SRYAHWA PUBLICATIONS

Journal of Sports and Games
ISSN: 2642-8466 | Volume 7, Issue 2, 2025
https://doi.org/10.22259/2642-8466.0702005

RESEARCH ARTICLE

Comparison of Projected Wins of Three Projection Systems in Major League Baseball

David P. Chu*, Kajal, Gurdeepak Sidhu

Department of Mathematics and Statistics, University of the Fraser Valley, 33844 King Road, Abbotsford, BC, Canada V2S 7M8.

Received: 14 October 2025 Accepted: 29 October 2025 Published: 12 November 2025

Corresponding Author: David P. Chu, Associate Professor, Department of Mathematics and Statistics, University of the Fraser Valley, 33844 King Road, Abbotsford, BC, Canada V2S 7M8.

Abstract

Player Empirical Comparison and Optimization Test Algorithm (PECOTA), sZymborski Projection System (ZiPS), and FanGraphs are three well known projection systems in Major League Baseball (MLB). In this article, we will compare the effectiveness of these three projection systems using data from the seasons of 2013-2024. With different assumptions on correlation or covariance structure of the total numbers of games won by MLB teams in a season, three models are developed. Based on these models, we test the null hypothesis that the projected winning percentages are plausible values of the actual winning percentages for MLB teams. P-values of the Mahalanobis distance between the observed wins and projected wins are computed to evaluate the effectiveness of these three projection systems. Bonferroni confidence intervals, confidence ellipsoids, and Benjamini-Hochberg procedure for multiple hypothesis testing are also used to compare these three systems. Simulations are generated as well. The checking of the validity of normality assumption is also given.

Keywords: PECOTA, ZiPS, FanGraphs, Mahalanobis Distance, Matchup Games.

1. Introduction

In Major League Baseball (MLB), there are many projection systems attempting to predict the numbers of wins achieved by teams in a season. These projections are usually made prior to the start of the season. Player Empirical Comparison and Optimization Test Algorithm (PECOTA), sZymborski Projection System (ZiPS), and FanGraphs are three well known projection systems in MLB. We wish to compare the predicted wins of these three projection systems for the seasons of 2013-2024. Thus the projected wins as well as the observed wins of MLB teams are compiled for this period to compare the effectiveness of the predictions of these three projection systems. Chu and Wang (2019) suggested that the preseason projected wins could be used to help assess whether a team's belief in analytics has a positive impact on the team.

We first consider the sum of squares of the difference between the projected wins and observed wins for each of these three systems. However, this squared mathematical distance does not take into account various random factors affecting the numbers of games won by teams in a season. Rather, the squared statistical distance or Mahalanobis distance between the projected wins and observed wins will be examined here. Three models are proposed based on which we assess the Mahalanobis distance of these three projection systems.

As a first approximation, we assume in Model 1 that the numbers of wins obtained by teams are independent. Each number of wins can be regarded

Citation: David P. Chu, Kajal, Gurdeepak Sidhu. Comparison of Projected Wins of Three Projection Systems in Major League Baseball. Journal of Sports and Games. 2025; 7(2): 35-45.

as a binomial random variable. As there are 30 teams in MLB, the Mahalanobis distance follows approximately a chi-square distribution with 30 degrees of freedom. Under the null hypothesis that the projected wins are plausible values of the actual wins for all 30 MLB teams in a given year, we are able to calculate the observed Mahalanobis distance and the corresponding p-value. This p-value allows us to reach the conclusion whether all the differences between projected and actual wins are statistically significant or not for a particular projection system.

Model 2 imposes the correlation structure for the numbers of wins obtained by teams. Three different kinds of correlation structures are considered: linear, squared, and logarithmic. However, only the squared and logarithmic structures are suitable for computing the Mahalanobis distance because the associated variance-covariance matrices are invertible.

The number of wins obtained by a team is further broken down into the sum of numbers of wins in the matchup games against each team. This approach in Model 3 gives us a more precise assessment for each number of wins in the matchup games. The covariance between the numbers of wins obtained by two different teams will be estimated. A new variance-covariance matrix can then be formulated to recalculate a more accurate Mahalanobis distance between the projected wins and observed wins. These three models are shown in Section 2.

In professional sports such as National Hockey League (NHL) and National Basketball Association (NBA), there are matchup games for one team playing against another team. Hence the idea of the matchup games model introduced here could be extended to these kinds of sports for assessing the accuracy of teams' projected wins.

Bonferroni confidence intervals, confidence ellipsoids in higher dimensional spaces, and Benjamini-Hochberg procedure for multiple hypothesis testing are used to compare the effectiveness of these three projection systems. These results are presented in Section 3.

In Section 4, simulations are implemented to generate 1,000 realizations to test the null hypothesis mentioned above by observing how many Mahalanobis distances falling outside the 95% confidence interval. Benjamini-Hochberg procedure for multiple hypothesis testing is employed again to compare these three projection systems. All the calculations and simulations are done via the statistical software R.

The checking of the validity of normality assumption will be discussed in Section 5. Finally, conclusion and comments are given in Section 6.

2. Modeling

Let $\mathcal{W}_1$, $\mathcal{W}_2$, ..., $\mathcal{W}_{30}$ be the projected wins for the 30 MLB teams in a season. PECOTA, ZiPS, and FanGraphs will be one of the three projection systems that generates these projected wins. Suppose that x_1, x_2, ..., x_{30} be the corresponding observed wins for these 30 MLB teams in that particular season. To measure the accuracy of the projected wins generated by a projection system, one may consider the squared mathematical distance

$$D_0 = (x_1 - \mathcal{W}_1)^2 + (x_2 - \mathcal{W}_2)^2 + \ldots + (x_{30} - \mathcal{W}_{30})^2.$$

Then one may compare the three different values of D_0 produced by PECOTA, ZiPS, and FanGraphs. The smaller the value of D_0, the better the projection system is to generate the projected wins for MLB teams. However, where do we draw the line on the value of D_0 beyond which the projection system is deemed to be not effective in generating the projected wins for teams. We need to develop some models to help answer this question.

