作業研究 期末專題書面報告
中華職棒野手先發名單最佳化

第 J 組
經濟四-錢紫翎 (B07303129)　生傳四-盧家雯 (B07610058)
經濟四-柯逸萱 (B07303045)　經濟四-陳宛婷 (B07303106)

June 2022

1 簡介

我們使用 2022 年中華職棒的官方賽程，假定自己為某隊總教練，透過最佳化方式的整數規劃模型排出每場比賽的最佳先發野手陣容。整數規劃模型將會分成兩階段，第一階段先根據我方27人球員名單以及對方先發投手的數據，利用自定義的攻擊與守備分數，選出各守備位置攻守分數總和最高的球員。再於第二階段將第一階段得到的守備名單，根據球員在各棒次的加權上壘率來安排棒次，最後會得到一個禮拜中每場比賽的最佳陣容。最後將此模型實踐於中職各球隊四、五月的實際賽程，並與真實排定的先發名單對比，針對我們模型進行結果評量與分析。

2 問題敘述與解決辦法

2.1 問題敘述

先發陣容是整場棒球比賽的核心部分，如何綜合球員的攻防能力派出最强陣容往往是很多球迷關心的重要議題。現行安排先發的方法多是依據經驗法則，選手會被分配到自己最常守的位置，然後再選出每個位置打擊能力最突出者作為先發；棒次安排的部分，傳統上前兩棒具腳程及高上壘能力、中心三棒有高長打能力、後面棒次則是攻擊較弱的選手。除了這些原則外，有時還要考慮輪休、養成或傷病等問題，而且教練幾乎獨攬先發名單排定權，因此先發選擇非常容易受到個人偏好的影響。

我們假設自己是一個職業球隊的總教練，在給定賽程表和對方先發投手的情況下，模擬真實職棒，納入球員輪休的考量，使用最佳化的方式排出每場比賽的最佳先發名單，希望能提供一種客觀可行的模型解決安排先發的問題。如果能透過數據和模型優化陣容安排，既可以提升勝率、降低個人主觀意見的影響，球迷也比較不會質疑球隊安排的陣容。

2.2 解決辦法

我們將陣容安排分為兩個階段。首先根據球員2021年的數據為每個位置選出攻守綜合能力最好的人選，並最大化所有先發選手的攻守能力之和，接著為每場比賽選出的先發球員分配棒次。我們將在模型跟變數介紹的部分，詳細說明本專案如何定義球員的攻擊與守備分數。

第一部分可視為背包問題（knapsack problem）的變體，只需要選擇哪些成員為先發選手，本專案會透過整數規劃（Integer Programming）找出相應的最佳決策，並以真實比賽的先發名單作為整數規劃模型的比較基準。以此模型來決定本週各個場次，各球員「會」或「不會」成為該比賽的先發選手，以及會上場球員的守備位置。考慮到輪休問題與預先安排名單的可行性，我們會以一週為單位，一次安排整週所有賽事的先發陣容。

第二部分則是依照前一部分所選的球員進行棒次排序，若在第一部分有 N 場比賽，第二部分的模型就會執行 N 次。本階段目標是最大化每場比賽九位打者的加權上壘率（wOBA, Weighted On Base Average）總和，這是目前數據棒球派在棒次排序問題中所參照的指標之一。

3 資料蒐集

本專案所需的數據和資料皆從中職官網取得，透過 Selenium 自動爬蟲的方式來獲取資料，以下將對我們欲使用的資料進行描述：

1. 賽程表

   爬取 2022 年 4 月至 5 月的賽程，再爬取每場賽事的比賽場地和先發陣容（每隊 9 名打者和 1 名投手）做為後續模型比較用的評量基準。2022 年 4 月至 5 月總共進行 79 場比賽，中信兄弟參與 33 場、味全龍參與 27 場、富邦悍將參與 34 場、統一獅參與 32 場以及樂天桃猿參與 32場。

