

統計数理

第65巻 第2号

(通巻126号)

PROCEEDINGS OF THE INSTITUTE OF STATISTICAL MATHEMATICS

目次

特集「スポーツ統計科学の新たな挑戦」

「特集 スポーツ統計科学の新たな挑戦」について		
田村 義保・酒折 文武	183
ストレートに着目した空振りに影響を与える要因の定量的分析 [原著論文]		
永田 大貴・南 美穂子	185
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 [原著論文]		
酒折 文武・圓城寺 啓人・竹森 悠渡・西塚 真太郎・保科 架風	201
Covariate Balancing Propensity Scoreを用いた、スクイズ作戦の有効性の解析 [原著論文]		
中村 知繁・南 美穂子	217
項目反応理論を用いた野球選手の能力評価指標の提案 [原著論文]		
阿部 興・作村 建紀・鎌倉 稔成	235
バレーボール各国代表チームのレーティング手法の提案および 結果予測・大会形式評価への応用 [原著論文]		
小中 英嗣	251
重力モデルを用いたサッカー選手の動きの定量化 [原著論文]		
土田 潤・宿久 洋	271
トラッキングデータを用いたサッカーの試合における戦況変化の抽出 [研究ノート]		
神谷 啓太・中西 航・泉 裕一郎	287
ドローン分割と階層的クラスタリングを用いた集団スポーツにおける フォーメーション解析手法の提案 [研究ノート]		
成塚 拓真・山崎 義弘	299
サッカーの攻撃におけるプレーの最適化アルゴリズムの開発 [研究ノート]		
徐 広孝・大澤 啓亮・見汐 翔太・安藤 梢・鈴木 宏哉・西嶋 尚彦	309

整数値自己回帰モデルの最近の発展 [研究詳解]		
中嶋 雅彦・酒折 文武・川崎 能典	323

2017年12月

大学共同利用機関法人 情報・システム研究機構 統計数理研究所

〒190-8562 東京都立川市緑町10-3 電話 050-5533-8500(代)

本号の内容はすべて <http://www.ism.ac.jp/editsec/toukei/> からダウンロードできます

ISSN 0912-6112

統
計
数
理

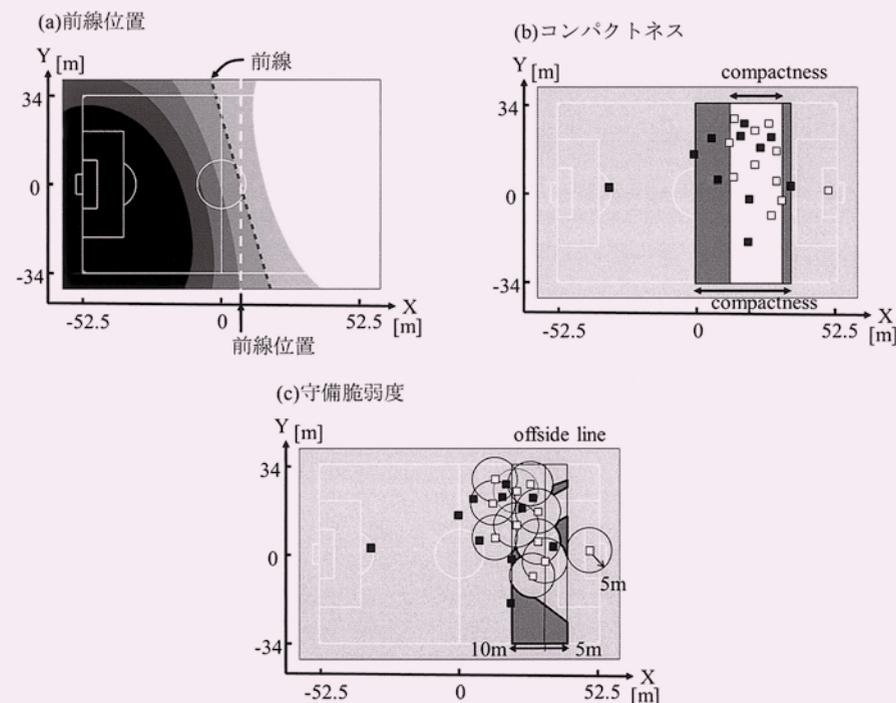
PROCEEDINGS OF THE INSTITUTE OF STATISTICAL MATHEMATICS

第65巻 第2号 2017

統計数理

Vol. 65, No.2

PROCEEDINGS OF THE INSTITUTE OF STATISTICAL MATHEMATICS



統計数理研究所

統計数理

(年2回発行)

編集委員長 加藤 昇吾

編集委員 足立 淳

小山 慎介

武田 朗子

野間 久史

南 和宏

特集担当編集委員 酒折 文武 (中央大学)

田村 義保

編集室

池田 広樹

長嶋 昭子

脇地 直子

渡邊 百合子

「統計数理」は、統計数理研究所における研究成果を掲載する統計数理研究所「彙報」として1953年に歴史を始め、1985年に誌名を変更し今の形となりました。現在は、統計数理研究所の研究活動に限らず、広く統計科学に関する投稿論文を掲載し、統計科学の深化と発展、そして統計科学を通じた社会への貢献を目指しています。

投稿を受け付けるのは、次の6種です。

- a. 原著論文
- b. 総合報告
- c. 研究ノート
- d. 研究詳解
- e. 統計ソフトウェア
- f. 研究資料

投稿された原稿は、編集委員会が選定・依頼した査読者の審査を経て、掲載の可否を決定します。投稿規程、執筆要項は、本誌最終頁をご参照ください。

また、上記以外にも統計科学に関して編集委員会が重要と認める内容について、編集委員会が原稿作成を依頼することがあります。

その他、「統計数理」に関するお問い合わせは、各編集委員にお願いします。

All communications relating to this publication should be addressed to associate editors of the Proceedings.

大学共同利用機関法人 情報・システム研究機構

統計数理研究所

〒190-8562 東京都立川市緑町10-3 電話050-5533-8500(代)

<http://www.ism.ac.jp/>

© The Institute of Statistical Mathematics 2017

印刷：笹氣出版印刷株式会社

PROCEEDINGS OF THE INSTITUTE OF STATISTICAL MATHEMATICS

Vol. 65, No. 2

Contents

Special Topic : New Challenges to Statistical Science in Sports

On the Special Topic “New Challenges to Statistical Science in Sports”

Yoshiyasu TAMURA and Fumitake SAKAORI 183

Factors Affecting Batters’ Contact with a Four-seam Fastball

Daiki NAGATA and Mihoko MINAMI 185

A Statistical Analysis of Medial Collateral Ligament Injury Using Baseball Tracking Data in MLB

Fumitake SAKAORI, Hiroto ENJOJI, Yuto TAKEMORI, Shintaro NISHIZUKA and Ibuki HOSHINA 201

Effectiveness of the Squeeze Play Using Covariate Balancing Propensity Scores

Tomoshige NAKAMURA and Mihoko MINAMI 217

Measurements of Baseball Players’ Batting Abilities

Ko ABE, Takenori SAKUMURA and Toshinari KAMAKURA 235

Statistical Rating Method for Volleyball National Teams to Predict Results and Determine Competition Format Design

Eiji KONAKA 251

Quantitative Evaluation of Soccer Players’ Movements

Jun TSUCHIDA and Hiroshi YADOHISA 271

Tracking Data to Extract Changes in Football Game Situation

Keita KAMIYA, Wataru NAKANISHI and Yuichiro IZUMI 287

Characterization of the Formation Structure in Team Sports

Takuma NARIZUKA and Yoshihiro YAMAZAKI 299

Development of Optimization Algorithm for Attack Play in Football

Hiroataka JO, Keisuke OOSAWA, Syota MISHIO, Kozue ANDO, Koya SUZUKI and Takahiko NISHIJIMA 309

Research Review

Recent Development of Integer-valued Autoregressive Models

Masahiko NAKAJIMA, Fumitake SAKAORI and Yoshinori KAWASAKI 323

December, 2017

Research Organization of Information and Systems

The Institute of Statistical Mathematics

10-3 Midori-cho, Tachikawa, Tokyo 190-8562, JAPAN

表紙の図は本誌 292 ページを参照

「特集 スポーツ統計科学の新たな挑戦」について

田村 義保¹・酒折 文武² (オーガナイザー)

ICT, GPS, センサー技術などの発達により, スポーツの世界においても選手やボールのトラッキングデータをはじめとして, これまで以上に大規模かつ複雑なデータが収集されるようになった. そして, 以前から活用されてきたボックススコアやスタッツ, あるいはプレイ・バイ・プレイの行動履歴データ, あるいは試合の動画等と合わせて, チーム編成や戦略評価, さらにリアルタイムな戦術判断へのさらなる活用が待たれている. 2019年のラグビーワールドカップ日本開催や2020年東京オリンピックなどビッグイベントを控え, 2014年には日本スポーツアナリスト協会が設立されるなど, その機運も高まっている.

欧米諸国では, 野球, サッカーやアメリカンフットボール, バスケットボールなどを始めとしたメジャーなスポーツにおいて先進的にデータ収集や高度な統計分析が行われ, 科学的な見地から戦術解析や選手評価がなされてきた. また, ASAでは1992年より毎回スポーツ統計のセッション(SIS, Statistics in Sports)が生まれ, 数々のシンポジウム開催やJQAS(Journal of Quantitative Analysis in Sports)とJSA(Journal of Sports Analytics)の発行, さらにJASA等の論文誌への論文投稿も多く見られるなど, 学術的な場においてもこうした研究は盛んに行われており, 統計学の専門家による積極的な関与が見られる.

一方, 日本では, スポーツの場における統計学の活用は必ずしも十分であるとはいえなかった. このような状況を鑑みて, 2009年から日本統計学会でスポーツ統計分科会の活動を開始した. さらに, 2011年からは統計数理研究所において共同研究を開始するとともに, 裾野の拡大とこの分野の研究推進を目指して「スポーツデータ解析コンペティション」を開催してきた. コンペティションでは, 日本プロ野球・アメリカメジャーリーグ・Jリーグ・Bリーグのデータについてデータスタジアム株式会社から提供を受け, 中央大学や立教大学の協力を得て, 大学院生・大学生を含むコンペ参加者へのデータ提供, 発表の審査, 優秀者発表会を行っている. 成果の一部は統計数理研究所共同研究リポート「スポーツデータ解析における理論と事例に関する研究集会」として発行してきている.

このコンペティションは, 大学院生や大学生のデータ解析経験の場の提供という意味で, データサイエンス教育・データサイエンティスト育成の役割も担っている. スポーツを含むあらゆるビジネス領域におけるデータサイエンティストの必要性とその不足を受け, 2016年12月に文部科学省は「数理及びデータサイエンスに係る教育強化」の拠点校として, 北海道大学, 東京大学, 滋賀大学, 京都大学, 大阪大学, 九州大学の6校を選定している. また, 平成29年度「データ関連人材育成プログラム」の取組機関として2017年8月に東京医科歯科大学, 電気通信大学, 大阪大学, 早稲田大学を代表機関とする4組織を選定している. さらに, 2017年4月からは滋賀大学データサイエンス学部で, 大学としては日本で初めてのデータサイエンティスト教育が始まっている. 今後は, こうした機関におけるデータサイエンス教育とも協力し,

¹ 統計数理研究所: 〒190-8562 東京都立川市緑町10-3

² 中央大学 理工学部: 〒112-8551 東京都文京区春日1-13-27

コンペティションのさらなる活性化が期待される。

コンペティションを含む活動を通じて、スポーツ統計科学としての発展と、スポーツ業界やスポーツ現場への貢献や交流も進んできた。本特集は、こうしたスポーツ統計科学の現状や今後を広く知らせるために計画した。

本特集には、野球、バレーボール、サッカーと 3 つの競技に関わる計 9 編の論文が掲載されている。6 編が原著論文、3 編が研究ノートである。これらの論文の一部はスポーツデータ解析コンペティションの成果を論文としてまとめたものである。4 編は野球に関する論文であり、永田論文、酒折論文、中村論文はアメリカのメジャーリーグ (MLB) に関する分析、阿部論文は日本プロ野球に関する分析である。同じ野球ではあるが、永田論文と酒折論文は PITCHf/x と呼ばれる投球のトラッキングデータに基づくものであるのに対し、中村論文と阿部論文ではプレイ・バイ・プレイのデータを用いている。また、小中論文はバレーボールの試合結果からのレーティングに関する研究である。そして 4 編がサッカーに関する論文である。これらは J リーグのトラッキングデータやボールに関係するプレイ・バイ・プレイのデータを用いている。なお、これらのデータのうち、メジャーリーグの PITCHf/x データとバレーボールの試合結果のデータはインターネットを通じて自由に入手可能である。また、野球やサッカー、そしてそれ以外の競技においても、ある程度集約されたデータであれば多くのものは入手可能である。スポーツに興味のある統計学研究者のかたがたのさらなる参入も期待したいと思っている。

最後に、この特集「スポーツ統計科学の新たな挑戦」の査読者の方々、並びに編集担当の方々、本特集でもデータ活用を許諾いただけたデータスタジアム株式会社に、この場をお借りして感謝を申し上げたい。

ストレートに着目した空振りに影響を与える 要因の定量的分析

永田 大貴¹・南 美穂子²

(受付 2016 年 12 月 28 日；改訂 2017 年 5 月 17 日；採択 5 月 24 日)

要 旨

PITCHf/x は投球の軌道を追尾することによってボールの座標や変化量などのデータを計測できるシステムである。本稿では、PITCHf/x データを用いてノビについて分析を行った。ノビとは空振りしやすいストレートに対して用いられる言葉であり、ノビのあるストレートは初速と終速の差が小さいという定説がある。しかし実際の PITCHf/x データを眺めると定説とは逆の関係が見て取れる。そこで打者のボールへのコンタクトを定義した上で、コンタクトを球速差で説明するロジスティック回帰モデルを適用した。それにより、球速差はコンタクトに対して負の関係性を有するという結果が得られた。また本稿では、ボールの変化量に着目し、変化量とコンタクトとの関係性を評価するために多変量スプライン平滑法を用いた一般化加法モデルによる分析を行い、縦変化量の大きさが重要である事が分かった。さらに、ボールの質以外の各投手ごとの打ちにくさを変量効果として追加したモデルについても解析を行い、その予測値を比較する事により上原は MLB(メジャーリーグベースボール)2014 シーズンにおいて最も打ちづらい特徴を有した投手であるという結果を得た。

キーワード：PITCHf/x データ，ストレート，ノビ，ボールの変化量，一般化加法モデル，変量効果。

1. はじめに

近年、スポーツにおいてデータ活用による戦術分析や選手のパフォーマンス向上を図ろうという動きが日本でも活発化してきている。野球やサッカーを始め、国内においてはバレーボール・ラグビーなどを中心に戦術的または要因的分析を行うことにより、勝利に焦点を置いたデータ活用が行われている。野球において統計学的な見地から分析を行い、選手の評価や戦略を考える分析手法であるセイバーメトリクスはまさにその代表格と言えるであろう。野球においては様々な価値基準や選手の能力を示す指標が存在するが、セイバーメトリクスではこれらの重要性を数値から客観的に分析し、それによってプレー戦術に対し統計学的根拠を与えた。そしてスポーツアナリティクスの分野において、今最も注目を集めているのがトラッキングシステムによって得られたデータを用いた解析である。トラッキングシステムとは主に野球やサッカーなどの球技において、選手個別の動作やボールの軌跡を追跡・記録・分析するためのシステムであり、それによって取得されたデータのことをトラッキングデータと呼ぶ。本研究

¹ 慶應義塾大学大学院 理工学研究科：〒223-8522 神奈川県横浜市港北区日吉 3-14-1

² 慶應義塾大学 理工学部：〒223-8522 神奈川県横浜市港北区日吉 3-14-1

では、野球における投球に対するトラッキングシステムである PITCHf/x により取得されるデータに着目し、ストレートのノビについて解析を行った。

1.1 PITCHf/x データ

PITCHf/x は米国 SPORTVISION 社によって開発されたシステムで、球場に設置した複数台のカメラの映像を基にして投球におけるボールの座標や軌道、速度や変化量など様々な情報を自動的に取得する。メジャーリーグベースボール (MLB) においてはこのシステムが全 30 スタジアムに設置されており、チーム内での分析やトレード、またファン向けのコンテンツとしても活用されている。PITCHf/x システムによって取得できるデータは以下のようにまとめられる。

- 座標に関するデータ：リリース点、プレート到達点
- 速度に関するデータ：初速、終速、加速度
- 変化に関するデータ：総回転数、変化量、球種

PITCHf/x データの特徴として、3次元の座標軸と原点を定めボールの位置を計測することがあげられる。ホームプレートを原点とし、 x 軸を水平方向 (サード方向を負、ファースト方向を正)、 y 軸を前後方向 (投手方向であれば正)、 z 軸を垂直方向と各軸をフィート単位で定めている。また、投球のリリース時におけるボールの速度 (初速) と、プレート到達時における速度 (終速)、さらにリリースからプレートまでの平均的な加速度がデータとして得られる。変化量は、ボールに回転がないという仮定のもとで到達する点と実際の (主に回転などによって引き起こされた変化による) 到達点との偏差としている。ただし、ここでのプレート到達点は原点から 1.417 フィート離れた x - z 平面、リリース到達点は原点から 50 フィート離れた x - z 平面における座標点の近似値である。

今回分析に用いたデータベース内の変数 pfx は重力等の加速度を含めた軌道偏差と定義されている。変化量の計算を以下に示す (Kagan, 2009)。投手から投げられたボールの到達点は

$$(1.1) \quad \mathbf{x}_t = \mathbf{x}_0 + t\mathbf{v}_0 + \frac{t^2}{2}\mathbf{a}$$

を用いて計算される。ここではボールの動きに対して等加速 (減速) 度運動を仮定している。ここで、

$$(1.2) \quad \mathbf{x}_t = \begin{pmatrix} \text{時刻 } t \text{ におけるボールの } x \text{ 座標} \\ \text{〃} & y \text{ 座標} \\ \text{〃} & z \text{ 座標} \end{pmatrix}$$

であり、 \mathbf{x}_0 はリリース点の座標、 \mathbf{v}_0 はリリース時における速度ベクトル、 \mathbf{a} はリリースからプレート到達時までの平均的な加速度ベクトルである。つまり、プレート到達時刻 t^* における回転のない場合の予測到達点 \mathbf{x}_{t^*} は

$$(1.3) \quad \mathbf{x}_{t^*} = \mathbf{x}_0 + t^*\mathbf{v}_0 + \frac{t^{*2}}{2}\mathbf{a}$$

と表される。しかし、ボールに対して回転などの変化が加えられるため、予測到達点 \mathbf{x}_{t^*} と実際のプレート到達点は異なる。したがって、変化量は

$$(1.4) \quad \begin{pmatrix} x \text{ 方向の変化量} \\ 0 \\ z \text{ 方向の変化量} \end{pmatrix} = \begin{pmatrix} \text{プレート到達時のボールの } x \text{ 座標} \\ 1.417 \\ \text{プレート到達時のボールの } z \text{ 座標} \end{pmatrix} - \mathbf{x}_{t^*}$$

と計算される。ここで、 x 方向の変化量を横変化量、 z 方向の変化量を縦変化量と呼ぶことにする。本稿では MLB の公式オンラインサイト Gameday から 2014 年レギュラーシーズンの $PITCHf/x$ データを取得し、解析に用いた。

本稿では、ストレートの変化量や各投手の打ちづらさなどの要因が空振りに与えている影響を定量的に分析する。空振りしやすいストレートに対してはノビという言葉が用いられるが、ここではコンタクトに対して影響を与えている要因を探ることでノビについて議論を行った。まず第 2 節ではコンタクトを定義した上で、「初速と終速の差が小さいストレートがノビのあるボールである」という定説に着目し、球速差を説明変数としたロジスティック回帰分析を行い定説について議論する。第 3 節では球速差ではなく変化量に着目した分析の必要性について主張し、第 4 節では解析に用いたスプライン法による一般化加法モデリング手法について紹介した上で解析を行い、その結果に対する考察を行う。第 5 節は計測されるデータでは記述できない要因として各投手の打ちづらさを考え、変量効果としてモデルに取り入れた解析を行い各投手の打ちづらさを評価した。第 6 節ではまとめと今後の課題について述べる。

2. 球速差に着目したノビの定説の検証

本稿ではコンタクトに着目した解析を行う。日本におけるストレートは MLB ではフォーシームファストボールという名称であり、 $PITCHf/x$ データにおいては各データから球種が自動判別され記録されている。 $PITCHf/x$ データに基づいて判別された球種ラベルが FF (フォーシームファストボール) のみを対象に分析を行うこととする。

ここでコンタクトとは、打者が投球に対してバットを振りに行って当てられたかどうかを示すものである。ボールに対するコンタクトを表 1 のように定める。

打者が投球に対してバットを振りにいって空振りした時を非コンタクト、凡打・ファウル・ヒットなどボールをバットに当てることができた時をコンタクトとする。ストレートに対するコンタクトを考えることは、空振りを考える事と等しい。打者がバットに当てることが難しいストレートを投じることができるとは投手にとって最大の強みともいえる。ここでは、コンタクトしにくいストレートとはどのような特徴を持つボールなのかを定量的に明らかにしたい。

2.1 ノビの定説と日本人投手の比較

ノビとは空振りしやすいストレートに対して用いられる言葉であり、初速と終速の差が小さいストレートがノビのあるボールであるという定説が存在する。しかし、 $PITCHf/x$ システムにより観測されるデータからは、定説とは逆の関係が見て取れる(金沢, 2015)。

各投球におけるボールの球速差を、球速差 = 初速 - 終速 と定める。ただし、単位はマイル/時である。また、球速差を初速で除したものを減速率とする。コンタクト率を以下のように定める。

$$(2.1) \quad (\text{コンタクト率}) = \frac{(\text{コンタクト数})}{(\text{コンタクト数}) + (\text{非コンタクト数})}$$

表 1. コンタクトの定義.

投球結果	変数の定義
空振りストライク	非コンタクト
凡打, ファウル, エラー出塁, ヒット	コンタクト

表 2. 日本人投手のストレートにおける速度に関する特徴量とコンタクト率.

投手名	初速	球速差	減速率	コンタクト率
田澤	93.9	7.86	8.4%	84%
ダルビッシュ	92.6	7.20	8.3%	86%
田中	91.3	7.15	7.8%	88%
黒田	90.9	7.52	8.3%	86%
藤川	90.6	7.78	8.6%	68%
岩隈	89.6	7.60	8.5%	81%
和田	88.9	7.08	8.0%	82%
上原	88.0	7.76	8.8%	67%

表 3. 球速差を説明変数とするロジスティック回帰モデルの推定結果.

パラメータ	推定値	標準誤差	z 値	p 値
α	3.3	0.076	43.07	$< 2 \times 10^{-16}$
β	-0.21	0.010	-21.19	$< 2 \times 10^{-16}$

PITCHf/x データから実際の投手の球速差とコンタクトとの関係を確認したい. 表 2 は日本人投手の MLB2014 シーズンにおける初速, 球速差の標本平均とそれを用いて計算した減速率, およびコンタクト率の表である.

これらの投手の中で, 上原や藤川は一般的にノビのあると言われている部類の投手であり, 実際にストレートのコンタクト率が他の投手に比べ極めて小さいことがわかる. しかし, 両投手の球速差を見てみると他の投手や MLB 平均(7.40 マイル)と比べ決して小さいとは言えないどころか, むしろ大きい傾向にある. また, 初速に対してどれほど減速したかを減速率として表しているが, 減速率が最大となったのは上原であった. これは, これまで考えられてきた球速差の小さいストレートがノビのあるストレートであるという定説とは正反対の事実をデータが示していることになる.

2.2 球速差とコンタクトの解析

ここで, コンタクトと球速差の関係を解析するため以下のロジスティック回帰モデルを用いた解析を考える. 目的変数 Y_i を i 番目の投球に対して打者がボールに対してコンタクトできたかを表す二値変数とする. MLB2014 年シーズンにおける投手が投じたストレートに対して投球結果がコンタクト・非コンタクトに該当するものを対象データとして解析を行った. モデル式と推定結果を以下に示す.

$$(2.2) \quad \log \frac{p_i}{1-p_i} = \alpha + \beta(\text{球速差})_i, Y_i \sim \text{Bernoulli}(p_i).$$

表 3 はモデル(2.2)における回帰係数パラメータの推定値, 標準誤差, z 値, p 値をまとめたものである.

球速差の回帰係数推定値 $\hat{\beta}$ は負であり p 値が十分に小さく有意であるという結果が得られた. この結果をそのまま解釈すると, 球速差が大きいストレートほど空振りが取りやすいということは否定できないという結論が与えられる. 次に球速差ではなく, 変化量に着目した解析を行う.