2.1 Model 1

As each MLB team usually plays n = 162 games in a season, the projected winning percentage for Team i is $\mathcal{W}_i/n$, i = 1, 2, ..., 30. Note that only 60 games were played by each team during the pandemic year in 2020. The observed wins x_i can be regarded as an observed value from a random variable X_i representing the number of games won by Team i in a season. For the time being, let us assume the independence of the winning of games for each team, i.e., the winning of one game for a team does not affect its winning of another game. Hence the random variable X_i can be treated as a binomial random variable with parameters n and p_i, where n is the number of games played in a season and p_i is the probability of winning a game for Team i. (If a team does not play 162 games in a season, then n will be adjusted accordingly.) Let μ_i and σ_i^2 be the mean and variance of X_i, respectively.

It is well known that X_i can be approximated by a normal distribution when n is large. In practice, this approximation is adequate whenever $n*p_i \geq 15$ and $n*(1 - p_i) \geq 15$. These conditions imply that the number of wins and number of losses are both at least 15 games in a season. Since 162 games are played by each team in a season, these two conditions are certainly satisfied by each team.

Suppose we further assume the independence of X_i and X_j for all i and j, $i \neq j$. We then consider the squared statistical distance or Mahalanobis distance of $X = [X_1, X_2, ..., X_{30}]'$ that is to measure the distance between X and its mean while taking into account the shape of the distribution. In this case, we have

$$D_1 = \left(\frac{X_1-\mu_1}{\sigma_1}\right)^2+\left(\frac{X_2-\mu_2}{\sigma_2}\right)^2+\ldots+\left(\frac{X_{30}-\mu_{30}}{\sigma_{30}}\right)^2 \quad (1)$$

Each term $((X_i - \mu_i)/\sigma_i)^2$ in (1) has approximately a chi-square distribution with 1 degree of freedom. Based on the assumption of the independence of X_i, it seems that D_1 has approximately a chi-square distribution with 30 degrees of freedom. Since there is rarely a tie in an MLB game, the sum of X_i can be treated as a constant that is the total number of games played in a season, i.e., 162*30/2 = 2430. Hence the number of degrees of freedom is adjusted to 29.

We wish to test H_0: Projected wins are plausible values of the actual wins for all 30 MLB teams in a given year versus H_a: Projected wins are not plausible values of the actual wins for all 30 MLB teams in that given year (i.e., at least one projected win is not plausible). We will consider three sets of hypothesis testing, one for each projection system: PECOTA, ZiPS, FanGraphs. Since the winning percentage is the number of wins/n, the above hypothesis testing is equivalent to testing the projected winning percentages ($w_i/n = \tilde{p}_i$) are plausible values of the actual winning percentages (p_i) for all MLB teams. Hence H_0 is changed to $p_i = \tilde{p}_i$, $i = 1, 2, ..., 30$. The mean and variance of X_i, under H_0, can be estimated by

$$\hat{\mu}_i = n\tilde{p}_i = w_i \quad (2)$$

$$\hat{\sigma}_i^2 = n\tilde{p}_i(1 - \tilde{p}_i) = w_i(1 - w_i/n) \quad (3)$$

Table 1 shows the distance D_1 between the observed wins of MLB teams and the projected wins of PECOTA, ZiPS, and FanGraphs for 2013-2024. The corresponding p-values are also given in the parentheses. We notice that all p-values, except PECOTA (2020), ZiPS (2016, 2020), and FanGraphs (2013, 2016, 2020), were less than 0.05. Note that Year 2020 was an unusual MLB season during the pandemic. Other than these six instances, with 5% level of significance, there was sufficient evidence to show that the projected wins produced by each of these three projection systems were not plausible values of the actual wins for all 30 MLB teams for 2013-2024. It implies that at least one projected win was significantly different from the corresponding actual win. For these six instances, however, there was insufficient evidence to show any significant difference of at least one projected win and the corresponding actual win.

Table 1. *Using Model 1 to calculate D_1, with p-value in parentheses, between the observed wins of MLB teams and the projected wins of PECOTA, ZiPS, and FanGraphs for 2013-2024.*

Year	PECOTA	ZiPS	FanGraphs
2024	70.5 (2.5E-5)	56.2 (1.8E-3)	61.0 (4.6E-4)
2023	93.3 (1.1E-8)	89.0 (5.1E-8)	84.8 (2.2E-7)
2022	68.7 (4.6E-5)	74.0 (8.3E-6)	71.4 (2.0E-6)
2021	123.4 (1.3E-13)	100.8 (7.3E-10)	102.8(3.5E-10)
2020	39.9 (8.6E-2)	36.8 (1.5E-1)	42.2 (5.4E-2)
2019	68.3 (5.1E-5)	61.5 (4.0E-4)	75.6 (5.1E-6)
2018	67.5 (6.2E-5)	77.9 (2.3E-6)	81.6 (6.8E-7)
2017	60.0 (6.2E-4)	66.5 (9.1E-5)	64.6 (1.6E-4)
2016	53.1 (4.1E-3)	35.2 (1.9E-1)	40.2 (8.1E-2)
2015	74.3 (7.5E-6)	64.1 (1.9E-4)	63.2 (2.4E-4)
2014	57.5 (1.3E-3)	55.2 (2.3E-3)	49.5 (1.0E-2)
2013	63.5 (2.2E-4)	65.7 (1.2E-4)	38.5 (1.1E-1)

During the period of 2013-2024, PECOTA had the smallest values of D_1 (or larger p-values) for 3 years, ZiPS for 5 years, and FanGraphs for 4 years. A smaller value of D_1 implies a shorter statistical distance between the projected wins and actual wins after variances are taken into account. So the smaller the value of D_1, the better the projection system performs.

