Scientific Research Publishing

Open Journal of Genetics, 2025, 15(2), 47-61
https://www.scirp.org/journal/ojgen
ISSN Online: 2162-4461
ISSN Print: 2162-4453

Improving Prediction of Genotypic Values in Historical Crop Trial Data via Stepwise Adjustment Method

Jixiang Wu[1*], Linghe Zeng[2], Johnie N. Jenkins[1], Jack C. McCarty[1]

[1]Genetics and Sustainable Agriculture Research Unit, USDA-ARS, Mississippi State, USA
[2]Crop Genetics Research Unit, USDA-ARS, Stoneville, MS, USA
Email: *Jixiang.wu@usda.gov

How to cite this paper: Wu, J.X., Zeng, L.H., Jenkins, J.N. and McCarty, J.C. (2025) Improving Prediction of Genotypic Values in Historical Crop Trial Data via Stepwise Adjustment Method. *Open Journal of Genetics*, **15**, 47-61.
https://doi.org/10.4236/ojgen.2025.152005

Received: April 25, 2025
Accepted: June 15, 2025
Published: June 18, 2025

Abstract

Improving prediction of genotypic values from long-term historical crop trial data will enhance the utilization for both genetic study and crop improvement. However, because many long-term historical crop trial data are highly unbalanced due to the frequent changes in test entries and locations, it is statistically challenging to analyze the long-term historical data simultaneously without proper adjustment. In this study, we proposed a stepwise method that can be used to adjust the differences caused by environmental conditions among years. First, this method was evaluated by Monte Carlo simulation, which showed that this stepwise adjustment method can consistently improve the prediction impacted by environmental conditions among years. Second, the stepwise adjustment method was applied to a 16-year soybean trial data set in South Dakota and showed that model fitness for genetic gain over these 16 years was improved compared to the model fitness using the non-adjusted data (0.85 vs 0.48). The annual genetic gain estimated from non-adjusted data was 1.35 bushel/ac while the adjusted annual genetic gain was 0.72 bushel/ac, which was more in line with annual state soybean production from 1987-2011.

Keywords

Annual Genetic Gain, Genotypic Effect Prediction, Historical Crop Trial Data, Stepwise Adjustment, Overlapped Genotypes

1. Introduction

Plant researchers have been spending decades of efforts to conduct regional and national crop trials before entries of interest are released to the market. However, these historical crop trial data have been underutilized because processing long-

term crop trial data simultaneously is still statistically challenging due to high unbalanced data structures. Revisiting long-term historical crop trial data with appropriate statistical methods will help us reveal more desirable genetic information that can be used for crop improvement.

Multi-environment crop trial aims to determine those entries with high yield potential and stable performance across a wide range of environmental conditions. With multi-environment trials, genotype-by-environment (GE) interaction, which is relevant to yield stability, is one of most important parameters of interest to many statisticians [1]-[4]. Many approaches have been proposed and applied to multi-environment trial data analyses. Linear regression-based yield stability analysis has been a major focus in multi-environment crop yield trials [5]-[9]. Additive main and multiplicative interaction (AMMI) method [10] has gained more popularity because it targets both mean performance and environment-specific performance. GGEbiplot, a graphical tool, can be used for multi-environment trial data for yield stability or GE interaction [11]. Linear mixed model (LMM) approaches have also been applied to analyze multi-environment trial data to explore genotypic effects, environmental effects, and GE interaction effects with its flexibility for unbalanced data structure and/or missing data points [12].

It is unlikely that the above-mentioned methods can be used to conduct yield stability or GE interaction across years for a crop trial data set spanning years because of serious imbalance in the data [13] [14]. Thus, a different focus such as genetic progress among years rather than GE or yield stability investigation could be more appropriate [4] [15]. Estimation of historical genetic gain/progress will be impacted by the degree of confounding of genotypic and environmental effects. It is statistically impossible to compare genotypes developed today and 20 years ago without adjusting environmental conditions among years or without the same checks being used. Likely, this is one of key factors preventing extensive investigation on long-term crop trial data across years. One way to investigate genetic progress is to select several varieties released from different time periods and evaluate them under one or a few environments [15]-[17] so that these varieties can be evaluated under the same environmental conditions. However, with this approach, there could be several potential issues: 1) Only a small portion of varieties from the historical trials could be evaluated; 2) The representative environmental conditions could be narrow, not representing a wide range of crop growing regions where crop trials have been conducted for decades; 3) additional huge cost for labor and land is needed. The second approach is to use overlapped/carryover standards (or checks) to conduct environmental adjustment from year to year [18]. For example, in the national cotton variety trial, the same three or four standards are used within each cycle (equivalent to every three years, while there is one or two standards that roll over to the next cycle). With this approach, it is possible to re-visit the long-term crop trial data without adding field trial cost. However, with only one standard rolled over to the following cycle such as in cotton trial [18], it is statistically unknown if a small number of standards like one or two are

sufficient for adjustment. Additionally, if one standard in a particular growing season or location is missing, then the adjustment could cause a critical issue.