3. 変化量に着目した分析

3.1 変化量に着目した分析の必要性

先ほどのロジスティック回帰モデルの解析から、初速と終速の差が大きいストレートは空振りしやすいボールであるということを否定できない結果が得られた。以下の表4は対象データにおけるボールの減速率の平均値・中央値・第1四分位数・第3四分位数・標準誤差をまとめたものである。これによると減速率の平均値8.0%ほどで、第1四分位数・第3四分位数は平均から0.6%ほど差がないことがわかる。つまりストレートに関してはボールの減速率はそこまで大きな違いはなく、この差が打者のコンタクトに対して大きく影響しているとは考えづらい。

以上の理由から、ここでは球速差ではなく球速差とトレードオフの関係にあるボールの変化量に着目することでコンタクトを説明することを考える。変化量や球速差はボールの回転数と回転軸によって決まり互いに関係しあっているため、実際にコンタクトに対して影響を有している変数は変化量であると考え、ボールの変化量に着目しコンタクトとの関係性を分析する。

3.2 各投手の変化量の比較

日本人6投手のストレートの変化量を図1に示した。図1は縦・横の変化量を2平面にプロットした図である。各軸はインチ単位であり原点から離れた位置にある点は大きく変化している事になる。横変化量が負の値を取っている場合にはサード方向にボールが変化しており右投手であればシュートしていることになる。また、基本的にストレートの縦変化量は正の値

表4. 減速率の平均値・中央値・第1四分位数・第3四分位数・標準誤差(パーセント単位).

第1四分位数	中央値	平均値	第3四分位数	標準誤差
7.486	8.051	8.058	8.621	0.851

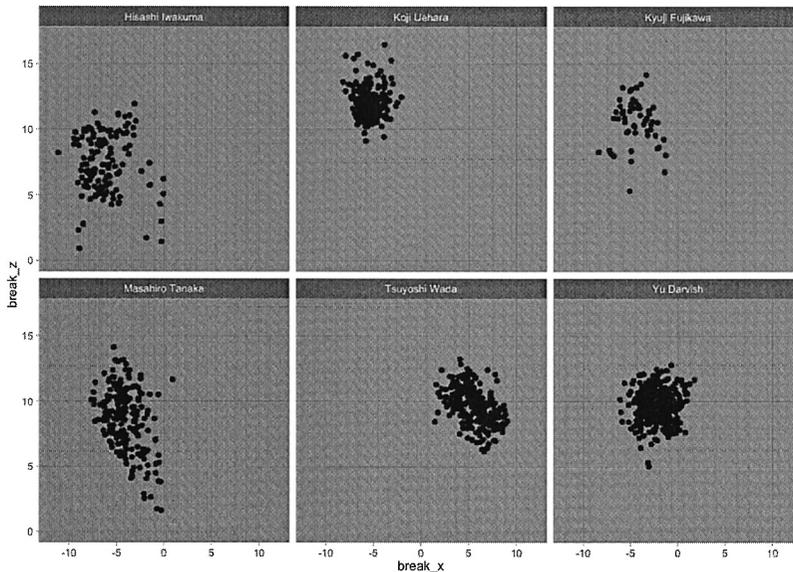


図1. 日本人投手のストレートの変化量.

をとる。これはボールに対してバックスピがかかるとによりボールに対して揚力が働くためである。

図 1 を見ると変化量は各投手それぞれに特徴を有することがわかる。上原や藤川は比較的縦変化量が大きく、一方で岩隈は横変化量が大きい。また、この中で唯一左投手である和田は、全く異なる変化量(特に横変化量)を有する。ダルビッシュや田中は比較的平均的な変化量である。

図 1 において上原・藤川などのノビのあると言われていた投手の変化量をその他の投手と比較すると縦変化量が大きく、また実際のコンタクト率(表 2)も小さいことが分かった。そこで、PITCHf/x データにおける変化量に着目してコンタクトとの関係性を明らかにしたい。変化量は縦・横の二方向に対してデータが得られ、各変化量はコンタクトに対して単調な線形関係で影響するものではない(変化量の僅かな差がコンタクトに対して大きく影響を及ぼす可能性がある)と考えられる。また、これらとコンタクトの関係を適切に評価するには 2 変量間の交互作用を柔軟にモデリングを行う必要があるため、ここでは多変量間と目的変数との関係を柔軟にモデリングすることができる一般化加法モデルによる解析を行う。

4. スプライン平滑法を用いた解析と解析結果の考察

ロジスティック回帰モデルにおいて、変化量などの変数に対してスプライン関数 f を適用したモデルを考える。スプライン関数は局所的な特徴を捉えることを可能にする多数の基底関数の線形和で表されるなめらかな関数であり、変化量とコンタクトとの関係を柔軟にモデリングすることを可能にする。4.1 節では解析に用いたスプライン法による一般化加法モデリング手法について概要を示す。

4.1 平滑化関数を用いた一般化加法モデルによるモデリング

一般化加法モデルとは、一般化線形モデルの線形予測子に非線形関数を含むように拡張したものである(Hastie and Tibshirani, 1986)。ロジスティック回帰モデルは一般化線形モデルに含まれるモデルであり、ここでは、線形予測子に変化量などのスプライン関数を含むロジスティック回帰モデルを用いている。Thin plate regression spline 法は Wood (2006) によって提案された平滑化手法で、自然 3 次スプライン法、thin plate spline 法の柔軟性を保持しつつ計算量を抑えるように工夫されている。

ここでは、説明変数に対する非線形な関数として解析に用いた自然 3 次スプライン法とそれを多変量に拡張した thin plate spline 法についての表現とパラメータの推定方法について示す。ただし自然 3 次スプライン法は thin plate spline 法の単変量の場合を指す手法である。

まず、目的変数の平均構造を 1 次元の説明変数 x を関数 f で表すモデル

$$(4.1) \quad y_i = f(x_i) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2),$$

$$f(x) = \sum_{j=1}^q \beta_j b_j(x)$$

を用いて自然 3 次スプライン法について説明する。ここで、 ϵ_i は互いに独立な正規誤差であり、 $b_j(x)$ はパラメータを含まない基底関数である。

関数 f は $\beta = (\beta_1, \dots, \beta_q)^T$ の線形な関数として表されているので、目的変数ベクトルを $\mathbf{y} = (y_1, \dots, y_n)^T$ とした時に

$$(4.2) \quad \mathbf{y} = X\beta + \epsilon$$

と表現できる。ここで、 X の i 行 j 列成分を $X_{(ij)}$ とした時に、 $X_{(ij)} = b_j(x_i)$ である。

3 次スプライン関数は、3 次多項式を 2 階微分までが連続であるようにつなぎ合わせたものであり、各区間において 3 次多項式のつなぎ目を節点 (knot) という。節点の数を $q-2$ 個とし、節点を $x_1^* < x_2^* < \dots < x_{q-2}^*$ とする。

3 次スプライン関数の基底関数の表現としては様々なものがあるが、例えば Wood (2006) や Gu (2002) で詳細が示されているような次の表現がある。 $b_1(x) = 1, b_2(x) = x, b_{j+2}(x) = R(x, x_j^*)$ であり、 $R(x, z)$ は以下のように表される ($j = 1, 2, \dots, q-2$)。

$$(4.3) \quad R(x, z) = [(z-1/2)^2 - 1/12] [(x-1/2)^2 - 1/12] / 4 \\ - [(|x-z| - 1/2)^4 - 1/2(|x-z| - 1/2)^2 + 7/240] / 24.$$

3 次スプライン関数に対して以下の端点での 2 次微分がゼロという制約

$$(4.4) \quad f''(x_1^*) = 0, f''(x_{q-2}^*) = 0$$

を付け加えたものが自然 3 次スプライン関数である。

関数の柔軟性は節点 (基底) の数によって変化するため、節点を多くすると柔軟な関数を表現できる一方、最小二乗法による推定ではデータに当てはまりすぎて複雑な関数を選んでしまう。そこですべてのデータ点を節点とすることによって十分な柔軟性を保ちつつ、当てはまりすぎを抑えるために関数の複雑さに対して罰則を与えることで関数のなめらかさを制御することとする。つまり、罰則付き二乗誤差、

$$(4.5) \quad V(\beta) = \|\mathbf{y} - X\beta\|^2 + \lambda \int_{\Omega} f''(x)^2 dx$$

の最小化によってパラメータ β の推定を行う。ここで、 $\lambda (> 0)$ は平滑化パラメータであり、 Ω は関数を定義する空間とする。 $\int_{\Omega} f''(x)^2 dx$ は関数の複雑さを表しており、 λ は複雑さに対する罰則を調整するパラメータである。

関数 f は基底関数で $f(x) = \sum_j \beta_j b_j(x)$ と表されるので、罰則項は β の 2 次形式であり、罰則付き誤差二乗和はある半正定値行列 S を用いて、

$$(4.6) \quad V(\beta) = \|\mathbf{y} - X\beta\|^2 + \beta^T S \beta$$

と表すことができる。ここで先ほどの基底関数に対しては行列 S の各成分は $S_{(i+2, j+2)} = R(x_i^*, x_j^*)$ と表される ($i, j = 1, 2, \dots, q-2$)。 $V(\beta)$ を β について最小化を行うことで推定値

$$(4.7) \quad \hat{\beta} = (X^T X + \lambda S)^{-1} X^T \mathbf{y}$$

を得る。

次に適切な平滑化パラメータ λ の値を選択することが必要である。 λ の値を大きくとれば、推定における罰則を重くするため比較的直線に近づき、 λ の値を小さくとれば推定結果は複雑な曲線となる。平滑化パラメータの選択については一般化交差検証法 (GCV; Wood, 2008)、制約付き最尤法 (REML; Wood, 2011) を用いた選択などがある。本研究では一般化交差検証法を用いて選択を行った。

次に、自然 3 次スプラインと同様な考え方に基づいた多変量平滑法である thin plate spline について、ここでは簡単のため 2 変量の場合に限定して述べる。実際のモデリングでは、変換量など 2 次元の変数のコンタクトに与える影響を表すことを考えている。いま 2 次元の説明変数ベクトルを $\mathbf{x} = (x_1, x_2)^T$ とし、観測されたデータを $(x_i, y_i), i = 1, 2, \dots, n$ とする。ここで、モデル

$$(4.8) \quad y_i = f(\mathbf{x}_i) + \epsilon_i$$

を考える. f は x_1, x_2 について 2 階微分まで連続な関数とする. このとき, f の推定における罰則項 $J(f)$ を

$$(4.9) \quad J(f) = \int \int \left(\frac{\partial^2 f}{\partial x_1^2} + 2 \frac{\partial^2 f}{\partial x_1 \partial x_2} + \frac{\partial^2 f}{\partial x_2^2} \right) dx_1 dx_2$$

と定義すると, これを最小にする関数は, $\eta(r) = r^2 \log(r)/(8\pi)$ としたとき

$$(4.10) \quad f(\mathbf{x}) = \alpha_1 + \alpha_2 x_1 + \alpha_3 x_2 + \sum_{i=1}^n \delta_i \eta(\|\mathbf{x} - \mathbf{x}_i\|)$$

と表せる (Wood, 2006; Green and Silverman, 1994). また, 改めて半正定値行列 E の成分を $E_{(ij)} = \eta(\|\mathbf{x}_i - \mathbf{x}_j\|)$, $T_i = (1, x_{1i}, x_{2i})$, $T = (T_1, T_2, \dots, T_n)^T$, $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3)^T$, $\boldsymbol{\delta} = (\delta_1, \delta_2, \dots, \delta_n)^T$ と定める. 制約 $T^T \boldsymbol{\delta} = \mathbf{0}$ を満たすとき, f を thin plate spline と呼び, 当てはめは罰則付き誤差二乗和,

$$(4.11) \quad S(\boldsymbol{\alpha}, \boldsymbol{\delta}) = \|\mathbf{y} - E\boldsymbol{\delta} - T\boldsymbol{\alpha}\|^2 + \lambda \boldsymbol{\delta}^T E \boldsymbol{\delta}$$

の $T^T \boldsymbol{\delta} = \mathbf{0}$ という条件のもとでの最小化問題となる. ここで罰則項 $J(f)$ は f の二階の微分 (偏微分) により求められるため罰則は $\boldsymbol{\delta}$ にのみ依存する.

Thin plate spline は, 2 階微分までが連続な関数の中で

$$(4.12) \quad \|\mathbf{y} - g(\mathbf{x})\|^2 + \lambda J(g)$$

を最小にするという点において最良の平滑法であり, また節点や基底関数の選択が不要であるという利点があるが, 計算負荷が高くデータ点が多くなると計算時間が大きな問題となる. Thin plate regression spline (Wood, 2003) は式 (4.11) における行列 E を, 固有値分解により求めた固有値の大きい成分のみで構成されるランク k の行列 E_k に置き換えることにより細かい変動を除去し, 基底の次元を低くして計算量を抑えている.

一般化加法モデルは, 一般化線形モデルの線形予測子にスプライン項を含めるように拡張したものであり, 平均構造はリンク関数 g を用いて,

$$(4.13) \quad g(\mu_i) = X_i \boldsymbol{\theta} + f_1(x_{1i}) + f_{23}(x_{2i}, x_{3i}) + \dots$$

のように表される. スプライン項は, 基底関数の線形和として表すことができるので, 対数尤度関数は一般化線形モデルと同様に表せ, これに罰則項を加えた罰則付き対数尤度関数を最小化することによってモデルの当てはめを行う.

統計解析ソフト R の mgcv パッケージは thin plate regression spline を含む様々な平滑化関数を用いた一般化加法モデルによる解析を行うためのものであり, 本研究ではこのパッケージを用いて thin plate regression spline を用いたロジスティック回帰モデルの当てはめを行った.

4.2 変化量に着目した解析結果の考察

本節では, 変化量とコンタクトの関係に着目し, 変化量に対して柔軟なモデリングを行うために thin plate regression spline を適用した解析を行う.

コンタクトを目的変数とするロジスティック回帰モデルにおいて, 説明変数においては変化量に加えてコンタクトと関係があると思われるボール・ストライクカウントや球速, またプレート到達点やリリース点などの変数に対してはスプライン関数を用いることにする. まず,

変化量とコンタクトとの関係を明らかにしたい。しかし、投手の利き手によって横変化方向が異なってくるため、ここでは右投手を基準として左投手の投じた投球の横変化量の正負を反転させ、解析の対象データとした。また、コンタクトに対しては当然のようにプレート到達点に関係しているため、データの均一性を保つ目的でストライクゾーンに到達したボールのみを対象に分析を行った。ここでストライクゾーンの横幅はホームプレートの幅であり、縦幅としては $PITCHf/x$ によって取得される各投球ごとのストライクゾーンの上限と下限のデータの平均値を用いた。

MLB2014 シーズンにおける全投手が投じたストレートで対象となったものの標本サイズは 90774 であった。目的変数はコンタクト(2 値変数)、説明変数としては、カウント(ボール・ストライクカウント, 12 水準のカテゴリカル変数)、打者の対角フラグ(投手と利き手が異なる場合を 1, 同じ場合を 0 とする 2 値変数)、球速(リリース時の速度で連続変数)、変化量(縦・横の 2 次元連続変数)、プレート到達点(縦・横の 2 次元連続変数)、リリース点(縦・横の 2 次元連続変数)を用いる。球速には自然 3 次スプライン法、変化量、プレート到達点、リリース点には 2 次元 thin plate regression spline を用いたスプライン項としてモデルに含めた。

スプライン関数を用いたロジスティック回帰モデルはコンタクト確率 p_i が以下のように表せるモデルである。

$$(4.14) \quad \log\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + f_1(z_{i1}) + f_{23}(z_{i2}, z_{i3}) + \dots$$

ここで、 $x_{i1}, x_{i2}, \dots, z_{i1}, z_{i2}, z_{i3}, \dots$ は説明変数であり、 f_1, f_{23}, \dots はスプライン関数である。

表 5 は線形項の係数推定値とその t 値, p 値である。カウントの各水準に対する推定値は、カウント 0-0 を基準としたものである。2 ストライクであったときのカウントは有意であり推定値は正であった。これは、追い込まれたカウントにおいて打者は三振したくないという意識により、できる限りボールに対してコンタクトしにいくようなバッティング傾向になることの表れであると解釈できる。また、対角打者フラグの値は正であり、有意である。これは、野球の一般論として投手は対角の打者に対して不利であることと整合性が取れており、コンタクトにおいても、右投手に対しては左打者の方が一定量有利であると解釈できる。

表 6 はスプライン項に対する有効自由度とカイ二乗値, その p 値である。コンタクトに対して非線形な関係を有していると思われる変数に対してはスプライン関数(球速に対しては単変

表 5. スプライン項を含むロジスティック回帰モデルによる推定結果(線形項).

変数	推定値	t 値	p 値
(Intercept)	1.99	72.54	$< 2.0 \times 10^{-16}$
カウント 0-1	-0.020	-0.54	0.58
カウント 0-2	0.166	3.45	5.4×10^{-4}
カウント 1-0	-0.014	-0.36	0.71
カウント 1-1	-0.024	-0.64	0.51
カウント 1-2	0.171	4.16	3.1×10^{-5}
カウント 2-0	0.108	1.92	0.05
カウント 2-1	0.063	1.43	0.15
カウント 2-2	0.296	7.15	8.2×10^{-13}
カウント 3-0	0.412	2.07	0.03
カウント 3-1	0.340	5.43	5.5×10^{-8}
カウント 3-2	0.494	10.76	$< 2.0 \times 10^{-16}$
対角打者フラグ	0.188	8.923	2×10^{-16}

表 6. スプライン項を含むロジスティック回帰モデルによる推定結果(スプライン項).

変数	有効自由度	カイ二乗値	p 値
球速	6.0	353.0	$< 2 \times 10^{-16}$
リリース点 (横)・リリース点 (縦)	17.4	463.4	$< 2 \times 10^{-16}$
横変化量・縦変化量	14.6	650.9	$< 2 \times 10^{-16}$
プレート到達点 (横)・プレート到達点 (縦)	32.2	4529.2	$< 2 \times 10^{-16}$

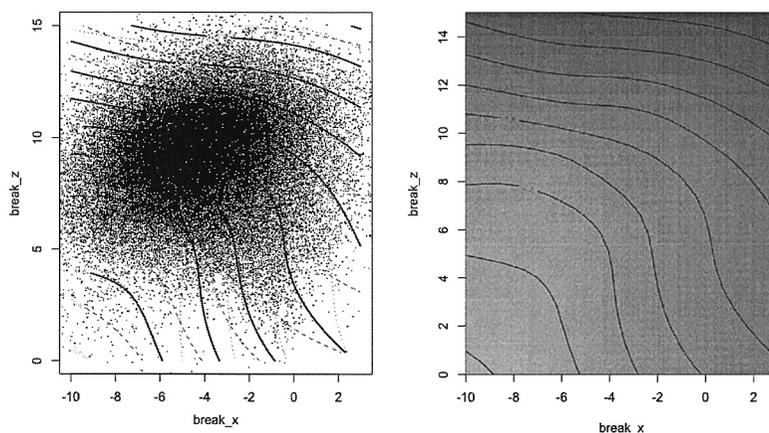


図 2. 変化量のスプライン関数の等高線とサンプル点(左), モノクロ等高線(右).

量の自然 3 次スプライン, リリース点・プレート到達点・変化量に対しては thin plate regression spline) を適用しモデルを構築した. また, スプライン項の各関数における回帰係数に対して検定統計量を構成しカイ二乗検定を行っている (Wood, 2006). 変化量に対する p 値は十分に小さく有意である. つまり, 変化量がコンタクトに対して影響を有していると言える. 他の変数についても p 値は十分に小さく, 有意水準 5% で棄却できるという結果が得られた. また, モデル全体の null deviance と deviance の差(カイ二乗値)は 3636.1 であった. この値はモデルの有効自由度 76.38 のカイ二乗分布に従う. カイ二乗値は自由度に対して十分大きく, このことからモデル全体でも有意であるという結果が得られた.

次にモデルの AIC の比較を行う. ここでは, さきほどの変化量に対してスプライン関数を適用し推定を行ったモデルの AIC と, それぞれの変化量の変数をそのままロジスティック回帰モデルの線形項に当てはめたモデルの AIC を比較した(スプライン項を含む場合の AIC は自由度として有効自由度を用いている). その結果, スプライン項を適用したモデルは, 変化量に対して線形性を仮定したモデルに比べ AIC が 65009.8 から 64943.7 へと減少するという結果が得られた. この結果から変化量に対しては線形なモデルよりも柔軟なスプライン関数を用いたモデリングの方が, AIC の観点からは適切であると判断できる.

図 2 における左図は横変化量と縦変化量の散布図に推定した 2 次元スプライン関数の等高線図を描いたもので, 右図はスプライン関数のモノクロの等高線図である. モノクロ等高線図は, 色が濃いほど対数オッズ/コンタクト率が低い. 図 2 を見ると, 縦変化量が大きい領域 (y 軸の値が 10~15) ではコンタクト確率(対数オッズ)が比較的低く縦変化量の変化に対して対数オッズも大きく変化することがわかる. その一方で縦変化量が小さく (y 軸の値が 5~10) かつシュート方向に変化 (x 軸の値が -5~-10) するボールはコンタクト確率(対数オッズ)が比較的

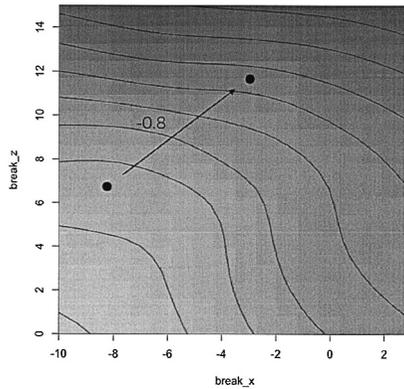


図 3. 変化量の変化による対数オッズの変化.

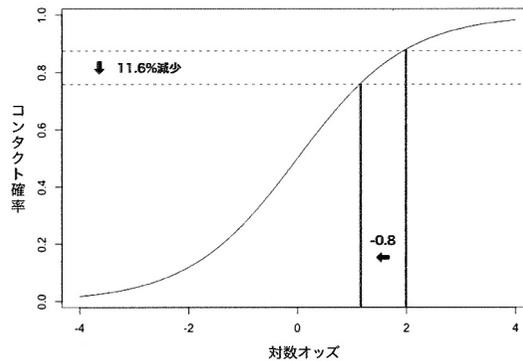


図 4. ロジスティック関数における変化.

高い。つまり、縦変化量の大きいストレートはコンタクトしにくいボールであり、縦変化量が小さくかつシュートするストレートはコンタクトしやすいボールであると解釈できる。推定結果から、縦変化量が増加するにつれてコンタクト確率は減少するという関係が見受けられ、空振りを考えた時、特に縦変化量が非常に重要であることがわかる。また単に縦変化量だけでなく、縦変化量が比較的小さい時に限って横変化量(シュート変化が小さいこと)が重要であることも分かった。これは変数に対して柔軟な関数を仮定し推定・視覚化することにより得られた結果であり、単純なロジスティック回帰モデルからはこのような解釈を得ることは難しい。

次に、変化量の変化に対する対数オッズの変化を確認することによって、コンタクトしにくさがどれほど変化するか考察を行いたい。図 3 では変化量のスプライン関数の値が大きい領域から小さい領域に対数オッズの値が変化した場合の例を示している。

図 3 のようにボールの変化量が変わった場合、対数オッズは 0.8 ほど減少すると推定された。仮に対数オッズ推定値の平均値から対数オッズが 0.8 減少したとすると、コンタクト確率は 0.875 から 0.758 へ減少し 11.6% ほど減少するという結果が得られた(図 4)。以上より、変化量はコンタクトに対して影響を有しており、変化量の変化に対してコンタクト確率が大きく変化することもわかった。

図 5 は対象データにおいて 100 球以上のストレートを投じた投手のコンタクト率をヒストグラムにしたものである。これによるとストレートの平均のコンタクト率で多いのは 86% 近辺

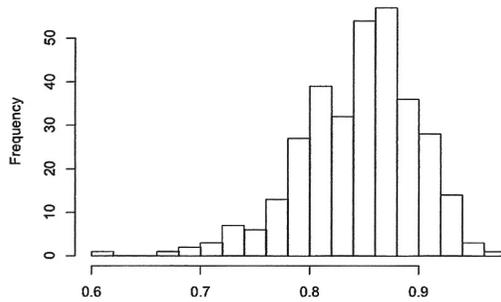


図 5. 投手ごとのコンタクト率のヒストグラム.

