CO-004

日本プロ野球における混合分布モデルを用いた野手の分類
Classification of fielders in Nippon Professional Baseball using a Gaussian mixture clustering model

織田 大志　　廣津 信義
Taishi Oda　Nobuyoshi Hirotsu
順天堂大学大学院スポーツ健康科学研究科

1. はじめに

「野球」というスポーツは，幅広い国々でアマチュアからプロまで存在するメジャーなスポーツである．日本においても，NPB（日本野球機構）傘下の 12 球団が公式戦を行い，毎年優勝を争っている．

その公式戦で選手が残した記録（長打率，出塁率，OPS など）をデータとして統計学的見地から分析する手法のことをセイバーメトリクス[1]と呼ぶ．セイバーメトリクスでは，代表的な統計手法の一つ，多変量解析も行われている．多変量解析[2]とは，複数の変数に関するデータを元に，これらの変数間の相互関連を分析する統計的技法である．

その中でも，後述する代表的な手法，クラスター分析と主成分分析を用いて，選手の記録をデータとして分析した先行事例はいくつかある．

クラスター分析において最も多いのが k-means 法（c-平均法）[3]による選手の分類である．k-means 法は，非階層型クラスタリングの一つで，クラスターの平均を用い，対象を与えられたクラスター数 k 個に分類する．田中 成典・鳴尾 丈司・山本 雄平・西藤 怜(2021)は，関西大学体育会野球部に所属する選手を，スイング計測装置から得られるスイング特性によって分類した．また，同様に実際の打撃成績によっても分類し，両者のクラスタリング結果を比較した．この際，クラスター数 k を決定するのに，両者のクラスタリング結果が最も近くなる値を選択している．しかし，比較対象が存在しないデータを対象にした場合，k-means 法は，クラスター数を主観的に決定しなければならない．

一方で，クラスター数を客観的に決定するクラスタリング手法も存在する．混合分布モデル(Gaussian Mixture Model)によるクラスター分析[5]では，適切なクラスター数，分散共分散行列の型を BIC 値によって決定することができる．またこのモデルは，複数の正規分布の重ね合わせで表現される多峰型の分布を持つデータに対して有効な点も魅力である．

酒折 文武・圓城寺 啓人・竹森 悠渡・西塚 真太郎・保科 架風(2017)は，2010 年から 2014 年に投球した日本プロ野球の投手に混合分布モデルによるクラスター分析を適用した．分類に当たって使用した 1 試合当たりの投球数の分布が多峰型であり，15 球と 95 球あたりに明確にピークがあったからである．BIC 値によって適切なクラスター数は 5 となった．

このように，混合分布モデルによるクラスター分析によって，対象のデータを柔軟に分類することが可能である．しかしながら，クラスター分析を行う際，分類に使用する変数が複数の場合，一つの問題が生じる．それは，分類結果の解釈が困難になるということである．酒折他(2017)は 1 試合当たりの投球数のみを使用しているので，クラスターの特徴は投球数の違いでのみ決定されることは明白である．逆に，変数の数が莫大，例えば 10 や 20，100 といった高次元の場合，変数の値でクラスターの特徴を決定する作業は煩雑になる．

この問題を解決できるのが主成分分析である．主成分分析とは，相関のある多数の変数から，相関の無い少数で，全体のばらつきを最もよく表す主成分と呼ばれる変数を合成する手法である．蔭山 雅洋・田中 成典・山本 雄平・鳴尾 丈司(2021)は，対象者の大学野球選手のスイングデータに基づき，複数の要因を集約してスイングデータ 7 変数を総合的に評価するため，主成分分析を行った．7 変数は二つの主成分に縮約され，第 1 主成分（ボールの高低）にはスイングの空間的なコンパクトさ，打球角度や方向に影響するバットの角度，打球速度に影響するヘッドスピードやローリングが，第 2 主成分（ボールの内外角）にはスイングの時間的なコンパクトさと適切な打撃位置でボールインパクトを迎えているかを判断できるインパクト加速度が影響すると解釈した．このように，元の変数の情報から縮約された主成分の情報を新しく解釈することができる．

そこで，本研究では，多数の変数の情報を縮約し，選手を柔軟にグループ化する一つの「枠組み」を提示できるのではないかと考えた．ここでいう「枠組み」とは，酒折，蔭山らのように，主成分分析・クラスター分析を単体で実施するのではなく，主成分分析の結果を利用してクラスター分析を行うという「流れ」を指す．主成分分析で高次元の変数を少数の変数に縮約した後クラスター分析を実施すれば，元の変数の情報をできるだけ損なうことなく，かつクラスター分析の結果の解釈も煩雑にならずに済む．