It is well-known that good performers from the current year trial will likely remain the following year. In other words, these good performers could be repeated for at least one more year. We observed that a range of genotypes were overlapped/repeated between every two consecutive years in various national and regional crop trial data (interested readers can refer to published crop trial reports from various institutes). Therefore, it will enhance the value of long-term crop trial data when a new method can be provided with its efficiency validated.

In this study, our first objective was to propose a stepwise method to adjust environmental effects among years using overlapped entries between every two consecutive years. Our second objective was to evaluate the adjustment efficiency of this new method by Monte Carlo simulation. Our third objective was to integrate this method to adjust a 16-year soybean trial data set from South Dakota and to re-estimate annual genetic progress accordingly. The purpose of this study was to provide a new method to better predict genotype performance from long-term historical crop trial data so that desirable genetic information can be further utilized for genetic study and crop improvement

2. Methods and Simulations

2.1. Stepwise Adjustment Method

Assuming there is at least overlapping entries between any two consecutive years, the procedure for the stepwise adjustment method can be described as follows,

Step 1: Assume that $o_{(1,2)}$ is overlapped entries between years 1 and 2. Calculate the mean values of $o_{(1,2)}$ overlapped entries for years 1 and 2, $\bar{m}_{1(1,2)}$ and $\bar{m}_{2(1,2)}$, respectively.

Step 2: Calculate the mean difference between the first year and the second year using the equation (1)

$$\Delta_{(1,2)} = \bar{m}_{2(1,2)} - \bar{m}_{1(1,2)} \tag{1}$$

where, $\Delta_{(1,2)}$ is the mean difference between the first year and the second year based on $o_{(1,2)}$ overlapped entries. If $\Delta_{(1,2)}$ is positive, it indicates that year 2 is higher than year 1. In the same manner, if $\Delta_{(1,2)}$ is negative, it indicates that year 2 is lower than year 1.

Step 3: Adjust phenotypic values for year 2 with the equation (2)

$$\boldsymbol{y}_{2(adj)} = \boldsymbol{y}_2 - \Delta_{(1,2)} \tag{2}$$

where, $\boldsymbol{y}_{2(adj)}$ is the vector of the adjusted phenotypic values for year 2 and $\boldsymbol{y}_2$ is the vector of the non-adjusted phenotypic values for year 2. The adjusted values $\boldsymbol{y}_{2(adj)}$ are comparable to the phenotypic performance under the mean environmental conditions in year 1.

Step 4: Adjust phenotypic values for the year $(h+1)$ when $h \geq 2$.

Given $h \geq 2$ and there are $o_{(h,h+1)}$ overlapped entries between consecutive year h and year $(h+1)$, the adjustment for year $(h+1)$ can be conducted us-

ing the following equation 3:

$$\boldsymbol{y}_{(h+1)(adj)} = \boldsymbol{y}_{(h+1)} - \Delta_{(h,h+h)} \tag{3}$$

where, $\boldsymbol{y}_{(h+1)(adj)}$ is the vector of the adjusted phenotypic values for year $(h+1)$; $\boldsymbol{y}_{(h+1)}$ is the vector of the non-adjusted phenotypic values for year h; and $\Delta_{(h,h+1)}$ can be calculated by the following equation (4),

$$\Delta_{(h,h+1)} = \bar{m}_{(h+1)(h,h+1)} - \bar{m}_{(h)(h,h+1)(adj)} \tag{4}$$

where, $\bar{m}_{(h+1)(h,h+1)}$ is the non-adjusted mean value of $o_{(h,h+1)}$ overlapped genotypes for year $(h+1)$; $\bar{m}_{(h)(h,h+1)(adj)}$ is the adjusted mean value of $o_{(h,h+1)}$ overlapped genotypes for year h. $\Delta_{(h,h+1)}$ is the mean difference of phenotypic values of $o_{(h,h+1)}$ overlapped genotypes between years h (adjusted mean) and $h+1$ (non-adjusted mean). The adjustment process is recursive until the last year/season in the trial. All adjusted phenotypic value vectors from the second year are statistically comparable to the mean environmental conditions under year 1.

2.2. Simulation Study

It will be helpful to numerically evaluate the effectiveness of the above stepwise adjustment method. An effective way to evaluate a statistical method and/or model is Monte Carlo simulation technique [3] [19] [20]. However, it is important that the simulated data should be generated from a linear model representing a multi-year and multi-location crop trial that can be described as follows:

$$y_{hijk} = \mu + Y_h + L_i + G_j + YL_{hi} + YG_{hj} + LG_{ij} + YLG_{hij} + B_{k(hi)} + e_{hijk} \tag{5}$$

where y_{hijk} = phenotypic value for jth genotype for kth block under ith location and hth year; μ = population mean; Y_h = year effect; L_i = location effect; G_j = genotypic effect; YL_{hi} = year-by-location interaction effect; YG_{hj} = genotype-by-year interaction effects; LG_{ij} =genotype-by-location interaction effect; YLG_{hij} = genotype-by-year-by-location interaction effects; $B_{k(hi)}$ = block effect within each location and year; and e_{hijk} = random error. In our simulation study, we set all effects as random except for population mean.