2. 防守成績

守備率是中職唯一可以衡量選手守備能力的數據，我們在防守資料欄位紀錄所有野手 2021 年各守備位置的守備機會和守備率（有些隊員可能守備過多個位置，不管守備次數我們均會紀錄）。包含中信兄弟 26 位球員、味全龍 32 位球員、富邦悍將 33 位球員、統一獅 32 位球員以及樂天桃猿 22 位球員。

3. 打擊成績

   爬取 2021 年有一軍例行賽出賽紀錄的野手之賽季打擊數據與分項打擊數據。賽季數據有打席、上壘率、長打率等等，分項數據則包含球員在各月份、各場地以及各棒次的表現（包含打席、打數、安打、二安、三安、全壘打、壘打數、四壞、（故四）、死球以及上壘率）。

4. 投打對決

   與分項成績相似，我們爬取 2021 年有一軍例行賽出賽紀錄的野手對戰不同球隊各投手的資料（包含打席、打數、安打、二安、三安、全壘打、壘打數、四壞、（故四）、死球以及上壘率）。

5. 球員異動紀錄

   為了更加貼近真實賽季情況，我們爬取中職官網所列的四、五月球員異動紀錄，動態更新每場比賽可以選擇的 27 人球員名單。

以上資料將用於整數規劃模型的計算中，此外，對於沒有 2021 年資料的球員，我們會使用他們更早以前的資料或是直接使用 2022 年賽季的數據。

4 整數規劃模型

以下先說明模型將用到的集合與參數之定義，並簡單解釋參數定義的由來及根據，接著說明決策變數以及兩階段模型的目標式和限制式。

4.1 集合

1. G：一週所有比賽的集合，$G = \{1, ..., N\}$，N 代表一週的比賽數量。
2. S_1：包含球隊的明星球員，他們實力堅強、人氣極高，每場比賽皆需要他們先發，否則會引起球迷不滿。

3. S_2：包含球隊的潛力新秀，通常是球隊在選秀會以高順位選進的球員，每週至少要先發一場累積經驗值，否則會引起球迷抗議。

4. S_3：包含隊上較年長或是有傷勢疑慮的球員，為了保護他們的身體，一週至少一場不能先發。

5. I：包含一週內所有可上場的球員們，若某球員該週某幾天被降到二軍，則由模型的決策變數控制他無法上場。

6. J：場上各個守位代號，不包含投手，$J = \{2, ..., 9\}$；而 $J' = \{1, ..., 9\}$ 則包含投手。

7. K：棒次一到九棒，$K = \{1, ..., 9\}$。

S_1、S_2 與 S_3 都是可客製化的虛擬集合，供總教練排定名單時自行定義集合中的球員，本專案因難以定義集合的篩選標準，執行模型時暫不考慮此三個集合。

4.2 參數

1. A_{ij}：如果球員 $i \in I$ 可以守位置 $j \in J$，$A_{ij} = 1$，否則為 0。

2. F_{ij} ：球員 $i \in I$ 守位置 $j \in J$ 的守備分數。

3. B_i：球員 $i \in I$ 的攻擊分數。

4. w_F：守備分數的係數，依守備位置不同。

5. w_B：攻擊分數的係數，依守備位置不同。（不同守備位置對攻擊能力的要求不一樣）

6. V_{ij}：球員 $i \in I$ 守位置 $j \in J$ 的守備分數與攻擊分數的總和，計算公式：

$$V_{ij} = A_{ij}(w_B B_i + w_F F_{ij}).$$

7. P_{jk}：先發守位是 $j \in J'$ 的球員打第 $k \in K$ 棒的打席數（PA, Plate Appearance)。

8. O_{jk}：先發守位是 $j \in J'$ 的球員打第 $k \in K$ 棒的加權上壘率（wOBA, Weighted On Base Average）。

10. Z_g：第 $g \in G$ 場比賽的權重值，對手球隊越强，權重值越高。

4.3 數值定義與資料處理

- 攻擊分數 B_i 是球員 2021 年整個賽季、場地、月份、對戰投手的標準化攻擊指數（OPS^+）乘上相應打席的加權總和值：

$$B_i = \frac{OPS^+_{season}PA_{season}+OPS^+_{month}PA_{month}+OPS^+_{field}PA_{field}+OPS^+_{vsP}PA_{vsP}}{PA_{season}+PA_{field}+PA_{month}+PA_{vsP}}$$