であり、80%を下回る投手は限られていることがわかる。推定した対数オッズの平均値から対数オッズが0.8減少するという仮定はあまり現実的ではないが、変化量の変化によってコンタクト確率が大きく変化するという事は明らかであろう。

ここでは、変化量がコンタクトという現象に影響を有しており、変化量の変化に対するコンタクト確率の変化を定量的な観点から解釈することができた。

5. 投手の打ちづらさの評価

5.1 変量効果を用いた解析の必要性

前節では、コンタクトに関係のあると思われる変数を用いてロジスティック回帰モデルを当てはめ、それらの関係性を評価した。ここで一般的にトラッキングシステムにより計測可能な変数を用いたが、投手と打者との対戦を考えた時、それらの他に考えるべき要素が存在すると考えられる。例えば、同じボールを打者に対して投球したとしても、投手のフォームやその他の持ち玉(投手が有している変化球)によってコンタクトのしづらさは異なる。

図6は先ほどのモデル(式(4.14))から算出された当てはめ値と実際のコンタクト率を二人の投手について示したものである。右図のダルビッシュに関しては当てはめ値がコンタクト率付近に分布しているが、左図の上原に関してはコンタクト率が当てはめ値から大きく乖離している。このことから上原においては各投球のコンタクト確率を低下させている要因があると考え

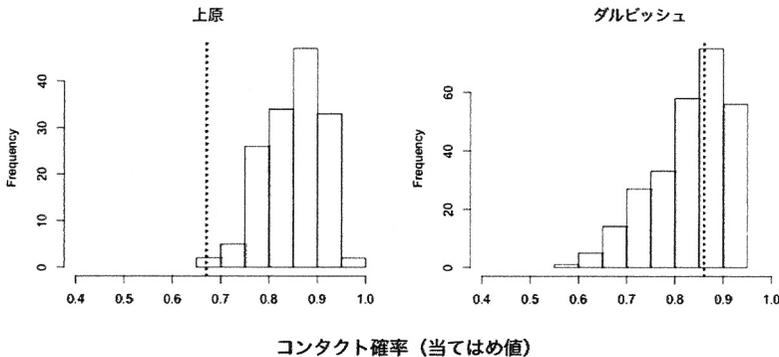


図 6. 上原(左図)とダルビッシュ(右図)におけるコンタクト率の当てはめ値のヒストグラムとデータにおけるコンタクト率(破線).

られる。

上原投手を例にとってみると、ストレートの他にスプリット・フィンガー・ファストボール (split-finger fastball) のような縦に急激に落ちるボールを有している。また、上原のフォームの特徴に、同じフォームから異なる球種のボールを投げられること、投球時のテイクバックが小さくボールの出所がわかりにくいこと、テイクバックしてからボールが手から離れるまでの時間が短いことなどがあげられる (<http://www.tokyo-sports.co.jp/sports/baseball/485297/>)。これらの要因は打者に対してはストレートをコンタクトしづらくさせるため、同じ球質のボールであったとしてもそれらがコンタクトに与える影響は大きいと思われる。

5.2 変量効果を加えたモデリングと予測値についての考察

本節では、各投手が有する打ちづらさを個体差のように捉えそれらの能力を変量効果としてモデルに組み込み推定を行うことにより各投手の打者対戦における優位性を評価することを試みる。以下のような、線形予測子の中に変量効果を線形で加えたモデルを考える。

$$(5.1) \quad \log\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + f_1(z_{i1}) + f_{23}(z_{i2}, z_{i3}) + \cdots + W_i \gamma$$

このモデルは式(4.14)の線形予測子に投手の変量効果 γ を加えたものであり、 γ は 2014 年シーズンにおいて投球を行った全投手の変量効果のパラメータベクトルである。 W_i はその投球に対応する投手を表す変数である。 i 番目の投球が k 番目の投手によるものであれば、 k 番目の要素のみが 1 であるような $W_i = [0, \dots, 1, \dots, 0]$ であり、 $W_i \gamma$ は変量効果 γ_k を表すものとする。また、それぞれの変量効果パラメータは $\gamma_k \sim N(0, \sigma_k^2)$ に従うとする。mgcv パッケージの gam 関数ではスプライン関数内の引数を指定することにより単純な変量効果をモデルに組み込むことができる。モデルの推定に関しては再パラメータ化を行うことにより、単純な変量効果を罰則付き回帰モデルとして表すことで変量効果の予測を行うことができる。

ここで投手ごとの変量効果パラメータをモデルに加えることの妥当性を示すため、変量効果をモデルに加えなかったモデル(式(4.14))と変量効果を加えたモデル(式(5.1))とで AIC を比較した。変量効果を加えたモデルは加えなかったモデルに比べ AIC が 64943.7 から 64470.0 に大きく減少するという結果が得られた。この結果からコンタクトに対しては個々の投手の打ちづらさによる要因が大きく影響しており、変量効果を加えることによるモデリングが妥当であることが言える。

次に各投手に対しての変量効果をモデルに組み込むことで投手の打ちづらさを定量的に評価したい。特に投手の打ちづらさに興味があるため、各投手に対する変量効果を予測し、それらの予測値を投手ごとに比較することで各投手の打ちづらさの考察を行う。表 7 は予測した投手ごとの変量効果の値に関して標本サイズが 100 球以上の投手を対象に最も値が小さい 6 名について示したものである。ストレートの平均球速、一試合当たりの平均投球数(全球種)、そして主に用いている変化球(持ち球)についてまとめた。

表 7 に示したように上原は対象シーズンの MLB においてコンタクトへの投手固有の影響を示す変量効果の予測値が最も低く、つまりモデルに含めた変数以外にコンタクト率を最も大きく下げる特徴を持っているという解析結果を得た。その他の投手に関しても、スプリット・フィンガー・ファストボールやチェンジアップなど、ストレートと逆の縦変化を起こすボールを有する投手はストレートを打ちづらくさせる傾向にあることがわかった。また全体的に、一試合当たりの平均投球数の少ない中継ぎや抑えピッチャーは打ちづらいと言う結果も得られた。

鶴岡(2016)にも、上原のストレートは鉛直方向に対して極めて大きく変化しており、それが打ちづらさの要因であると記述されている。しかし、先ほども述べたように上原においてはス

表 7. 各投手の変量効果の予測値(下位 6 名).

投手名	予測値	平均球速	投球数/試合	主な持ち球 (投球数が多い順)
上原浩治	-0.650	88.02	14.86	ストレート, SFF, ツーシーム
Bradley Boxberger	-0.588	93.24	16.76	ストレート, チェンジアップ, カットボール
Charlie Furbush	-0.580	91.92	9.98	ストレート, スライダー, ツーシーム
Jose Valverde	-0.498	93.07	17.80	ストレート, SFF
Nick Vincent	-0.472	90.07	13.53	スライダー, ストレート
Brandon McCarthy	-0.459	93.62	94.74	ツーシーム, カットボール, カーブ, ストレート

* SFF はスプリット・フィンガー・ファストボール

トレートの球質自体はもちろん容易に打てるものではないが、ストレートが有している特徴以外にもコンタクト確率を低下させている要素があり、それは並の投手とは比べものにならないほどのものであると考えられるため極めて低いコンタクト率が実現している。

他にも特徴的な予測値を有する投手をいくつかあげたい。日本人投手の中で予測値が低かったのは岩隈(-0.327)であった。岩隈に関しても縦に落ちるスプリット・フィンガー・ファストボールを有しており、またサイドハンドから投げる投手としては非常に大きな縦変化を有するストレートを投じることができると考えられる。また、ダルビッシュ(0.410)は平均的な投手よりも大きな値であった。

6. まとめ

本稿では PITCHf/x データを用いて、ストレートの各特徴量と空振りとの関係性を分析した。このデータの特徴は、座標や速度・変化量など多次元の変量が各球ごとに得られることであり、これらのデータを分析する上で、目的変数との多次元変数の関係を柔軟に分析できるモデルを用いたモデリングが求められる。その点において多変量スプライン平滑法を用いた解析は、有用な手法であると考えられる。今回の分析においては、ストレートの特に変化量に着目をし、コンタクトとの関係の解析を行った。空振りを考えた時、球速やコース・高さが重要であることは直感的に理解できると思うが、ストレートの変化量の違いによってコンタクトのしやすさが変わることはこれまであまり考えられていなかったように思える。そもそも、ストレートとは変化しないボールであり、変化を起こすボールは変化球であるといった認識が日本においては一般的である。しかし、トラッキングシステムが導入され普及するにつれてストレートの変化に着目することでより適切な投手の評価を行うことが期待される。また、ストレートの一つの表現としてノビがあるといった言葉が使われてきたが、その定義は曖昧であり明確な議論はされてこなかった。今回の分析において、一般的にノビのある投手と言われていた上原や藤川といった類の投手は他の投手に比べストレートの縦変化量が大きく、実際のコンタクト率も低いことがわかった。また、ノビの感覚的な理解として、バットがボールの下を通過するといった表現が用いられる。もしノビが縦変化量であると仮定すれば、ボールが回転によって物理的に到達する点よりも上であればこの現象との対応関係は取れており、そういった感覚をバッターボックスで感じることは縦変化量によるところが大きいのかもしれない。

では、はたして縦変化量の大きいストレートは良いものなのだろうか。その答えについてはさらに詳細な分析を行う必要があると考えられる。特に、飛翔する(ホームランになりやすい)ストレートやゴロアウトの取れるストレートは、おそらくボールの変化量との関係性を有しており、単に空振りといった観点からのみでなく飛翔やゴロアウトといった観点から変化量との関係性を分析する必要があるため、それはこれからの課題としたい。

さらに本稿では、各投手の打ちづらさを変量効果を用いてモデリングすることを試みた。予

測値の比較を行うことで上原や岩隈はストレートの質以外にも打ちづらい要素を持っておりそれを定量的に評価することができた。また、その他の投手においてもストレートと逆の変化をするチェンジアップやスプリット・フィンガー・ファストボールのような縦に落ちる球種を有している投手は予測値の値が小さくなる傾向にあることがわかった。

今回は PITCHf/x データに着目し、統計的な手法を用いて分析を行った。トラッキングデータは日本においてはあまり馴染みのあるものとは言えないが、徐々に様々なスポーツ・分野で導入が進んでおりこれから活用が行われていくものと思われる。また、データのみから解釈を行うだけでなく、統計的な解析により定量的な評価や、戦術的な分析を行うことが求められるであろう。

謝 辞

本論文を執筆するにあたり、原稿を注意深くお読み頂き多くの重要な指摘をして下さった 2 名の査読者の方に感謝を申し上げます。また本論文は第 5 回スポーツデータ解析コンペティションにおける発表結果をもとに作成されたものである。コンペティションの主催者である日本統計学会スポーツ分科会とデータ提供者であるデータスタジアム株式会社様にも重ねて感謝申し上げます。なお本研究の一部は、先端研究拠点事業(日本学術振興会：JSPS Core-to-Core Program)の助成を受けたものである。

参 考 文 献

- Green, P. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models*, Chapman and Hall, New York.
- Gu, C. (2013). *Smoothing Spline ANOVA Models*, Springer, New York.
- Hastie, T. and Tibshirani, R. (1986). Generalized additive models, *Statistical Science*, **1**, 297–318.
- Kagan, D. (2009). The anatomy of a pitch: Doing physics with PITCHf/x data, *The Physics Teacher*, **42**, 412–416.
- 金沢慧 (2015). 『「初速」と「終速」の差が小さければ良いストレートなのか?』, <http://www.baseball-lab.jp/column/entry/194/> (閲覧日: 2016 年 11 月 30 日).
- 鶴岡弘之 (2016). 『上原のストレートはなぜ打たれない? ICT で明らかに最先端テクノロジーがスポーツ市場を活性化する』, <http://jbpres.ismedia.jp/articles/-/48463> (閲覧日: 2016 年 11 月 30 日).
- Wood, S. N. (2003). Thin plate regression splines, *Journal of the Royal Statistical Society, Series B*, **65**, 95–114.
- Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R*, Chapman and Hall, New York.
- Wood, S. N. (2008). Fast stable direct fitting and smoothness selection for generalized additive models, *Journal of the Royal Statistical Society, Series B*, **70**, 495–518.
- Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation semi-parametric generalized linear models, *Journal of the Royal Statistical Society, Series B*, **73**, 3–36.

Factors Affecting Batters' Contact with a Four-seam Fastball

Daiki Nagata¹ and Mihoko Minami²

¹Graduate School of Science and Technology, Keio University

²Department of Mathematics, Keio University

In baseball, “nobi” is a four-seam fastball in which a batter has trouble making contact. Our research aims to understand the origin of nobi. It has been speculated that the velocity a four-seam fastball with nobi does not change much from the time it leaves the pitcher's hand to when it crosses the plate. Our previous analysis of nobi using PITCHf/x, which is a system that measures data such as the coordinates and break of a pitch by tracking the ball's trajectory, revealed the opposite relation. Consequently, we applied a logistic regression model to explain bat contact by the difference in the ball speed after defining the batter's contact with a pitch. A negative relation was obtained.

This study focuses on the break of a pitch. We analyzed the relationship between the break of a pitch and contact quantitatively. Additionally, we investigated the break of the ball by a generalized additive model using a multivariate spline smoothing method to evaluate the relationship between the break of the ball and bat contact. Vertical breaks are important. Moreover, adjusting the model to replace pitch quality as a random effect with hitting difficulty by pitcher revealed that in the 2014 MLB (Major League Baseball) season, Uehara was the most difficult pitcher for batters to face.

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

酒折 文武¹・圓城寺 啓人²・竹森 悠渡²・
西塚 真太郎²・保科 架風³

(受付 2017 年 1 月 31 日；改訂 3 月 22 日；採択 3 月 30 日)

要 旨

野球の投手における肘内側側副靭帯の損傷は近年増加しており、大きな問題となっている。予防の重要性にもかかわらず、そのリスク要因に関する科学的なコンセンサスが得られているとは言い難い。そこで本論文では、アマチュア野球の投手経験者とスポーツドクターの意見を参考にして、肘内側側副靭帯損傷のリスク要因の候補を再検討した。そして、先発投手とリリーフ投手とに層別してそれぞれロジスティック回帰モデルを立て、AIC を用いた変数選択により選択されたリスク要因について、調整オッズ比を算出した。その結果、先発投手については、球種数が少ないこと、リリース位置が体から横に離れていること、1 試合当たりの投球数が多いことがリスク要因であることがわかった。またリリーフ投手に関しては、球種数が少ないこと、リリース位置が体から横に離れていること、ファストボールの球速が速いこと、登板間隔が短いことがリスク要因であることがわかった。これらの結果は、他の研究成果の一部を肯定しているとともに、先発投手やリリーフ投手における 1 試合の投球数や登板間隔に関する重要な示唆を与えているといえる。

キーワード：オッズ比，ロジスティック回帰，スパースロジスティック回帰，Lasso。

1. はじめに

野球の投手における肘内側側副靭帯の損傷は近年増加しており、大きな問題となっている。それに伴い、肘内側側副靭帯再建手術、いわゆるトミー・ジョン手術がとりわけ米国のメジャーリーグ(MLB)では一般的となっている。我々日本人にも馴染みの深い、テキサス・レンジャーズに所属するダルビッシュ有投手も、2015 年にトミー・ジョン手術を受け 1 年以上の欠場を余儀なくされたことは記憶に新しい。

肘内側側副靭帯損傷の予防は非常に重要である一方で、そのリスク要因に関する科学的なコンセンサスが得られているとは言い難い(Whiteside et al., 2016)。例えば、アメリカスポーツ医学研究所は若年時からの蓄積によって故障が引き起こされると述べており、日本の全国高等学校野球選手権大会等での連戦連投、多投へのアメリカでの批判は多い。また、前述のダル

¹ 中央大学 理工学部：〒 112-8551 東京都文京区春日 1-13-27

² 中央大学大学院 理工学研究科：〒 112-8551 東京都文京区春日 1-13-27

³ 滋賀大学 データサイエンス教育研究センター：〒 522-8522 滋賀県彦根市馬場 1-1-1

ビッシュ有投手をはじめとする日本人選手や野球関係者の間では、1試合での投球数ではなく登板間隔が問題であり、MLBにおける先発投手の中4日での登板が故障の大きな原因であるとの意見がある。さらに、青少年に対する研究ではあるが、1試合の投球数やシーズンでの投球数が大きなリスク要因との指摘もある(Lyman et al., 2002)。その他、Olsen et al. (2006)では、青少年の投手に対する調査の結果をロジスティック回帰モデルにより分析し、怪我(多くは肘)のリスク要因として1年間に競技として投球する月数が多いこと、1試合あたりの投球数が多いこと、ファストボール(速球。球速の速い投球のこと)の球速が速いこと、疲労時に投球することが多いこと、の4つを挙げている。トミー・ジョン手術に限定しても、Keller et al. (2016)では、トミー・ジョン手術を受けた投手83名について、対照群83名との比較において、手術2年前の投球における各球種の球速には2群に有意差がなかったが、ファストボールの投球割合は手術群のほうが有意に高いと報告しており、ファストボールを多く投げる投手が肘内側側副靭帯損傷となる傾向があることを示唆している。

これらの研究に対して、Whiteside et al. (2016)では、近年の計測技術の向上により得られるようになった、投球に関する詳細な変数を用いてリスク要因を探っている。彼らは、手術群および対照群各104名の投手について、肘内側側副靭帯損傷のリスク要因となりうるデモグラフィック変数、セイバーメトリクス(野球における数理科学的なアプローチ)の指標FIPに加え、PITCHf/xという、球場に設置されたカメラの映像から、球速、球種、手からボールが離れたリリースポイントの座標、本塁を通過したときの座標、変化量、スピン量などを算出するシステムによる変数を用いてロジスティック回帰モデルによる分析を行っている。その結果、登板間隔が短い、投げる球種が少ない、リリースポイントの横位置が小さい(体に近い)、身長が低い、平均球速が速い、1試合あたりの投球数が多い、という6つをリスク要因として挙げている。しかしながら、モデルの立て方やリスク要因の候補の選定には一考の余地が残されている。例えば、投手の役割(先発投手かリリーフ投手か)によってこれらの要因の影響は大きく異なることが想定されるにもかかわらず、手術群と対照群のマッチングのみに使用されている。また、球速に関しては、球種の違いを考慮せずに全投球の平均球速のみしか使用していない。そして何よりも、リリースポイントが体に近いことは、Aguinaldo and Chambers (2009)などの他の研究と矛盾した結論である。

そこで本論文では、肘内側側副靭帯損傷のリスク要因についてさらなる検討を行った。まず、日本のアマチュア野球での投手経験者、スポーツドクター、トレーナーの意見を参考にし、改めてリスク要因の候補を再検討した。さらに、先発投手とリリーフ投手とに層別し、それぞれロジスティック回帰モデルを立て、AICを用いた変数選択により選択されたリスク要因の調整オッズ比を算出した。その結果、先発投手については、球種数が少ないこと、リリース位置が体から横に離れていること、1試合あたりの投球数が多いことがリスク要因であることがわかった。リリーフ投手に関しては、球種数が少ないこと、リリース位置が体から横に離れていること、ファストボールの球速が速いこと、登板間隔が短いことがリスク要因であることがわかった。これらの結果は、他の研究成果の一部を肯定しているとともに、先発投手やリリーフ投手における1試合の投球数や登板間隔に関する重要な示唆を与えているといえる。

本論文の構成は以下のとおりである。2節では分析に用いたデータとモデルについて説明する。3節では詳細な分析結果を述べる。それを受けて4節ではまとめと考察、そして今後の課題について述べる。

2. 研究手法

本論文では、2012年から2016年の間に肘内側側副靭帯損傷によるトミー・ジョン手術を行っ

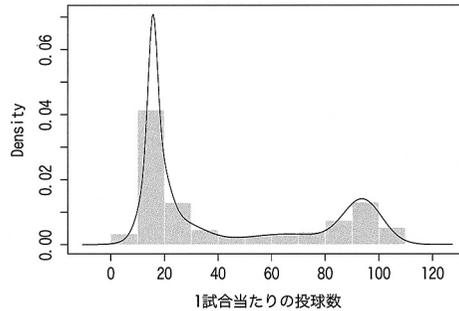


図 1. 全投手の 1 試合当たりの投球数の分布.

た投手(手術群)と、損傷していない(トミー・ジョン手術を行っていない)投手(対照群)に対して、ケースコントロール研究によって肘内側副靭帯損傷の要因分析を行った。手術同年あるいは前年のように故障時期に近いデータを用いた場合、故障の兆候がすでに投球に現れている可能性がある。一方で、あまり過去のシーズンのデータを用いても、故障への影響が不明でありノイズを多く含んでしまい、さらに投手の投球スタイル自体が変化している可能性もある。そこで、手術群については、手術年の 2 年前のシーズンのデータを用いることとした。また、対照群は対応する手術群投手と同年のシーズンのデータを用いた。

2.1 手術群

Baseball Heat Maps (2016) の Disabled List Data を参照し、2012 年から 2016 年の間に肘内側副靭帯損傷によるトミー・ジョン手術を行った投手の中で、前々年の全投球数が 100 球以上の投手全 74 名を手術群とした。なお、複数回手術を行った投手は最初に手術を行った年を選んだ。

2.2 1 試合あたり投球数での分類

先発投手やロングリリーフ投手などの長いイニングを投げる投手と、中継ぎ投手や抑え投手のような短いイニングを投げる投手では、リスク要因が異なることが想定される。そこで、1 試合あたり投球数の大小によって 2 群にわけるため、混合正規分布モデルによるクラスタリングを行った。

まず、2010 年～2014 年に投球した全投手における 1 試合当たりの投球数を確認したところ、図 1 のヒストグラムのように多峰型の分布であり、15 球と 95 球あたりに明確にピークがあることがわかった。そこで、これに混合正規分布を当てはめることとし、特異モデルである混合正規分布への適用については諸論があるところだが、BIC を用いてモデル選択を行った。各クラスターのサイズはすべて異なるとし、クラスター数(混合要素数)は 1 から 9 まで、各クラスターが等分散のモデル(E)と不等分散のモデル(V)を候補とした。各モデルの BIC の値は図 2 のようになり、5 クラスで不等分散のモデルが選択された。このモデルにおける各クラスターのパラメータの推定値は表 1 の通りである。

この結果を踏まえ、クラスター 1, 2, 3 が中継ぎ投手や抑え投手(以下リリーフ投手)、クラスター 4, 5 が先発投手やロングリリーフ投手(以下先発投手)と考え、クラスター 3 と 4 の分類の境界値、すなわち両クラスターへの帰属確率が等しくなる点を解析的に求め、この値(42.34)を超える選手を先発投手、それ以外をリリーフ投手と定めた。この基準をもとに上述の 74 名を先発とリリーフに振り分けたところ、ちょうど 37 名ずつであった。

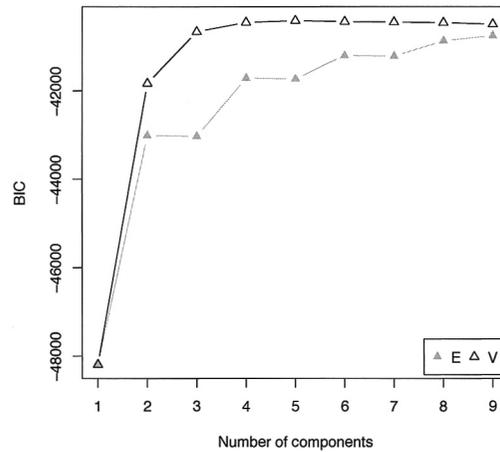


図 2. 混合正規分布モデルの BIC.

表 1. 混合正規分布モデリングの結果(混合要素数 5).