この「流れ」を提示した先行研究には，César Soto-Valero(2017)がある．彼は，FIFA の公式ウェブサイトから入手できる 7,705 人のヨーロッパのサッカー選手の指標に主成分分析を施し，情報を 2 つの変数に縮小した．次に，その変数を使用して混合分布モデルによるクラスター分析を行い，類似した選手からなる 4 つのクラスターを生成した．主成分分析とクラスター分析を用いて，サッカー選手をグループ化するための枠組みを示すことが目的であった．

また，西内啓(2012)は，2011 年 J1 に出場したサッカー選手の指標 200 以上に主成分分析を施し，5 つの変数に縮約した．さらに，クラスター分析を行い，類似した選手からなる 8 つのクラスターを生成した．同じクラスターに所属している選手同士は，お互い代替選手となれる可能性についても言及している．

本研究はサッカーではなく日本プロ野球の選手に同様な分析を実施する．つまり，日本プロ野球選手が残した成績に主成分分析を施し，少数の変数に縮約した後，クラスター分析によって選手を分類する．さらに，ランダムフォレスト[10]によってどの成績がクラスター分析の結果に寄与しているのか後付けで調査する．最後に，同じクラスターに所属している選手同士は，トレードができる，またはお互い代替選手となれる可能性について考察する．

以上の分析には，プロ野球に関するデータ分析を扱う株式会社 DELTA が提供するサービス”1.02 Essence of

Baseball”(https://1point02.jp/op/index.aspx)から入手できるデータ，2020 年の公式戦に出場した野手 327 人の 115 種類の打撃指標を分析する．株式会社 DELTA は，野球を客観視して既存の視点とは違う角度から野球を考える組織で，優れた分析家と協力して，球界に多くの発見をもたらすことを目指す企業である．ちなみに，日本プロ野球機構公式サイト(https://npb.jp/)では，22 種類公開されている．

2. 手順

以下の手順に従って分析を行う．すべての計算は，統計解析向けプログラミング言語 R(version 3.5.1)で実行されている．

2.1 主成分分析

多変量解析を行う際，変数の削減・選択を行うことがある．本研究では，主成分分析を使用することにした．主成分分析とは，相関のある多数の変数から，相関の無い少数で，全体のばらつきを最もよく表す主成分と呼ばれる変数を合成する手法である．

$X_{n\times p}$をn個の個体，p個の変数から成るデータセットとする．合成変数は，p次元のデータをより低いk次元($k \le p$)に縮約された線形結合式

$$z_j = a_{1,j}x_1 + a_{2,j}x_2 + a_{3,j}x_3 + \cdots + a_{p,j}x_p,\ (j = 1, \cdots, k)$$

で表される．このときの係数$a_{i,j}(i = 1, \cdots, p)$を主成分と呼ぶ．主成分分析においては，この主成分を，z_jの分散が最大となるように$\sum_{i=1}^{p} a_{i,j} = 1$という制約の下で求める．主成分に縮約されている元の変数の情報は，主成分負荷量によって確認することができる．

本研究では，115個の打撃指標に主成分分析を実施，累積寄与率70%前後を目安[3]として，主成分を選択する．そして，各主成分と元の変数との相関を表す主成分負荷量から，主成分の特徴を解釈する．

2.2 混合分布モデルによるクラスター分析

データによっては，頻度のピークが二箇所以上あるような多峰型の分布を持つ場合がある．このようなデータに対して，正規分布のような分布の山が一つだけである単峰型の分布のみを仮定するモデルは適切でなく，その代わりに，二つ以上の異なる単峰型の分布の合成も仮定した混合分布モデルが用いられる．

今任意の混合分布に含まれる$\boldsymbol{G}$個の正規分布の確率密度関数を$\boldsymbol{f_1(x;\theta_1)}, \ldots, \boldsymbol{f_G(x;\theta_G)}$, これらの混合比を$\boldsymbol{\pi_1, \ldots, \pi_G}$とする．$\boldsymbol{\theta_g(g = 1, \ldots G)}$は確率（密度）関数$\boldsymbol{f_g(x;\theta_g)}$に含まれるパラメータからなるベクトルである．また，混合比$\boldsymbol{\pi_1, \ldots, \pi_G}$については$\boldsymbol{0 \le \pi_g \le 1 (g = 1, \ldots G)}, \boldsymbol{\sum_{g=1}^{G} \pi_g = 1}$を満たすものとする．このとき，混合分布モデルの確率（密度）関数は次で与えられる．

$$\boldsymbol{f(x;\theta) = \sum_{g=1}^{G} \pi_g f_g(x;\theta_g)}$$

このモデルに含まれるパラメータ$\theta = (\theta_1^T, \ldots, \theta_G^T, \pi_1, \ldots, \pi_{G-1})^T$を推定するには，EM アルゴリズムを用いる．