2.2 Model 2

The assumption of independence of X_i and X_j, for $i \neq j$, may not hold true in Model 1. It is because the sum of all variables $X_1 + X_2 + ... + X_{30} = 2430$, which is a constant. When a team wins a baseball game, it means another team loses a game. This is due to the fact that it is a zero-sum game and there is (almost) no tie for

a game. Hence some correlation may exist between X_i and X_j. Let the correlation between X_i and X_j, $i, j = 1, 2, \ldots, 30$, $i \neq j$, be

$$\rho_{i,j} = \frac{\sigma_{i,j}}{\sigma_i \sigma_j},$$

where $\sigma_{i,j}$ is the covariance between X_i and X_j, and σ_i is the standard deviation of X_i. Thus,

$$\sigma_{i,j} = \rho_{i,j} \sigma_i \sigma_j.$$

Negative (or non-positive) correlation is expected to exist between X_i and X_j because of the zero-sum games. Let $m_{i,j}$ be the number of baseball games played between Teams i and j, where $i \neq j$. Note that $m_{i,j} = m_{j,i}$ and $m_{i,i} = 0$ for all i. For the extreme cases: when $m_{i,j} = 0$, it implies that $\rho_{i,j} = 0$; when $m_{i,j} = 162$, it implies that $\rho_{i,j} = -1$. Furthermore, when $0 < m_{i,j} < 162$, we have $-1 < \rho_{i,j} < 0$. As $m_{i,j}$ increases from 0 to 162, the value of $\rho_{i,j}$ decreases from 0 to −1.

Here we consider three models for $\rho_{i,j}$, $i \neq j$, satisfying all conditions mentioned above.

Model 2a: $\rho_{i,j} = -m_{i,j}/162$ (a linear model);

Model 2b: $\rho_{i,j} = -(m_{i,j}/162)^2$ (a squared model);

Model 2c: $\rho_{i,j} = -\log_2((m_{i,j}/162) + 1)$ (a logarithmic model).

Figure 1 displays the graphs of functions $\rho_{i,j}$'s in Models 2a-2c.

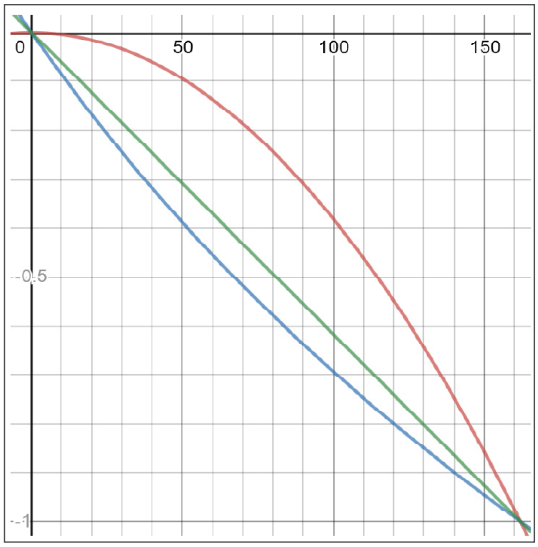

Figure 1. *Graphs of functions in Models 2a-2c with the top graph from the squared model, middle graph the linear model, and bottom graph the logarithmic model.*

The correlation matrix is given by

$$\rho = \begin{bmatrix} 1 & \rho_{1,2} & \cdots & \rho_{1,30} \\ \rho_{2,1} & 1 & \cdots & \rho_{2,30} \\ . & . & \cdots & . \\ . & . & \cdots & . \\ \rho_{30,1} & \rho_{30,2} & \cdots & 1 \end{bmatrix} \quad (4)$$

The variance-covariance matrix is given by

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho_{1,2}\sigma_1\sigma_2 & \cdots & \rho_{1,30}\sigma_1\sigma_{30} \\ \rho_{2,1}\sigma_2\sigma_1 & \sigma_2^2 & \cdots & \rho_{2,30}\sigma_2\sigma_{30} \\ . & . & \cdots & . \\ . & . & \cdots & . \\ \rho_{30,1}\sigma_{30}\sigma_1 & \rho_{30,2}\sigma_{30}\sigma_2 & \cdots & \sigma_{30}^2 \end{bmatrix} = \Lambda \rho \Lambda \quad (5)$$

where Λ is a 30×30 diagonal matrix with diagonal elements $\sigma_1, \sigma_2, \ldots, \sigma_{30}$.

Suppose that $X = [X_1, X_2, \ldots, X_{30}]'$ follows a multivariate normal distribution $N_{30}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where the mean vector is $\boldsymbol{\mu} = [\mu_1, \mu_2, \ldots, \mu_{30}]'$ and variance-covariance matrix $\boldsymbol{\Sigma}$ is given by (5) with $\rho_{i,j}$ being one of the values shown in Model 2a, 2b or 2c, and σ_i^2 estimated by (3) under H_0. Let

$$D_2 = (X - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (X - \boldsymbol{\mu}) \quad (6)$$

be the Mahalanobis distance between X and $\boldsymbol{\mu}$ while taking into account the covariances among those variables in X. Since the sum of X_i can be treated as a constant, Johnson and Wichern (2019) shows that D_2 has a chi-square distribution with 29 degrees of freedom.

When $\rho_{i,j} = 0$, $i \neq j$, D_2 is reduced to D_1 that involves no correlations among the variables. Hence D_2 is a generalization of D_1 when correlations are taken into account. In our situation, there are 30 variables involved as 30 MLB teams play in a season. So each correlation between any two variables might not be too large. Would D_2 generate a significantly different value from that of D_1 when all correlations

are included in calculating D_2? Or would D_1 be able to provide a good approximation for D_2?

Let's first consider Model 2a. For fixed j, $\sum_{i=1,i\neq j}^{30} \rho_{i,j} = -\sum_{i=1,i\neq j}^{30} \rho_{i,j}\, m_{i,j}/162 = -1$. It implies that $\sum_{i=1}^{30} \rho_{i,j} = 0$ as $\rho_{i,i} = 1$ for all i. In this case, the determinant of $\boldsymbol{\rho}$ in (4) is zero. Consequently, the inverse of $\boldsymbol{\rho}$ does not exist. It implies that the inverse of the corresponding variance-covariance matrix $\boldsymbol{\Sigma}$ in (5) does not exist either. As a result, we will not be able to compute D_2 in (6) when $\rho_{i,j} = -m_{i,j}/162$ as shown in Model 2a.

However, the correlation matrix (4) with $\rho_{i,j} = -(m_{i,j}/162)^2$ given in Model 2b will not generate zero determinant. Therefore, the inverse of the correlation matrix exists and so does the corresponding variance-covariance matrix $\boldsymbol{\Sigma}$ in (5). Likewise, with the condition $\rho_{i,j} = -\log_2((m_{i,j}/162)+1)$ given in Model 2c, both the inverse of correlation matrix and the inverse of variance-covariance matrix exist. Therefore, we will compute D_2 only for the values of $\rho_{i,j}$ given in Models 2b and 2c.