- 守備分數 F_{ij} 是球員守該位置的守備率除以全聯盟守該位置的平均守備率（皆使用 2021 年數據）：

$$F_{ij} = \frac{FPCT_{season}}{FPCT_{\mu_{pos}}}$$

- 守備分數與攻擊分數的係數 w_F、w_B 分別由各位置的守備率除以中職全聯盟平均守備率以及各位置的標準化攻擊指數（OPS^+）除以全聯盟平均（$OPS^+ = 100$）而來（皆使用 2021 年數據）。

$$w_F = \frac{FPCT_{\mu_{pos}}}{FPCT_{\mu}} \quad , \quad w_B = \frac{OPS^+_{\mu_{pos}}}{100}$$

- 每一場比賽權重 Z_g 是根據對手球隊於 2021 全年戰績的勝率除以所有球隊勝率平均，在此不詳細贅述。

- 若球員去年賽季打擊資料的打席數小於 50，標準化攻擊指數會根據打席數線性調整。

- 守某一個位置的守備機會小於 20，此球員守這個位置的守備率會調成全聯盟於該首位的平均守備率（$FPCT_{\mu_{pos}}$）。

- 加權上壘率（$wOBA$）與標準化攻擊指數（OPS^+）皆遵照既定公式計算：

$$OPS^+ = \frac{\text{上壘率}}{\text{聯盟平均上壘率}} + \frac{\text{長打率}}{\text{聯盟平均長打率}} - 1$$

$$wOBA = \frac{0.72\text{保送}+0.75\text{觸身球}+0.9\text{一壘安打}+1.24\text{二壘安打}+1.56\text{三壘安打}+1.95\text{全壘打}}{\text{打席}}$$

4.4 決策變數

1. x_{ijg}：如果球員 $i \in I$ 在第 $g \in G$ 場比賽守位置 $j \in J$，$x_{ijg} = 1$，否則為 0。

2. y_{jk}：如果先發守位是 $j \in J'$ 的球員打第 $k \in K$ 棒，$y_{jk} = 1$，否則為 0。

4.5 模型一：選擇先發球員

最大化一週所有比賽的上場選手綜合分數之和，且每場比賽根據對手球隊强弱乘以相應的權重值。

$$\max \quad \sum_{g\in G}\sum_{j\in J} Z_g V_{ij} x_{ijg}$$

s.t. 每個守位只能有一人上場：

$$\sum_{i\in I} x_{ijg} = 1 \qquad \forall j \in J, g \in G$$

每個球員最多只能上場守一個位置：

$$\sum_{j\in J} x_{ijg} \le 1 \qquad \forall i \in I, g \in G$$

每個球員只能守去年有守備紀錄的位置：

$$x_{ijg} \le A_{ij} \qquad \forall i \in I, j \in J, g \in G$$

捕手最多連續出賽三場（其他野手無輪休限制）：

$$\sum_{g=n-4}^{n} x_{i2g} \le 3 \qquad \forall i \in I, n \in G, \text{ if } n \ge 4$$

S_1 的選手每場都要先發：

$$\sum_{g\in G} x_{ijg} = N \qquad \forall i \in I, j \in J, \text{ if player } i \in S_1$$

S_2 的選手至少要先發一場：

$$\sum_{g\in G} x_{ijg} \ge 1 \qquad \forall i \in I, j \in J, \text{ if player } i \in S_2$$