この EM アルゴリズムの E ステップで用いられる条件付き期待値

$$\gamma_{ig} = E(Z_{ig}|x_i) = Pr(Z_{ig} = 1|x_i)$$
$$Pr(Z_{ig} = 1|x_i) = \frac{\boldsymbol{\pi_g f_g(x_i;\theta_g)}}{\sum_{h=1}^{G} \boldsymbol{\pi_h f_h(x_i;\theta_h)}}$$

の推定値が最大となる成分へ第i観測値を分類すれば，混合分布を用いてクラスター分析を実施することができる．

また，クラスター数及び分散共分散行列の型の BIC[10]によって，最適なモデルを選択できる．本研究では，選択した主成分によって計算される主成分得点を用いて混合分布モデルによるクラスター分析を実施する．

2.3 ランダムフォレスト

ランダムフォレストとは，母体となる全てのデータを段階的に分割し，決定木と呼ばれるツリー状の分析結果を出力する方法である．条件に応じて分岐を設定し，ルートから辿り，条件に最適なものに分割するアルゴリズムである．クラスター分析の結果を目的変数として，説明変数が分類結果にどの程度影響するかを知ることができる．

2.4 データの絞り込みと分割

野手 327 人，115 個の打撃指標に対して 2.1 から 2.3 を適用すると，変数が多量なため主成分の特徴を解釈しきれない可能性がある．クラスター分析の結果も同様である．極端に出場が少ない選手がいることから，分類結果も G（試合数）や PA（打席数）によって大きく左右されてしまうことが考えられる．

そこで本研究では，選手を WAR[11]が 1.0 以上の 72 人に絞り込む．WAR とは，セイバーメトリクスによる打撃・走塁・守備・投球を総合的に評価して選手の貢献度を表す指標である．G や PA 自体によって絞り込まないのは，明確な評価基準が存在しないからである．

表 1　WAR の評価基準

評価	値
MVP	6.0-
スーパースター	5.0-6.0
オールスター	4.0-5.0
好選手	3.0-4.0
固定レギュラー	2.0-3.0
レギュラー	1.0-2.0
補欠	0.0-1.0

また，主成分分析の結果の解釈を容易にするために株式会社 DELTA が分別した 8 つの項目に属する指標ごとに分析を実施する．

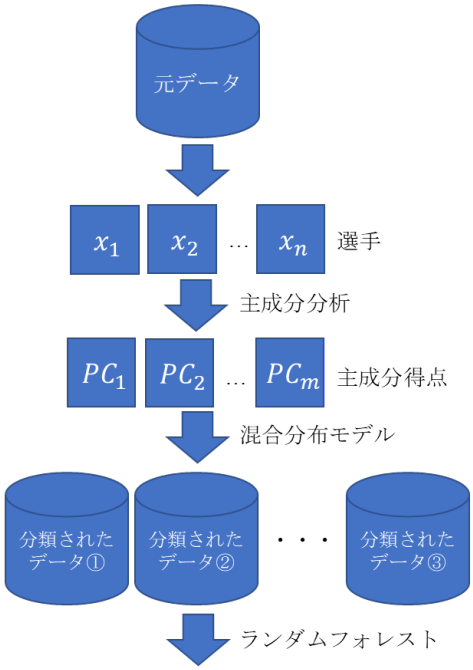

図 1　分析の流れ

図 1 に分析の流れを示した．野手 327 人，115 個の打撃指標に対して上記を適用した後，8 つの項目に分けた際も項目ごとに上記の流れを適用する．

3. データ

株式会社 DELTA が提供している 2020 年の公式戦に出場した野手 327 人の 115 種類の打撃指標を用いる．打撃指標はその意味合いによって詳細な項目 8 つに分かれている．

表 2　株式会社 DELTA 指標一覧

項目	概要
Standard	セイバーメトリクス系指標ではない標準的な成績．PA（打席），AVG（打率）など．
Advanced	一般的なセイバーメトリクス系指標．SLG（長打率），OBP（出塁率）など．
Batted Ball	打球について統計的に纏められたデータ．ゴロ，フライの割合（GB%，FB%）など
Win Probability	勝利貢献度を表す指標．WPA+，WPA-（勝利期待値を上昇，下落させた合計値）など．
Pitch Type	投球された球種の割合や球速のデータ．FAv，CTv（ストレート，カットボールの平均球速）など．
Pitch Value	投球された球種毎の得点増減．wCB（カーブに対する得点増減の合計）など．
Plate Discipline	打者の選球眼についてまとめられたデータ．Contact%（対戦打者がスイングした際，打球が発生した割合）など．
Value	WAR（Wins Above Replacement）の算出に必要な指標．