Table 2 shows the distance D_2 between the observed wins of MLB teams and the projected wins of PECOTA, ZiPS, and FanGraphs for 2013-2024, using Model 2b. All p-values, except PECOTA (2020), ZiPS (2016, 2020), and FanGraphs (2013, 2016, 2020), were less than 0.05. Other than these six instances, with 5% level of significance, there was sufficient evidence to show that the projected wins produced by these three projection systems were not plausible values of the actual wins for all 30 MLB teams for 2013-2024.

For the period of 2013-2024, PECOTA had the smallest value of D_2 for 3 years, ZiPS for 5 years, and FanGraphs for 4 years.

Table 2. *Using Model 2b to calculate D_2, with p-value in parentheses, between the observed wins of MLB teams and the projected wins of PECOTA, ZiPS, and FanGraphs for 2013-2024*

Year	PECOTA	ZiPS	FanGraphs
2024	70.3 (2.8E-5)	55.9 (1.9E-3)	60.8 (4.9E-4)
2023	92.9 (1.3E-8)	88.7 (5.7E-8)	84.5 (2.5E-7)
2022	68.0 (5.7E-5)	73.4 (1.0E-5)	70.7 (2.4E-5)
2021	122.2 (2.1E-13)	100.0(1.0E-9)	102.1 (4.5E-10)
2020	24.0 (7.3E-1)	22.4 (8.0E-1)	25.4 (6.6E-1)
2019	67.7 (6.3E-5)	60.9 (4.8E-4)	74.8 (6.6E-6)
2018	67.2 (7.2E-5)	77.5 (2.7E-6)	81.1 (8.0E-7)
2017	59.4 (7.4E-4)	65.7 (1.2E-4)	64.1 (1.9E-4)
2016	52.6 (4.6E-3)	34.9 (2.1E-1)	39.8 (8.8E-2)
2015	74.0 (8.4E-6)	63.5 (2.2E-4)	62.7 (2.8E-4)
2014	57.2 (1.4E-3)	54.8 (2.6E-3)	49.3 (1.1E-2)
2013	62.8 (2.7E-4)	64.9 (1.5E-4)	38.1 (1.2E-1)

Table 3 shows the distance D_2 between the observed wins of MLB teams and the projected wins of PECOTA, ZiPS, and FanGraphs for 2013-2024, using Model 2c. The p-values of PECOTA (2014, 2020), ZiPS (2014, 2016, 2020), and FanGraphs (2013-2016, 2018, 2020) were greater than 0.05. With 5% level of significance, there was insufficient evidence to reject the null hypothesis that the projected wins of the 30 MLB teams were plausible values of their actual wins for these years. Besides these eleven instances, there was sufficient evidence to show that at least one projected win was not a plausible value of the actual win of these 30 MLB teams.

PECOTA, ZiPS, and FanGraphs had the smallest value of D_2 for 3, 4, and 5 years, respectively, during the period of 2013-2024.

Table 3. *Using Model 2c to calculate D_2, with p-value in parentheses, between the observed wins of MLB teams and the projected wins of PECOTA, ZiPS, and FanGraphs for 2013-2024*

Year	PECOTA	ZiPS	FanGraphs
2024	68.6 (4.7E-5)	53.0 (4.2E-3)	59.7 (6.8E-4)
2023	88.4 (6.4E-8)	84.0 (2.9E-7)	83.4 (3.7E-7)
2022	58.8 (8.6E-4)	62.1 (3.4E-4)	62.6 (2.9E-4)
2021	50.1 (8.9E-3)	69.6 (3.4E-5)	83.0 (4.2E-7)
2020	23.0 (7.7E-1)	21.5 (8.4E-1)	24.6 (7.0E-1)
2019	57.5 (1.2E-3)	53.5 (3.7E-3)	64.1 (1.8E-4)

2018	60.9 (4.8E-4)	72.1 (1.5E-5)	9.3 (9.9E-1)
2017	53.6 (3.6E-3)	58.0 (1.1E-3)	54.4 (2.9E-3)
2016	47.7 (1.6E-2)	31.2 (3.6E-1)	33.9 (2.4E-1)
2015	49.2 (1.1E-2)	56.5 (1.6E-3)	39.3 (9.7E-2)
2014	39.5 (9.2E-2)	38.5 (1.1E-1)	13.6 (9.9E-1)
2013	50.5 (8.0E-3)	57.4 (1.3E-3)	33.9 (2.4E-1)

2.3 Model 3

To improve the covariance structure of $\mathbf{\Sigma}$, we decompose each X_i (the total number of games won by Team i in a season) into $X_{i,j}$ (the number of games won by Team i over Team j in that season). More specifically,

$$X_1 = X_{1,1} + X_{1,2} + \ldots + X_{1,30} \quad (7)$$

$$X_2 = X_{2,1} + X_{2,2} + \ldots + X_{2,30} \quad (8)$$

$$\vdots \qquad \vdots$$

$$X_{30} = X_{30,1} + X_{30,2} + \ldots + X_{30,30} \quad (9)$$

Note that $X_{i,i} = 0$, $i = 1, 2, \ldots, 30$. Because of zero-sum games, we have $X_{i,j} + X_{j,i} = m_{i,j}$ (or $m_{j,i}$) that is the number of games played between Teams i and j, where $i, j = 1, 2, \ldots, 30$ and $i \neq j$. For example, Arizona and Colorado played 19 games against each other in 2022; however, Arizona and Baltimore did not play any game against each other in that year. The numbers of matchup games between teams might also be changed from one year to another year, e.g., Arizona and Colorado played only 13 games against each other in 2023.