	クラスター 1	クラスター 2	クラスター 3	クラスター 4	クラスター 5
サイズ	0.179	0.276	0.167	0.140	0.238
平均	15.67	16.49	24.62	66.08	94.00
分散	2.933	19.428	91.547	299.907	51.538

2.3 対照群

先発およびリリーフそれぞれについて、選択した手術群の各投手に対して、同年に投球した同じ利き腕の投手 1 名ずつを対照群の投手として選定した。

選定にあたっては、過去に一度もトミー・ジョン手術を行っていない投手を対象とした。また、同姓同名の選手がおり、これらの選手は抽出対象から除外した。その上で、登板数や総投球数、投球スタイルなどの調べたい要因以外に故障に影響を与える可能性がある、体格や年齢・経験年数などをコントロールするために

- 手術前々年のシーズンにおける年齢
- 身長
- 体重
- メジャー経験年数
- 投球イニング

という 5 変数を Baseball Reference (2016) のウェブサイトから取得し、手術群の各投手と標準化ユークリッド距離が最も近い投手を選んだ。投球イニングを加えたのは、投球イニング自体の影響を除いた上で、1 試合の投球数と登板間隔のどちらが故障に影響を与えるのを見るためである。なお、対照群の投手で同じ投手が選ばれた場合には、年度が違った場合はそのまま使用し、年度まで同じ場合は次に距離の近い選手を選んだ。また、2.4 節での分析における欠測を避けるため、フォーシーム・ツーシーム・カットボールなどファストボールを 1 球も投げていない投手も除外した。以上の流れにより、先発とリリーフ合わせて各 74 名の手術群と対照群の投手をリストアップした。

表 2. 説明変数一覧.

x_1	登板間隔の中央値 (日)
x_2	登板数
x_3	1 試合当たりの平均投球数
x_4	球種数
x_5	ファストボールの投球割合
x_6	カーブの投球割合
x_7	チェンジアップの投球割合
x_8	ファストボールの平均球速 (マイル毎時)
x_9	ファストボールの最大球速と平均球速の差 (マイル毎時)
x_{10}	ファストボールの横方向の変化量の平均 (インチ)
x_{11}	ファストボールの縦方向の変化量の平均 (インチ)
x_{12}	ファストボールの回転数の平均
x_{13}	リリース位置の横座標の平均 (ライン)
x_{14}	リリース位置の横座標の標準偏差 (ライン)
x_{15}	リリース位置の縦座標の平均 (ライン)
x_{16}	リリース位置の縦座標の標準偏差 (ライン)

2.4 変数

リスク要因の検討には、まず、先行研究である Whiteside et al. (2016) で用いている変数を参考にした。その上で、日本のアマチュア野球の経験者と、スポーツドクター、トレーナーへのインタビューを行い、以下のような意見を得た：

- 登板数や 1 試合当たりの投球数が多い投手ほど、疲労が蓄積されるため怪我に繋がるのではないか。
- 疲労によって体が開き肘が下がること、逆に肘の位置が高すぎるにより負担がかかるため、リリース位置も考慮すべきである。
- 中 4 日の登板間隔が短いというダルビッシュ有投手らの意見は尤もであり、短い登板間隔で疲労回復ができていないのかは疑問が残る。
- 球種によってボールの握りや肘の捻りが様々なので、球種数が多いと怪我しやすい可能性がある。
- 力を入れて投球すれば、きれいな投げ方でも肘に負担がかかり怪我に繋がると考えられるため、ファストボールの球速や、最大球速との球速差などを考慮すべきである。
- カーブ、チェンジアップなどの変化球も肘に負担がかかると考えられるので含めるべきである。

これらを踏まえて、2.1 節で抽出した手術群と対照群各 74 名、計 148 の投手に対し、肘内側副韌帯損傷の要因と考えられる説明変数として表 2 にまとめた 16 変数を用いた。いずれも PITCHf/x データから算出した。本研究では R の pitchRx パッケージ (Sievert, 2014, 2015) を用いて PITCHf/x データを取得した。以下、これらの 16 変数について説明する。

まず、日本のプロ野球よりもトミー・ジョン手術が頻繁に行われている MLB では、中 4 日での登板が主流であり、登板間隔が短いことが怪我の原因であるという意見が聞かれる。そこで、登板間隔に関する変数として、登板間隔(日)の中央値 x_1 と、登板数 x_2 を用いた。 x_1 で中央値を用いた理由は、故障者リスト(DL)入りやマイナー落ちなどの影響を除くためである。次に、投球数や登板回数の多い投手ほど怪我につながりやすいと考えられるため、1 試合当たりの平均投球数 x_3 を用いた。そして、球種が少ないほど故障しやすいという Whiteside et

al. (2016) の指摘を受けて、投球した球種数(最大で 10 種類) x_4 を用いた。さらに、各球種の影響を見るために、フォーシーム・ツーシーム・カットボールなどを合わせたファストボール、カーブ、チェンジアップの 3 球種の投球割合 x_5, x_6, x_7 を用いた。スライダの投球割合も検討したが、ファストボールの投球割合 x_5 との相関が -0.9 程度と非常に強かったため変数には加えなかった。

また, Keller et al. (2016) で指摘されるファストボールの影響をさらに細かく見るため, ファストボールの平均球速(マイル毎時) x_8 , 最大球速と平均球速の差(マイル毎時) x_9 , 横方向および縦方向の変化量の平均(インチ) x_{10}, x_{11} , 回転数の平均 x_{12} を用いた。投球フォームによる影響を考慮するため, 投球時にボールをリリースした位置の横座標, 縦座標の平均(ライン = $1/12$ マイル) x_{13}, x_{15} をそれぞれ用いた。加えて, 疲労による体の開きやそれに伴う肘下がりなどによる平均的なリリース位置からのずれの具合を測るために, リリース位置の横方向, 縦方向それぞれの標準偏差(ライン) x_{14}, x_{16} を用いた。カーブやチェンジアップについても同様の変数群を加えたいところであるが, そもそもこれらの球種を投げない投手も少なくないため, 今回は変数に加えなかった。

2.5 分析方法

上記のようにして選んだりリスク要因の候補全 16 変数を説明変数とし, トミー・ジョン手術の有無(手術群 or 対照群)を目的変数として, 先発とリリーフとに分けてそれぞれロジスティック回帰モデル

$$(2.1) \quad \text{logit}(p_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{16} x_{i16} = \beta^T \mathbf{x}_i$$

により各説明変数による故障のリスクに関する分析を行った。最尤法により推定した係数 $\hat{\beta}_j$ および正規近似による β_j の信頼区間 $[l_j, u_j]$ から, 各説明変数の調整オッズ比 $e^{\hat{\beta}_j}$, および母オッズ比の 95% 信頼区間 $[e^{l_j}, e^{u_j}]$ を求めた。

説明変数の数が 16 と多いため, AIC を基準とするステップワイズ法(変数増減法)によって変数選択を行い, 選ばれた変数に対する調整オッズ比と 95% 信頼区間を求めた。

これらの 2 つのロジスティック回帰モデルでは判別を目的とはしていないが, 参考のために確率 0.5 を境界としたときの判別結果と正判別率を求めた。また, モデル M の適合度を評価するために, スルモデル M_0

$$(2.2) \quad \text{logit}(p_i) = \beta_0$$

との尤度比検定の結果と, Nagelkerke の疑似決定係数(Nagelkerke, 1991)

$$(2.3) \quad R^2 = \frac{1 - \left\{ \frac{L_{M_0}(\hat{\beta}_0)}{L_M(\hat{\beta})} \right\}^{2/n}}{1 - L_{M_0}(\hat{\beta}_0)^{2/n}}$$

の値を求めた。ここで, $L_M(\beta)$ はモデル M の下での尤度, ハットはそのモデルの下での最尤推定量, n はサンプルサイズである。後者は重回帰における決定係数の考え方を一般化し, 上限が 1 となるように調整したものである。

AIC を用いた変数選択では, 説明変数の数が多いため総当たり法ではなくステップワイズ法を用いたが, 別のアプローチとして考えられるのは L_1 正則化法である。 L_1 正則化法を用いることにより, 実質的に L_0 での正則化と考えられる総当たり法よりも高速に分析を行うことができる。ここでは, モデル(2.1)に Lasso (Friedman et al., 2010; Hastie et al., 2009)による L_1 正則化を加えたスパースロジスティック回帰モデルによる分析も合わせて行った。Lasso による

表 3. 手術群と対照群における 5 変数に関する平均の差の検定. 上：先発投手, 下：リリーフ投手.

	手術群	対照群	t 値	P 値
年齢	25.62 ± 3.29	26.00 ± 2.76	-0.536	0.59
身長	190.50 ± 5.22	190.50 ± 4.90	0.000	1.00
体重	98.93 ± 8.15	98.20 ± 7.89	0.394	0.69
経験年数	4.19 ± 3.47	4.24 ± 3.16	-0.070	0.94
イニング	117.07 ± 65.95	118.28 ± 70.77	-0.076	0.94

	手術群	対照群	t 値	P 値
年齢	27.22 ± 3.61	27.00 ± 3.06	0.278	0.78
身長	189.54 ± 5.78	189.95 ± 5.59	-0.312	0.76
体重	99.36 ± 9.80	98.38 ± 8.24	0.466	0.64
経験年数	3.97 ± 2.69	4.14 ± 2.51	-0.268	0.79
イニング	46.56 ± 20.61	46.37 ± 22.17	0.038	0.97

推定量は

$$(2.4) \quad \hat{\beta}_{\text{lasso}} = \arg \min_{\beta} \left\{ -\frac{1}{n} \log L(\beta) + \lambda \sum_{j=1}^{16} |\beta_j| \right\}$$

により与えられ, 正則化パラメータ λ の値によっては縮小推定とともにいくつかの係数が正確に 0 と推定され, 変数選択の役割も果たす. ここでは, λ の選択には 10-fold クロスバリデーションを用いた.

3. 分析結果

分析結果は以下の通りである.

3.1 対照群選定の妥当性の確認

対照群の投手を選定する際に用いた 5 変数が手術群と対照群において差がないことを確認するため, 変数ごとに等分散を仮定した平均の差の検定を行った. P 値は表 3 の通りとなり, 各変数に有意な差がないことがわかった. なお, これら 5 変数は手術群と対照群とで調整されているため, 以降の分析には用いていない.

3.2 各リスク要因の分布

リスク要因の候補となる各説明変数の要約統計量は表 4 の通りである. 先発投手とリリーフ投手では当然ながら, 登板間隔 x_1 , 登板数 x_2 , 1 試合当たり投球数 x_3 が大きく異なることがわかる. さらに, 球種数 x_4 , 変化球の投球割合 x_6, x_7 , ファストボールの球速 x_8 , 球速差 x_9 , 横変化量 x_{10} , 回転数 x_{12} , リリース位置 x_{13}, x_{15} などほとんどの変数においても先発投手とリリーフ投手に差がみられることがわかる. また, 手術群と対照群を比較すると, 先発投手ではリリース位置横 x_{13} 以外には大きな差が認められないのに対し, リリーフ投手では登板間隔 x_1 , 1 試合当たり投球数 x_3 , ファストボール球速 x_8 , リリース位置横 x_{13} などに差がみられる.

なお, 説明変数間の相関はほとんど無く, 相関係数の絶対値が 0.5 を超えるものは先発投手では登板数 x_2 と球種数 x_4 (0.53), ファストボールの縦変化量 x_{11} と回転数 x_{12} (0.59), リリー

表 4. 各リスク要因の要約統計量(平均 ± 標準偏差).

変数	先発投手 ($n = 74$)		リリーフ投手 ($n = 74$)	
	手術群 ($n = 37$)	対照群 ($n = 37$)	手術群 ($n = 37$)	対照群 ($n = 37$)
登板間隔 x_1	5.49 ± 0.49	5.62 ± 2.68	2.57 ± 0.85	3.41 ± 1.36
登板数 x_2	21.57 ± 10.48	22.14 ± 11.73	51.32 ± 23.17	46.35 ± 23.23
1 試合当たり投球数 x_3	90.23 ± 12.31	86.63 ± 17.53	16.87 ± 5.39	19.19 ± 7.27
球種数 x_4	6.49 ± 1.15	6.73 ± 1.48	5.49 ± 1.15	5.97 ± 1.09
ファストボール割合 x_5	0.53 ± 0.21	0.49 ± 0.22	0.54 ± 0.24	0.48 ± 0.24
カーブ割合 x_6	0.08 ± 0.08	0.08 ± 0.08	0.04 ± 0.09	0.06 ± 0.10
チェンジアップ割合 x_7	0.11 ± 0.08	0.12 ± 0.08	0.06 ± 0.06	0.09 ± 0.08
ファストボール球速 x_8	91.15 ± 2.25	91.14 ± 2.13	93.32 ± 2.45	92.02 ± 2.91
ファストボール球速差 x_9	3.88 ± 1.05	3.87 ± 1.12	3.50 ± 1.38	3.27 ± 1.08
ファストボール横変化量 x_{10}	-1.91 ± 5.45	-2.08 ± 5.15	-1.88 ± 5.63	-0.84 ± 6.14
ファストボール縦変化量 x_{11}	7.93 ± 1.64	7.8 ± 1.77	7.75 ± 2.72	7.92 ± 2.03
ファストボール回転数 x_{12}	2028.47 ± 351.28	2013.13 ± 260.14	2126.23 ± 276.44	2092.68 ± 297.8
リリース位置横 x_{13}	22.77 ± 5.55	19.64 ± 6.82	24.15 ± 8.49	20.68 ± 8.67
リリース位置縦 x_{14}	71.6 ± 3.91	72.62 ± 4.37	70.3 ± 4.08	70.46 ± 5.41
リリース位置横ばらつき x_{15}	3.14 ± 1.05	3.15 ± 0.90	3.67 ± 1.37	3.78 ± 1.44
リリース位置縦ばらつき x_{16}	2.52 ± 0.70	2.42 ± 0.49	2.65 ± 0.95	2.52 ± 0.84

ス位置の横のばらつき x_{15} と縦のばらつき x_{16} (0.55), リリーフ投手では, 登板間隔 x_1 と登板数 x_2 (-0.62), 登板間隔 x_1 と 1 試合当たり投球数 x_3 (0.60), 登板数 x_2 と 1 試合当たり投球数 x_3 (-0.52), ファストボールの球速 x_8 と横変化量 x_{10} (-0.58)のみであった.

3.3 各リスク要因のオッズ比

表 5 は, リスク要因ごとに 1 変量のロジスティック回帰分析を行い, 各変数のオッズ比を算出したものである. 係数の分布の正規近似により求めた 95% 信頼区間を併記した.

登板間隔 x_1 , 1 試合当たり投球数 x_3 , カーブの投球割合 x_6 については, 先発投手とリリーフ投手でオッズ比が大きく異なることがわかる. しかし, 多くの変数においては差がそれほど見られず, さらにオッズ比の信頼区間が 1 を含んでいることもわかる.

なお, ファストボール等の投球割合 x_5, x_6, x_7 は 0 から 1 のみを取りうるため, オッズ比の値自身は解釈しにくい. 逆に, 2000 程度の大きな値を取るファストボール回転数 x_{12} も同様である.

3.4 ロジスティック回帰の結果と調整オッズ比

全変数を用いたロジスティック回帰分析の結果と, 各変数の調整オッズ比および信頼区間を示したものが表 6 である.

5% 有意となった変数は, 先発投手においてはリリース位置横 x_{13} のみ, リリーフ投手については登板間隔 x_1 のみであった. 説明変数間の相関関係の影響から, 球種数 x_4 , リリース位置縦ばらつき x_{16} の調整オッズ比の値が元のオッズ比とは大きく異なっている. また, 各球種の投球割合 x_5, x_6, x_7 のオッズ比についても大きく異なっているが, これらの値は前述の通り非常に小さな値をとるため解釈しづらい.

リリーフ投手においても, 各球種の投球割合 x_5, x_6, x_7 を除いては, 登板間隔 x_1 , 1 試合当たり投球数 x_3 など他変数との相関が比較強い変数ではオッズ比と調整オッズ比は多少異なるが, ファストボール球速差 x_9 やリリース位置横ばらつき x_{15} などのその他の一部の変数に

表 5. 各リスク要因のオッズ比.

変数	先発投手 ($n = 74$)		リリーフ投手 ($n = 74$)	
	オッズ比	95%信頼区間	オッズ比	95%信頼区間
登板間隔 x_1	0.9625	[0.6981, 1.2511]	0.4519	[0.2391, 0.7532]
登板数 x_2	0.9953	[0.9543, 1.0377]	1.0095	[0.9896, 1.0304]
1 試合当たり投球数 x_3	1.0164	[0.9858, 1.0503]	0.9411	[0.8620, 1.0139]
球種数 x_4	0.8672	[0.6008, 1.2293]	0.6739	[0.4317, 1.0193]
ファストボール割合 x_5	2.2649	[0.2558, 22.0998]	2.8578	[0.4318, 20.9748]
カーブ割合 x_6	1.9658	[0.0045, 899.2441]	0.1232	[0.0005, 16.7727]
チェンジアップ割合 x_7	0.1602	[0.0004, 53.2924]	0.0017	[0.0000, 0.9381]
ファストボール球速 x_8	1.0009	[0.8087, 1.2390]	1.2007	[1.0093, 1.4536]
ファストボール球速差 x_9	1.0105	[0.6559, 1.5588]	1.1726	[0.8059, 1.7657]
ファストボール横変化量 x_{10}	1.0064	[0.9217, 1.0996]	0.9696	[0.8947, 1.0486]
ファストボール縦変化量 x_{11}	1.0446	[0.7962, 1.3750]	0.9706	[0.7907, 1.1811]
ファストボール回転数 x_{12}	1.0002	[0.9987, 1.0017]	1.0004	[0.9988, 1.0021]
リリース位置横 x_{13}	1.0872	[1.0075, 1.1838]	1.0495	[0.9942, 1.1130]
リリース位置縦 x_{14}	0.9406	[0.8353, 1.0516]	0.9928	[0.8993, 1.0947]
リリース位置横ばらつき x_{15}	0.9869	[0.6063, 1.6016]	0.9453	[0.6664, 1.3206]
リリース位置縦ばらつき x_{16}	1.3689	[0.6309, 3.2856]	1.1808	[0.6997, 2.1017]

おいても差が見られる。

また、AIC を基準としたステップワイズ法(変数増減法)によって変数選択を行った結果が表 7 であり、確率 0.5 を境界としたときの判別結果や正判別率と、尤度比検定の結果、疑似決定係数の値が表 8 にまとめられている。

先発投手に関しては、1 試合当たり投球数 x_3 、球種数 x_4 、リリース位置横 x_{13} という 3 変数が選択された。まず、球種数が少ないほど故障しやすく、球種が 1 つ少ないと故障のオッズが $1/0.7509 = 1.33$ 倍(95% 信頼区間は $1/1.1013 = 0.91$ 以上 $1/0.4952 = 2.02$ 以下)となることがわかる。これは、Whiteside et al. (2016) とほぼ同様の結果である。また、リリース位置が体から横に離れるほど故障しやすく、横に 1 インチ(= 12 ライン)離れると故障のオッズが $1.0999^{12} = 3.14$ 倍(95% 信頼区間は $1.0160^{12} = 1.21$ 以上 $1.2039^{12} = 9.27$ 以下)となることがわかる。これは、Whiteside et al. (2016) と正反対の結果である。しかしながら、Whiteside et al. (2016) の結果は横手投げの投手のほうが上手投げの投手よりも肘への負荷が有意に大きいという指摘(Aguinaldo and Chambers, 2009)と矛盾するため、今回の結果は妥当であるといえる。そして、1 試合当たりの投球数が多いほど故障しやすく、1 試合当たり投球数が 1 球多いと故障のオッズが 1.03 倍(95% 信頼区間は 0.99 以上 1.06 以下)となることがわかる。これも Whiteside et al. (2016) とほぼ同様の結果である。一方、登板間隔は選択されなかった。このことは、主に日本人投手の言う、1 試合での投球数ではなく登板間隔こそがリスクであるという意見に反する結果であり、MLB での投手起用の方針を支持するものである。とはいえ、尤度比検定の結果は有意ではあるが、正判別率や疑似決定係数の値からは、リスク要因の選定に改善の余地が残されている。また、あくまでも MLB の投手における結果であり、これが日本人投手にも同じことが言えるかはさらなる議論の余地がある。

リリーフ投手に関しては、登板間隔 x_1 、球種数 x_4 、ファストボール球速 x_8 、リリース位置横 x_{13} という 4 変数が選択された。まず、先発投手と同様、球種数が少なく、リリース位置が体から横に離れるほど故障しやすいことがわかる。そのオッズ比は、球種数は $1/0.6697 = 1.49$ (95% 信頼区間は $1/1.0776 = 0.93$ 以上 $1/0.4034 = 2.48$ 以下)、リリース位置横(インチ)は

表 6. 先発投手およびリリーフ投手におけるロジスティック回帰分析の結果(全変数). 上: 先発投手, 下: リリーフ投手.

	推定値	標準誤差	z 値	P 値	調整オッズ比	95% 信頼区間
切片	-1.3358	19.9426	-0.0670	0.9466		
登板間隔 x_1	0.1757	0.1508	1.165	0.2440	1.1921	[0.8496, 1.6416]
登板数 x_2	-0.0051	0.0363	-0.1415	0.8875	0.9949	[0.9244, 1.0682]
1 試合当たり投球数 x_3	0.0436	0.0236	1.8473	0.0647	1.0446	[0.9988, 1.0971]
球種数 x_4	-0.5173	0.2845	-1.8182	0.0690	0.5961	[0.3276, 1.0159]
ファストボール割合 x_5	2.9403	1.6738	1.7566	0.0790	18.922	[0.8132, 639.926]
カーブ割合 x_6	-0.2055	3.7229	-0.0552	0.9560	0.8142	[0.0005, 1342.5372]
チェンジアップ割合 x_7	-3.7282	4.5957	-0.8113	0.4172	0.0240	[0.0000, 176.7731]
ファストボール球速 x_8	-0.0325	0.2040	-0.1592	0.8735	0.9680	[0.6426, 1.4448]
ファストボール球速差 x_9	-0.1262	0.3016	-0.4184	0.6756	0.8814	[0.4799, 1.5977]
ファストボール横変化量 x_{10}	0.0419	0.0682	0.6145	0.5389	1.0428	[0.9117, 1.1953]
ファストボール縦変化量 x_{11}	0.1739	0.2377	0.7316	0.4644	1.1899	[0.7476, 1.9285]
ファストボール回転数 x_{12}	-0.0003	0.0015	-0.1739	0.8619	0.9997	[0.9968, 1.0027]
リリース位置横 x_{13}	0.1181	0.0536	2.205	0.0275	1.1253	[1.0185, 1.2598]
リリース位置縦 x_{14}	-0.0435	0.0788	-0.5517	0.5811	0.9575	[0.8161, 1.1168]
リリース位置横ばらつき x_{15}	0.2240	0.3522	0.6360	0.5248	1.2511	[0.6186, 2.5490]
リリース位置縦ばらつき x_{16}	0.6418	0.6174	1.0396	0.2985	1.8999	[0.5974, 7.1152]

	推定値	標準誤差	z 値	P 値	調整オッズ比	95% 信頼区間
切片	-16.0479	14.4209	-1.1128	0.2658		
登板間隔 x_1	-1.2147	0.5081	-2.3904	0.0168	0.2968	[0.0942, 0.7129]
登板数 x_2	-0.0124	0.0192	-0.6464	0.5180	0.9877	[0.9499, 1.0254]
1 試合当たり投球数 x_3	0.0358	0.0703	0.5084	0.6112	1.0364	[0.8978, 1.1918]
球種数 x_4	-0.3389	0.2872	-1.1800	0.2380	0.7126	[0.3962, 1.2422]
ファストボール割合 x_5	0.5391	1.5732	0.3427	0.7318	1.7145	[0.0791, 42.3171]
カーブ割合 x_6	-1.9066	3.7203	-0.5125	0.6083	0.1486	[0.0001, 259.6404]
チェンジアップ割合 x_7	-3.0066	4.4531	-0.6752	0.4996	0.0495	[0.0000, 3060.9778]
ファストボール球速 x_8	0.2065	0.1558	1.3256	0.1850	1.2293	[0.9118, 1.6961]
ファストボール球速差 x_9	-0.0711	0.3130	-0.2272	0.8202	0.9313	[0.5046, 1.7648]
ファストボール横変化量 x_{10}	0.0381	0.0599	0.6363	0.5246	1.0389	[0.9250, 1.1743]
ファストボール縦変化量 x_{11}	-0.0437	0.1755	-0.2489	0.8035	0.9573	[0.6644, 1.3503]
ファストボール回転数 x_{12}	0.0005	0.0011	0.4724	0.6367	1.0005	[0.9984, 1.0027]
リリース位置横 x_{13}	0.0430	0.0404	1.0649	0.2869	1.0440	[0.9654, 1.1336]
リリース位置縦 x_{14}	0.0166	0.0687	0.2417	0.8090	1.0167	[0.8879, 1.1670]
リリース位置横ばらつき x_{15}	-0.2086	0.2526	-0.8256	0.4090	0.8117	[0.4868, 1.3383]
リリース位置縦ばらつき x_{16}	0.2091	0.4650	0.4497	0.6529	1.2326	[0.5024, 3.2197]

$1.0528^{12} = 1.85$ (95% 信頼区間は $0.9596^{12} = 0.88$ 以上 $1.1252^{12} = 4.12$ 以下)である. 次に, ファストボールの球速が速いほど, そして登板間隔が短いほど故障しやすいことがわかる. オッズ比はそれぞれ, 1.23 (95% 信頼区間は 1.00 以上 1.08 以下), $1/0.4211 = 2.37$ (95% 信頼区間は $1/0.7453 = 1.34$ 以上 $1/0.2047 = 4.89$ 以下)である. これは, Whiteside et al. (2016)における, (ファストボールに限らない)平均球速が速いほど, そして登板間隔が短いほど故障しやすいという結果に対応している. 判別結果や尤度比検定, 疑似決定係数の結果から, リリーフ投手に関する分析結果は妥当であるといえる.

表 7. 先発投手およびリリーフ投手におけるロジスティック回帰分析の結果(変数選択後).
上：先発投手, 下：リリーフ投手.