If all entries are grown in all years and all locations, then all effects in the above linear model can be dissected either by ANOVA (analysis of variance) or linear mixed model approaches. However, as mentioned above, many genotypes are not repeated among years, genotypic effects cannot be separated from environmental effects among years or genotype-by-environment interaction effects, which could impact the efficiency of the stepwise adjustment method. Therefore, our simulated crop trial data were generated reflecting real-world scenarios from the linear model (Eq. 1) and nine settings of variance components (**Table 1**) were applied to reflect the potential impacts from year effects and genotype-by-environmental effects. However, interested readers may use other settings to evaluate the outcomes for other particular purposes.

Other parameters used for simulation study are as follows. Within each year 20 entries are repeated four times with an RCB design under each location. The num-

bers of overlapping entries for simulation are 1, 2, 4, 6, 8, 10, and 15. The process is repeated five (5), 10, and 15 years, respectively. The simulation for each of 189 ($9 \times 7 \times 3 = 189$) settings was repeated 100 times.

Table 1. Nine (9) sets of variance components used for simulation study.

Variance component†	Set1	Set2	Set3	Set4	Set5	Set6	Set7	Set8	Set9
V_G	40	40	40	40	40	40	40	40	40
V_Y	40	40	40	60	60	60	80	80	80
V_L	20	20	20	20	20	20	20	20	20
V_{GY}	20	10	0	20	10	0	20	10	0
V_{GL}	20	10	0	20	10	0	20	10	0
V_{YL}	20	10	0	20	10	0	20	10	0
V_{GYL}	20	10	0	20	10	0	20	10	0
V_B	20	20	20	20	20	20	20	20	20
V_e	20	20	20	20	20	20	20	20	20

†: V_G = variance component for genotype; V_Y = variance component for year; V_L = variance component for location; V_{YL} = variance component for year-by-location interaction; V_{GY} = variance component for genotype-by-year interaction; V_{GL} = variance component for genotype-by-location interaction; V_{GYL} = variance component for genotype-by-year-by-location interaction; V_B = variance component for block; and V_e = variance component for random error respectively.

Once data were generated, genotypic values for all entries were calculated from control group data ($\widehat{GV_{ck}}$), non-adjusted data ($\widehat{GV_0}$), and adjusted data ($\widehat{GV_A}$). Correlation coefficients were calculated between $\widehat{GV_{ck}}$ and $\widehat{GV_0}$ and between $\widehat{GV_{ck}}$ and $\widehat{GV_A}$, respectively. The higher a correlation coefficient is, the better predicted genotypic values are. The function lmm. simudata available in the R library, minque [21] was used to generate simulated trial data. The simulation study was conducted by the R scripts, which were developed by the senior author of this study and all data analyses including simulation and application were conducted at the platform RStudio [22] [23].

3. Results

3.1. Simulation Study

In our simulation study we focused on determining factors that could impact prediction of genotype values. The first factor is numbers of overlapped entries (1, 2, 4, 6, 8, 10, and 15 used in this study) between every two consecutive years/seasons. The second factor is length of trial years (5, 10, and 15 used in this simulation study). Because both year effects and genotype-by-environment interaction effects could impact the prediction of genotypic effects, we also consider variation of years (sets 1 - 3 → 4 - 6 → and 7 - 9, low → high) and variances

of two- and three-way interactions (sets 1, 4, 7→ 2, 5, 8 → 3, 6, 9, high → low) (Table 1). As mentioned in Methods, genotypic values were calculated: $\widehat{GV_{ck}}$, GV_0, and $\widehat{GV_A}$ for control group data, non-adjusted trial data, and adjusted trial data, respectively. Parameters r and r_A are defined as correlation coefficients between $\widehat{GV_{ck}}$ and $\widehat{GV_0}$ and between $\widehat{GV_{ck}}$ and $\widehat{GV_A}$, respectively. A high correlation coefficient with control group $\widehat{GV_{ck}}$ implies the genotypic values are better predicted to those for the control group. The simulation results provided in Tables 2-4 are mean correlation coefficients based on 100 simulations of data for each setting.

Table 2. Summarized simulation results for nine (9) sets of variance components settings with seven (7) different overlapped entry numbers with five (5) years of trial.

		Overlapped entry number							
	Correlation coefficient†	1	2	4	6	8	10	15	Mean
Set1	r	0.693	0.699	0.710	0.730	0.748	0.775	0.829	0.741
	r_A	0.689	0.757	0.795	0.838	0.860	0.870	0.915	0.818
Set2	r	0.739	0.713	0.728	0.743	0.771	0.786	0.868	0.764
	r_A	0.772	0.837	0.867	0.888	0.905	0.919	0.954	0.877
Set3	r	0.773	0.755	0.768	0.777	0.805	0.834	0.863	0.796
	r_A	0.974	0.983	0.987	0.989	0.991	0.993	0.995	0.987
Set4	r	0.652	0.663	0.657	0.691	0.728	0.748	0.821	0.709
	r_A	0.698	0.751	0.807	0.824	0.859	0.885	0.917	0.820
Set5	r	0.677	0.677	0.678	0.739	0.719	0.764	0.839	0.728
	r_A	0.782	0.832	0.876	0.891	0.900	0.920	0.953	0.879
Set6	r	0.701	0.707	0.701	0.722	0.752	0.793	0.847	0.746
	r_A	0.973	0.981	0.986	0.989	0.991	0.993	0.995	0.987
Set7	r	0.639	0.636	0.635	0.643	0.679	0.710	0.792	0.676
	r_A	0.715	0.742	0.803	0.829	0.848	0.872	0.920	0.818
Set8	r	0.652	0.643	0.675	0.669	0.710	0.713	0.815	0.697
	r_A	0.781	0.841	0.866	0.890	0.900	0.915	0.952	0.878
Set9	r	0.670	0.663	0.674	0.701	0.743	0.759	0.804	0.716
	r_A	0.971	0.981	0.987	0.989	0.991	0.992	0.995	0.987
Mean	$\bar{r}$	0.690	0.686	0.693	0.714	0.741	0.767	0.832	0.732
	$\bar{r}_A$	0.833	0.869	0.896	0.912	0.924	0.935	0.959	0.904