Let us first consider the covariance between X_1 and X_2. Due to the independence of games Teams 1 and 2 played against Teams 3, 4, ..., 30 (i.e., $Cov(X_{1,2}, X_{2,j}) = 0$, $j = 3, 4, \ldots, 30$ and $Cov(X_{1,j}, X_{2,j}) = 0$, $i, j = 3, 4, \ldots, 30$), the covariance between X_1 and X_2 becomes

$$\begin{aligned} Cov(X_1, X_2) &= Cov(X_{1,1}+X_{1,2}+\ldots+X_{1,30}, X_{2,1}+X_{2,2}+\ldots+X_{2,30}) \\ &= Cov(X_{1,2}, X_{2,1}) \\ &= Cov(X_{1,2}, m_{1,2} - X_{1,2}) \\ &= -Cov(X_{1,2}, X_{1,2}) \end{aligned} \quad (10)$$

$Cov(X_{1,2}, m_{1,2})$ is zero because $m_{1,2}$ is a fixed number. The covariance term $Cov(X_{1,2}, X_{1,2})$ is simply the variance term $Var(X_{1,2})$. $X_{1,2}$ can be regarded as a binomial random variable with parameters $m_{1,2}$ and $p_{1,2}$, where $m_{1,2}$ is the number of games played between Teams 1 and 2, and $p_{1,2}$ is the probability that Team 1 will win over Team 2 in a game. Thus the variance of $X_{1,2}$ is $m_{1,2} * p_{1,2} * (1 - p_{1,2})$. Similarly, the variance of $X_{2,1}$ is $m_{2,1} * p_{2,1} * (1 - p_{2,1})$, where $m_{2,1} = m_{1,2}$ and $p_{2,1}$ is the probability that Team 2 will win over Team 1 in a game. Note that $p_{1,2} + p_{2,1} = 1$. Since $Cov(X_1, X_2) = Cov(X_2, X_1)$, it implies that $Var(X_{1,2}) = Var(X_{2,1})$. Thus $p_{1,2} * (1 - p_{1,2}) = p_{2,1} * (1 - p_{2,1})$, and this equation is always true since $p_{1,2} + p_{2,1} = 1$.

Under the equivalent percentage version of H_0, i.e., $p_i = \mathcal{W}_i/n$, $i = 1, 2, \ldots, 30$, $p_{1,2}$ can be estimated by $p_1/(p_1 + p_2) = (\mathcal{W}_1/n)/((\mathcal{W}_1/n) + (\mathcal{W}_2/n)) = \mathcal{W}_1/(\mathcal{W}_1 + \mathcal{W}_2)$. Similarly, $p_{2,1}$ can be estimated by $p_2/(p_1 + p_2) = (\mathcal{W}_2/n)/((\mathcal{W}_1/n) + (\mathcal{W}_2/n)) = \mathcal{W}_2/(\mathcal{W}_1 + \mathcal{W}_2)$. With these estimations, $p_{1,2} + p_{2,1}$ is always 1. Hence, $Cov(X_1, X_2)$ in (10) can be estimated by $-m_{1,2} * \mathcal{W}_1/(\mathcal{W}_1 + \mathcal{W}_2) * \mathcal{W}_2/(\mathcal{W}_1 + \mathcal{W}_2)$ under H_0. Similarly, $Cov(X_2, X_1)$ can be estimated by $-m_{2,1} * p_{2,1} * (1 - p_{2,1}) = -m_{1,2} * \mathcal{W}_2/(\mathcal{W}_1 + \mathcal{W}_2) * \mathcal{W}_1/(\mathcal{W}_1 + \mathcal{W}_2)$, which is the estimated value of $Cov(X_1, X_2)$. By following the above procedure for the general terms $i \neq j$, $Cov(X_i, X_j) = Cov(X_j, X_i)$ can be estimated by

$$-m_{i,j} * \mathcal{W}_i/(\mathcal{W}_i + \mathcal{W}_j) * \mathcal{W}_j/(\mathcal{W}_i + \mathcal{W}_j) \text{ under } H_0. \quad (11)$$

The mean and variance of X_i, $i = 1, 2, \ldots, 30$, are estimated by (2) and (3), respectively, under H_0. Therefore, the variance-covariance terms $Cov(X_i, X_j)$, $i, j = 1, 2, \ldots, 30$, shown in (3) and (11) can be computed directly to form the entries of $\mathbf{\Sigma}$ in (6). With this matrix $\mathbf{\Sigma}$, we are able to compute its inverse. The value of D_2 in (6) can then be calculated to compare the distance between the observed wins of MLB teams and the projected wins of PECOTA, ZiPS, and FanGraphs for 2013-2024 under H_0. The corresponding p-values can also be evaluated using the chi-square distribution with 29 degrees of freedom. The results are given in Table 4.

The p-values of PECOTA (2016, 2020), ZiPS (2016, 2020), and FanGraphs (2013, 2016, 2020) were greater than 0.05. With 5% level of significance, there was insufficient evidence to reject the null hypothesis that the projected wins of the 30 MLB teams were plausible values of their actual wins for these years. Besides these seven instances, there was sufficient evidence to show that at least one projected win was not a plausible value of the actual win of the 30 MLB teams.

PECOTA had the smallest value of D_2 for 3 years, ZiPS 4 years, and FanGraphs 5 years during the period of 2013-2024.

Table 4. *Using Model 3 to calculate D_2, with p-value in parentheses, between the observed wins of MLB teams and the projected wins of PECOTA, ZiPS, and FanGraphs for 2013-2024*

Year	PECOTA	ZiPS	FanGraphs
2024	67.9 (5.8E-5)	52.8 (4.4E-3)	57.4 (1.3E-3)
2023	89.0 (5.2E-8)	85.0 (2.1E-7)	82.6 (4.8E-7)
2022	62.1 (3.4E-4)	67.1 (7.5E-5)	64.0 (1.9E-4)
2021	117.7 (1.2E-12)	94.4 (7.5E-9)	97.3 (3.0E-9)
2020	39.7 (8.9E-2)	23.5 (7.5E-1)	5.4 (9.9E-1)
2019	63.5 (2.2E-4)	56.3 (1.7E-3)	67.1 (7.0E-5)
2018	63.6 (2.2E-4)	70.6 (2.5E-5)	83.0 (4.2E-7)
2017	55.2 (2.3E-3)	60.5 (5.4E-4)	60.8 (4.9E-4)
2016	31.6 (3.4E-1)	29.7 (4.3E-1)	33.6 (2.5E-1)
2015	70.8 (2.4E-5)	59.3 (7.5E-4)	53.8 (3.4E-3)
2014	56.3 (1.7E-3)	53.0 (4.2E-3)	47.5 (1.7E-2)
2013	55.8 (2.0E-3)	59.3 (7.5E-4)	34.6 (2.2E-1)

3. Confidence Regions

3.1 Bonferroni Confidence Intervals

From Model 1, X_i can be treated as a binomial random variable with parameters $n = 162$ and p_i. The unknown p_i can be estimated by $\hat{p}_i = x_i/n$, where x_i is the number of observed wins for Team i in a season. Hence an approximately 95% confidence interval for p_i, $i = 1, 2, ..., 30$, is

$$\hat{p}_i \pm z_{0.025}\sqrt{\hat{p}_i(1-\hat{p}_i)/n}, \tag{12}$$

where $z_{0.025} = 1.960$ is the upper 2.5% critical value of the standard normal distribution.