	推定値	標準誤差	z 値	P 値	調整オッズ比	95% 信頼区間
切片	-2.3455	1.887	-1.2430	0.2139		
1 試合当たり投球数 x_3	0.0250	0.0174	1.4382	0.1504	1.0253	[0.9919, 1.0633]
球種数 x_4	-0.2864	0.2011	-1.4246	0.1543	0.7509	[0.4952, 1.1013]
リリース位置横 x_{13}	0.0952	0.0428	2.2258	0.0260	1.0999	[1.0160, 1.2039]

	推定値	標準誤差	z 値	P 値	調整オッズ比	95% 信頼区間
切片	-15.0922	10.1447	-1.4877	0.1368		
登板間隔 x_1	-0.8650	0.3276	-2.6401	0.0083	0.4211	[0.2047, 0.7453]
球種数 x_4	-0.4010	0.2480	-1.6168	0.1059	0.6697	[0.4034, 1.0776]
ファストボール球速 x_8	0.2019	0.1080	1.8690	0.0616	1.2237	[0.9982, 1.5333]
リリース位置横 x_{13}	0.0514	0.0323	1.5904	0.1117	1.0528	[0.9896, 1.1252]

表 8. 変数選択後のモデルにおける先発投手およびリリーフ投手の判別結果. 左：先発投手, 右：リリーフ投手.

手術群と判別	対照群と判別	合計	手術群と判別	対照群と判別	合計
手術群	24	13	28	9	37
対照群	16	21	11	26	37
合計	40	34	39	35	74

正判別率: 0.61	正判別率: 0.73
$\chi^2 = 8.025, P$ 値: 0.0455	$\chi^2 = 20.422, P$ 値: 0.0004
Nagelkerke の疑似 R^2 : 0.137	Nagelkerke の疑似 R^2 : 0.322

3.5 スパースロジスティック回帰

Lasso を用いた正則化ロジスティック回帰モデルによるソリューションパスと, 10-fold クロスバリデーションにより選ばれた正則化パラメータの値 $\hat{\lambda}$ を示したのが図 3 である. 図の横軸は正則化パラメータ λ の対数, 縦軸は偏回帰係数を表す. 各曲線が一つの変数に対応しており, 正則化パラメータの値に対応する係数の推定値を表している. また, 図中の縦の点線が, クロスバリデーションによって選ばれた正則化パラメータの値 $\hat{\lambda}$ を表しており, そのときの各曲線の値が, この分析での最終的な推定値である.

クロスバリデーションで選択された λ の値 $\hat{\lambda}$ に基づくと, 先発投手についてはすべての係数が 0 と推定されてしまったことが図からわかる. リリーフ投手に関する係数の推定値 $\hat{\beta}_{\text{lasso}}$ およびそれを元に計算した“調整オッズ比” $e^{\hat{\beta}_{\text{lasso}}}$ は表 9 のようになった.

先発投手については, すべての変数が選択されず, 3.4 節のモデルでの説明力の弱さが強調される結果となった. またリリーフ投手については, 3.4 節の変数選択後のモデルで選ばれた変数は全て選ばれており, lasso の縮小推定の効果によりオッズ比はいずれも 1 に近づいていることがわかる. その近づき具合は概ね, 元のオッズ比の平方根程度である. すなわち, 例えばファストボールの球速 x_8 が 1 マイル増えたときのオッズ比を 0.5 マイル増えたときのオッズ比程度に過小評価していることになる. もちろん, この過小評価の度合いは λ に依存するため, 常にこのような結果が得られるわけではない. また, 上記変数に加えてチェンジアップの投球割合 x_7 が影響を及ぼしているとの結果であった.

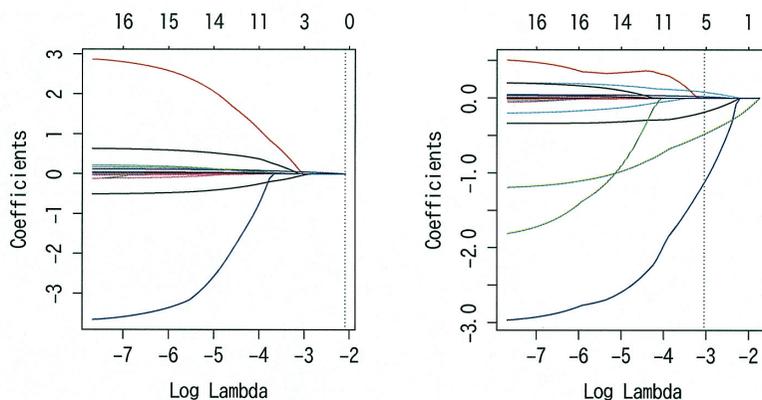


図 3. Lasso によるソリューションパス. 左: 先発投手, 右: リリーフ投手. 各折れ線が変数に対応しており, クロスバリデーションで選ばれた正則化パラメータの値(縦点線)を用いると, 先発投手では全ての変数の係数が 0 と推定され, リリーフ投手では上から順にファストボール球速(水色), リリース位置横(青色), 球種数(黒色), 登板間隔(緑色), チェンジアップ割合(青色)以外は 0 と推定されている.

表 9. Lasso による推定値(リリーフ投手).

	推定値	調整オッズ比
登板間隔 x_1	-0.4830	0.6169
球種数 x_4	-0.1866	0.8297
チェンジアップ割合 x_7	-1.1275	0.3238
ファストボール球速 x_8	0.0801	1.0834
リリース位置横 x_{13}	0.0185	1.0187

以上を踏まえると, リリーフ投手についてはステップワイズ法での結果と(オッズ比は過小評価しているものの)ほぼ類似した結果であると言えよう. 別の見方をすれば, ステップワイズ法では局所解に落ちている可能性もあるが, いずれにしても結果が妥当であることが確認できたとも言える.

4. おわりに

本研究では, 先発投手やロングリリーフ投手など長いイニングを投げる投手(先発投手)と, 中継ぎ投手や抑え投手など短いイニングを投げる投手(リリーフ投手)とに分けて, それぞれ肘内側側副韌帯損傷のリスク要因について検証した.

その結果, 先発投手については, 球種数が少ないこと, リリース位置が体から横に離れていること, 1 試合当たりの投球数が多いことがリスク要因であることがわかった. また, これらのリスクを定量的に評価することができた. 分析結果は, MLB での投手起用の方針を支持し, 主に日本での論調となっている「1 試合での投球数は多くてもよい, 登板間隔が短いのが問題である」という意見に反する結論となった. とはいえ, 先発投手に関してはリスク要因の選定に改善の余地が残されており, また, これが日本人投手にも同じことが言えるかはさらなる議論の余地がある.

またリリーフ投手に関しては, 球種数が少ないこと, リリース位置が体から横に離れている

こと、ファストボールの球速が速いこと、登板間隔が短いことがリスク要因であることがわかった。すなわち、先発とは異なり、リリーフ投手では1試合の投球数よりも登板間隔が重要であるとの結果が得られた。

また、本研究では変数選択法としてAICを用いたステップワイズ法と、Lassoによるスパースロジスティック回帰モデルとを用いた。リリーフ投手については非常に類似した結果が得られた一方で、先発投手についてはスパースロジスティック回帰モデルではよい推定結果が得られなかった。これは、ある程度の説明力がある説明変数がある場合にはステップワイズ法でもスパースロジスティック回帰モデルでも同様の結果が得られるが、説明力が十分に高くない場合にはスパースロジスティック回帰モデルではその影響をうまく捉えることができないことを示唆している。本研究の主たる目的ではないが、両方法の結果の違いを副産物的に理解することができた。

謝 辞

適切かつ有益な助言を頂いたことに対し、2人の匿名査読者と編集委員の先生に感謝する。本研究の着想に至るには、日本統計学会スポーツ統計分科会主催のスポーツデータ解析コンペティション、および情報・システム研究機構の新領域融合研究プロジェクト『社会コミュニケーション』データ中心科学リサーチコモンズ事業『人間・社会データ』の支援が不可欠であった。また、著者の1名は青山学院大学経営学部グローバルビジネス研究所の支援を受けた。本研究の一部は科学科研究費(基盤研究(C)(一般)No.15K00054)の助成を受けている。

参 考 文 献

- Aguinaldo, A. L. and Chambers, H. (2009). Correlation of throwing mechanics with elbow valgus load in adult baseball pitchers, *The American Journal of Sports Medicine*, **37**, 2043–2048.
- Baseball Heat Maps (2016). <http://www.baseballheatmaps.com/disabled-list-data/>, 2016年12月19日アクセス。
- Baseball Reference (2016). <http://www.baseball-reference.com>, 2016年12月19日アクセス。
- Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent, *Journal of Statistical Software*, **33**, 1–22.
- Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., Springer, New York.
- Keller, R. A., Marshall, N. E., Guest, J. M., Okoroha, K. R., Jung, E. K. and Moutzouros, V. (2016). Major League Baseball pitch velocity and pitch type associated with risk of ulnar collateral ligament injury, *Journal of Shoulder and Elbow Surgery*, **25**, 671–675.
- Lyman, S., Fleisig, G. S., Andrews, J. R. and Osinski, E. D. (2002). Effect of pitch type, pitch count, and pitching mechanics on risk of elbow and shoulder pain in youth baseball pitchers, *The American Journal of Sports Medicine*, **30**, 463–468.
- Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination, *Biometrika*, **78**, 691–692.
- Olsen, S. J., Fleisig, G. S., Dun, S., Loftice, J. and Andrews, J. R. (2006). Risk factors for shoulder and elbow injuries in adolescent baseball pitchers, *The American Journal of Sports Medicine*, **34**, 905–912.
- Sievert, C. (2014). Taming PITCHf/x Data with pitchRx and XML2R, *The R Journal*, **6**, <http://journal.r-project.org/archive/2014-1/sievert.pdf>.

- Sievert, C. (2015). pitchRx: Tools for Harnessing ‘MLBAM’ ‘Gameday’ Data and Visualizing ‘pitchfx’, R package version 1.8.2.
- Whiteside, D., Martini, D. N., Lepley, A. S., Zernicke, R. F. and Goulet, G. C. (2016). Predictors of ulnar collateral ligament reconstruction in major league baseball pitchers, *The American Journal of Sports Medicine*, **44**, 2202–2209.

A Statistical Analysis of Medial Collateral Ligament Injury Using Baseball Tracking Data in MLB

Fumitake Sakaori¹, Hiroto Enjoji², Yuto Takemori²,
Shintaro Nishizuka² and Ibuki Hoshina³

¹Faculty of Science and Engineering, Chuo University

²Graduate School of Science and Engineering, Chuo University

³The Center for Data Science Education and Research, Shiga University

The incidence of ulnar collateral ligament (UCL) reconstruction surgeries among baseball pitchers has increased in recent decades. Despite the importance of preventing UCL injuries, there is as yet no scientific consensus regarding the risk factors for such injuries.

In this paper, we reconsidered candidate risk factors for UCL injuries, referring to the opinions of an amateur pitcher and a sports doctor, and then obtained adjusted odds ratios for selected risk factors via a logistic regression model and stepwise variable selection using AIC. The results revealed the following risk factors: for starting pitchers, smaller repertoire of pitch types, horizontal release location farther from the body, and a greater mean pitch count per game; and for relief pitchers, smaller repertoire of pitch types, horizontal release location farther from the body, greater mean pitch speed of fast balls, and fewer days between consecutive games. These results support previous studies of the risk factors of UCL injuries, and provide important suggestions regarding pitch count per game and mound interval for both starter and relief pitchers.

Covariate Balancing Propensity Score を 用いた, スクイズ作戦の有効性の解析

中村 知繁¹・南 美穂子²

(受付 2017 年 1 月 15 日; 改訂 4 月 30 日; 採択 8 月 28 日)

要 旨

近年, スポーツではデータの取得が盛んに行われており, 特にアメリカの大リーグの試合データは全打席一球ごとの記録が公開されている. 本論文では, このデータを用いて野球におけるスクイズ作戦が得点する確率にあたる影響について Covariate Balancing Propensity Score (CBPS, Imai and Ratkovic, 2014) を用いて解析を行った結果を報告する. 特に, 本解析においては, どうしても 1 点が欲しい状況である得点差が 0 点または 1 点の場合のスクイズ作戦の有効性に焦点を当てる (ただし満塁の場合は除く). スクイズ作戦が得点する確率にあたる影響を推定するためには, スクイズ作戦をとった場面と, とらなかった場面を比較するというのが簡単な方法である. しかし, スクイズ作戦をとるかとらないかはランダムではなく, 打者や投手の情報といった共変量に依存して決定されるのが一般的であるため, 単純な 2 群の比較を行ってしまうと, 推定されたスクイズ作戦の影響に共変量の影響が含まれてしまい, 適切に推定を行うことができない. そこで, 本論文では傾向スコアを用いて共変量調整を行った上で, スクイズ作戦が得点する確率に与える影響を推定した. さらに傾向スコアの推定をロジスティック回帰モデルを用いてスクイズ作戦の影響を推定した場合と, Covariate Balancing Propensity Score を用いてスクイズ作戦の影響を推定した場合の結果を比較した. その結果, CBPS を用いて影響を推定する方が, ロジスティック回帰モデルを用いて傾向スコアを推定した場合に比べて, 比較する 2 つの群の共変量の分布の平均がよく釣り合い, スクイズ作戦が得点する確率に与える影響を安定して推定できることがわかった. また, CBPS を用いた解析からスクイズ作戦は平均の意味で 18.2% 得点する確率を上昇させることがわかった.

キーワード: 野球, スクイズ作戦, 因果推論, 共変量調整, Covariate Balancing Propensity Score.

1. はじめに

近年スポーツの世界では, プレイデータの取得が盛んに行われるようになり, データを用いた作戦や戦術の評価を行うことができるようになりつつある. 特に, 大リーグのデータは, 過去の試合記録が 20 年分以上公開されており, これらのデータを用いて野球における戦略や戦術について, 統計的に評価をすることが可能になりつつある.

野球における戦略には, 手段の有効性の検証が十分になされていないものも多く, 野球ファ

¹ 慶應義塾大学大学院 理工学研究科: 〒 223-8522 神奈川県横浜市港北区日吉 3-14-1

² 慶應義塾大学 数理科学科: 〒 223-8522 神奈川県横浜市港北区日吉 3-14-1

ンや、解説者などの間でも意見が割れる戦術が存在する。そのうちの代表的な作戦が、バント作戦である。バント作戦にはいくつかの種類があり、よく知られたものでは、ノーアウト、または1アウトかつランナー1塁の状況において、バントを行い、打者はアウトになるが、ランナーを得点圏に進めるという犠牲バント作戦や、ノーアウト・1アウトでランナーが3塁にいる状況で、バントをし、ランナーを生還させるスクイズバント作戦(以下、スクイズ作戦)などがある。これらの作戦は、成功した時の華やかさがあり、とくにスクイズ作戦については、成功すると歓声上がる。一方で、そのような華やかな側面があるために、実際には失敗しているケースも少なからずあるが、スクイズ作戦の有効性について過剰に評価している可能性もある。そこで、本論文では、スクイズ作戦に焦点を絞り、その有効性について解析した結果を報告する。

野球におけるバント作戦の評価を行った文献には、Albert and Bennet (2001)や、仁志・鳥越(2012)などがある。これらの文献では、選手の打率と犠牲バントの成功確率を計算した上で、得点する確率の期待値を計算し、犠牲バントの有効性について解析している。また、Lee (2011)では確率的フロンティア分析の手法を用いて、犠牲バントや盗塁などを包括したスモールベースボール戦略の有効性を、得点数を犠牲バントや盗塁数で回帰することによって推定している。この他にも、論文という形式は取られていないが、日本経済新聞が「日本でも「脱バント」が浸透するか」(丹羽, 2011)という連載を行っていたり、アメリカにおいても、Turner Sportsというメディアが「Explaining Why the Bunt Is Foolish in Today's MLB」(Zachary, 2013)という記事を掲載していることから、バント作戦については興味関心が高いことがうかがえる。これらの記事で話題にしているのは、犠牲バント作戦であり、バント作戦をとった場面と、取らなかった場面での進塁率や、生還率について単純な標本平均を用いて比較し、その結果、犠牲バントは有効な作戦とは言えないという結論を導いている。

これらの手法におけるバント作戦の有効性の評価は、単純な2群比較をもとにした期待値の比較や、回帰係数を因果効果として解釈し行われている。では、果たしてこれらの解析によって犠牲バント作戦や、今回の対象であるスクイズ作戦の有効性を解析することはできるだろうか？

バントの有効性について議論するためには、有効性とは何かを定義する必要がある。そのために以下の記号を導入する。サンプル(場面) i に対して、 T_i をバント作戦を行ったかどうかの処置変数、 Y_i を得点したかどうかの2値の結果変数、 X_i をバント作戦及び、結果変数に関係する状況を表す共変量ベクトルとする。例えば、共変量としては、打者や投手の情報が挙げられる。さらに、 $Y_i(1)$ をバントをした時に観測される結果変数、 $Y_i(0)$ をバントをしなかった時に観測される結果変数とすると、結果変数 Y_i は

$$(1.1) \quad Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$$

と表すことができる。実際、同一の標本 i から $Y_i(1), Y_i(0)$ の両方が観測されることはなく、バント作戦をとったかどうかで、いずれか一方が観測される。ここで、バント作戦の有効性を、バントをした場合と、バントをしなかった場合に観測された結果の差の期待値であると考えることにすれば、バント作戦の有効性は以下のように定義することができる。

$$(1.2) \quad \mu = E[Y_i(1) - Y_i(0)]$$

この μ を平均処置効果(Average Treatment Effect; ATE)と呼ぶ。この考え方は統計的因果推論の枠組みの1つであり、Rosenbaum and Rubin (1983), Hirano et al. (2003), 星野 (2009)などが詳しい。

ここで、先ほどの単純な標本平均を用いた比較は、上記の記号を用いて以下のように表さ

れる。

$$(1.3) \quad \frac{1}{N_1} \sum_{i=1}^N T_i Y_i - \frac{1}{N_0} \sum_{i=1}^N (1 - T_i) Y_i$$

ここで、 N は観測されたデータのサンプルサイズである。また、 N_1 はスクイズ作戦をとった群のサンプルサイズであり、 N_0 はスクイズ作戦をとらなかった群のサンプルサイズである（ただし、 $N = N_1 + N_0$ である）。この推定値は漸近的 ($N_0, N_1 \rightarrow \infty$) に、 $E[Y_i(1)|T_i = 1] - E[Y_i(0)|T_i = 0]$ へと収束するため、これはバント作戦の有効性を、平均処置効果の意味で表していると考えすることはできない。このような理由から、単純平均を用いた比較では、バントの有効性を正しく推定できていない。

そこで、中村・南 (2016) では、このような問題へ対処するために、傾向スコア (Rosenbaum and Rubin, 1983) を用いて、大リーグのデータから犠牲バントの因果効果の推定を行った。傾向スコア $\pi(X_i)$ とは、共変量が与えられた元で処置を受ける確率であり以下で定義される。

$$(1.4) \quad \pi(X_i) = P(T_i = 1|X_i)$$

本論文では、投手や打者の情報(共変量)が与えられたもとで、スクイズ作戦をとる確率に対応する。

さて、Rosenbaum and Rubin (1983) では、以下の強く無視可能な割り付けの仮定(1.5)が成立するとき、傾向スコアを用いることで、 μ の推定が可能であることを示した。

$$(1.5) \quad (Y_i(1), Y_i(0)) \perp\!\!\!\perp T_i | X_i$$

これは、スクイズ作戦をとるか、とらないかに関係する変数 X_i を条件付けたもとで、得点のインジケータである潜在結果変数と、バント作戦を取るか取らないかの割り付け変数が独立になることを意味する。強く無視可能な割り付けの仮定が成立するもとは、傾向スコアの真値が既知であるとき、式(1.6)で定義される推定量は μ に対する一致推定量となる。式(1.6)で定義される推定量を、傾向スコアによる逆重み付け推定量 (Inverse Probability Weighted estimator, IPW 推定量) と呼ぶ。

$$(1.6) \quad \hat{\mu} = \frac{\sum_{i=1}^N \frac{T_i Y_i}{\pi_i}}{\sum_{i=1}^N \frac{T_i}{\pi_i}} - \frac{\sum_{i=1}^N \frac{(1 - T_i) Y_i}{1 - \pi_i}}{\sum_{i=1}^N \frac{1 - T_i}{1 - \pi_i}}$$

ただし、この推定量は傾向スコアが既知であるもとで定義されているが、解析では傾向スコアは未知であることが多い。そのため、実際にはデータから推定された傾向スコアを用いて推定量を計算する。一般的には、傾向スコア π_i を p 次元のパラメータベクトル $\beta \in \mathbb{R}^p$ を用いてパラメータ化し、データからパラメータを推定することで傾向スコアの推定値を得る。ここで、 β でパラメータ化された傾向スコアを $\pi_\beta(X_i)$ と表すことにする。

$\pi_\beta(X_i)$ を推定する際には、 $\pi_\beta(X_i)$ に対して式(1.7)を仮定したロジスティック回帰モデルが用いられることが多い。以降、ロジスティック回帰モデルを用いた傾向スコアの推定値とは、最尤法を用いて得られたパラメータの推定値を $\pi_\beta(X_i)$ に代入したものを意味することとする。

$$(1.7) \quad \pi_\beta(X_i) = \frac{\exp(X_i^T \beta)}{1 + \exp(X_i^T \beta)}$$

中村・南 (2016) においても、傾向スコアをロジスティック回帰モデルを用いて推定し、犠牲バントが得点する確率に与える影響について解析を行っている。しかしながら、傾向スコアを用いた解析においては、傾向スコアに対するモデルを誤特定した場合には、IPW 推定量の一致

性が失われるという点に注意が必要である。特に、Kang and Schafer (2007)では、傾向スコアに対するモデルの誤特定が、推定値に大きな影響を与えることをシミュレーションによって確認している。実際、野球の試合においてスクイズ作戦をとるかどうかを決定する際、投手や打者の情報に依存していることは先ほども指摘したが、どのような関数の形で依存しているかを正確に特定するのは困難である。そのため、傾向スコアに対するモデルは多少の誤特定を含むと考えるのが自然であり、傾向スコアに対するモデルを誤特定をしているならば IPW 推定量の一致性が損なわれる可能性がある。

このような問題に対処するため、本論文では傾向スコアに対するモデルが、真のモデルからそれほど大きくずれていない場合に、結果を頑健に推定することができる Covariate Balancing Propensity Score (Imai and Ratkovic, 2014; CBPS)を用いる。CBPS は、傾向スコアの推定で用いられる方程式が、両群の共変量分布を釣り合わせる重みであると解釈できる点に着目して傾向スコアを推定する手法である。CBPS の頑健な性質は、シミュレーションによっても確かめられている。

本論文における解析では、CBPS の釣り合いの関数として、Imai and Ratkovic (2014)においても使用される、ロジスティック回帰モデルのスコア方程式から導出される釣り合いの関数と、共変量の 1 次のモーメントの関数の両方を用いて推定を行った。その結果、ロジスティック回帰モデルを用いて傾向スコアを推定した場合に比べて、CBPS を用いた場合のほうが、推定された傾向スコアで重み付けされた 2 群の共変量の分布が平均の意味で釣り合った。よって、今回の解析で推定されたスクイズ作戦が得点する確率に与える影響は CBPS を用いる方が、バイアスが小さく推定されていると考えることができる。

本論文の構成は以下の通りである。まず、第 2 節でスクイズの有効性を推定するために用いたデータと、その加工手順について説明したうえで、スクイズ作戦をとった場面と、とらなかった場面の単純な比較を行った結果について述べる。第 3 節では、ロジスティック回帰モデルを用いて傾向スコアを推定し、スクイズ作戦の有効性を推定した結果と、その問題点について指摘する。続く、第 4 節では第 3 節での問題点を踏まえて、CBPS を用いてスクイズ作戦の有効性を解析した結果と、第 3 節の結果との比較を行い、第 5 節で議論をまとめる。

2. 解析対象のデータと、単純な標本平均による比較

本節では、解析に用いたデータの加工手順について述べた後で、スクイズ作戦の得点する確率に与える影響を、スクイズ作戦をとった場面と、とらなかった場面の単純な 2 群の比較を行った結果について述べる。

2.1 解析対象となるデータの前処理

まず本論文で扱うデータは、2006 年～2014 年の大リーグの試合データである(全サンプルサイズは 15 万 8491 であった)。このデータは、<http://www.retrosheet.org/>から取得することが可能である。

特に本論文ではスクイズ作戦の得点する確率に与える影響を解析する場面を、対戦チーム同士の点差が僅差であるような場合、すなわち得点差が $-1, 0, 1$ の状況のもとでのスクイズの有効性を解析するため、まずこのような場面にデータを絞り込んだ(このような場面は 7 万 7096 であった)。

スクイズ作戦には、状況の異なる 2 種類の作戦がある。それは、満塁におけるスクイズ作戦と、満塁ではない場面におけるスクイズ作戦である。これらの 2 つは、作戦としてのリスクが大きく異なるため、今回の解析においては、後者を対象とする。具体的には、「ランナー三塁」

表 1. スクイズをした場合としなかった場合で、得点した場面と、しなかった場面を集計した結果.

