance, offering insights into the application of Bayesian methods in agricultural data science, especially in scenarios characterized by data scarcity.

Optimal Prior Selection: The best performing model will be identified by analyzing the MSE and R^2 across all 70 combinations — indicative of the smallest MSE and highest R^2. The coefficients from this model will be considered the most suitable set of prior information for subsequent Bayesian regression analysis.

The table on the previous page is part of the result, and as we can compare the MSE and R^2 of all 70 combinations, we can find that the data1,3,4,5 (Austria, Belgium, Germany, and France) is the best combination as it has the largest R^2. As we can see, all these countries are in the middle of Europe. This may explain why the above prior in Section 2.2(Spain, Italy, and Poland) is not a good choice. They are far away in location, so their annual temperature and rainfall vibrate too magnificently. In the next step, we choose the three combinations with the largest R^2 (in the bottom three rows of Table 3) to conduct the informative Bayes regression and compare it with non-informative prior to see whether those new combinations offer suitable prior information for the maize yield prediction.

4 Bayesian Model

Two Bayesian models with distinct prior assumptions were developed.

1. Non-Informative Prior Model: We have chosen the flat prior as the base prior to testing the performance of the model with no prior information on the countries.

2. Informative Prior Model: It will allow us to introduce empirical data from an OLS regression linked to a subset of countries we have selected before. We extract the mean and the covariance matrix to incorporate them into the prior.

4.1 MCMC Details

The Bayesian inference was run where the MCMC simulations were applied by the ‘emcee‘ package. By generating the samples from the posterior distribution gives a full understanding of all the possible parameter values and also the probabilities for these parameter values. A 5000-step MCMC simulation was conducted for each model to ensure sufficient parameter space explorations were made. The parameters estimated for the intercept, temperature, rainfall, and year by applying the MCMC procedure estimated the posterior distribution of all the parameters in the models. We also used 32 walkers to enhance the reliability of convergence.

4.2 Simulation Studies on Synthetic Data

The study employed synthetic data from a subset of countries for 2014-2020 to evaluate the models. This data serves as a test set to simulate real-world applications and assess predictive performance. The choice of synthetic data allowed for controlled experimentation and validation of the models under known conditions.

For convenience, we set the model for data2, data6, data7, and data8 as Model 1, the model for data2, data4, data6, and data7 as Model 2, the model for data1, data2, data6, and data7 as Model 3, and the remaining four data sets that are not chosen are set as the priors from their corresponding OLS results for each model, respectively. We first formulated three models from 1990 to 2013, and then we used the models to predict the yield data for 2014 to 2020. The tables are listed in the next section.

4.3 Tables

The model parameter results and the model performance results are shown in the tables. Note that for convenience, we will display the minimum and maximum values of MSE out of 8 predicted data sets, with the counts of R^2 greater than 0 for each model.

Table 4: Model Parameter Results

Model	β_0	β_1	β_2	β_3
Model 1 Non-informative	-2873729.24	1435.59	3570.53	52.68
Model 1 Informative	-1841048.99	935.63	2452.99	30.37
Model 2 Non-informative	-2337720.99	1165.45	1977.73	95.21
Model 2 Informative	-1689800.56	872.77	1169.15	19.31
Model 3 Non-informative	-2677530.12	1332.06	3589.07	68.16
Model 3 Informative	-1714064.13	884.20	1572.77	14.49

Table 5: Model Performance Results

Model	Min MSE	Max MSE	$R^2 > 0$ Counts
Model 1 Non-informative	64574971.54	1342154571.36	1
Model 1 Informative	24683449.62	683473032.33	3
Model 2 Non-informative	55679641.41	3516922539.48	1
Model 2 Informative	24024127.23	985257197.89	2
Model 3 Non-informative	52844851.54	1787546977.09	1
Model 3 Informative	24784424.90	981522550.06	1

4.4 Data Analysis and Results Interpretation

4.4.1 Prediction Results Analysis

The Bayesian models' predictive performance was assessed using the Mean Squared Error (MSE) and R-squared (R^2) metrics across different data sets representing various countries.

Model 1 (Non-Informative Prior) shows a relatively high variance in MSE spread across all eight data sets, meaning different prediction performances between each dataset. Moreover, 7 out of 8 countries represented a negative R^2 value, signifying that this model lacks consideration of the data variance effectively. However, positive R^2 occurred once in this model, which indicates the model has some degree of predictability.

The MSEs have reduced significantly compared to the non-informative model, indicating generally an improvement in accuracy. For Model 1 (Informative Prior), three countries showed positive R^2 values, indicating a significant improvement to the non-informative counterpart benefiting from the employment of OLS parameters from the prior.

Model 2 and Model 3 showed the same trend as Model 1, showing more negative R^2 values of all the populous country's predictions. For all data sets, one positive R^2 value has been exhibited in both non-informative prior models. In comparison, informative prior models have two positive R^2 values and one positive R^2 value for Model 2 and Model 3, respectively. From the results, it's observable that the models have a varying level of predictability, and the data sets with more positive R^2 values tend to suggest that some country data are more accurately predictable based on the chosen variables.

4.4.2 Interpretation of Results

Generally, the country-based performance difference of the models greatly varied with different countries to reflect the variance of each country's agricultural and climatic conditions. Generally, from the model prediction results, the informative prior models continually outperformed the non-informative models. This suggests that the blend of OLS results as prior increased explanatory power of the model to a certain level. On the other hand, assuming most R^2 values remain negative across all the datasets, it means very poor borrowing of the models, possibly due to over-simplification of the models or exclusion by choice of other relevant variables. The point is illustrated that there is a need for more complex or tailored models to capture nuances of agricultural yield determinants better.

5 Conclusion

In conclusion, the best prior information is constructed by Austria, Belgium, Germany, and France in the case where only four groups can be selected. The annual average temperature and rainfall lead to an inaccurate prediction since the previous information would lead to incorrect coefficients. Geographical differences of each country determine variations in coefficients of how rainfall and average annual temperature affect corn yields. So, the varying effectiveness of the models across separate data sets highlights a need for continual refinement and adaptation to suit particular agricultural contexts and data characteristics. And the model accuracy is significantly improved when we use the prior information properly. The informative models generally tended to perform better with smaller MSE values and more positive R^2 values in general. This implies that the integration to the model from OLS results may show how the model improvement should be done under some context. On the other hand, high values for MSE in informative models indicate that this approach is not equally competent for all datasets outlining different complexities in agricultural yield predictions and influence from region-specific factors.

References

[1] Grzenda, Wioletta. "Informative Versus Non-Informative Prior Distributions and their Impact on the Accuracy of Bayesian Inference." *Statistics in Transition. New Series* 17.4 (2016): 763-780.

[2] Malik, Tahir Abbas, and Muhammad Aslam. "Bayesian Inference for Logit-Model using Informative and Non-informative Priors." *Journal of Statistics* 21.1 (2014)