Consider the projected winning percentages $\tilde{p}_i = \mathcal{W}_i/n$, $i = 1, 2, ..., 30$, and see how many of them fall in the corresponding confidence interval for p_i shown in (12). If any one of the $\tilde{p}_i$'s does not fall in the corresponding confidence interval for p_i, then we can say that at least one of the projected wins is different from one of the actual wins with probability approximately $1 - (0.95)^{30} \approx 0.7854$. Therefore, the chances that all 30 $\tilde{p}_i$'s fall in the corresponding confidence interval simultaneously are approximately 0.2146.

In order to adjust the overall confidence level from 21.46% to 95%, we use the Bonferroni confidence interval as follows:

$$\hat{p}_i \pm z_{0.025/30}\sqrt{\hat{p}_i(1-\hat{p}_i)/n}, \tag{13}$$

where $i = 1, 2, ..., 30$ and $z_{0.025/30} = 3.144$. When $z_{0.025/30}$ is used in (13) instead of $z_{0.025}$, there are 95% chances that all 30 projected winning percentages $\tilde{p}_i$'s fall in the corresponding Bonferroni confidence interval simultaneously. Table 5 shows the number of projected winning percentages for PECOTA, ZiPS, and FanGraphs, falling in the corresponding 95% Bonferroni confidence interval for 2013-2024.

Table 5. *Number of projected winning percentages (out of 30 teams) for PECOTA, ZiPS, and FanGraphs, falling in the corresponding 95% Bonferroni confidence interval for 2013-2024*

Year	PECOTA	ZiPS	FanGraphs
2024	29	29	29
2023	29	27	28
2022	29	29	30
2021	28	29	28
2020	30	30	30
2019	29	29	29
2018	27	28	29
2017	29	29	29
2016	28	30	30
2015	29	30	30
2014	29	28	30
2013	30	29	30
Mean	28.8	28.9	29.3
St Dev	0.84	0.90	0.78

It is desirable to see all 30 projected winning percentages simultaneously fall in the corresponding 95% Bonferroni confidence interval. However, only PECOTA (2013, 2020), ZiPS (2015, 2016, 2020), and FanGraphs (2013-2016, 2020, 2022) had achieved this. We see that PECOTA and ZiPS produced the similar mean (28.8, 28.9) and comparable values (0.84, 0.90) for standard deviation. It seems that FanGraphs produced more accurate and precise results with higher mean (29.3) and smaller standard deviation (0.78). Nevertheless, the One-way Analysis of Variance (ANOVA) test shows that there is insufficient evidence at 5% level of significance to support that the average numbers of projected wins simultaneously falling in the corresponding 95% Bonferroni confidence interval are not the same for these three projection systems.

3.2 Confidence Ellipsoids

Under Models 2b-2c and 3, $X = [X_1, X_2, ..., X_{30}]'$ follows a multivariate normal distribution $N_{30}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with the mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$. Recall that D_2 in (6) has a chi-square distribution with 29 degrees of freedom. Hence a 95% confidence ellipsoid for $\boldsymbol{\mu}$ is

$$D_2 = (X - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(X - \boldsymbol{\mu}) \le \chi^2_{29}\,(0.05) \qquad (14)$$

where χ^2_{29} (0.05) = 42.56 is the upper 5% critical value of the chi-square distribution with 29 degrees of freedom. Under H_0, $\boldsymbol{\mu}$ can be estimated by $[\mathcal{W}_1, \mathcal{W}_2, ..., \mathcal{W}_{30}]'$ shown in (2). With confidence 95%, the vector of observed wins $x = [x_1, x_2, ..., x_{30}]'$ should fall in the above ellipsoid given in (14).

When D_2 is less than or equal to 42.56, this is equivalent to the associated p-value greater than 0.05 as seen in Tables 2-4. Consequently, we obtain the same results and conclusions as presented in Sections 2.2 and 2.3.

3.3 Multiple Hypothesis Testing

The usual naive method of statistical testing on a single hypothesis may not be suitable for testing multiple hypotheses with the same significance level. The Bonferroni method, however, is usually more conservative and results in more acceptance of the status quo H_0. The Benjamini-Hochberg procedure for multiple hypothesis testing can be used to test the significance of multiple statements. This procedure tends to balance the effect of the previous two approaches and is helpful in reducing false positives (type I error). For more details about the Benjamini-Hochberg procedure, see their paper (1995) or Tan et. al (2019).

Suppose we wish to test $H_0 : p_1 = w_1/162$ versus $H_a : p_1 \ne w_1/162$. Under H_0, the test statistic is

$$Z \approx \frac{x_1/n_1 - p_1}{\sqrt{p_1 * (1 - p_1)/n_1}} = \frac{x_1/n_1 - w_1/162}{\sqrt{w_1/162 * (1 - w_1/162)/n_1}}, \qquad (15)$$

where n_1 (usually 162) is the number of games played by Team 1 in a season. We calculate the observed test statistic z and then find the corresponding p-value called PV_1. Repeat the same process for the other 29 teams to obtain PV_2, PV_3, ..., PV_{30}. Rearrange these p-values in descending order to obtain

$$PV_{(30)} \ge PV_{(29)} \ge ... \ge PV_{(1)}. \qquad (16)$$

Compare these ordered p-values term-wise with significance levels

$$\alpha > \frac{29}{30}\alpha > ... > \frac{1}{30}\alpha, \qquad (17)$$

i.e., compare $PV_{(i)}$ with $(i/30)\alpha, i = 1, 2, ..., 30$. Choose the largest K such that $PV_{(K)} \le (K/30)\alpha$ to declare that K of the projected winning percentages are statistically different from the actual winning percentages with significance level approximately α, say 5%. The smaller the value of K, the fewer the projected winning percentages are different from the actual ones and hence the better the projection system is. It is desirable to have $K = 0$, indicating that all 30 pairs of projected and actual winning percentages are not statistically significantly different.