	得点が入らなかった	得点がいった	得点率
スクイズをした群	189	222	0.544
スクイズをしなかった群	12846	15343	0.540

表 2. スクイズをした群と、しなかった群の単純平均の差のばらつきをブートストラップ法で計算した結果.

	2.5%	25%	50%	75%	97.5%
スクイズをした群と、しなかった群の得点する確率の差	-0.054	-0.021	-0.004	0.013	0.043

「ランナー一塁・三塁」「ランナー二塁・三塁」の3つの場面を解析の対象とする(このような場面は5万9640であった).

ここからは、解析対象としたデータの加工の手順について具体的に述べる. 本解析で対象としたデータでは、前年に試合に出場していない選手(例えば、新人選手)が含まれている. このような選手に対しては、前年の試合記録を利用することができないため、解析の対象とすることが難しい. そこで、今回はこのようなデータは解析の対象からは取り除くこととした. 続けて、解析の対象となるデータに含まれている選手の中でも、前年度の出場記録が少ない選手を解析の対象から除外した. 具体的には、打席数が30打席に満たない打者や、投球イニング数が10イニングに満たない投手が含まれているデータを解析の対象から取り除いた(このような場面は3万8458であった).

さらに、野球の試合においては延長戦があるが、延長戦は1点で勝負が決まるため、通常のイニングとは性質が異なっている. そのため、延長戦のデータも解析の対象から取り除いている(このような場面は3万3718であった).

最後に、ボールカウントが3の場合には、打者はスクイズ(フォアボールを狙うほうが良い)をしないため、このような場面も解析対象から取り除いた. 以上のようなデータの加工を行った結果、解析の対象となったサンプルサイズは2万8600であった.

2.2 単純な2群の比較による、スクイズ作戦の得点する確率に対する影響の解析

ここでは、スクイズ作戦が得点する確率に与える影響を、単純な2群の比較を行って推定した結果について述べる. まず、スクイズ作戦をとった場合と、とらなかった場合のそれぞれでデータを分割し、その中で得点が入った場面数と、入らなかった場面数をまとめたものが表1である.

表1から、スクイズ作戦をとった場合の方が、スクイズ作戦をとらなかった場合に比べて、得点する確率は0.4%上昇することがわかる. 次に、ブートストラップ法を用いて、この推定値のパーセントイル点を計算した結果が、表2である.

表2より、スクイズ作戦が得点する確率に与える影響の95%信頼区間は $[-0.054, 0.043]$ となり、信頼区間は0を含むことがわかる. この結果、単純な2群の比較を用いた解析では、スクイズ作戦は得点する確率には有意水準5%で影響を与えていないということになる.

ところで、ランナーが3塁にいる場面では、得点する確率に影響を与える変数は、スクイズ作戦をとるかどうかだけではない. 例えば、ランナー3塁の状況での打者の打率や、長打率は、得点する確率(結果変数)に影響を与えることは野球経験者なら想像がつく. また、打率の高い打者であれば、打者にスクイズ作戦を取らせるよりも、バッティングを行わせたほうが成功確

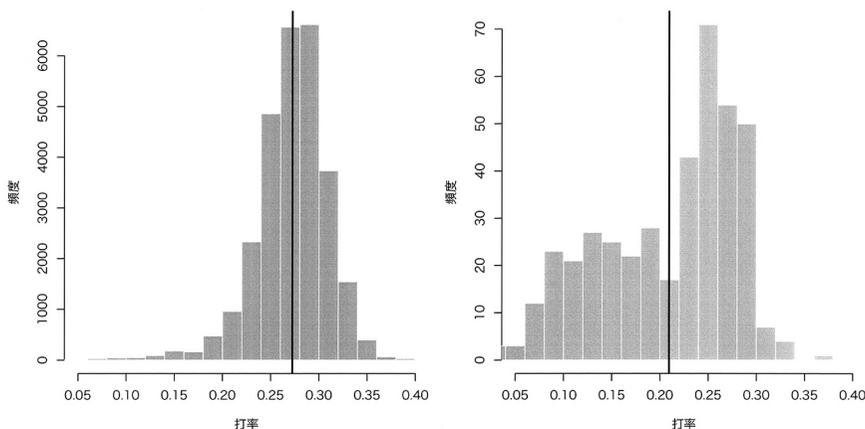


図 1. スクイズ作戦をとった群と、とらなかった群の打率の分布。右：スクイズ作戦をとった群，左：スクイズ作戦をとらなかった群。実線はそれぞれの群の打率の平均値。

率が高くなると考え、スクイズ作戦を取らせないようにするかもしれない。つまり、打率の変数は処置にも関連している可能性がある。ここで、例に上げた打率と長打率のスクイズ作戦を取った群と取らなかった群のヒストグラムは図 1 である。

図 1 から、スクイズした場面と、スクイズを行っていない場面では、打率の分布が異なっていることが視覚的に確認でき、平均での差を計算すると 0.063 であった。よって、単純な 2 群の比較によるスクイズ作戦の影響の推定では、スクイズ作戦の選択と、得点する確率の両方に影響を与えていると考えられる変数が存在しているにもかかわらず、その影響を無視して推定が行われている可能性が高く、スクイズ作戦のみの影響を推定できていないと考えられる(ここでは、スクイズ作戦の選択に関係する代表的な変数の 1 つである打率の分布について紹介したが、これら以外の変数のヒストグラムについては、補足資料として本論文の最後に記載した URL に掲載している)。

これらの問題に対処するためには、スクイズ作戦をとるかとらないかと、得点する確率の両方に影響を与える変数の影響を調整する(共変量調整を行う)必要がある。次節以降の解析では、傾向スコアを用いて共変量の影響を取り除くことで、スクイズ作戦の影響を推定した結果を述べる。

3. ロジスティック回帰モデルを用いた傾向スコアの推定と、その解析結果

本節では、まず傾向スコアで用いた共変量の選択について述べる。そのあと選択した共変量を用いて傾向スコアをロジスティック回帰モデルで推定した結果について述べ、最後に推定した傾向スコアを用いてスクイズ作戦の因果効果を推定した結果と、ロジスティック回帰モデルを用いた傾向スコアの推定の問題点を指摘する。

3.1 傾向スコアの推定に用いた共変量の選択

ここでは、傾向スコアを推定する際にモデルに含めた共変量について述べる。第 1 節で述べたようにモデルに含める共変量は、強く無視可能な割り付けの仮定 (1.5) が成立するように選択する必要がある。解析では、観測されている変数のうち、以下の変数を調整する変数として選択した。

共変量の選択においては、まず得点する確率を上昇させる(もしくは下降させる)可能性のあ

表 3. 傾向スコアの推定に用いた共変量.

基本情報	打者に関する共変量	投手に関する共変量
イニング	代打かどうか	被安打率
イニングの表・裏	打順	セーブ率
アウトカウント	打率	失点率
走者の位置	長打率	奪三振数
得点差	次の打者の打率	
打者と投手の利き腕	次の次の打者の打率	
	ホームラン率	
	打点率	
	四死球率	
	敬遠率	
	三振率	
	犠打率	
	犠飛率	
	盗塁率	
	死球率	

る変数を選択した。例えば打者の情報に関しては、打点や長打率といった変数は、高いほど打席でヒットを打つ確率が上昇するため、結果変数に関係する変数であると考えられる。また、打率や長打率は結果変数に対して影響を与えることはよく知られた事実であり、スクイズ作戦を指示する監督もよく知っているはずであるから、この変数はスクイズ作戦を取るかどうかを決定する際に考慮されていると考えるのが自然である。よって、打率と長打率はスクイズ作戦を取るかどうかの処置へも影響を与える変数である。

また対戦投手に関する情報も重要である。被安打率が高い打者であれば、打率が低い打者でも打てる可能性が高まり、一方で奪三振率が高い投手では打率が高い打者でもヒットを打てる可能性が小さくなると思えることができる。よって、結果変数に影響を与える変数である。一方で、このような情報も監督は考慮してスクイズ作戦を取るかどうかを選択するため、これはスクイズ作戦をとるかどうかの処置変数へも影響を与えると思えることができる。

上記で上げていない表 3 に掲載した共変量も、結果変数へと影響を与えると思えられる変数であり、その情報をもとにスクイズ作戦をとるかどうかを選択していると思えると、これらの変数についても調整するのが妥当であると判断し、解析においては調整することとした。また、これらの変数は昨年度のプレイ成績に基づいているため、処置後変数ではない点をここで指摘しておく。

最後に、因果推論における共変量の選択は、強く無視可能な割り付けを成立させるように選択することが必要であるが、この仮定が潜在結果変数を含むため検証することはできない。よって、ここで述べた共変量を調整することによって交絡を完全に除去することができると思えるのは考えにくい。しかしながら、上記で述べた変数は結果変数にも、処置変数へも影響を与えていると思える変数であり、これらが比較する 2 群で異なることが原因で平均処置効果がバイアスを持って推定されてしまう (Greenland et al., 1999) ため、これらについて調整することは重要であると思え、解析では表 3 にあげる共変量について調整を行うこととした。

3.2 ロジスティック回帰モデルを用いた傾向スコアの推定

ここでは、選択した共変量を用いたロジスティック回帰モデルをデータに当てはめて推定された傾向スコアの推定結果について述べる。一般的に傾向スコアを推定するにはロジスティック回帰モデルがよく用いられる。ここでも、同様にロジスティック回帰モデルを用いて

表 4. ロジスティック回帰モデルを用いて推定した、各群の傾向スコアの逆数(重み)のパーセンタイル点. 基準化した重みは、重みの和が各群サンプル数になるように調整をかけている.

パーセンタイル点	5%	25%	50%	75%	95%	99%	99.9%
スクイズした群	1.78	3.15	12.29	36.66	216.85	637.28	9277.42
スクイズしなかった群	1.00	1.00	1.00	1.01	1.06	1.19	6.24
スクイズした群 (基準化)	0.03	0.05	0.18	0.53	3.12	9.17	133.54
スクイズしなかった群 (基準化)	0.99	0.99	0.99	0.99	1.04	1.18	6.15

表 5. ロジスティック回帰モデルを用いて傾向スコアを推定した場合の因果効果の推定量のブートストラップ法によるパーセンタイル点の計算結果.

	2.5%	25%	50%	75%	97.5%
平均処置効果 (ATE)	-0.293	-0.092	0.043	0.261	0.330

傾向スコアの推定を行った. すなわち, 傾向スコア $\pi(X_i)$ に対して, 以下のモデルを仮定した.

$$(3.1) \quad \pi(X_i) = \pi_{\beta}(X_i) = \frac{\exp(X_i\beta)}{1 + \exp(X_i\beta)}$$

ここで, X_i は選択した共変量のベクトルである.

ここで, データにモデル(3.1)を当てはめて得られた β の最尤推定値を $\hat{\beta}$ とする. このとき, $\hat{\beta}$ を代入して得られる傾向スコアの推定値 $\pi_{\hat{\beta}}(X_i)$ の逆数の, パーセンタイル点をまとめたものが表 4 である.

ここで, スクイズを行わなかった群の傾向スコアの推定値の逆数が, スクイズしなかった群の傾向スコアの推定値の逆数より全体として大きくなっているが, これは 2 つの群のサンプルサイズの差によるものである. また, 傾向スコアとは与えられた共変量 X_i のもとでスクイズ作戦をとる確率であるから, その推定値の逆数が大きな値をとるとするのは「ほとんどスクイズ作戦を取らないところでスクイズ作戦をとった」と解釈することができる. 例えば, スクイズ作戦をとった群で, 99.9 パーセンタイル点の標本の傾向スコアの推定値の逆数が 9000 を超えているが, ほとんどスクイズ作戦をとらない場面で, スクイズ作戦をとった標本である. 本節の最後に, 少数の標本の傾向スコアの逆数が, 大きな値をとる場合の問題点について述べる.

3.3 推定した傾向スコアを用いたスクイズ作戦の影響の推定

ここでは, 推定した傾向スコアを用いてスクイズ作戦が得点する確率に与える因果効果を推定した結果について述べる. まず, 第 1 節で述べたように, IPW 推定量は強く無視可能な割り付けが成立するもとの, バントの有効性, 式(1.2), に対する一致推定量となる. 推定した傾向スコアに基づいて IPW 推定量を計算した結果, スクイズ作戦の平均処置効果は -0.007 となった. この推定量はすべての場面でスクイズ作戦をとった場合と, とらなかった場合に得点する確率の差の推定値であるから, 全場面でスクイズ作戦をとったとすると平均して約 0.7% 得点する確率が低下すると解釈することができる. また, この推定値に対する信頼区間を求めるためブートストラップ法を用いたところ, 結果は表 5 のようになった.

この結果から, 95% 信頼区間が $[-0.293, 0.330]$ となり信頼区間が 0 を含むため, 5% の有意水準では影響があるとはいえないことがわかった.

さて統計的因果推論を行う際には, 平均処置効果の他にも, 処置群における平均処置効

表 6. ロジスティック回帰モデルを用いて傾向スコアを推定した場合の、スクイズ作戦の ATT 及び、ATU のブートストラップによるパーセンタイル点の計算結果.

	2.5%	25%	50%	75%	97.5%
スクイズした群における平均処置効果	0.065	0.102	0.120	0.140	0.174
スクイズしなかった群における平均処置効果	-0.297	-0.094	0.042	0.264	0.334

果 (Average Treatment Effect on the Treated; ATT) と、対照群における平均処置効果 (Average Treatment Effect on the Untreated; ATU) にも興味があることが多い. ATT と ATU はそれぞれ式 (3.2), (3.3) で定義される.

$$(3.2) \quad \mathbb{E}[Y_1 - Y_0 | T = 1]$$

$$(3.3) \quad \mathbb{E}[Y_1 - Y_0 | T = 0]$$

ATT は、処置が行われたケースで、もし処置を行っていなかったら、どの程度結果に変化があったかに対応する推定量であり、ATU は処置が行われなかったケースで、もし処置を行っていなかったらどの程度結果に変化があったかに対応する推定量である. つまり、ATT はスクイズを行った場面で、もし行っていなかったら、得点する確率がどの程度変化するかに対応する. ここで、傾向スコアの真値が既知で、強く無視可能な割り付けの仮定が成立するとき式 (3.4) と式 (3.5) はそれぞれ、ATT 及び ATU に対する一致推定量となる.

$$(3.4) \quad \mathbb{E}[Y_1 - Y_0 | T = 1] = \bar{y}_1 - \frac{\sum_{i=1}^N \frac{(1 - T_i)\pi_i Y_i}{1 - \pi_i}}{\sum_{j=1}^N \frac{(1 - T_j)\pi_j}{1 - \pi_j}}$$

$$(3.5) \quad \mathbb{E}[Y_1 - Y_0 | T = 0] = \frac{\sum_{i=1}^N \frac{T_i(1 - \pi_i)Y_i}{\pi_i}}{\sum_{j=1}^N \frac{T_j(1 - \pi_j)}{\pi_j}} - \bar{y}_0$$

ここで、 \bar{y}_1, \bar{y}_0 は、それぞれ処置群 $T = 1$ と対照群 $T = 0$ における結果変数の標本平均である. これらの推定量を傾向スコアの推定値を用いて計算した結果、スクイズ作戦をとった場面で、スクイズをとらなかった場合の平均処置効果は 0.119 であり、逆にスクイズ作戦をとらなかった場面で、スクイズ作戦をとった場合の平均処置効果は 0.005 であった.

この結果から、もしスクイズ作戦をとった場面でとらなかったとすると、平均の意味で 11.9% の得点する確率の低下となることがわかる. 一方で、もしスクイズ作戦をとらなかった場面でとったとすると得点する確率は平均の意味で 0.5% 上昇することがわかる. さらに、先ほどと同様に、これらの推定量の信頼区間を 1500 回のブートストラップ法によって求めたものをまとめたのが表 6 である.

この結果から、スクイズ作戦をとった場面で、もしスクイズ作戦をとらなかった場合には、スクイズ作戦が得点する確率に与える影響の 95% 信頼区間は [0.065, 0.174] となり、5% の有意水準で影響があるといえる. 一方で、スクイズ作戦をとらなかった場面で、もしスクイズ作戦をとった場合には、スクイズ作戦が得点する確率に与える影響の 95% 信頼区間は [-0.297, 0.334] となり、5% の有意水準では影響があると結論付けることはできないという結果となる.

ここで、傾向スコアを用いた解析を行ったモチベーションは、比較する 2 群の共変量が得点する確率に与える影響を、傾向スコアの逆数で各標本に重み付けすることで取り除くことであった. 図 1 で確認した通り、スクイズ作戦をとった群と、スクイズ作戦を取らなかった群の打率の分布は視覚的に確認可能な程度には異なっている. 傾向スコアによって、この分布の差が調整されたと考えると、傾向スコアの逆数で重み付けした打率の分布は比較する 2 群で同じ

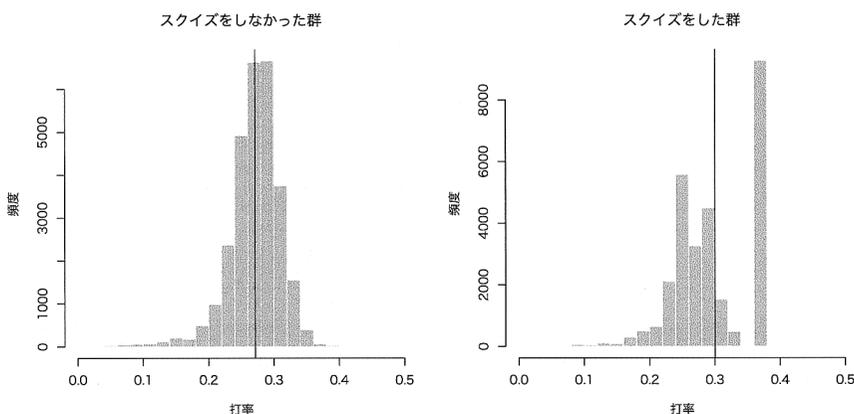


図 2. スクイズ作戦をとった群と、とらなかった群の打率の分布を傾向スコアの推定値で重み付けしたもの. 右: スクイズ作戦をとった群, 左: スクイズ作戦をとらなかった群. 黒の実線は重み付けられた打率の平均値である.

ような形状となるのが望ましい. 実際, 重みをつけたヒストグラムは図 2 となる(ここでは, スクイズ作戦の選択に関する代表的な変数の 1 つである打率の分布について紹介したが, これら以外の変数の重み付けヒストグラムについては, 補足資料として本論文の最後に記載した URL に掲載している).

図 2 より, ロジスティック回帰モデルによって推定された傾向スコアを用いて重み付けされた打率の分布は平均の意味で釣り合っているとは言い難い. また, スクイズ作戦をとった群のヒストグラムでは右端が大きく跳ねており, この周辺の標本に非常に大きな重みがかかっているが, スクイズ作戦をとらなかった群のヒストグラムの右端はなだらかに減少していることが確認できる. 実際, スクイズ作戦をとった群の右端で大きく跳ねているのは 1 つの標本に大きな重みがかかっているためである. この 1 つの標本にかかる重みは, スクイズを行った群の傾向スコアの逆数の和のうち 32.5% に対応する重みとなっている. このように 1 つの標本に大きな重みがかかるとき, 式 (1.6) より IPW 推定量の推定値は 1 つの標本の影響を強く受けることになる. ここでのスクイズの解析では, スクイズ作戦をとった場合の結果の 32.5% を 1 つの標本の結果によって説明されることになり, 小さな重みのかかった場面の結果が過少に評価されて結果が推定されることになる. 以上の議論から, ロジスティック回帰モデルを用いて傾向スコアを推定した上で, 推定した傾向スコアを用いてスクイズ作戦の因果効果を推定した場合には, 少なくとも打率の影響が平均の意味で取り除かれていない可能性が示唆された. また, ロジスティック回帰モデルを用いた場合には, 稀なケースに大きな重みがかかりすぎるという問題があることがわかった.

ロジスティック回帰モデルを用いた場合に, 稀なケースに大きな重みをかけすぎるという問題はブートストラップ法を用いて推定量の信頼区間を計算した際にもわかる.

図 3 は, ブートストラップ標本から計算された ATE の推定値のヒストグラムである. このヒストグラムからは 2 つの峰があることが確認できる. これはブートストラップ標本に, 稀にしかスクイズしない場面でスクイズしたかしなかったかで, 傾向スコアの推定値が変化し, 結果として得られるスクイズ作戦の因果効果が大きく変化することを示している. すなわち, ロジスティック回帰モデルで傾向スコアを推定するというのは, スクイズ作戦の因果効果を推定する場合には結果が安定的に得られないということを示唆している.

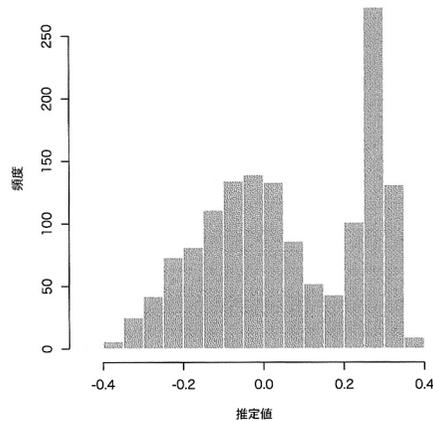


図 3. 傾向スコアの推定にロジスティック回帰モデルを用いた場合の平均処置効果の推定値のブートストラップ標本のヒストグラム。

ここまでの議論から、因果効果を推定する際には、スクイズ作戦及び得点する確率に影響する共変量の分布の平均は揃っている方が望ましく、結果はリサンプリングに対して安定的である方が望ましい。そこで、これらのロジスティック回帰モデルを用いた傾向スコアの推定から生じる問題点を解決するために、次節では Covariate Balancing Propensity Score を用いることで、これらの問題を解消・緩和して因果効果を推定した結果について述べる。

4. CBPS を用いた解析

本節では、Covariate Balancing Propensity Score (以下、CBPS) を用いてスクイズ作戦の因果効果を推定した結果について述べる。傾向スコアをロジスティック回帰モデルで推定した場合には、少数の標本に大きな重みが与えられ過ぎるという問題と、平均の意味での両群の共変量の釣り合いが取れない問題などがわかった。本節では CBPS について述べた後で、CBPS を用いることでロジスティック回帰モデルを用いて傾向スコアを推定した場合に指摘した問題点を解消しつつ、スクイズ作戦が得点する確率へ与える影響を推定することができることを指摘する。

4.1 Covariate Balancing Propensity Score とは

CBPS (Imai and Ratkovic, 2014) とは、式 (4.1) を満たす $\pi_\beta(X_i)$ として定義される。

$$(4.1) \quad \mathbb{E} \left[\left(\frac{T_i}{\pi_\beta(X_i)} - \frac{(1-T_i)}{1-\pi_\beta(X_i)} \right) f(X_i) \right] = 0$$

ここで関数 f は解析者が選択する共変量の関数である。この $\pi_\beta(X_i)$ が Covariate Balancing (共変量釣り合い) と呼ばれる理由は、式 (4.1) を式 (4.2) のように変形することにより $\pi_\beta(X_i)$ が、比較する 2 つの群の共変量の分布のモーメントを釣り合わせていると解釈できるためである。

$$(4.2) \quad \mathbb{E} \left[\frac{T_i}{\pi_\beta(X_i)} f(X_i) \right] = \mathbb{E} \left[\frac{1-T_i}{1-\pi_\beta(X_i)} f(X_i) \right]$$

実際、ロジスティック回帰モデルによって傾向スコアを推定する際には、 $\pi_\beta(X_i)$ に対して式 (1.7) を仮定し、 f に対して $\partial \pi_\beta(X_i) / \partial \beta$ をおくことに対応する。すなわち、ロジスティック回

帰モデルを用いて傾向スコアを推定した場合には、推定された傾向スコアは、 $\partial\pi_\beta(X_i)/\partial\beta$ を2つの群で平均の意味で釣り合わせる重みであると解釈できる。

次に、式(4.1)を満たす β の推定について述べる。パラメータ β をデータから推定する際には、式(4.3)を β について解く必要がある。

$$(4.3) \quad g(\beta) = \sum_{i=1}^N \left[\left(\frac{T_i}{\pi_\beta(X_i)} - \frac{(1-T_i)}{1-\pi_\beta(X_i)} \right) f(X_i) \right] = 0$$

しかしながら、この方程式(4.3)は β の次元と方程式の数が一致するかどうかは f の選択に依存し、一般的には解が存在しない。また、CBPSによる傾向スコアの推定では、関数 f の選択によって、パラメータの次元よりも、方程式の数が多くなることが多い。そこで、この方程式を直接的に解くのではなく、式(4.4)によって β の推定値を計算する。

$$(4.4) \quad \hat{\beta} = \underset{\beta}{\operatorname{argmin}} g(\beta)^T \Sigma(\beta) g(\beta)$$

ここで、 $\Sigma(\beta)$ は共変量 X に依存する正定値行列であり、解析ではImai and Ratkovic (2014)の式(14)を用いた。このようにすることで、 β をデータから推定することが可能となり、傾向スコアの推定値を計算することができる。このようにして得られる推定量の性質については、Hansen (1982)やHansen et al. (1996)が詳しい。

4.2 CBPSを用いたスクイズ作戦の解析

ここまで、CBPSが共変量の関数 $f(X_i)$ を比較する2つの群で釣り合わせるという性質を持つことを述べた。この性質を利用して、第3節で傾向スコアの逆数で重み付けされた打率の分布の平均が釣り合わなかった問題を解消する。すなわち、共変量 X_i が与えられた元でスクイズ作戦をとる確率 $\pi_\beta(X_i)$ は、共変量釣り合いの関数 f として式(4.5)を選んでモデル化されると仮定する。

$$(4.5) \quad f(X_i) = \begin{pmatrix} s_\beta(X_i) \\ X_i \end{pmatrix}$$

ここで、

$$(4.6) \quad s_\beta(X_i) = \frac{1}{1 + \exp(X_i^T \beta)} \left(1 - \frac{1}{1 + \exp(X_i^T \beta)} \right)$$

である。 $s_\beta(X_i)$ はロジスティック回帰モデルを用いて傾向スコアを推定する際に2つの群で釣り合わせる関数である。ここで、 f として X_i を加えることは、スクイズ作戦をとるかとならないかと、得点する確率の両方に影響をあたえる共変量の分布の平均を釣り合わせる重みとして、傾向スコアを推定していることに対応する。 f として、式(4.5)を選択したもとの、データから傾向スコアを計算した結果を表7にまとめた。

この結果(表7)と、ロジスティック回帰モデルを用いて推定された傾向スコアの表4を比較すると、スクイズしなかった群の傾向スコアの逆数にはそれほど大きな違いが見取れない。一方で、スクイズした群の傾向スコアの逆数については、CBPSの方が5%点は大きく、50%点以降ではロジスティック回帰モデルを用いた方が大きく推定されていることがわかる。すなわち、CBPSの方が傾向スコアのばらつきが小さいことがわかる。

推定したCBPSを用いた平均処置効果(ATE)の推定結果は0.182となった。また第3節と同様に、スクイズ作戦をとった場面でスクイズ作戦をとらなかった場合の因果効果(ATT)を推定した結果は0.106、逆にスクイズ作戦をとらなかった場面でスクイズ作戦をとった場合の因果

表 7. 推定された、各群の CBPS の逆数(重み)のパーセンタイル点. 基準化した重みは、重みの和が各群サンプル数になるように調整をかけている.