†: r and r_A are correlation coefficients between predicted genotype values from control group and data without adjustment and between predicted genotype values from control group and adjusted data.

Table 3. Summarized simulation results for nine (9) sets of variance components settings with seven (7) different overlapped entry numbers with 10 years of trial.

		Overlapping entry number							
	Correlation coefficient†	1	2	4	6	8	10	15	Mean
Set1	r	0.652	0.669	0.685	0.696	0.721	0.761	0.833	0.717
	r_A	0.597	0.675	0.735	0.771	0.813	0.835	0.910	0.762
Set2	r	0.686	0.686	0.715	0.729	0.752	0.790	0.856	0.745
	r_A	0.682	0.754	0.835	0.854	0.886	0.892	0.943	0.835
Set3	r	0.720	0.741	0.744	0.763	0.772	0.822	0.872	0.776
	r_A	0.953	0.971	0.979	0.985	0.988	0.989	0.995	0.980
Set4	r	0.619	0.642	0.645	0.654	0.672	0.720	0.801	0.679
	r_A	0.579	0.669	0.727	0.774	0.804	0.838	0.908	0.757
Set5	r	0.630	0.630	0.663	0.675	0.696	0.742	0.807	0.692
	r_A	0.692	0.781	0.828	0.859	0.877	0.894	0.947	0.840
Set6	r	0.654	0.665	0.700	0.699	0.736	0.758	0.823	0.719
	r_A	0.953	0.969	0.979	0.986	0.988	0.990	0.995	0.980
Set7	r	0.574	0.586	0.600	0.621	0.649	0.703	0.785	0.645
	r_A	0.577	0.675	0.745	0.775	0.798	0.837	0.905	0.759
Set8	r	0.589	0.625	0.611	0.627	0.651	0.697	0.788	0.655
	r_A	0.684	0.770	0.831	0.844	0.879	0.896	0.946	0.836
Set9	r	0.590	0.618	0.633	0.656	0.670	0.712	0.794	0.668
	r_A	0.952	0.969	0.980	0.985	0.987	0.990	0.995	0.980
Mean	$\overline{r}$	0.637	0.653	0.670	0.682	0.705	0.746	0.818	0.702
	$\overline{r}_A$	0.762	0.820	0.862	0.882	0.901	0.915	0.954	0.871

†: r and r_A are correlation coefficients between predicted genotype values from control group and data without adjustment and between predicted genotype values from control group and adjusted data.

Table 4. Summarized simulation results for nine (9) sets of variance components settings with seven (7) different overlapped entry numbers with 15 years of trial.

		Overlapped entry number							
	Correlation coefficient†	1	2	4	6	8	10	15	Mean
Set1	r	0.695	0.715	0.719	0.719	0.743	0.777	0.822	0.741
	r_A	0.688	0.759	0.799	0.829	0.848	0.873	0.915	0.816
Set2	r	0.728	0.740	0.740	0.762	0.764	0.812	0.858	0.772
	r_A	0.801	0.842	0.870	0.886	0.902	0.925	0.952	0.883

Continued

Set3	r	0.768	0.757	0.791	0.786	0.818	0.826	0.868	0.802
	r_A	0.973	0.981	0.987	0.989	0.991	0.993	0.995	0.987
Set4	r	0.647	0.649	0.671	0.675	0.732	0.767	0.811	0.707
	r_A	0.707	0.752	0.792	0.839	0.851	0.879	0.914	0.819
Set5	r	0.685	0.679	0.689	0.711	0.742	0.770	0.816	0.727
	r_A	0.799	0.841	0.868	0.884	0.909	0.922	0.949	0.882
Set6	r	0.711	0.697	0.716	0.740	0.751	0.768	0.841	0.746
	r_A	0.972	0.983	0.987	0.989	0.991	0.993	0.995	0.987
Set7	r	0.621	0.631	0.649	0.661	0.680	0.713	0.775	0.676
	r_A	0.702	0.761	0.794	0.831	0.845	0.884	0.914	0.819
Set8	r	0.639	0.663	0.655	0.675	0.679	0.736	0.784	0.690
	r_A	0.783	0.841	0.872	0.894	0.905	0.924	0.949	0.881
Set9	r	0.658	0.627	0.681	0.694	0.714	0.739	0.793	0.701
	r_A	0.970	0.982	0.986	0.989	0.991	0.992	0.995	0.986
Mean	$\bar{r}$	0.686	0.686	0.703	0.716	0.737	0.767	0.821	0.731
	$\bar{r}_A$	0.837	0.873	0.894	0.912	0.922	0.938	0.957	0.905

†: r and r_A are correlation coefficients between predicted genotype values from control group and data without adjustment and between predicted genotype values from control group and adjusted data.