4. 結果

115 種類の打撃指標に主成分分析を実施した結果，第 18 主成分までの累積寄与率が約 7 割であることがわかった．第 18 主成分までの主成分負荷量から主成分の特徴の解釈を試みたが，全選手，全指標に対して分析を適用すると，主成分の特徴を解釈しきれなかった．よって，2 節で示した手順通り株式会社 DELTA が分類する 8 項目に属する指標ごとに分析を実施した．

4.1 主成分分析

8 つの項目毎に主成分分析を実施し，累積寄与率を纏めた結果が図 2 である．累積寄与率が約 7 割を目安に主成分を選択すると，表 3 のようになった．

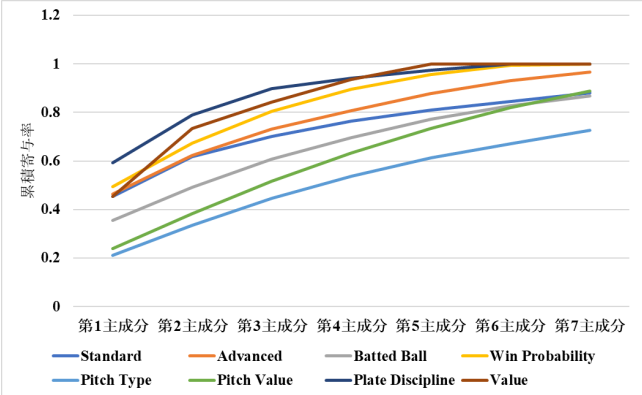

図 2　各項目の累積寄与率

表 3　選択した主成分の数

Standard	Advanced	Batted Ball	Win Probability
3	3	4	2
Pitch Type	Pitch Value	Plate Discipline	Value
6	5	2	2

続いて，主成分負荷量から主成分の特徴の解釈を試みた．表 4 は，各項目の主成分で負荷量が大きい指標上位 3 つを纏めたものである．ここでは簡単のため，主成分の数が同じだった Win Probability，Plate Discipline，Value のみ示す．

表 4　負荷量の大きい指標上位

項目	第 1 主成分	第 2 主成分
Win Probability	REW	WPA-
	RE24	Clutch
	WPA	PH
Plate Discipline	SwStr%	Swing%
	Z-Swing%	O-Swing%
	Swing%	O-Contact%
Value	RAR	WAR
	Defense	Replacement
	Replacement	Batting

これらの指標から，主観的に各主成分の特徴を解釈する

と表 5 のようになる．

表 5　主成分の解釈

項目	第 1 主成分	第 2 主成分
Win Probability	得点貢献	チャンスに強い
Plate Discipline	ストライク球への積極的スイング	ボールに手を出す
Value	守備総合評価	打撃総合評価

Win Probability の第 1 主成分であれば，REW，RE24 のように，各選手がどれだけ得点期待値を増減させたかによって貢献度を評価する指標の主成分が高い．つまり，第 1 主成分によって計算される主成分得点の値が高いほどその選手は得点貢献度が高いと判断して，第 1 主成分の特徴を「得点貢献」とした．

同様に，Plate Discipline の第 1 主成分であれば，SwStr%（スイングストライク率＝空振り率．全投球に対し，打者が空振りしストライクとなったケース割合），Z-Swing%（ストライクゾーンのスイング率．全投球に対し，ストライクゾーンに投球された球をスイングした割合）の主成分が高い．つまり，第 1 主成分によって計算される主成分得点の値が高いほどその選手はストライク球へ積極的にスイングしていると判断して，第 1 主成分の特徴を「ストライク球への積極的スイング」とした．

4.2 混合分布モデルによるクラスター分析

主成分得点を用いて，混合分布モデルによるクラスター分析を実施した．表 6 は，それぞれの項目で最も BIC が高かったクラスター数である．

表 6　BIC が高いクラスター数

Standard	Advanced	Batted Ball	Win Probability
3	2	2	2
Pitch Type	Pitch Value	Plate Discipline	Value
2	2	1	2

選択した主成分の数が 2 だった Win Probability と Value において，色でクラスター分割がわかるように散布図を作成したのが図 3,4 である．

