Jinsoo Han (40827)

Traveling Salesperson Problem and its Applications for the Optimum Scheduling

Introduction

When you think of the sports, what word do you imagine first? Competition? Winner? Soccer? Those words pretty well describe the sports. For me, however, fairness is the first word. For the competition, fairness is a necessary and sufficient factor. This is why there are many rules that restrict and prohibit any unfair behaviors. However, some rules are not overwhelming enough to reach the shaded area of fairness which can greatly influence the results of the sports: scheduling. After one season is over, all sports leagues' associations announce the next year's schedule. It is highly likely that teams are not fully satisfied with their schedule for many reasons including more traveling distances than other teams during the season. In this project, I will present the current problems of Major League Baseball (MLB) schedule and how to improve in various criteria: distance, the number of home/away games, and circadian advantage.

The structure of the MLB

There are now 30 teams in MLB that can be divided into American League (AL) and National League (NL). Each league has 3 regional sub-leagues, East, Central, and West. Teams are classified into regional sub-leagues according to the locations of their home stadiums. Table 1 shows the good description of how teams are distributed into each league.

Major League Baseball					
American League			National League		
East	Central	West	East	Central	West
- Baltimore Orioles -Boston Red Sox -New York Yankees - Tampa Bay Rays -Toronto Blue Jays	- Chicago White Sox - Cleveland Indians - Detroit Tigers - Kansas City Royals - Minnesota Twins	- Houston Astros - LA Angels - Oakland Athletics - Seattle Mariners - Texas Rangers	- Atlanta Braves - Miami Marlines - New York Mets - Philadelphia Phillies - Washington Nationals	- Chicago Cubs - Cincinnati Reds - Milwaukee Brewers - Pittsburgh Pirates - St. Louis Cardinals	- Arizona Diamondbacks - Colorado Rockies - LA Dodgers - San Diego Padres - San Francisco Giants

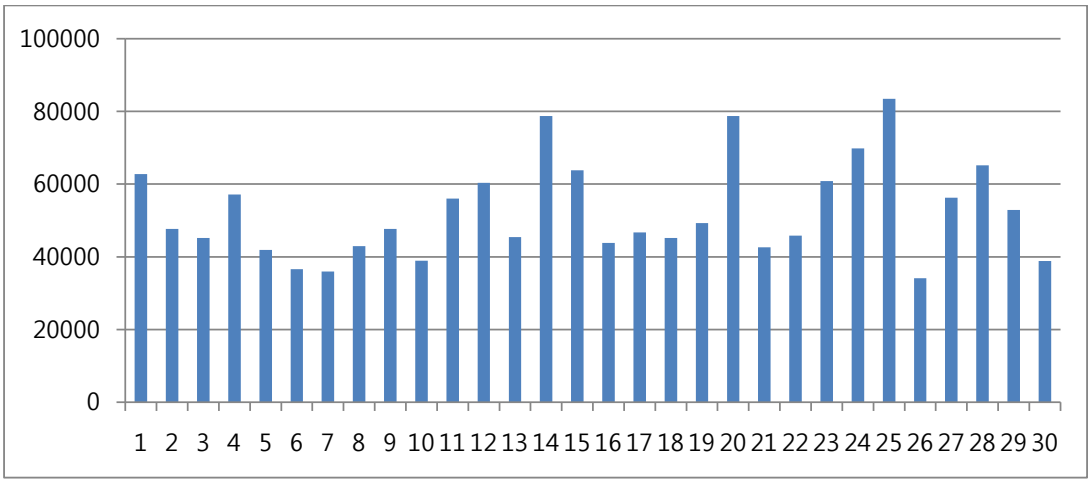
Table 1 describes the classification of MLB teams into each league.

How schedules are made and their problems

According to 2013 Schedule, on which I will focus, all teams will have 52 rounds. Teams will play 25 rounds against each side within their sub-regional division and 22 rounds against league opponents not in their sub-regional division. The rest rounds will be done with teams in the other league. The problem, here, is that not every teams has a match with all the teams. Each team has its own opponent pools while the other team has other opponent pools; for example, St. Louis Cardinals does not have any match with Baltimore Orioles while Toronto Blue Jays has several matches with Orioles. This complicates laying out the schedule perfectly fair to all teams.