Overall, as the number of overlapping entries between consecutive years/seasons increases, both correlation coefficients increase with a few exceptions for non-adjusted group when overlapping number is two or small (last two rows in Tables 2-4), suggesting that overlapping entry number is a key factor to increase the accuracy of predicted genotypic values. Predicted genotypic values from the adjusted data had higher correlation values with genotypic values from control groups than predicted genotypic values from non-adjusted data with genotypic values from control groups with a few exceptions (last two rows and last columns in Tables 2-4), suggesting that the difference caused by the environmental conditions among years can be adjusted by the method proposed in this study. For example, based on the simulation data on five years of trials, the difference in correlation coefficients ranged from 0.127 to 0.203, peaked at 6 to 8 overlapped entries, resulting in 15.3 to 29.3% increase compared to the non-adjusted group (last two rows in Table 2). Similar results can be found for the simulated data with 10 and 15 years of trial (Table 3 and Table 4). The predicted efficiency is related to environmental variations among years. With adjustment, prediction efficiency is better when variance is high (80) compared to smaller year variation (40 and 60, refer to the last columns in Tables 2-4). In addition, as an example shown in Figure 1,

the adjusted predicted genotypic values were highly correlated with the preset genotypic values compared to the non-adjusted ones with a year variation of 80, five years of trials, and 15 overlapped genotypes. It is also noticeable that low GE interaction variation results in improved prediction when adjustment is applied (set 3 > 2 > 1; set 6 > 5 > 4, or set 9 > 8 > 7, the last columns in Tables 2-4). Additionally, it appears that prediction efficiency remains about the same for different numbers of trial years when adjustment is applied, suggesting that this method is suitable to adjust long-term historical trial data to predict genotype values across trial years.

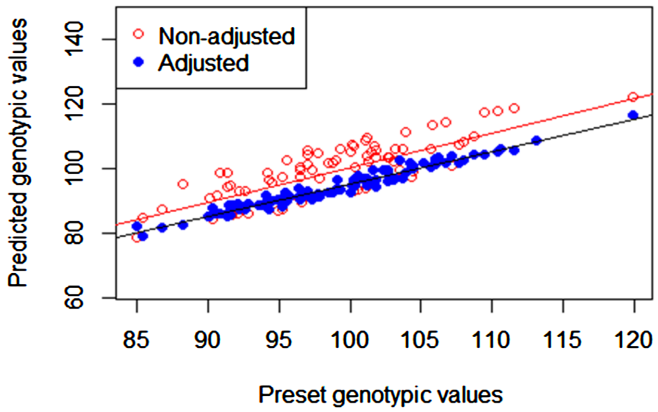

Figure 1. Predicted genotypic values from preset, non-adjusted, and adjusted groups (one case from the simulated data based on set 9 in Table 1).

Overall, with our simulation data, we can conclude that this stepwise adjustment method can help adjust the difference caused by environmental differences among years. Its efficiency is related to the number of overlapping entries between two consecutive years and year variance. As expected, high GE interaction will decrease the adjustment efficiency.

3.2. Demonstration

Though this stepwise adjustment method could be applied to different crop trial data, we choose to use the soybean trial data from six locations in eastern South Dakota covering 2001 through 2016 as our application because soybean production was significantly impacted by the weather conditions among years. The soybean trial data are available at the website http://igrow.org for more detailed information regarding the trial data.

A total 2946 different soybean genotypes were grown over 16 years (2001-2016) [4]. The numbers of overlapping entries between 15 pairs of consecutive years are provided in Table 5. The overlapped entries ranged from 36 to 119 among these 15 pairs of consecutive years (Table 5). The proportion of overlapped genotypes between each two consecutive years varied from 9 to 26%. These overlapped soybean genotypes could be a useful leverage for our new method to adjust the predicted genotypic values.

Table 5. Numbers of overlapping entries between two consecutive years for soybean trials in South Dakota from 2001-2016.

Year	Overlapped†	Combined‡	Proportion*
2001	416	416	1.0000
2002	110	699	0.1574
2003	119	595	0.2000
2004	98	567	0.1728
2005	92	546	0.1685
2006	95	465	0.2043
2007	81	424	0.1910
2008	90	388	0.2320
2009	78	395	0.1975
2010	36	387	0.0930
2011	60	338	0.1775
2012	77	329	0.2340
2013	60	303	0.1980
2014	50	290	0.1724
2015	76	289	0.2630
2016	41	301	0.1362

†: Number of overlapped genotypes between current and previous years; ‡: total genotypes between current and previous years; *: proportion = overlapped genotypes/total genotypes.