Table 6 displays the multiple hypothesis testing for PECOTA, ZiPS, and FanGraphs, showing K distinct projected winning percentages from the actual winning percentages. PECOTA (2013, 2020), ZiPS (2013-2016, 2020, 2022), and FanGraphs (2015, 2016, 2020) have $K = 0$, indicating that there was no significant difference between the projected and actual winning percentages for these years. By comparing the mean of K, it seems that ZiPS is preferable. Nevertheless, the One-way ANOVA test shows that there is insufficient evidence at 5% level of significance to support that the true average numbers of projected winning percentages distinct from the actual winning percentages are not the same for these three projection systems.

Table 6. *Multiple hypothesis testing for PECOTA, ZiPS, and FanGraphs, showing K distinct projected winning percentages from the actual winning percentages for 2013-2024*

Year	PECOTA	ZiPS	FanGraphs
2024	1	1	1
2023	3	3	4

2022	1	0	3
2021	7	7	6
2020	0	0	0
2019	3	1	1
2018	3	1	3
2017	1	1	1
2016	2	0	0
2015	1	0	0
2014	3	0	2
2013	0	0	1
Mean	2.08	1.17	1.83
St Dev	1.93	2.04	1.85

4. Simulations

Instead of having only one instance/realization of the observed wins for each team to be compared with the projected wins, simulations are implemented to generate 1,000 realizations of the observed wins. This allows us to have more extensive comparison of the observed wins with the projected wins. By doing so, we may be able to achieve more reliable result on the effectiveness of the projection systems in forecasting the actual wins of MLB teams.

From Model 3, $X_{i,j}$ (the number of games won by Team i over Team j in a season), $i \neq j$, can be regarded as a binomial random variable with parameters $m_{i,j}$ and $p_{i,j}$, where $m_{i,j}$ is the number of games played between Teams i and j, and $p_{i,j}$ is the probability of winning for Team i over Team j. The parameter $p_{i,j}$ can be estimated by $\hat{p}_{i,j} = x_{i,j}/n_{i,j}$, where $x_{i,j}$ is the observed wins for Team i over Team j in $n_{i,j}$ games. Note that $n_{i,j} = m_{i,j}$ if no game is cancelled and no extra game is added in a regular season. If $x_{i,j} = 0$ or $n_{i,j}$, $\hat{p}_{i,j}$ will be modified to $(x_{i,j} + 1)/(n_{i,j} + 2)$, i.e., adding one extra win and one extra loss to the original outcome. Thus $\hat{p}_{i,j}$ will not be 0 or 1 in the simulated binomial distribution. Searching from the MLB record books, we are able to find all the observed wins $x_{i,j}$ for Team i over Team j in $n_{i,j}$ games for each season of 2013-2024.

Now we run 1,000 simulations using R to generate the simulated values for the binomial distribution of $X_{i,j}$ with parameters $m_{i,j}$ and $\hat{p}_{i,j}$. We only need to generate the simulated values of $X_{i,j}$ for $1 \leq i < j \leq 30$. It is not necessary to generate $X_{j,i}$ because $X_{j,i} = m_{i,j} - X_{i,j}$. Note that $X_{i,i} = 0$, $i = 1, 2, ..., 30$. Using equations (7)-(9) and the simulated values $X^*_{i,j}$, we obtain the simulated value X_i^* for X_i, $i = 1, 2, ..., 30$.

Let $X^* = [X_1^*, X_2^*, ..., X_{30}^*]'$ be the vector of simulated values for X. Then the simulated Mahalanobis distance is

$$D_3^* = (X^* - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(X^* - \boldsymbol{\mu}) \quad (18)$$

that follows approximately a chi-square distribution with 29 degrees of freedom. Under the equivalent percentage version of H_0, $\boldsymbol{\mu}$ in D_3^* can be replaced by $[\mathcal{W}_1, \mathcal{W}_2, ..., \mathcal{W}_{30}]'$ shown in (2). Applying the estimates of $\text{Var}(X_i)$ in (3) and $Cov(X_i, X_j)$, $i \neq j$, in (11), we are able to compute the inverse of $\boldsymbol{\Sigma}$ in (6 or 18) and hence the value of D_3^* in (18) using the simulated value of X^*.

We obtain 1,000 observed values of D_3^*. First, we compare them with $\chi^2_{29}(0.05) = 42.56$ to see how many of them falling outside the 95% confidence interval. Second, we do a multiple hypothesis testing using Benjamini-Hochberg procedure to test the significance of multiple statements based on the simulated X_i^*, $i = 1, 2, ..., 30$, over 1,000 times. Following the procedures in (15)-(17), we will obtain a value of K for each simulation. For 1,000 simulations, we have 1,000 values of K and then take the average of these values of K. The results are given in Table 7.

Table 7. *Numbers of observed $D_3^* > \chi^2_{29}(0.05)$ and average K instances showing significant difference between 1,000 simulated observed wins and projected wins of PECOTA, ZiPS, and FanGraphs for 2013-2024 (* calculating means and standard deviations without the simulated results of Year 2020)*

Year	PECOTA		ZiPS		FanGraphs	
	$D_3^* > \chi^2_{29}(.05)$	K	$D_3^* > \chi^2_{29}(.05)$	K	$D_3^* > \chi^2_{29}(.05)$	K
2024	435	3.5	262	2.1	238	2.5
2023	516	5.6	423	4.7	333	4.2
2022	426	3.7	357	4.1	348	3.9

2021	671	7.7	576	5.9	522	6.4
2020	5	1.0	8	0.9	6	1.1
2019	446	3.7	434	2.9	441	4.4
2018	419	3.4	413	3.7	473	4.0
2017	274	3.0	327	3.5	317	3.5
2016	244	2.6	130	1.1	166	4.4
2015	303	3.9	213	2.8	222	2.8
2014	285	3.2	268	3.0	215	2.7
2013	350	3.5	428	3.9	195	1.2
Mean	364.5	3.7	319.9	3.2	289.7	3.4
St Dev	164.5	1.6	154.3	1.4	145.9	1.5
Mean*	397.2	4.0	348.3	3.4	315.5	3.6
St Dev*	125.1	1.4	124.8	1.3	120.9	1.3