Now, let's see the schedule of the year 2013. As you can see figure1 in the next page, the total distances that teams should travel during the season are unevenly distributed. Teams are noted by the alphabetical orders from Arizona Diamondbacks (1) to Washington Nationals (30). The longest distance is 83448 km and the team is Seattle Mariners while St. Louis Cardinals needs to travel only 34068 km during the season, which means that the Mariners should travel twice more than the Cardinals. The distribution of each team' total traveling distance is well expressed by variance and standard deviation: 1.7636×10^{8} and 1.328×10^{4} respectively.

Jinsoo Han (40827)



Traveling Salesman Problem and Formulation

Traveling Salesman Problem (TSP) focuses on finding the itinerary with the minimized total distance traveled under the constraint of given locations. Optimizing the schedule is very similar to the TSP in the big picture; however, it necessitates some small changes in detail. First of all, while the TSP does not consider the existences of other salesmen, creating one team's schedule needs to take the itineraries of other teams into consideration. Furthermore, the main goal of the TSP is to minimize the total distance, but the main goal of this project is to decrease the unfairness in several criteria: total distance, the number of home/away games, and circadian advantage; therefore, I will aim at reducing the variance. As I am not trying to diminish the total amount of the three categories, the optimized schedule might produce the increased total amount but with smaller variances which represent the fairness.

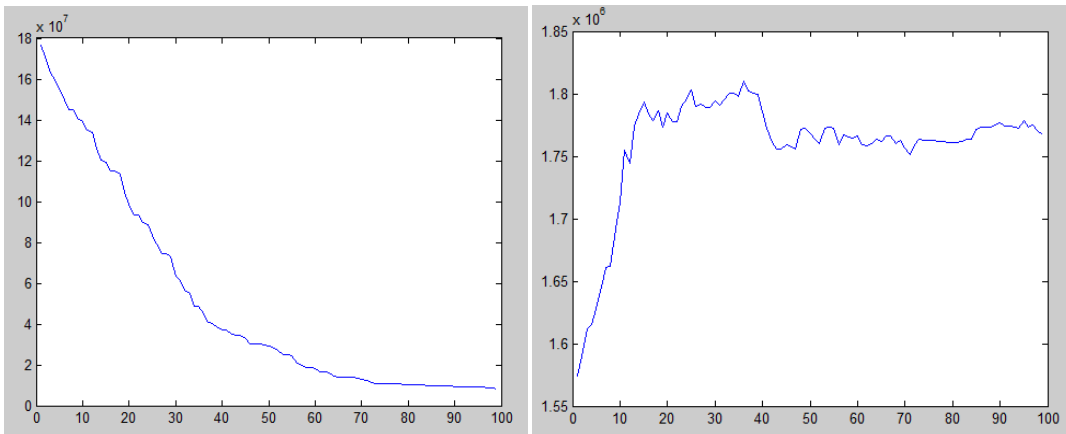
As I am attempting to minimize the variance, I will call it the energy E of the system. In terms of mathematical equation, our E will be expressed as $E = \frac{1}{30}\sum_{1}^{30}(d_i - m)^2$, ($d_i$ is the total distance traveled by team i). Also, I will update the local state according to the

Metropolis Hastings algorithm. By combining the simulated annealing technique and the Metropolis Hastings algorithm, local trials will be accepted with the probability min $(e^{\beta(E(\varepsilon)-E(\varepsilon'))},1)$, ($E(\varepsilon)$ is the previous energy and $E(\varepsilon')$ is the present energy). The remaining issue for the formulation is the rules of local updates. In common TSPs, the rule of transition to an adjacent permutation is picking two different locations and reversing the interval between the two. In my project, I will take two ways of random local updates for this Monte Carlo simulation. First update is SwapRounds. It involves picking two random rounds among 52 rounds and exchanging the two chosen rounds. For example, after choosing round 2 and round 6, I will swap the two rounds entirely. The next update is SwapHomes. The process of SwapHomes is also simple. I will randomly choose two matches of which home team and away team are converted. For example, I will randomly pick a match in which team A is a home team and team B is an away team. Then, I will pick another match in which team A is an away team and team B is a home team and swap the two matches entirely. Therefore, SwapRounds changes the whole MLB teams' schedule and SwapHomes modifies two team's schedule.