S_3 的選手至少一場不能先發：

$$\sum_{g\in G} x_{ijg} < N \qquad \forall i \in I, j \in J, \text{ if player } i \in S_3$$

x_{ijg} 是二元變數：

$$x_{ijg} \in \{0, 1\} \qquad \forall i \in I, j \in J, g \in G.$$

完成模型一之後，每場從剩下沒選為先發的球員中挑選攻擊分數 B_i 最高的選手作為該場比賽指定打擊 (DH)，然後使用模型一的結果於模型二排定每場比賽的先發棒次。

4.6 模型二：排定棒次

每場比賽最大化所有棒次的加權上壘率總和。

$$
\begin{aligned}
\max \quad & \sum_{k\in K}\sum_{j\in J'} O_{jk}y_{jk} \\
\text{s.t.} \quad & \text{每個棒次只有一人：} \\
& \sum_{j\in J'} y_{jk} = 1 \qquad \forall k \in K \\
& \text{每個人一棒：} \\
& \sum_{k\in K} y_{jk} = 1 \qquad \forall j \in J' \\
& y_{jk} \text{ 是二元變數：} \\
& y_{jk} \in \{0,1\} \qquad \forall j \in J', k \in K.
\end{aligned}
$$

5 結果分析

將中華職棒五個隊伍依照 2022 年真實的賽程表排程，使用我們定義的數值計算每一場比賽真實先發名單的目標值作為模型評量基準，與我們最佳化模型兩個階段得到的目標值一同比較，以下五張表顯示各隊伍在今年賽季開始兩個月內的真實目標值與模型最佳化目標式值平均以及模型的優化效果。

	第一階段	第一階段加權平均	第二階段
真實資料	2234.426	579.390	2.570
模型最佳化	2636.034	672.458	3.150
優化效果	+18 %	+16 %	+23 %

Table 1: 富邦悍將的真實資料與最佳化模型結果

	第一階段	第一階段加權平均	第二階段
真實資料	2234.767	636.787	2.145
模型最佳化	2511.643	732.388	3.106
優化效果	+12 %	+15 %	+45 %

Table 2: 中信兄弟的真實資料與最佳化模型結果

	第一階段	第一階段加權平均	第二階段
真實資料	1949.993	538.761	2.550
模型最佳化	2340.635	643.857	2.885
優化效果	+20 %	+20 %	+13 %

Table 3: 味全龍的真實資料與最佳化模型結果

	第一階段	第一階段加權平均	第二階段
真實資料	2119.307	536.794	2.499
模型最佳化	2867.298	718.776	3.089
優化效果	+35 %	+34 %	+24 %

Table 4: 統一獅的真實資料與最佳化模型結果

	第一階段	第一階段加權平均	第二階段
真實資料	3058.428	746.088	2.585
模型最佳化	3547.264	885.993	3.277
優化效果	+16 %	+19 %	+27 %

Table 5: 樂天桃猿的真實資料與最佳化模型結果

第一階段真實資料平均值最接近模型結果的是中信兄弟隊，代表他們每場比賽選擇的先發球員最接近最佳化結果，但是第二階段棒次安排的目標值卻與模型相差 45%，遠高於其他四隊的優化效果，代表中信兄弟本季實際棒次安排有相當大的進步空間。中信是去年的總冠軍，而且今年陣容跟去年相差無幾，然而 2022 年目前的戰績不如去年，可從我們的模型推論出棒次安排是可能的原因之一。

第二階段真實資料平均值與模型結果最接近的是味全龍，然而他們第一階段的總平均不管是真實資料或是模型跑出來的值，皆是五隊最低。這與味全龍才剛加入中職一軍第二年，球員陣容不若其他四隊完整有關。不過目前味全以相對較弱的戰力排在聯盟中段，原因可從模型與真實結果的對照發現出他們棒次安排是五隊中最接近最佳結果的。