The adjusted and non-adjusted annual means (bushel/acre) are provided in Figure 2. The difference between these two means for a particular year suggests the impact of weather conditions for soybean production. A lower adjusted mean indicates more suitable weather conditions compared to the weather conditions in 2001. The results showed that the non-adjusted annual yield was higher than the adjusted annual yield for most years, for example 2005, 2007, 2009, and 2013 to 2016, indicating that the weather conditions in these years were more suitable for soybean production compared to the weather conditions in 2001 (Figure 2). However, the non-adjusted annual yield means were much lower than the adjusted annual yield means for 2003 and 2012, suggesting that weather conditions in these two years were not suitable for soybean production as compared to those in 2001. One major reason for these two low yielding years (2003 and 2012) was caused by early frost-kill in September while weather conditions were better for soybean seed growth and development maturement in other years. The adjusted and non-adjusted means were similar in 2006, indicating that the weather conditions were similar between 2001 and 2006.

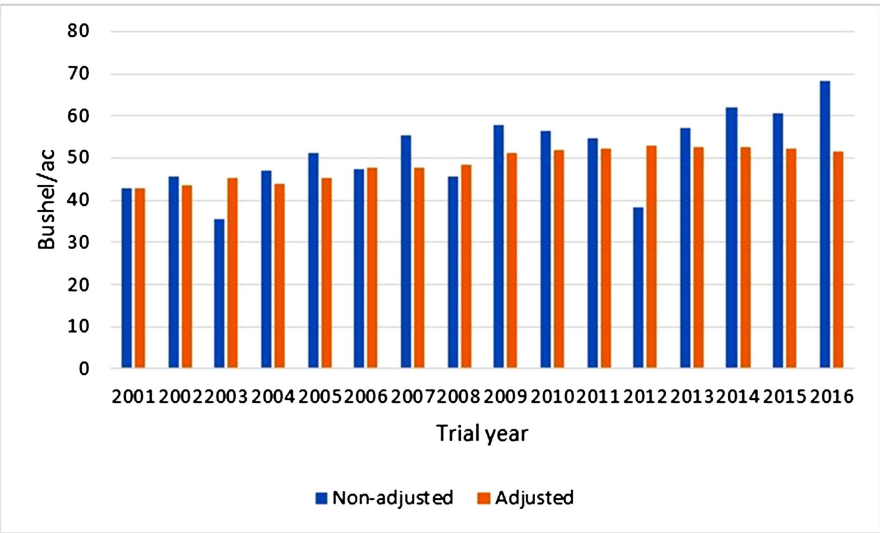

Figure 2. Non-adjusted and adjusted annual means for soybean yield from South Dakota soybean yield trial in 2001-2016.

After data adjustment, the trend for annual means was smoother than that for the non-adjusted data. The adjustment also resulted in an improved model fitness for the genetic gain model over these 16 years compared to that for non-adjusted data (**Table 6**). The annual genetic gain estimated from non-adjusted data was 1.35 bushel per acre (one bushel is equivalent to 60 lbs.) while the annual genetic gain for adjusted data was 0.72 bushel per acre, equivalent to about 10 bushels per acre over 16 years. It appears that favorable weather conditions in 2014-2016 led to historically high yield and thus inflated the annual genetic gain if data were not adjusted. The adjusted annual gain was more consistent to the nation-wide annual increase 0.53 bushel per acre based on soybean production from 1981 to 2024 [24]. On average, the annual increase in South Dakota was 0.47 bushel per acre based on the soybean production in the similar areas during the period of 1987 to 2011 [25].

Table 6. Annual genetic gain and model fitness expressed as adjusted coefficient of determination for non-adjusted and adjusted soybean trial data from eastern south Dakota State environment (2001-2016).

	Unadjusted	Adjusted
Genetic gain	1.35 (bushel/ac)	0.72 (bushel/ac)
R^2	0.48	0.85

Based on the annual soybean production reports [24] [25], the weather conditions among years in South Dakota played a major impact on soybean production, suggesting the degree of adjustment on genotypic values would be considerably large. The results were consistent with our simulations as shown in **Tables 2-4**.

4. Discussion

National crop variety trial is a dynamic process with test entries and locations varying from year to year. However, such a dynamic process causes a serious need to better predict genotypic values which were not repeated across years and thus limits full utilization of long-term historical trial data. Because of this data structure issue, within-year data or balanced data across a few years were more popularly analyzed and utilized for selection and stability determination [6] [12] while long-term crop trial data have remained underutilized [26]. To better predict genotype values, a proper adjustment is needed to make the long-term crop trial data statistically comparable to the phenotypic performance under a particular year condition. That was our major goal of this study.