Fair distance optimization

As you can see in the left picture, MLB teams are gathered around northeast and southwest parts of America. The uneven distribution naturally incurs the difficulties of scheduling. To solve this natural problem, I manipulated the 2013 schedule into matrices that can be easily interpreted into Matlab program. The execution of Monte Carlo Method followed the formulation that I described

above and I made 100 local updates to reduce the energy function of the system, the variance of the total traveling distances.



Left figure depicts how the energy function (variance of total distances) has changed for 100 trials and right figure shows how the sum of all the teams' travel distance has changed for the same number of steps.

As you can see in the left graph, the simulation succeeded in decreasing the variances of MLB teams' total distances. It first started from 1.76×10^{8} and reached 8.4×10^{6} by diminishing to $\frac{1}{21}$ of the original variance. However, there were also some side effects. One side effect was that some teams have seven consecutive home games while other teams have six consecutive away games. In the original MLB schedule, the maximum number of consecutive home/away game was three. Another side effect was the increased total distance traveled. The right graph above traces how the sum of 30 teams' total distances has changed. Its initial value was 1574356 km and increased to 1767457 km. I would like to explain this outcome as the pie has grown bigger but is divided more evenly. Table 2 in the next page shows the distribution of total distance of MLB teams. Compared to Table 1 based on the original schedule announced by MLB association, all bins in the Table 2 have similar heights, showing the smaller variances.

Furthermore, as a measurement to the increased total distance, I used dual energy

functions for the variance $E_1 = \frac{1}{30}\sum_{1}^{30}(d_i - m)^2$ and for the total distances $E_2 = \sum_{1}^{30} d_i$.

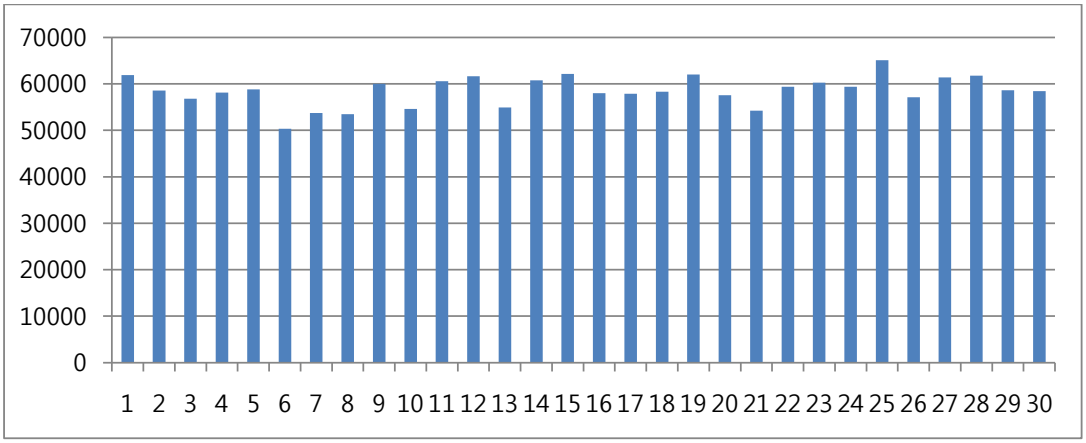
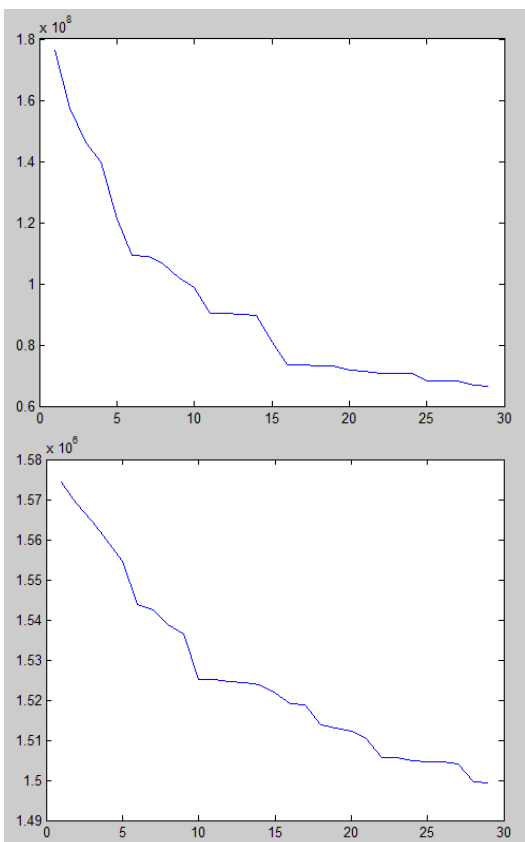


Table 2's bins are higher than those of Table 1 but they are more analogous to each other.