由第一階段平均值可觀察出今年整體球員戰力最強的是樂天桃猿，即使去年他們戰績並非位居前列，依然能從去年資料算出他們最強大的戰力值，不愧是近年中職火力最強大的球隊，因此目前是全聯盟第一。

統一獅因為開季至今傷兵不斷，可用選手多為新人球員，過往球員資料因為樣本較少，統計分析的結果會比較不精準，這也是統一獅第一階段真實資料平均值相差模型結果不少的理由之一。

富邦悍將貴為今年話題度最高的球隊，十分淒慘的戰績讓球迷紛紛怪罪於教練團，每每公布先發名單總是令人議論紛紛。但是根據上述圖表的結果可發現富邦不論是球員選擇或棒次安排，與最佳化結果相比都沒有太大的落差。球員能力值加總排名第三，被詬病的打擊貧弱部分，第二階段平均加權上壘率更僅次於樂天桃猿。因此從表面數據難以看出其戰績嚴重落後的原因，畢竟會影響球隊勝敗的因素絕不僅是野手攻守能力或棒次問題，投手能力、球隊氣氛、戰術執行、教練調度等等都是可能原因，這些超出本專題範圍就不詳細說明。

將五隊的最佳目標式值取平均，與由真實資料算出的目標式值平均做比較，可以得到如表 6 的結果。

	第一階段	第一階段加權平均	第二階段
真實資料	2319.384	607.564	2.470
模型最佳化	2780.575	730.694	3.101
優化效果	+20 %	+20 %	+26 %

Table 6: 全聯盟真實與最佳化結果平均

由模型算出的目標式值平均在任何個隊伍或整個聯盟的表現都比真實情況好，第一階段的加權平均值，也就是所有打者的平均守備及打擊能力比真實情況好 20%，而第二階段的平均值，也就是所有打者的平均加權上壘率也比真實值增加了 26%。

雖然模型算出的目標式值並不代表比賽得分數或失分數，也無法評估野手以外的因素，但可作為上場野手攻守綜合能力的衡量指標。如前述對

各隊的分析，模型結果與實際資料對照能用來評比今年每隊先發名單安排的優劣。即使我們使用的數據計算方式並非十分精確，卻能大致衡量上場球員的平均能力值，而且模型結果也都比實際資料表現更加優異，各球團或許能將我們的模型作為安排先發名單的參考之一。

6 結論

從結果分析可得知我們模型得到的最佳陣容相較於實際上場陣容，確實能更有效率、更客觀地安排先發名單。雖然數值計算方式不是十分精確，但我們相信這樣的最佳化方式仍具一定參考價值。比起以往較主觀、傳統的先發選擇方式，希望我們的模型能給予相對客觀的意見，讓球隊陣容的安排更加完善。

關於改善我們模型的方法，除了增加更多相關的參數計算球員的攻守能力以外，還可以使用敏感度分析（Sensitivity Analysis）或透過其他數據分析的方式， 研究不同參數、不同算法會對結果產生什麼影響，進而找出更適合的參數與計算方式。

此外，本次專案尚有三個未處理的虛擬集合，未來若能設計應用程式介面（API），在執行模型前先選擇每個集合所包含的球員，相信安排名單時會有更多彈性、更大的變化空間，也更貼近實際安排名單的情況。

有賴於中華職棒隊數少、球員流動率低的特性，我們模型才得以使用目前的數據進行整數規劃，若像大聯盟（MLB）年年有許多球隊交易、釋出球員，我們的模型與參數計算方式將需要更多調整，所以本專案的最佳化方式仍待改善。不過根據上述針對我們模型的分析，只要一個球隊具有充足完整的球員數據庫以及嚴謹、有效的量化方式，相信此最佳化方式能幫助球隊教練安排先發球員，進而提高球隊勝率、降低教練主觀偏好的因素，令更多球迷滿意球隊的先發陣容。