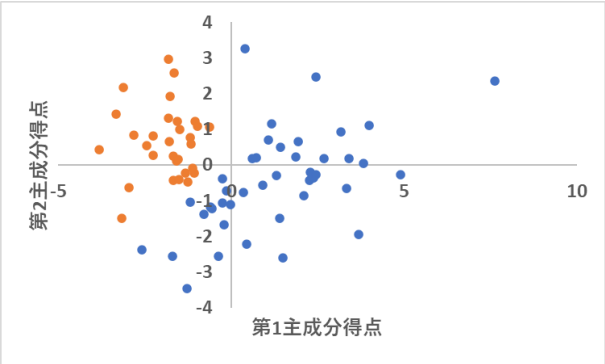

図 3　Win Probability クラスター別散布図

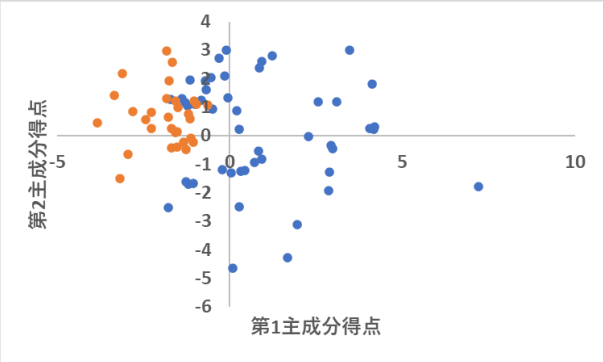

図 4　Value クラスター別散布図

横軸は第 1 主成分によって計算される主成分得点，縦軸は同様に第 2 主成分によって計算される主成分得点である．Win Probability であれば，おおよそ第 1 主成分得点が高い選手と低い選手で分かれている．第 1 主成分の特徴は「得点貢献」であったので，得点貢献度が高い選手とそうでない選手とでクラスターが分割されていると解釈できる．

続いて，各項目クラスターの主成分得点の平均を計算した結果を表 7 から表 13 に示す．

表 7　主成分得点平均（Standard）

クラスター番号	第 1 主成分	第 2 主成分	第 3 主成分
1	2.905	-1.297	-0.262
2	-0.002	1.665	0.757
3	-3.034	-0.687	-0.655

表 8　主成分得点平均（Advanced）

クラスター番号	第 1 主成分	第 2 主成分	第 3 主成分
1	-1.054	-0.143	1.010
2	0.997	0.135	-0.956

表 9　主成分得点平均（Batted Ball）

クラスター番号	第 1 主成分	第 2 主成分
1	0.640	0.730
2	-0.640	-0.730
クラスター番号	第 3 主成分	第 4 主成分
1	0.186	-0.208
2	-0.186	0.208

表 10　主成分得点平均（Win Probability）

クラスター番号	第 1 主成分	第 2 主成分
1	1.328	-0.462
2	-1.859	0.647

表 11　主成分得点平均（Pitch Type）

クラスター番号	第 1 主成分	第 2 主成分	第 3 主成分
1	-0.199	-0.175	0.231
	第 4 主成分	第 5 主成分	第 6 主成分

	0.332	0.292	-0.385
2	第 1 主成分	第 2 主成分	第 3 主成分
	0.273	-0.273	-0.005
	第 4 主成分	第 5 主成分	第 6 主成分
	-0.455	0.456	0.008

表 12　主成分得点平均（Pitch Value）

クラスター番号	第 1 主成分	第 2 主成分	第 3 主成分	第 4 主成分	第 5 主成分
1	0.411	-0.522	0.122	0.072	-0.292
2	-0.998	1.269	-0.297	-0.175	0.709

表 13　主成分得点平均（Value）

クラスター番号	第 1 主成分	第 2 主成分
1	0.844	0.137
2	-1.687	-0.274

各項目のクラスター毎の平均を計算することで，先の主成分の解釈と併せてそのクラスターの特徴を捉えることができる．Win Probability であれば，第 1 主成分得点の平均がクラスター2 より高いことから，クラスター1 には得点に貢献している選手がクラスター2 より多く所属していることがわかる．同様に第 2 主成分得点においても，平均がクラスター2 の方が高いことから，クラスター2 にはチャンスに強いタイプの選手が所属していることがわかる．

4.3 ランダムフォレスト

元の指標の中で，分類結果に大きく影響を及ぼしている変数を調べるために，クラスター分析の結果を目的変数，元の変数を説明変数とし，ランダムフォレストを適用した．その結果が図 5 である．それぞれの項目の中で，重要度が高かった指標上位 2 つを示す．