From many annual trial reports published online, we observed that it is a common pratice for a number of entries, especially good performers, to be available for two consecutive years though they vary from year to year. This study thus aimed to use the information of these overlapped entries between every two consective years to adjust environmental impacts aross years. As addressed in Methods section, such a process is a step-by-step (year-by-year) adjustment. The efficiency for this stepwise method was numerically evaluated through Monte Carlo simulation. Our simulation showed that the adjustment method can enhance prediction of genotypic values as evidenced by the improved correlation coefficient with control group compared to non-adjusted results (**Tables 2-4**). As expected, the adjustment is more significant when year variance is larger, suggesting a key objective in this study was achieved. The simulation also showed that a higher number of overlapped entries between two consecutive years tends to have a higher adjustment efficiency compared to a smaller overlapped entry number. Thus, using more overlapped entries every two consecutive years can improve adjustment efficiency compared to using standards only, given the same data set. The conclusion is desired by researchers. Number of years in trials doesn't show numerical impact on the adjustment efficiency, indicating that this method can be used for long-term historical trial data adjustment. The statistical model used for our simulation study was representative for a multi-year and multi-location trial, suggesting the adjustment method is practical. Thus, this adjustment method makes the simultanaous analysis of long-term historical data possible and provides a way to make different genotypes grown in different years comparable. However, our simulation showed that the adjustment efficiency can be negatively impacted by interaction effects especially by genotype-by-environment interaction effects but the impact can be reduced by using more overlapping entries. Such a conclusion is expected because interaction effects are confounded with environmental effects among years. Using more overlapped entries could help reduce the impact from GE interactions and thus increase adjustment efficiency.

Based on our Monte Carlo simulations, we conclude that this adjustment method can better predict genotypic values from historical trial data without

regrowing the entries in the past years of trials. Therefore, it can lead to several important applications. One major application is to adjust annual genetic gain over trial years. In the application to a 16-year soybean trial data collected in South Dakota environments, we observed that a large number of overlapped genotypes were available every two consective years (**Table 5**), these overlapped genotypes helped better predict genotypic values as evidenced by an improved genetic gain model fitness of 85% compared to that without adjustment with a low model fitness of 48%. In addition, the annual genetic gain from the trial data without adjustment was 1.35 bushel/ac and was overestimated because high yield in the last few years from favorable weather condition in September. However, the adjusted annua yield gain was 0.72 bushel/ac with much improved model fitness. The adjusted annual gain appeared to be more in line with the national and the state annual increase over decades [24] [25]. Similarly, this adjustment method can be applied to other trial data targeting a particular location, area, and/or region.

The entries used in long-term historical crop trial are a useful germplasm collection that can be used to identify QTLs associated with traits of importance. Because these entries could be very diversified, multiple desirable QTLs will be likely identified. With the rapid development of genotyping technologies, these entries can be genotyped at a low cost within a short timeframe. However, the cost for re-phenotyping on the traits of interest could be high, especially for those traits measured in the field. Using historical crop trial data will save a great cost associated with re-phenotyping [26]. Application of the method in this study to genetic association studies with historical data could improve the power of identifing QTLs and also reduce the false descovery rate. Predicted genotypic values can be further used to identify favorable loci associated with traits of interest. That is another way to better utilize long-term historical crop trial data.

Using overlapped entries between every two consecutive years to adjust long-term historical data has been validated to be statistically sound. Therefore, the method proposed in this study is a useful way to improve utilization of the long-term historical data. This method appears to be applicable to national cotton variety trials because the two or three same standards are unsually applied in every cycle (three years) with one standard rollovered to the following cycle [18]. We also noticed that the same standards/check have been used in national winter wheat trials for years. These same standards used among trial years could be another valuable resource to control the environmental effects among years. Therefore, there is a great need to numerically investigate the statistical properties regarding this type of crop trial design so a better crop trial design can be recommended with consideration of efficiency of both environmental control among years and land cost.

This study showed that overlapped geneotypes between each two consecutive years/seasons can be used to better predict genotype values in long-term historical trial data with the use of the method proposed in this study. However, this

stepwise adjustment method requires at least one overlaping genotype to be available between any two consecutive years.

Disclaimer

Mention of trade names or commercial products in this publication is solely for the purpose of providing specifc information and does not imply recommendation or endorsement by the U.S. Department of Agriculture. USDA is an equal opportunity provider and employer.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Gray, E. (1982) Genotype × Environment Interactions and Stability Analysis for Forage Yield of Orchardgrass Clones. *Crop Science*, **22**, 19-23. https://doi.org/10.2135/cropsci1982.0011183x002200010005x

[2] Kang, M.S. and Miller, J.D. (1984) Genotype×Environment Interactions for Cane and Sugar Yield and Their Implications in Sugarcane Breeding. *Crop Science*, **24**, 435-440. https://doi.org/10.2135/cropsci1984.0011183x002400030002x

[3] Zhu, J. (1993) Methods of Predicting Genotype Value and Heterosis for Offspring of Hybrids. *Journal of Biomathmatics*, **8**, 32-40.