Table 7 shows that, except for 2020 (pandemic year), the numbers of values of D_3^* falling outside the 95% confidence interval are much higher than 50 (5% of 1,000 simulations). As a result, the null hypothesis that PECOTA, ZiPS, and FanGraphs projected wins were plausible values of the actual wins are rejected at $\alpha = 0.05$ level of significance. Year 2020, however, had a distinct pattern from other years. The Benjamini-Hochberg procedure, a balance between the liberal naive approach and conservative approach, shows the average K (3.7, 3.2, 3.4 with simulated results of Year 2020; 4.0, 3.4, 3.6 without Year 2020) and standard deviation (1.6, 1.4, 1.5 with Year 2020; 1.4, 1.3, 1.3 without Year 2020) for PECOTA, ZiPS, and FanGraphs, respectively. With or without the simulated results of Year 2020, the One-way ANOVA test does not reject that the true average values of K for these three projection systems are the same at 5% level of significance.

5. Checking the Validity of Normality Assumption

The results obtained above is based on the multivariate normal assumption for the vector of actual wins $X = [X_1, X_2, ..., X_{30}]'$. Due to the limitation of the visual effect on higher dimensions, we have used the 95% probability plot to check the univariate normality for each component X_i, $i = 1, 2, ..., 30$. As well, we have used the 95% confidence contour plot to check the bivariate normality for each pair (X_i, X_j), $1 \le i \ne j \le 30$. Note that there are totally ${}^{30}C_2 = 435$ such pairs. There are no significant violations of the univariate normality for any component X_i, $i = 1, 2, ..., 30$ nor the bivariate normality for any pair (X_i, X_j), $1 \le i \ne j \le 30$, for the years of 2013-2024. To save space, we will not display the probability plots or confidence contour plots here.

6. Conclusion and Comments

With Models 1 and 2b, Tables 1-2 show that 30 out of all 36 projected wins were not plausible values of the actual wins of MLB teams at the 5% level of significance, except for PECOTA (2020), ZiPS (2016, 2020), and FanGraphs (2013, 2016, 2020). Note that 2020 is the pandemic year in which only 60 games were played in empty stadiums by MLB teams. Hence the winning percentages predicted in this special year may not reflect the true performance of teams in a usual regular season of 162 games played in fans-supported stadiums. Table 3 using Model 2c shows that 25 out of 36 projected wins were not plausible values of the actual wins of MLB teams at $\alpha = 5\%$ except for PECOTA (2014, 2020), ZiPS (2014, 2016, 2020), and FanGraphs (2013-2016, 2018, 2020).

Since the assumption of independence of the numbers of wins by teams is violated, the number of wins by a team is further decomposed into the sum of numbers of wins in the matchup games against each team. This approach in Model 3 gives a more precise assessment for each number of wins in the matchup games and hence provides more accurate results. Table 4 using Model 3 reveals that 29 out of 36 projected wins were not plausible values of the actual wins of MLB teams at $\alpha = 5\%$ except for PECOTA (2016, 2020), ZiPS (2016, 2020), and FanGraphs (2013, 2016, 2020). Furthermore, PECOTA had the smallest value of D_2 for 3 years, ZiPS for 4 years, and FanGraphs for 5 years, where D_2 measures the statistical distance between the projected wins and actual wins in the model of matchup games. Thus the smaller value of D_2 implies that the corresponding projection system performs better, i.e., producing more accurate projected wins for all 30 MLB teams as a whole.

The above results are based on the multivariate normal assumption for the vector of numbers of wins

of 30 MLB teams. The checking of the validity of normality doesn't show any significant violation of the assumption.

It is worth noting that the model of matchup games could also be applied to other professional sports such as NHL and NBA to assess the effectiveness of a projection system for its projected wins in a regular season.

It is extremely difficult to accurately predict the outcomes of the numbers of wins achieved by all MLB teams prior to a season of 162 games. There are so many unpredictable variables evolving in teams during the season. Some of these variables could be injuries of key players, adaptation of new players, errors made by players in games, etc. It seems that these three projection systems were not quite effective in predicting the numbers of wins achieved by MLB teams, although FanGraphs might look slightly more promising than the other two systems. Note that FanGraphs are the combination of Steamer and ZiPS. These projected wins, however, could serve as the expectation of the performance of each team prior to the start of a new season. As the season progresses, updates of the projected wins (as some projection systems are doing now) are necessary to reflect teams' momentum, injuries of key players, etc. to readjust previous predictions to more accurate predicted numbers of wins for all MLB teams.

7. References

1. Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological), 289-300*.
2. Chu, D., Wang, C., 2019. Empirical study on relationship between sports analytics and success in regular season and postseason in Major League Baseball. *Journal of Sports Analytics*, (5) 205-222.
3. FanGraphs Projection System Wins. https://web.archive.org/web/20240328204356/https://www.fangraphs.com/depthcharts.aspx?position=Standings (Substitute one of 2013-2023 for 2024 to redirect to the corresponding URL.)
4. Johnson, R., Wichern, D., 2019. Applied Multivariate Statistical Analysis, 6th ed. Pearson.
5. Matchup Games. https://www.baseball-almanac.com/teammenu.shtml
6. MLB Standings. https://www.baseball-reference.com/leagues/MLB/2024-standings.shtml (Substitute one of 2013-2023 for 2024 to get the corresponding URL.)
7. Observed Wins (Team i over Team j). https://www.baseball-reference.com/teams/ARI/2024-schedule-scores.shtml (Substitute one of 2013-2023 for 2024 and the corresponding team's abbreviation for ARI to get the corresponding URL.)
8. PECOTA Projection System Wins. https://web.archive.org/web/20240509182314/https://www.baseballprospectus.com /standings/ (Substitute one of 2013-2023 for 2024 to redirect to the corresponding URL.)
9. Tan, P., Steinbach, M., Karpatne, A., Kumar, V., 2019. Introduction to Data Mining, 2nd ed. Pearson, pp. 772-775.
10. ZiPS Projection System Wins. https://blogs.fangraphs.com/category/2024-zips-projections/(Substitute one of 2013-2023 for 2024 to redirect to the corresponding URL.)