パーセンタイル	5%	25%	50%	75%	95%	99%	99.9%
スクイズした群	1.85	3.16	14.10	34.19	149.51	409.61	1133.13
スクイズしなかった群	1.00	1.00	1.00	1.01	1.05	1.16	6.14
スクイズした群 (基準化)	0.05	0.09	0.36	0.89	3.85	11.79	28.01
スクイズしなかった群 (基準化)	0.98	0.99	0.99	1.00	1.04	1.16	5.89

表 8. CBPS を用いた場合の、スクイズ作戦の因果効果(ATE, ATT, ATU)のブートストラップによるパーセンタイル点の計算結果.

	2.5%	25%	50%	75%	97.5%
平均処置効果 (ATE)	0.092	0.162	0.203	0.240	0.300
処置群における平均処置効果 (ATT)	0.006	0.077	0.101	0.123	0.172
対照群における平均処置効果 (ATU)	0.093	0.165	0.206	0.244	0.303

効果(ATU)を推定した結果は0.185であった. さらに、それぞれの推定量の信頼区間を計算するために、1500回のブートストラップを行った結果を表8にまとめた.

表8より、平均処置効果の推定値の95%信頼区間は[0.092, 0.300]であることから、有意水準5%で得点する確率に与える影響は有意であるとわかる. よって、すべての場面でスクイズ作戦を行った場合にも得点する確率を上昇させることがわかる. 同様に、スクイズ作戦をとった場面でとらなかった場合の平均処置効果の95%信頼区間は[0.006, 0.172]であり、スクイズ作戦をとらなかった場面でスクイズ作戦をとった場合の平均処置効果の95%信頼区間は[0.093, 0.303]で、ともに5%の有意水準で有意であり、得点する確率を上昇させることがわかる. この結果は、第3節でロジスティック回帰モデルを用いた場合とは異なっている.

次に共変量の分布を傾向スコアの逆数で重み付けした分布について考える. 第3節での傾向スコアで重み付けされた共変量(打率)の分布は平均の意味で釣り合っていないかった. 図4は推

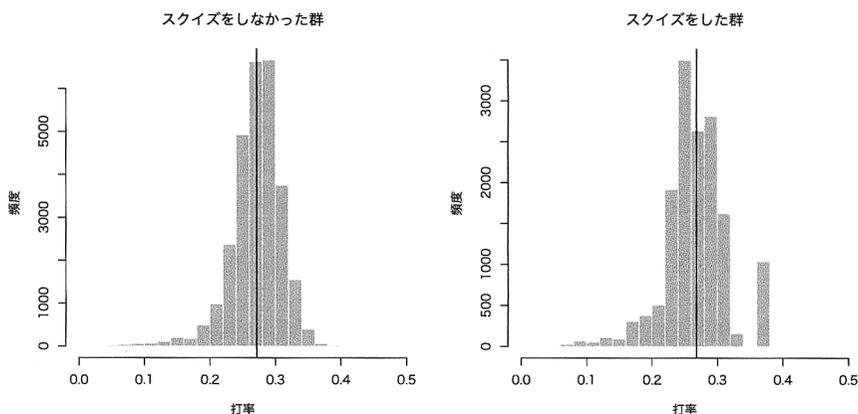


図 4. スクイズ作戦をとった群と、とらなかった群の打率の分布を CBPS の推定値で重み付けしたもの. 右:スクイズ作戦をとった群, 左:スクイズ作戦をとらなかった群. 黒の実線は重み付けられた打率の平均値である.

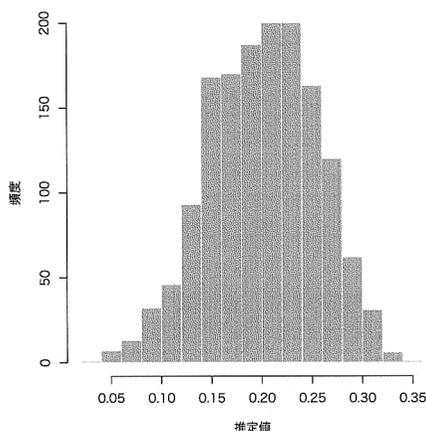


図 5. CBPS を用いた場合の平均処置効果の推定値のブートストラップ標本のヒストグラム。

定した CBPS で重み付けした共変量(打率)の分布である(ここでは、スクイズ作戦の選択に関係する代表的な変数の 1 つである打率の分布について紹介したが、これら以外の変数の CBPS による重み付けヒストグラムについては、補足資料として本論文の最後に記載した URL に掲載している)。

実際、スクイズ作戦をとった群と、とらなかった群の打率の重み付け分布の標本平均の差の絶対値は 0.003 であり、これは第 3 節で重み付けした分布の標本平均の差の絶対値 0.029 よりも小さいことがわかる。実際、視覚的にも図 2 よりも、図 4 の方が分布の形状が似ていることが確認できる。

ロジスティック回帰モデルを用いた場合には、ブートストラップ標本に基づいて計算されるスクイズの平均処置効果の推定値の分布が 2 つの峰を持つことが図 3 から確認できた。一方、CBPS を用いた場合のブートストラップ標本に基づく平均処置効果の推定量の分布は図 5 である。

2 つの図を比較すると、CBPS を用いた場合の分布の峰は 1 つになっているのがわかる。実際、ロジスティック回帰モデルを用いた場合、ブートストラップ標本に、スクイズ作戦をほとんどとらない場面をとったケースが含まれると 1 つの標本に大きな重みがかかり推定結果がブレやすくなる。しかし、CBPS では、スクイズ作戦をほとんどとらない場面をとったケースがブートストラップ標本に含まれたとしても、傾向スコアの値を大きくしすぎないように推定できていた。

以上の議論からスクイズ作戦の解析では、ロジスティック回帰モデルを用いると推定結果が不安定になるが、CBPS を用いることで影響を安定的に推定できることがわかった。さらに、ロジスティック回帰モデルを用いて傾向スコアを推定するとスクイズ作戦の有効性が得られなかったが、一方で CBPS を用いた場合には有効性が示された。また、スクイズ作戦をとらなかった場面で、もしスクイズ作戦をとった場合の有効性についてもロジスティック回帰モデルを用いて傾向スコアを推定した場合には示されなかったが、一方で CBPS を用いると有効性が示された。

5. まとめ

スクイズ作戦が得点する確率に与える影響は、得点差が 0 点または 1 点というどうしても 1

点が欲しい場面において、CBPS を用いた解析から得点する確率を上昇させる効果があることがわかった(ただし、満塁の場面を除く)。スクイズ作戦は、作戦の性質上、相手にスクイズ作戦をとることを察知されてしまうと、アウトカウントを増やしランナーを失うリスクのある作戦である。そのため、やらないであろうと想定される場面でやることで効果が生まれる。例えば、打率の高い打者が打席に立った際には「打つだろう」と想定するし、逆に打率が低い打者であれば1点を取るために「スクイズも警戒しよう」となると考えることができる。このような実際、スクイズを行った場面で傾向スコアが小さい、すなわちスクイズをほとんどしない場面での標本ではスクイズ作戦の成功率が高かった。このように考えれば、CBPS による解析の結果は直感的にも誤った結果ではない。ただし、この結果からすべての場面でスクイズをすれば良いという結論は早計である。なぜなら、スクイズ作戦は相手の意表をつく作戦であるから、すべての場面でスクイズ作戦をとるのは、もはやスクイズ作戦にはならないため、あくまでこの結果は「現状で行われているスクイズ作戦の選択」に対する評価であると考えるのが適切である。

次に、スクイズ作戦の解析において、第3節の結果と第4節の結果を比較することでCBPS を用いた方がよいと考えられる理由についてまとめる。まず、CBPS を用いて得られた結果は、ロジスティック回帰モデルを用いて傾向スコアを推定した場合には得られなかった結果である。CBPS の特に重要な性質は、第4節で述べたように、比較する2群で結果変数と処置の両方に影響を与える共変量の関数の重み付き分布のモーメントを釣り合わせる性質があることである。今回の解析では、共変量分布の1次モーメントが釣り合うように制約をつけることで、比較する2つの群の共変量の1次モーメントの差が得点する確率に与える影響を取り除いた。さらに、1次モーメントの釣り合いをとることで、ほとんどスクイズ作戦をとらない場面でスクイズ作戦をとった標本に大きすぎる重みを与えないようにすることができた。

また、ロジスティック回帰モデルを用いた共変量分布の調整結果である図2と、CBPS を用いた共変量分布の調整結果である図4を確認すると、CBPS を用いた場合のほうが2群の共変量の分布の形状が似ていることが視覚的に確認できる。Greenland et al. (1999)でも指摘されたように交絡が生じる原因は、比較する2群の背後にある共変量の分布が異なっていることである。このような理由からも、2群の共変量の分布をよく釣り合わせているCBPS を用いた場合のほうが、交絡の影響を取り除いてスクイズ作戦の効果を推定できていると考えられることができる。

さらに、傾向スコアを用いてデータからスクイズ作戦の因果効果を推定する際の課題として、共変量が観測されたもとでスクイズ作戦をとる確率に対するモデルは正確に特定できないという点である。実際、Kang and Schafer (2007)において、ロジスティック回帰モデルを用いて傾向スコアを推定した場合には、傾向スコアに対するモデルを誤特定すると、それが些細なモデルの誤特定であっても結果が安定しないことが指摘されている。一方で、CBPS は Imai and Ratkovic (2014)において小さなモデルの誤特定に対して結果を頑健に推定できることがシミュレーションで示されており、このような理由からも、スクイズ作戦の因果効果の推定にはCBPS を用いた方がよいことがわかる。

最後に、CBPS を用いて共変量を調整した上で推定した、処置群での平均処置効果(ATT)と対照群における平均処置効果(ATU)が異なっていることから、スクイズ作戦の効果は、共変量に依存する不均一性がある(すなわちスクイズ作戦の効果は共変量に依存している)可能性が示唆される。本論文で扱いきれなかったが、実際の戦術に因果推論の結果を反映させるためには、どのような状況のもとでスクイズ作戦がより効果的なのかについて解析を行う必要があると考える。例えば、周辺パラメトリックモデル(Rotnitzky and Robins, 1995; 星野, 2009)などを用いて、潜在結果変数の周辺構造を指定した上でスクイズの効果は推定すれば、どのような

状況のもとでスクイズ作戦が有効なのかについて更に深く解析できると考える。

また、本論文中で交絡の調整に用いた共変量である表3について、これらが調整すべきすべての交絡変数であるかの検討も行う必要がある。また、未観測の交絡が存在する場合の感度分析についても行う必要があるが、本論文では扱いきれていないので、これらの点については今後の課題としたい。

謝 辞

本稿を作成するに当たり、査読者の方から多くの詳細なコメントを頂いた。ここに記して感謝の意を表したい。なお本研究の一部は、2016年度KLL後期博士課程研究助成金(慶應義塾大学先端科学技術センター)及び、先端研究拠点事業(日本学術振興会: JSPS Core-to-Core Program)の助成を受けたものである。

付記

本論文におけるスクイズ作戦を行った場合と、行わなかった場合のそれぞれの共変量の分布、及び推定した傾向スコアの逆数によって重みづけし、調整したそれぞれの場合の共変量の分布についての詳細を<http://www.stat.math.keio.ac.jp/labs/mminami/research/>にこれを記載した。

参 考 文 献

- Albert, J. and Bennet, J. (2001). *Curve Ball: Baseball, Statistics, and the Role of Chance in the Game*, Springer, New York.
- Greenland, S., Pearl, J. and Robins, J. (1999). Confounding and collapsibility in causal inference, *Statistical Science*, **16**(1), 29–46.
- Hansen, P. (1982). Large sample properties of generalized method of moments estimators, *Econometrica*, **50**, 1029–1054.
- Hansen, P., Heaton, J. and Yaron, A. (1996). Finite-sample properties of some alternative GMM estimators, *Journal of Business & Economic Statistics*, **14**, 262–280.
- Hirano, K., Imbens, G. and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score, *Econometrica*, **71**, 1161–1189.
- 星野 崇宏 (2009). 『調査観察データの統計科学』, 岩波書店, 東京.
- Imai, K. and Ratkovic, M. (2014). Covariate balancing propensity score, *Journal of the Royal Statistical Society, Series B*, **76**, 243–263.
- Kang, Y. and Schafer, L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data, *Statistical Science*, **22**(4), 523–539.
- Lee, H. (2011). Is the small-ball strategy effective in winning games? A stochastic frontier production approach, *Journal of Productivity Analysis*, **35**, 51–59.
- 中村知繁, 南美穂子 (2016). 『傾向スコアを用いたバント効果の推定』, 岩波データサイエンス, Vol.3, 岩波書店, 東京.
- 仁志敏久, 鳥越規央 (2012). 『プロ野球のセオリー』, KKベストセラーズ, 東京.
- 丹羽政善 (2011). 日本でも「脱バント」が浸透するか, 日本経済新聞, 東京, <http://www.nikkei.com/article/DGXZZO23324240T10C11A2000000/>.
- Rosenbaum, P. and Rubin, D. (1983). The central role of propensity score in observational studies for causal effects, *Biometrika*, **70**, 41–55.

Rotnitzky, A. and Robins, M. (1995). Semiparametric regression estimation in the presence of dependent censoring, *Biometrika*, **82**, 805–820.

Zachary, R. (2013). Explaining Why the Bunt Is Foolish in Today's MLB, Bleacher Report, <http://bleacherreport.com/articles/1639658-explaining-why-the-bunt-is-foolish-in-todays-mlb>.

Effectiveness of the Squeeze Play Using Covariate Balancing Propensity Scores

Tomoshige Nakamura¹ and Mihoko Minami²

¹Graduate School of Science and Technology, Keio University

²Department of Mathematics, Keio University

Major League Baseball (MLB) has collected play-by-play data for the past 20 years. This data is available to the public. In this paper, we estimate the effect of a squeeze play on scoring using the covariate balancing propensity score (CBPS, Imai and Ratkovic, 2014) method. We focus on the case where the score difference is 0 or 1, except when the bases are loaded. A simple method is used to estimate the effect of a squeeze play on scoring. Specifically, sample averages are compared between two groups (attempting and not attempting a squeeze play). However, the decision to attempt a squeeze play is not random; it depends on the batter, pitcher, inning, etc. If these confounding variables are not considered, the estimated result will not represent the true effect of a squeeze play. In this paper, we estimate the effect of a squeeze play using a propensity score approach to adjust the effect of other variables. In the analysis, two types of estimation procedures for the propensity score are compared: the logistic regression model and the CBPS method. CBPS produces more balanced distributions of the covariates and the estimated effect of a squeeze play becomes more stable than using the logistic regression model to estimate the propensity score. CBPS indicates that a squeeze play has a positive effect on the scoring probability and increases the probability of scoring by 18.2%.

項目反応理論を用いた野球選手の 能力評価指標の提案

阿部 興¹・作村 建紀²・鎌倉 稔成²

(受付 2016 年 12 月 26 日; 改訂 2017 年 3 月 1 日; 採択 4 月 25 日)

要 旨

本研究は野球選手の、特に打撃の能力に関する新しい評価指標の提案を行う。打者の能力を評価する指標として、最も広く使用されるのは打率である。しかし、打率を一つの統計モデルと捉える場合、これは現実的でない仮定のもとに成立するものとも言える。打率はヒットを打つ確率がピッチャーの能力に関わらず常に一定であると見て計算されている。提案モデルはピッチャーによってヒットを打つ難しさが異なり、対戦するピッチャーが選手ごとに異なるという状況の下で、打撃の能力の選手間での比較を可能にする。我々が提案するモデルは、項目反応理論で用いられる 1 パラメータロジスティックモデル(ラッシュモデル)の拡張である。ラッシュモデルでは、潜在的な能力パラメータは各人一つとされているが、本論文では打者の能力に対応するマルコフパラメータを導入した。すなわち、我々は打席結果にマルコフ性を仮定して調子の波を表現する。これにより調子の波が前の打席の結果を受けて生じると解釈できる。モデルのパラメータの推定にはハミルトニアン・モンテカルロ法を用いる。パラメータ推定の安定性はシミュレーションによって評価する。提案手法の有益性は日本プロ野球の実際のデータを分析することで示す。

キーワード：階層ベイズモデル、MCMC、セイバーメトリクス、ロジスティックモデル。

1. はじめに

本研究は野球選手の、特に打撃の能力に関する新しい評価指標の提案を行う。選手の能力を公平かつ客観的に評価することは、球団を運営する上で重要な課題である。打者の能力を評価する指標として、最も広く使用されるのは打率である。次いで有名な指標として、長打率(SLG)、出塁率(OBP)、OPS などがある(Albert and Benett, 2003)。長打率は単打に 1、二塁打に 2、三塁打に 3、本塁打に 4 の重みをつけた、ヒット数の重み付き平均である。出塁率はヒットと四球と死球の数を足したものを、打数と四球と死球と犠飛の数を足したもので割った指標である。OPS は出塁率と長打率を足し合わせた指標である。Albert and Benett (2003)では、これらの指標をチームごとの平均的な打撃評価指標と得点との相関という観点から検討し、整理している。

より分析的な打者の能力評価指標として、Albert (2008)によるものがある。一般に打者には

¹ 中央大学大学院 理工学研究科：〒112-8551 東京都文京区春日 1-13-27

² 中央大学 理工学部：〒112-8551 東京都文京区春日 1-13-27

「調子の波」が存在すると言われている。Albert (2008)はベータ・二項モデルを用いて調子の波を表現した。このベータ・二項モデルは20打席ごとのヒットの数が二項分布に従うとし、二項分布の成功確率パラメータ p がベータ分布に従うとしたベイズモデルである。ここでのベータ分布は以下のようにパラメタライズされる。

$$(1.1) \quad \frac{1}{B(K\eta, K(1-\eta))} p^{K\eta-1} (1-p)^{K(1-\eta)-1} \quad (K > 0, 0 < \eta < 1).$$

ここで $B(\cdot)$ はベータ関数である。 η は分布の中心を表す。 K は精度パラメータで、大きいほど散らばりが小さい。すなわち、 K が小さく推定された選手ほど、調子の波が激しいと解釈できる。このモデルは成功確率 p の変化を表現し、これまで評価の難しかった調子の波を定量的に評価することを可能にした。

上述の指標はいずれもピッチャーの能力を考慮していない。ピッチャーによってヒットを打つ難しさが異なり、対戦するピッチャーが選手ごとに異なるにも関わらず、打撃の能力を選手間で比較したい場合、どのようにすればよいか。項目反応理論(item response theory; IRT)の研究成果は、この問題に一つの解を与える。IRTは、教育における達成度評価のような、対象となる人物(被験者)にある課題(負荷)が与えられたときに得られる反応から能力測定を行うために考案されたモデルである(Lord, 1952)。IRTの大きな特徴の一つは、被験者に課された課題への反応から、被験者の能力と課題の難しさを同時に評価することである。つまり、被験者の能力を示す能力パラメータと課された課題のレベルを示す項目パラメータを分離して評価するため、有益な情報を提供し得る。反応を表すモデルは、課題が持つパラメータの数および得られる反応値の種類によってさまざまなものが考案されている(Hambleton et al., 1991; De Ayala, 2008; Baker, 1992)。リンク関数にはロジスティック関数やプロビット関数が用いられる。またその推定手法についても、周辺最尤推定法(Bock and Aitkin, 1981)やマルコフ連鎖モンテカルロ法(MCMC)を用いたベイズ推定(Patz and Junker, 1999a, 1999b)など、さまざまな手法が確立されており、それを実装したソフトウェアやパッケージなども多い(Bilog-MG, 2005; Rizopoulos, 2006; Chalmers, 2012)。もともとは教育工学や心理学の分野で発展した理論であるが、近年ではマーケティングへの応用も報告されている(Jong et al., 2008; Raykov and Calantone, 2014)。

この手法を野球へ応用し、打席ごとのヒット・アウトという応答の確率を、打者の能力を表すパラメータと、投手の難易度(投球の打たれやすさ)を表すパラメータに分離して推定することが可能である。IRTにおける能力パラメータが打者の能力パラメータに、項目パラメータが投手の難易度パラメータにそれぞれ対応する。我々が提案するモデルは、IRTで用いられるラッシュモデル(1パラメータロジスティックモデル)の拡張である。ラッシュモデルでは、潜在的な能力パラメータは各人一つとされているが、本論文では打者の能力に対応するマルコフパラメータを導入した。ラッシュモデルはその単純さから、解釈が容易であるとされる。我々のモデルもオッズ比を用いることで、選手間の比較を容易なものにできる。

我々はAlbert (2008)と同様に、成功確率 p が変化することを仮定する。ただし、上述したベータ二項モデルにおける p の変化は、これまでの打席結果や、打席の状況に依存しないため、解釈がむずかしい。そこで我々は打席結果にマルコフ性を仮定して調子の波を表現する。これにより調子の波が前の打席の結果を受けて生じると解釈できる。3節では、MCMCを用いたパラメータの推定方法について述べる。4節で、パラメータ推定が可能であることを確かめる。5.1節では、2013年の日本プロ野球の実際のデータに対して分析を行う。

2. 提案モデル

$y_{i,j}$ を打者 j ($j = 1, \dots, n$) の打席 i ($i = 1, \dots, l_j$) での結果(アウトならば 0, ヒットならば 1 の値を取る)とする. ここで l_j は打者 j の合計の打席数である. $x_{i,j}$ を打者 j が打席 i で対戦した投手とする. $x_{i,j}$ は $(1, 2, \dots, m)$ のいずれかの値をとる. m はリーグ内の投手の数である. 表 1 は, 選手 j について記録されるデータを, 模式的に示したものである.

$y_{i,j}$ はパラメータ $p_{i,j}$ のベルヌーイ分布に従うと仮定し ($p_{i,j} = \Pr(y_{i,j} = 1)$), $p_{i,j}$ に以下のような回帰型の構造を仮定する.

$$(2.1) \quad \text{logit}(p_{i,j}) = \beta_{0,j} + \beta_{1,j}y_{i-1,j} + \tau(x_{i,j}).$$

ここで, $\tau(x)$ は,

$$(2.2) \quad \tau(x) = \begin{cases} b_1 & x = 1 \\ b_2 & x = 2 \\ \vdots & \vdots \\ b_m & x = m \end{cases}$$

なる関数である. 投手の難易度パラメータ b_k ($k = 1, \dots, m$) は平均が 0 になるよう基準化されているとする. 事前分布として, b_k は平均 0, 分散 σ^2 の正規分布に従うとする. $\beta_{0,j}$ は平均 μ , 分散 ξ^2 の正規分布を仮定する. $\beta_{1,j}$ には, 無情報階層事前分布として, 区間 $(-\infty, \infty)$ の一様分布を仮定する. 階層事前分布として, σ, ξ には区間 $[0, \infty)$, μ には区間 $(-\infty, \infty)$ の一様分布を仮定する.