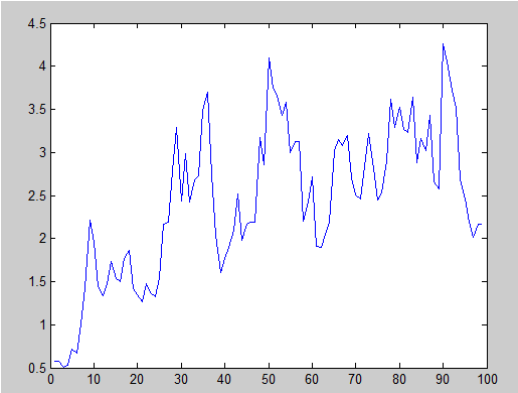


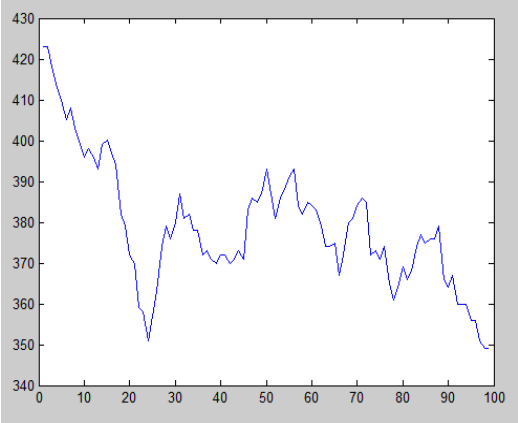
Because additional energy function is added to the simulation, it took a lot more time than the previous simulation. This is why the number of local updates is limited to 30. As you can see in the two left graphs (above is variance and the below is the total distance), the simulation achieved to produce distance optimized schedule. However, posing the dual energy functions could not solve the problem of successive home/away games, which is not the goal of distance-optimization. The problem of home/away games will be treated in the next part of the project.

Jinsoo Han (40827)

Home/away games optimization

In the fair distance optimization, the biggest problem was that some teams have the relatively large number of consecutive home games while other teams move around the continent. In this part, the simulation will focus on the fair distribution of consecutive home/away games. Having a match in the home stadium is a great advantage to the home team. This is why teams with a higher winning rate have more home games during the post season playoffs. Furthermore, having several matches in the home stadium in a row is a bigger problem. For the optimization, energy function will again be the variance, especially the variance of the number of consecutive home games. The structure of the simulation is exactly same with that of previous simulation. Its local update is also done through SwapHomes and SwapRounds.





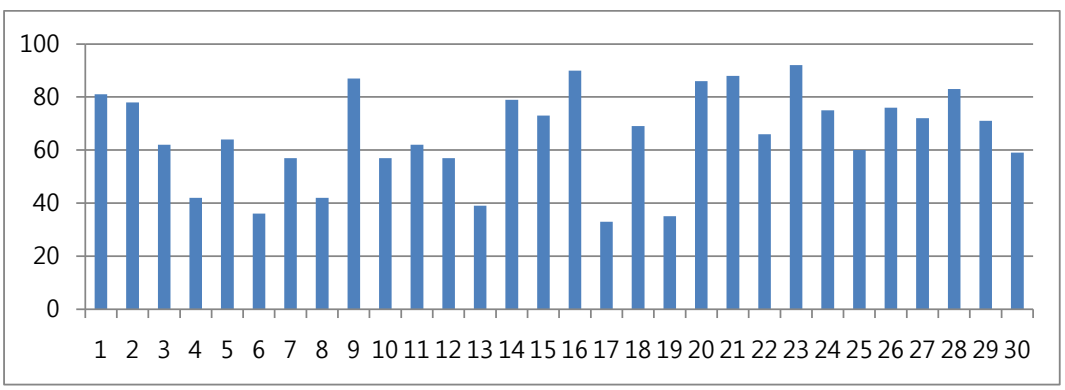
On interesting find in the left graph is that, contrary to the previous simulation, the energy gets higher than the initial state. From this, I could conclude that the MLB association focused on fair distribution of the number of home and away games when they first created the schedule. The second graph represents the total number of consecutive home games during the whole season. As the local state updates, the total number is decreasing, which means that teams have the least number of consecutive home games. This consequently

implies that the total distance traveled is increased because teams are not likely to stay in the home stadium in a row. With two simulations, I could expect that the total distance traveled by all MLB teams is inversely proportional to the total number of consecutive home games.