[4] Wu, J., Qi, J. and Kleinjan, J. (2017) Exploring Multi-Year Soybean Yield Trial Data in South Dakota Environments. New Prairie Press. https://doi.org/10.4148/2475-7772.1529

[5] Eberhart, S.A. and Russell, W.A. (1966) Stability Parameters for Comparing Varieties1. *Crop Science*, **6**, 36-40. https://doi.org/10.2135/cropsci1966.0011183x000600010011x

[6] Fan, X., Kang, M.S., Chen, H., Zhang, Y., Tan, J. and Xu, C. (2007) Yield Stability of Maize Hybrids Evaluated in Multi-Environment Trials in Yunnan, China. *Agronomy Journal*, **99**, 220-228. https://doi.org/10.2134/agronj2006.0144

[7] Finlay, K. and Wilkinson, G. (1963) The Analysis of Adaptation in a Plant-Breeding Programme. *Australian Journal of Agricultural Research*, **14**, 742-752. https://doi.org/10.1071/ar9630742

[8] Francis, T.R. and Kannenberg, L.W. (1978) Yield Stability Studies in Short-Season Maize. I. A Descriptive Method for Grouping Genotypes. *Canadian Journal of Plant Science*, **58**, 1029-1034. https://doi.org/10.4141/cjps78-157

[9] Lin, C.S., Binns, M.R. and Lefkovitch, L.P. (1986) Stability Analysis: Where Do We Stand. *Crop Science*, **26**, 894-900. https://doi.org/10.2135/cropsci1986.0011183x002600050012x

[10] Crossa, J., Gauch, H.G. and Zobel, R.W. (1990) Additive Main Effects and Multiplicative Interaction Analysis of Two International Maize Cultivar Trials. *Crop Science*, **30**, 493-500. https://doi.org/10.2135/cropsci1990.0011183x003000030003x

[11] Yan, W. and Hunt, L.A. (2001) Interpretation of Genotype × Environment Interaction for Winter Wheat Yield in Ontario. *Crop Science*, **41**, 19-25. https://doi.org/10.2135/cropsci2001.41119x

[12] Zhu, J., Xu, F. and Lai, M.G. (1993) Analysis Methods for Unbalanced Data from

Regional Trials of Crop Variety, Analysis for Single Trait. *Journal of Zhejiang Agricultural University*, **19**, 7-13.

[13] DeLacy, I.H., Basford, K.E., Cooper, M., Bull, J.K. and McLaren, C.G. (1996) Analysis of Multi-Environment Trials—An Historical Perspective. In: Cooper, M. and Hammer, G.L., Eds., *Plant Adaptation and Crop Improvement*, CAB International, 39-124.

[14] DeLacy, I.H., Redden, R.J., Butler, D.G. and Usher, T. (2000) Analysis of Line X Environment Interactions for Yield in Navy Beans. Pattern Analysis of Environments among years. *Australian Journal of Agricultural Research*, **51**, 619-628. https://doi.org/10.1071/ar97137

[15] Mackay, I., Horwell, A., Garner, J., White, J., McKee, J. and Philpott, H. (2010) Reanalyses of the Historical Series of UK Variety Trials to Quantify the Contributions of Genetic and Environmental Factors to Trends and Variability in Yield over Time. *Theoretical and Applied Genetics*, **122**, 225-238. https://doi.org/10.1007/s00122-010-1438-y

[16] Zhang, J., Abdelraheem, A. and Flynn, R. (2019) Genetic Gains of Acala 1517 Cotton since 1926. *Crop Science*, **59**, 1052-1061. https://doi.org/10.2135/cropsci2018.11.0686

[17] Campbell, B.T., Chee, P.W., Lubbers, E., Bowman, D.T., Meredith, W.R., Johnson, J., *et al*. (2011) Genetic Improvement of the Pee Dee Cotton Germplasm Collection Following Seventy Years of Plant Breeding. *Crop Science*, **51**, 955-968. https://doi.org/10.2135/cropsci2010.09.0545

[18] Todd Campbell, B., Boykin, D., Abdo, Z. and Meredith, W.R. (2015) Cotton. In: *Yield Gains in Major U.S. Field Crops*, CSSA Special Publications, 13-32. https://doi.org/10.2135/cssaspecpub33.c2

[19] Bondalapati, K.D., Jenkins, J.N., McCarty, J.C. and Wu, J. (2015) Field Experimental Design Comparisons to Detect Field Effects Associated with Agronomic Traits in Upland Cotton. *Euphytica*, **206**, 747-757. https://doi.org/10.1007/s10681-015-1512-2

[20] Wu, J., McCarty, J.C. and Jenkins, J.N. (2010) Cotton Chromosome Substitution Lines Crossed with Cultivars: Genetic Model Evaluation and Seed Trait Analyses. *Theoretical and Applied Genetics*, **120**, 1473-1483. https://doi.org/10.1007/s00122-010-1269-x

[21] Wu, J. (2019) Minque: An R Package for Linear Mixed Model Analyses. https://cran.r-project.org/web/packages/minque/index.html

[22] R Core Team (2023) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

[23] RStudio Team (2022) Rstudio: Integrated Development for R. R Studio, Inc.

[24] USDA-ERS (2025) Oil Crops Yearbook. https://www.ers.usda.gov/data-products/oil-crops-yearbook/documentation

[25] USDA-NASS (2012) South Dakota 2012 Annual Bulletin. https://data.nass.usda.gov/Statistics_by_State/South_Dakota/Publications/Annual_Statistical_Bulletin/2012/ab12019c.pdf

[26] Matthies, I.E., Malosetti, M., Röder, M.S. and van Eeuwijk, F. (2014) Genome-Wide Association Mapping for Kernel and Malting Quality Traits Using Historical European Barley Records. *PLOS ONE*, **9**, e110046. https://doi.org/10.1371/journal.pone.0110046