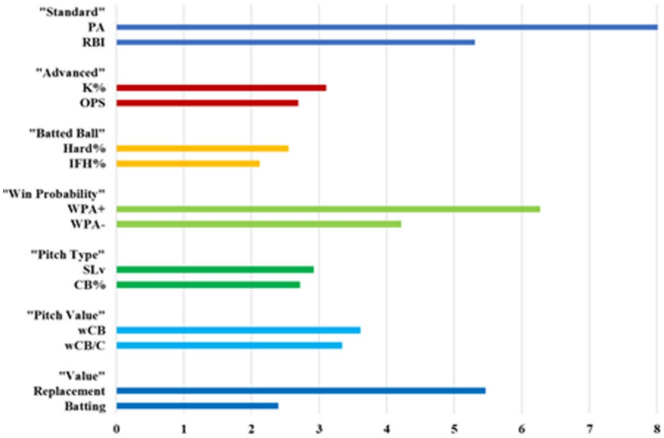

図 5　各指標の重要度（項目別）

Standard であれば，PA がクラスター分析の結果に大きく影響していることがわかる．選手を絞ったとしても，やはり打席数が多い選手と少ない選手で分かれていると推測できる．実際，各クラスターで重要度が高い指標の平均値を計算して比較すると表 14 のようになる．

表 14　重要度が高い指標の平均

項目	クラスター番号	PA	RBI
Standard	全体	373.444	44.375
	1	469.522	71.913
	2	403.630	34.407
	3	235.955	27.818
Advanced	クラスター番号	K%	OPS
	全体	18.322	0.789
	1	15.411	0.735
	2	21.076	0.841
Batted Ball	クラスター番号	IFH%	Hard%
	全体	6.361	36.008
	1	7.625	37.286
	2	5.097	34.731
Win Probability	クラスター番号	WPA-	WPA+
	全体	-5.880	7.110
	1	-7.202	9.241
	2	-4.030	4.127
Pitch Type	クラスター番号	SLv	CB%
	全体	128.964	7.374
	1	128.320	8.440
	2	130.037	5.596
Pitch Value	クラスター番号	wCB	wCB/C
	全体	0.900	1.045
	1	1.769	1.930
	2	-1.210	-1.103
Value	クラスター番号	Replacement	Batting
	全体	11.703	9.963
	1	13.810	14.679
	2	7.488	0.529

Standard において，PA の平均値が各クラスター間で違うことがわかる．他の指標についても同様なことが言える．

5. 考察

5.1 ダミー変数の導入

表 15 は，Advanced の指標を使ってクラスター分析を行った結果をダミー変数へ変換したもの（一部）である．Standard については 3 つのクラスターに分かれているため，他の項目とは異なり，表 16 のように 3 つに分けてダミー変数に変換した．どちらのクラスターに分かれているかで”0”または”1”の値へ変換し，Plate Discipline を除く 7 項目の分析結果の内，同じクラスターに所属している回数が多い選手のペアを特定した．

表 15　ダミー変数に変換後

選手	クラスター番号
吉田　正尚	0

安達　了一	0
S・モヤ	1
伏見　寅威	0

表 16　ダミー変数に変換後（Standard の場合）

選手	クラスター番号①	クラスター番号②	クラスター番号③
福田　周平	0	1	0
西川　龍馬	0	0	1
鈴木　誠也	1	0	0

5.2 階層的クラスター分析

次に，これらのダミー変数を使って，最近隣法による階層的クラスター分析[3]を改めて実施した．個体間の類似度あるいは非類似度（距離）に基づいて，最も似ている個体からグループを作っていく方法である．その結果は，図 6 から 8 に示すような樹形図で示される．末端のラベルが付いている部分を葉と言い，葉と葉の距離が近いほどその選手同士は同じクラスターへの分類頻度が高いということを表している．ここでいう距離とは，葉から上に延びている線が連結するまでの高さである．