Circadian advantage optimization

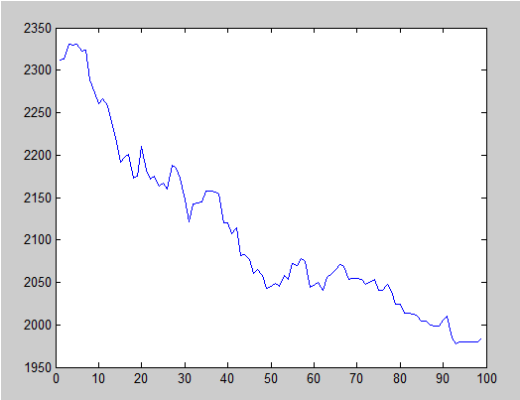
According to the research, "Measuring circadian advantage in Major League Baseball: a 10-year retrospective study," travels across different time zones would affect the athletic performances and influence the result of the game in that "teams with circadian advantage would be more likely to win." In addition to this research paper, there are so many academic publications supporting the effects of the circadian advantage that MLB even invested its funds to investigate the real influences.

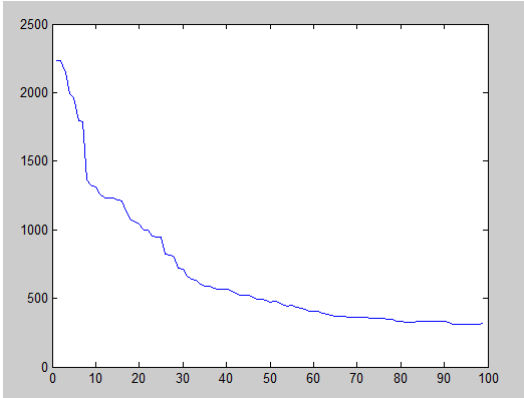
In the third simulation, I will replace the distances between each stadium in the previous simulations with time differences among them. The big picture would be same but the energy function should be changed to $E = \frac{1}{30}\sum_{1}^{30}(c_i - m)^2$, ($c_i$ is the total time differences that team i experiences).



The above figure describes the distribution of all teams' circadian advantages. The higher the bins, the more circadian disadvantage the team has. Teams in the central and east

part of the U.S. are likely to have lower circadian disadvantages while teams in the west seemed to have relatively higher circadian disadvantages.





The left above graph depicts the sum of total circadian disadvantages and the bottom graph illustrates the variance of circadian disadvantages of all MLB teams for 100 local MC steps. Unlike the first and second simulations, the third simulation was found to decrease the total circadian disadvantage as byproduct of annealing. I could ideally optimize the schedule in terms of the circadian advantage. In this simulation, the size of the pie got smaller and, at the same time, it was divided evenly.

Summary and suggestions

I approached to the 2013 MLB regular season schedule in three ways: distance-optimization, home/away game-optimization, and circadian advantage-optimization. Each optimization method has its own characteristics and results. How you define the fairness will determine which optimization would best work case by case. Surely, there would be other ways of optimizations such as airfare, which is proportional to the distance but negligible to the distance that can be covered by bus. I also admit that there could be some implicit constraints that the simulations could not reflect but necessary and important for making the

schedule.

REFERENCES

- Beom Jun Kim, *Traveling Baseball Players' Problem in Korea*, Journal of the Korean Physical Society, Vol 61, No. 3 August 2012

- Winter WC, Hammond WR, Green NH, Zhang Z, Bliwise DL, *Measuring circadian advantage in Major League Baseball: a 10-year retrospective study*, Int J Sports Physical Perform. 2009 Sep; 4(3):394-401

- Ronald W. Shonkwiler and Franklin Mendivil, *Explorations in Monte Carlo Methods*, Springer (2009)

- R. Feres, *Math 350 Fall 2012 – Homework 11*