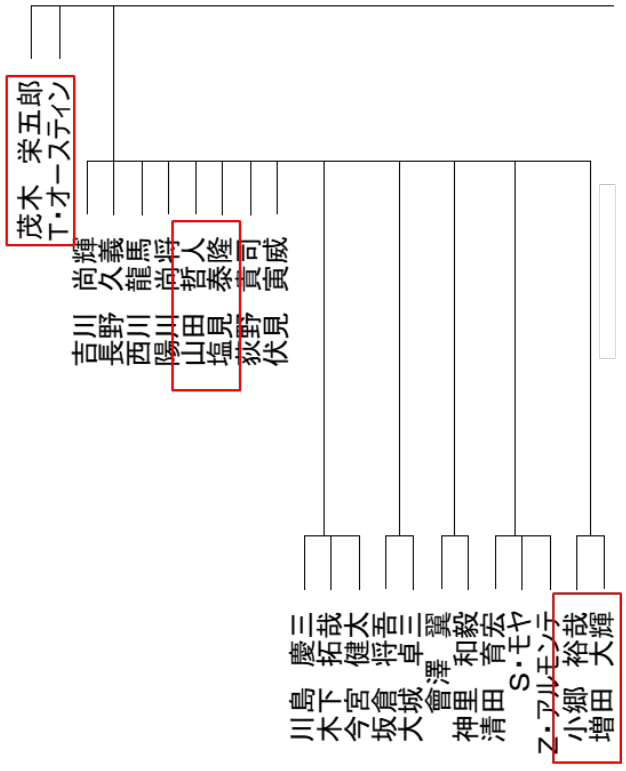

図 6　階層的クラスター分析結果①

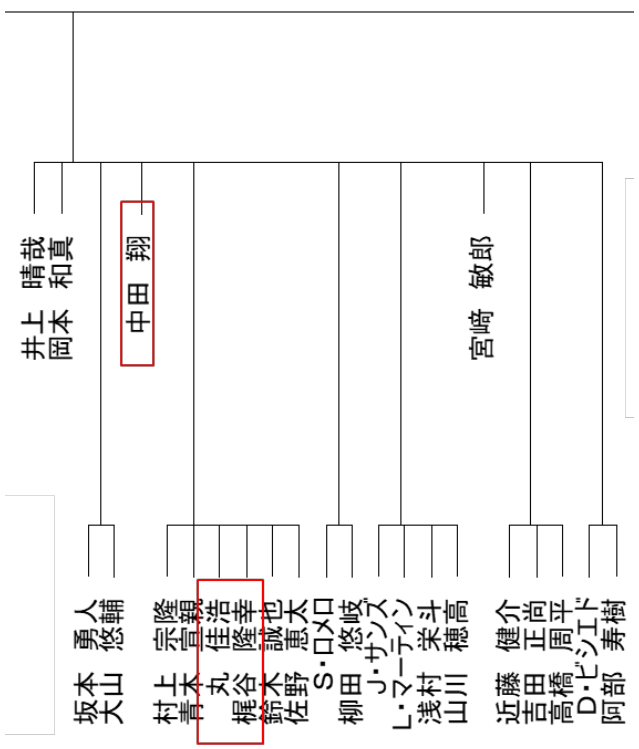

図 7　階層的クラスター分析結果②

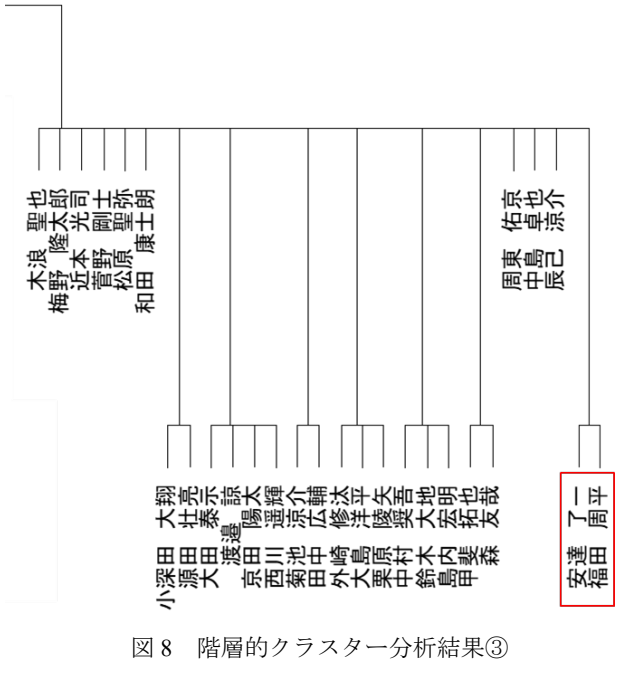

図 8　階層的クラスター分析結果③

図 6 で距離が近い小郷と増田であれば，7 項目のクラスター分析ですべて同じクラスターに分類されている．表 17 に示すように，ダミー変数がすべて同じ値だからである．ちなみに Standard は，①②③すべて同じ値だったときのみ同じクラスターに分類されているということを表す．

表 17　距離が近い選手のダミー変数パターン(1)

選手	Advanced	Batted Ball	Win Probability
増田 大輝	1	0	1
	Pitch Type	Pitch Value	Value
	0	0	1
	Standard①	Standard②	Standard③
	0	0	1
小郷 裕哉	Advanced	Batted Ball	Win Probability
	1	0	1
	Pitch Type	Pitch Value	Value
	0	0	1
	Standard①	Standard②	Standard③
	0	0	1

図 6 の茂木とオースティンも同様である．7 項目のクラスター分析の内 5 回同じクラスターに分類されている．反対に，距離が離れている茂木と増田は，7 項目の内 4 回同じクラスターに分類されている．このように，完全ではないが，距離の近い選手同士は，同じクラスターへの分類頻度が高かった．

表 18　距離が近い選手のダミー変数パターン(2)

選手	Advanced	Batted Ball	Win Probability
T・オースティン	1	0	0
	Pitch Type	Pitch Value	Value
	1	0	0
	Standard①	Standard②	Standard③
	0	0	1
茂木 栄五郎	Advanced	Batted Ball	Win Probability
	1	1	0
	Pitch Type	Pitch Value	Value
	0	0	0
	Standard①	Standard②	Standard③
	0	0	1

この同じクラスターへの分類頻度の多少は，選手の類似度と大きく関係している．項目は全部で 8 つあり，それぞれ意味合いが異なるから，分類頻度が高ければ高いほどその選手同士はあらゆる側面で類似しているということを表す．これは，球団がチーム強化を行う際に利用できる可能性がある．

5.3 代替選手

同じ球団の選手に着目すると，類似した選手同士は，お互い代替プレイヤーになれる可能性がある．図 6 であれば，ヤクルトスワローズの山田と塩見である．2021 年，クリーンアップを務めた山田が足のコンディション不良で離脱した際には，開幕戦で 6 番を務めていた塩見が代わりにクリーンアップを務めた．

同じ年のオリックスバファローズの安達と福田についても，図 8 に示すように距離が近い．2021 年，チームの方針で福田は二塁手に配置転換された安達と併用されている．離脱ではないものの，類似したタイプの選手の調子を見て同時に使っていた．

このように，現実のケースと比較することで，類似している選手が代替えに使われているように思われる．

5.4 トレード

逆に，葉と葉の距離が遠い選手同士は，トレード要員になれる可能性がある．異なるタイプの選手の獲得について，この結果を解釈すると，図 7 の中田翔に注目できる．中田は 2021 年のシーズン中，日本ハムから巨人へ無償トレードで移籍した．もし彼が有償だった場合，どの選手とのトレードが相応しかったのか．もちろん，お互いの球団が違うタイプの選手を求めていたという仮定の話である．異なるタイプという意味で巨人の選手で葉の距離が遠い丸や梶谷，坂本が挙げられる．

このように，トレード要員を決定することができる可能性もある．

6. 結論

本研究の目的は，株式会社 DELTA が提供する全ての指標を使って，野球選手をグループ化する一つの枠組みを提示することであった．2020 年の公式戦に出場した野手 327 人を WAR が 1.0 以上の選手に絞り込み，115 の打撃指標を 8 つの項目別で分けて分析を施した．その結果，主成分分析と混合分布モデルによるクラスター分析を使用して，その枠組みを提示することができた．

また，クラスター分析によって同じグループに分類された選手同士は，お互い代替選手となれる可能性について示唆することができた．逆に，クラスター分析で違うグループに分類されている選手同士は，お互いトレード要員になれる可能性についても示唆できた．今後の課題として，野手だけでなく投手を分類する枠組みを提示していきたい．また，守備評価・位置に関する指標を加えて分析していきたいと考えている．

参考文献

[1] データスタジアム株式会社 (2015) 「野球×統計は最強のバッテリーである」 pp. 26-49. 中央公論新社
[2] 大野 高裕 (2000) 「多変量解析入門」 pp. 1-108. 同友館
[3] 宮本 定明 (1999) 「クラスター分析入門 -ファジィクラスタリングの理論と応用-」 pp. 13-105. 森北出版株式会社
[4] 田中 成典, 鳴尾 丈司, 山本 雄平, 西藤 怜, “スイング計測装置を用いた大学野球選手の特性分析に関する研究”, 日本機械学会論文集, Vol.87, No.894, pp. 1-15, 2021
[5] 松井 秀俊, 小泉和之 (2019) 「統計モデルと推測」 pp. 169-191. 講談社
[6] 酒折 文武, 圓城寺 啓人, 竹森 悠渡, 西塚 真太郎, 保科 架風, “野球のトラッキングデータに基づいた肘内側側副靭帯損傷の要因解析”, 統計数理, Vol.65, No.2, pp. 201-215, 2017
[7] 蔭山 雅洋, 田中 成典, 山本 雄平, 鳴尾 丈司, “野球のスイング計測装置を用いた 9 分割コース別のスイング特性の分析”, 日本機械学会論文集, Vol.87, No.902, pp. 1-20, 2021
[8] C Soto-Valero. A Gaussian mixture clustering model for characterizing football players using the EA Sports' FIFA video game system. RICYDE. Revista Internacional de Ciencias del Deporte. Vol.13, No.49, pp. 244-259, 2017

[9] 西内 啓 (2012) 「遠藤保仁がいればチームの勝ち点は 117%になる」 pp. 74-122. ソフトバンク新書
[10] 金 明啓 (2007) 「R によるデータサイエンス」 pp. 271-275. 森北出版株式会社
[11] Fan Graphs, “What is WAR?” https://library.fangraphs.com/misc/war/, (参照 2022-06-24)