$\beta_{1,j}$ はヒットを打つ確率に対する, 直前の打席の影響を表すパラメータである. 我々は $\beta_{1,j}$ の 90% 信用区間を求め, これが 0 を含まない選手を「調子の波がある選手」, 0 を含む選手を「調子の波がない選手」とする.

2.1 オッズ比

ロジスティックモデルはオッズ比による解釈が容易である. 本研究では 3 種類のパラメータに対してオッズ比を評価する. 一つ目は打者同士のベースラインとなる能力を比較するためのオッズ比である. 直前の打席を失敗とし, ピッチャーを特定の選手に固定した場合にヒットを打つ確率に着目すると, 打者 j の打者 h に対するオッズ比は,

$$(2.3) \quad \exp(\beta_{0,j}) / \exp(\beta_{0,h})$$

である. 二つ目は投手の能力を比較するためのオッズ比である. 投手難易度パラメータは平均が 0 になるよう基準化されていることを仮定したため, ここでは 0 を基準としたオッズ比,

表 1. 解析対象となるデータの一例.

i	$y_{i,j}$	$x_{i,j}$
1	1	1
2	0	1
3	0	2
\vdots	\vdots	\vdots
l_j	1	4

$$(2.4) \quad \exp(b_j)$$

を考える。このオッズ比は平均的なピッチャーと比べて、どの程度ヒットという事象が起こりやすいかという指標であり、値が小さいほど能力の高いピッチャーと解釈できる。三つ目は打者の調子の波を把握するためのオッズ比である。

$$(2.5) \quad \exp(\beta_{1,j})$$

これは直前の打席結果がアウトだったときを基準として、直前の打席結果がヒットだったときはどの程度ヒットを打ちやすいかという指標である。5.1 節の事例研究ではこれらを使用して選手を評価する。

2.2 定常分布

本モデルでは、投手の難易度パラメータの平均が 0 になるよう基準化されていることを仮定した。打者 j が平均的な投手と対戦した場合のヒットの確率は、

$$P(y_{i,j} = 1) = \text{logit}^{-1}(\beta_{0,j} + \beta_{1,j}y_{i-1,j}).$$

である。一打席前でヒットを打てなかった場合にヒットを打つ確率を、

$$(2.6) \quad p_{0,j} = \frac{1}{1 + \exp(-\beta_{0,j})}$$

一打席前でヒットを打った場合にヒットを打つ確率を、

$$(2.7) \quad p_{1,j} = \frac{1}{1 + \exp(-(\beta_{0,j} + \beta_{1,j}))}$$

と表すことにする。これは遷移行列

$$(2.8) \quad \mathbf{P}_j = \begin{pmatrix} 1 - p_{0,j} & p_{0,j} \\ 1 - p_{1,j} & p_{1,j} \end{pmatrix}$$

の 2 状態マルコフ連鎖を形作る。このマルコフ連鎖は既約かつ非周期的 (Karlin, 1969) であり、以下を満たす行ベクトル π_j が定義できる。

$$(2.9) \quad \pi_j \mathbf{P}_j = \pi_j$$

$\pi_j = (\pi_{0,j}, \pi_{1,j})$ について解くと、

$$(2.10) \quad \pi_{0,j} = \frac{1 - p_{1,j}}{1 + p_{0,j} - p_{1,j}}, \quad \pi_{1,j} = \frac{p_{0,j}}{1 + p_{0,j} - p_{1,j}}$$

である。 π_j は定常分布と呼ばれる。 $\pi_{1,j}$ は十分な時間が経過したときの、選手 j がヒットを打つ割合と解釈でき、初期の状態には影響されない。これも指標の一つとして 5.1 節の事例研究で使用する。

3. パラメータの推定方法

式(2.1)で示したモデルの事後分布の統計量を求めるために、本研究ではハミルトニアン・モンテカルロ法(ハイブリッドモンテカルロ法ともいう)による MCMC を用いた (Bishop, 2006)。MCMC は推定の対象となる事後分布に従う乱数を、マルコフ連鎖を用いてシミュレートする方法である。近年、IRT の分野では MCMC が積極的に利用されており、MCMC による推定の性質は Fox (2010) で議論されているが、ハミルトニアン・モンテカルロ法を用いる方法に

ついては記載がない．MCMC 法のバリエーションとしてギブスサンプリング法やメトロポリス・ヘイスティングス法はよく知られており，IRT の分野でも盛んに利用されている．ギブスサンプリング法ではパラメータごとに全条件付き分布を導出し，変数ごとにサンプリングを行う．メトロポリス・ヘイスティングス法ではパラメータごとに適切な提案分布を選ぶ必要がある．本解析のようにパラメータが多い場合，ハミルトニアン・モンテカルロ法が便利である．

MCMC による IRT モデルのパラメータ推定法の中でも特に有名な手法として Albert (1992) により提案された，データ拡大 (data augmentation) アルゴリズムがある．0 または 1 の値を取る打席結果データ $y_{i,j}$ に対して，連続値を取る潜在変数 z を仮定する． z は平均 $\eta_{i,j} = \beta_{0,j} + \beta_{1,j}y_{i-1,j} + \tau(x_{i,j})$ ，分散 1 の正規分布に従うとし， $z_{i,j} > 0$ ならば $y_{i,j} = 1$ ， $z_{i,j} \leq 0$ ならば $y_{i,j} = 0$ となる．言い換えると， $y_{i,j} = 1$ という観測は 0 で右打ち切りされた $z_{i,j}$ の観測値であり $y_{i,j} = 0$ という観測は 0 で左打ち切り (Turnbull, 1976) された $z_{i,j}$ の観測値である．この場合尤度関数は

$$(3.1) \quad \prod_{j=1}^n \prod_{i=2}^{l_j} \Phi(0|\eta_{i,j}, 1)^{y_{i,j}} (1 - \Phi(0|\eta_{i,j}, 1))^{1-y_{i,j}}$$

と書ける．本研究ではリンク関数にロジット関数を用いる．この場合も Albert (1992) と同様の議論により，潜在変数 z を仮定すると z はロケーションパラメータが $\eta_{i,j} = \beta_{0,j} + \beta_{1,j}y_{i-1,j} + \tau(x_{i,j})$ のスケールパラメータが 1 のロジスティック分布に従うと考える． $F(z|\eta)$ をスケールパラメータ 1，ロケーションパラメータ η のロジスティック分布の分布関数とすると，尤度関数は，

$$(3.2) \quad \prod_{j=1}^N \prod_{i=2}^{l_j} F(0|\eta_{i,j})^{y_{i,j}} (1 - F(0|\eta_{i,j}))^{1-y_{i,j}}$$

と書ける．これはベルヌーイ分布にロジットリンク関数を仮定した場合の尤度と全く同一である．事後分布は，

$$(3.3) \quad \pi(\beta_0, \beta_1, b|y) = C \left(\prod_{i=1}^m \phi(b_i, 0, \sigma) \right) \left(\prod_{i=1}^m \phi(\beta_{0,i}, \mu, \xi) \right) \prod_{j=1}^n \prod_{i=2}^{l_j} F(0|\eta_{ij})^{y_{ij}} \{1 - F(0|\eta_{ij})\}^{1-y_{ij}}$$

となる．ここで ϕ は正規分布の密度関数， C は規格化定数である．Albert (1992) は各パラメータについての全条件付き分布を導出しているが，ハミルトニアン・モンテカルロ法ではその必要がなく，尤度関数と事前分布の積 (3.3) が分かればサンプリングできる．

4. シミュレーション

IRT の分野では，項目パラメータと能力パラメータが同時に推定される際，最尤推定法や条件付き最尤推定法を用いた場合，項目パラメータは一致推定量にならないことが知られている (豊田, 2005)．そのため，パラメータが安定して推定されるか検討が必要である．ここでは，シミュレーションを通じて，パラメータ推定の性質を調べる．シミュレーションでは打者を 75 人，投手を 165 人に設定し，打数は 250 で固定した．これらの設定は事例研究で用いるデータと近い数字にしている．打者能力パラメータ $\beta_{0,j}, \beta_{1,j}$ ($j = 1, \dots, 75$) と，投手難易度パラメータ b_k ($k = 1, \dots, 165$) を標準正規乱数で生成した．ピッチャー $x_{i,j}$ は各打者の各打席ごとに一樣乱数でランダムに選んだ．選択されたピッチャーに対応するパラメータ b_k から確率 $\text{logit}^{-1}(\beta_{0,j} + b_k)$ で 1，確率 $1 - \text{logit}^{-1}(\beta_{0,j} + b_k)$ で 0 を生成し，続けて成功確率

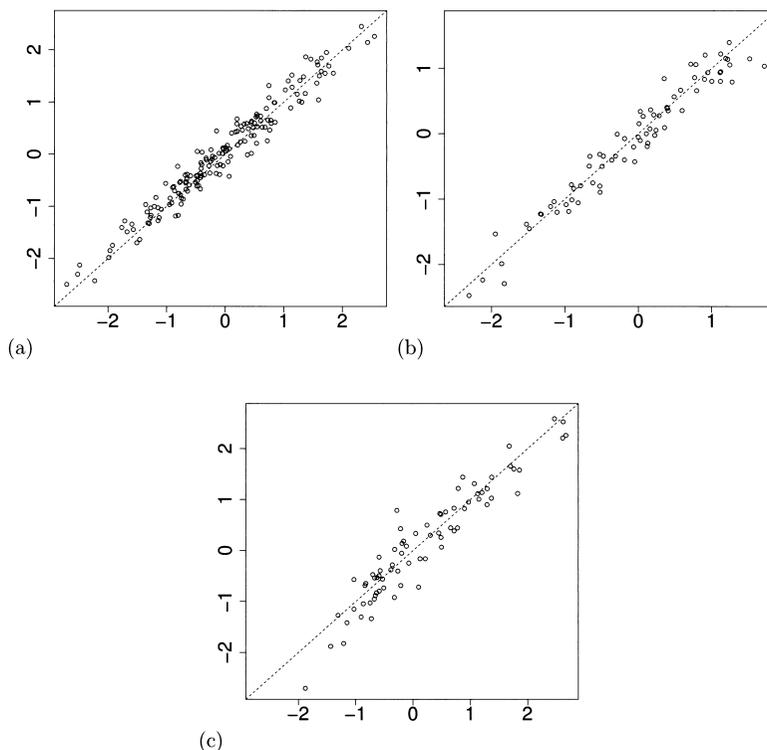


図1. 推定値と真値の比較. 横軸がシミュレーションで仮定した真値, 縦軸が推定値である.
(a) 投手難易度パラメータ b , (b) 打者能力パラメータ β_0 , (c) 打者能力パラメータ β_1 .

$\text{logit}^{-1}(\beta_{0,j} + \beta_{1,j}y_{i-1,j} + b_k)$ のベルヌーイ試行を繰り返して $y_{i,j}$ を生成した. こうして生成した $y_{i,j}$ と $x_{i,j}$ から元のパラメータを推定する. MCMC の試行回数は 2000 回とし, うち前半の 1000 をバーン・イン期間として捨てた. 連鎖の構成数は 3 とした. 事後期待値をパラメータの点推定値とした. 推定値と元のパラメータを比較したプロットを図 1(a)–1(c) に示す. これらの図は, 横軸にシミュレーションで生成した真値, 縦軸に推定値を取ってプロットした. 両方でパラメータが一致していれば, 散布図の各点が直線状に並ぶが, 点はおおむね直線状に分布している. 選手間の能力の相対的比較という観点からは, パラメータの大小の順序の関係が保たれていることが必要であり, そのことは相関係数で評価できる. 推定値と真値の相関係数は, b については 0.98, β_0 が 0.97, β_1 が 0.95 と高い値になった. 決定係数はそれぞれ 0.95, 0.94, 0.89 であった.

また, 区間推定には 90% 信用区間を使用した. 事後分布の両端から確率 $\alpha/2$ の部分を切り取って残った, 中央部の確率 $(1 - \alpha)$ に対応する区間を $100 \times (1 - \alpha)$ 信用区間と呼ぶ. ベイズモデルではパラメータが確率変数であると考えるので, 90% 信用区間は真値を含む確率が 90% となる区間である. 本解析では, MCMC サンプルの標本 5% 点と 95% 点を取り, それぞれを信用下限, 信用上限とすることで, 数値的な 90% 信用区間を得た. 表 2 より, 信用区間がその中にシミュレーションで仮定した真値を含む割合を計算すると, 90% に近い値になっている. このことが, 推定の信頼性を高めている.

投手難易度パラメータの事前分布のパラメータ σ については, 事後期待値が 1.05, 打者能

表 2. 推定された 90% 信用区間が真値を含んだ回数と割合.

	b	β_0	β_1
回数	151	70	68
割合	0.92	0.93	0.91

力パラメータの事前分布のパラメータについては μ が -0.11 , ξ が 0.98 であった. それぞれシミュレーションで仮定した真値 $\sigma = 1$, $\mu = 0$, $\xi = 1$ に近い値が推定された.

5. 事例研究

ここで扱うデータは, 2013 年のセ・リーグ, パ・リーグの試合の打席結果である. ただし, 四球, 死球, 敬遠, 犠打, 犠飛は打席から除いた. 年間を通じて 100 回以上打席に立った打者と, その選手に対して投球を行った投手を解析対象とした. セ・リーグとパ・リーグは分けて分析を行った. 対象となる打者はセ・リーグが 75 名, パ・リーグが 74 名, 対象となる投手はセ・リーグが 165 名, パ・リーグが 162 名である. 条件を満たす打者の打席数の平均はセ・リーグで 264.75, パ・リーグで 284.36 であった.

5.1 推定結果の概要

これまでに述べたように, 推定にはハミルトニアン・モンテカルロ法を用いた. MCMC の試行回数は 10000 回とし, うち前半の 5000 をバーン・イン期間として捨てた. 連鎖の構成数は 3 とした.

収束の判定はトレースラインの目視による確認と Gelman (1996) の収束判定指標 \hat{R} により行った. Gelman (1996) では, \hat{R} が 1.2 または 1.1 より小さければ収束したと判断してよいとされている. 本解析においてはすべてのパラメータについて, \hat{R} が 1.1 より小さいことが確認された.

投手難易度パラメータ b_k は値が小さいほど, 打つのが難しい投手である. 事後期待値の値が小さい順に上位 10 名の選手の推定結果を表 3 に示す. 90% 信用区間とオッズ比, 被安打率もあわせて表示した. 被安打率はピッチャーの能力を表す指標のひとつで, 被安打数を打数で割ったものである. また b_k の事後期待値を全投手についてヒストグラムで図示した(図 2 (a)–(b)). 図の点線は b_k の事後期待値の平均である. これらの図から b_k は 0 付近を中心に分布していることがわかる.

表 3 を見ると, 被安打率が 0.2 の金子選手が 1 位にランキングされ, 被安打率が 0.16 のウィリアムス選手が 10 位にランキングされるという逆転が見られる. ただし, ウィリアムス投手は打席数が少ないため, 信用区間の幅は広く推定されている. この結果がなぜ生まれたかを検討する. 金子選手とウィリアムス投手に関して, 打率三割以上の打者と対戦したときの被安打率と, 打率三割未満の打者と対戦したときの被安打率を集計した(表 4). ウィリアムス投手は打率の低い選手に対しての成績が好調な反面, 打率の高い選手には多くヒットを打たれている. 対照的に金子投手はウィリアムス投手と比べ, 打率の低い選手に対しての被安打率は低くないが, 打率の高い選手のヒットを少なく抑えている. そのため, 被安打率と比較したときに特異な値を取る. 本モデルは打者の能力と投手の能力を同時に評価しているため, このような場合, 能力の高い打者を多くアウトにしている投手が高く評価される. このように提案モデルは, 被安打率だけでは見落としてしまうおそれのある優秀な(能力の高い打者に強い)投手をピックアップできる.

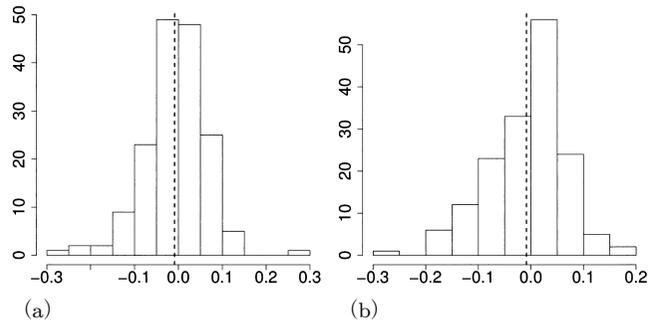


図 2. 投手難易度パラメータ b_i の事後期待値の分布 ((a)セ・リーグ, (b)パ・リーグ).

表 3. 投手難易度パラメータの推定結果.

		b	90%信用下限	90%信用上限	オッズ比	被安打率	打席数
セ・リーグ	前田 健太	-0.26	-0.44	-0.09	0.77	0.19	525
	マシソン	-0.24	-0.49	-0.03	0.78	0.17	182
	西村 健太郎	-0.21	-0.44	-0.01	0.81	0.19	217
	榎田 大樹	-0.19	-0.39	-0.00	0.83	0.19	247
	藤浪 晋太郎	-0.17	-0.35	-0.00	0.84	0.21	412
	ホールトン	-0.15	-0.34	0.02	0.86	0.21	297
	山口 鉄也	-0.15	-0.37	0.04	0.86	0.19	193
	岡田 俊哉	-0.14	-0.33	0.04	0.87	0.21	270
	加藤 康介	-0.12	-0.34	0.07	0.89	0.20	133
	久保 康友	-0.12	-0.32	0.06	0.89	0.23	184
パ・リーグ	金子 千尋	-0.29	-0.45	-0.13	0.75	0.20	652
	ファルケンボーグ	-0.18	-0.42	0.03	0.84	0.14	115
	田中 将大	-0.17	-0.32	-0.03	0.84	0.22	598
	森福 允彦	-0.17	-0.38	0.02	0.85	0.20	172
	攝津 正	-0.16	-0.32	-0.02	0.85	0.22	515
	嘉弥真 新也	-0.16	-0.37	0.03	0.85	0.20	165
	佐藤 達也	-0.15	-0.35	0.03	0.86	0.19	221
	サファテ	-0.15	-0.37	0.04	0.86	0.16	162
	則本 昂大	-0.14	-0.30	0.01	0.87	0.23	488
	ウィリアムス	-0.12	-0.34	0.07	0.88	0.16	126

表 4. 打率で層別した被安打率.

	打率 0.3 未満の打者	打率 0.3 以上の打者
ウィリアムス	0.16	0.35
金子 千尋	0.19	0.26

打者能力パラメータ $\beta_{0,j}$ は、直前の打席結果がアウトだったときのヒットを打つ能力を表す。野球ではヒットよりもアウトのほうが起こりやすいため、 $\beta_{0,j}$ は選手 j のベースラインとなる能力と解釈できる。解釈しやすくなるよう、逆ロジット関数で変換した $\beta_{0,j}$ の事後期待値に加えてその 90% 信用区間、オッズ比、2.2 節で検討した π_{1j} の $\beta_{0,j}$, $\beta_{1,j}$ からのプラグイン推定値を算出した。オッズ比は $\beta_{0,j}$ の値を降順に並べて 10 番目の選手を基準としたオッズ比である。表 5 がセ・リーグの選手、表 6 がパ・リーグの選手についての結果である。これらは

表 5. 打者能力パラメータの推定結果(セ・リーグ).

	$\text{logit}^{-1}(\beta_0)$	90%信用下限	90%信用上限	π_1	β_1	打率	オッズ比
ルナ	0.38	0.32	0.45	0.36	-0.25	0.35	1.30
松元 ユウイチ	0.36	0.28	0.45	0.33	-0.46	0.32	1.19
木村 昇吾	0.35	0.27	0.45	0.34	-0.25	0.32	1.16
バレンティン	0.35	0.30	0.41	0.34	-0.15	0.33	1.16
村田 修一	0.34	0.29	0.39	0.34	-0.04	0.33	1.09
廣瀬 純	0.34	0.27	0.40	0.31	-0.45	0.30	1.07
阿部 慎之助	0.33	0.28	0.39	0.30	-0.56	0.29	1.06
ブランコ	0.33	0.28	0.37	0.34	0.19	0.33	1.02
藤井 淳志	0.32	0.26	0.39	0.32	-0.06	0.31	1.01
梶谷 隆幸	0.32	0.26	0.39	0.37	0.58	0.36	1.00

表 6. 打者能力パラメータの推定結果(パ・リーグ).

	$\text{logit}^{-1}(\beta_0)$	90%信用下限	90%信用上限	π_1	β_1	打率	オッズ比
岡島 豪郎	0.35	0.28	0.41	0.33	-0.20	0.32	1.12
浅村 栄斗	0.34	0.29	0.38	0.33	-0.05	0.33	1.08
中村 晃	0.34	0.29	0.39	0.31	-0.35	0.31	1.07
長谷川 勇也	0.33	0.29	0.38	0.33	-0.05	0.33	1.06
今江 敏晃	0.33	0.28	0.38	0.33	-0.02	0.32	1.04
平野 恵一	0.33	0.27	0.39	0.31	-0.26	0.31	1.03
鶴岡 慎也	0.32	0.26	0.39	0.31	-0.26	0.30	1.01
柳田 悠岐	0.32	0.27	0.38	0.31	-0.16	0.31	1.01
角中 勝也	0.32	0.28	0.37	0.30	-0.35	0.29	1.00
藤田 一也	0.32	0.28	0.37	0.30	-0.35	0.29	1.00

いわば対戦した投手の効果を補正した打率である。

また打者能力パラメータ $\beta_{1,j}$ は値が正であれば、直前の打席でヒットを打つと続けてヒットを打つ確率が上がり、負であれば直前の打席でヒットを打てなかった場合にヒットを打つ確率が上がると解釈できる。MCMC サンプルから $\beta_{1,j}$ の 90% 信用区間を求め、これが 0 を含まないとき、調子の波が存在する選手とする。そのような選手を表 7-8 に示した。

表 5 の $\text{logit}^{-1}(\beta_0)$ を見ると、阿部選手は打率と比較して β_0 が大きく推定されており、梶谷選手に関しては打率と比較して β_0 が小さく推定されている。それに対応して、阿部選手は β_1 が小さい値を取り、梶谷選手は β_1 が大きい値を取っている。 β_0 と β_1 を総合した指標である π_1 は、打率に近い値を取っている。表 7 より、阿部選手、梶谷選手は「調子に波がある」プレイヤーである。

5.2 カイ二乗検定を用いた考察

ここでは、古典的なカイ二乗検定を用いて調子の波について分析を行い、提案モデルの方法が既存の結果と矛盾しないことを示す。調子の波を定量的に評価するもう一つの方法は、ある打席での成功・失敗と、その直前の打席での成功・失敗との関係を調べることである。このアプローチは Albert (2008) が用いたものである。いま 2×2 分割表の k 行 l 列目のセルを c_{kl} とする ($k = 1, 2, l = 1, 2$)。選手 j についての分割表を、 $c_{11} = \sum_i I\{y_{i-1,j} = 0, y_{i,j} = 0\}$, $c_{12} = \sum_i I\{y_{i-1,j} = 0, y_{i,j} = 1\}$, $c_{21} = \sum_i I\{y_{i-1,j} = 1, y_{i,j} = 0\}$, $c_{22} = \sum_i I\{y_{i-1,j} = 1, y_{i,j} = 1\}$ と

表 7. 打者能力パラメータ β_1 の推定結果(セ・リーグ).

	β_1	90%信用下限	90%信用上限	オッズ比	β_0	打率	π_1
福留 孝介	-1.12	-2.14	-0.22	0.33	0.23	0.20	0.20
堂林 翔太	-0.76	-1.43	-0.13	0.47	0.25	0.22	0.22
中村 悠平	-0.74	-1.53	-0.00	0.48	0.26	0.23	0.24
クラーク	-0.56	-1.13	-0.03	0.57	0.27	0.24	0.24
阿部 慎之助	-0.56	-1.02	-0.09	0.57	0.33	0.29	0.30
鳥谷 敬	-0.48	-0.89	-0.07	0.62	0.31	0.28	0.28
山田 哲人	0.46	0.00	0.93	1.59	0.26	0.28	0.29
梶谷 隆幸	0.58	0.10	1.06	1.79	0.32	0.36	0.37

表 8. 打者能力パラメータ β_1 の推定結果(パ・リーグ).

	β_1	90%信用下限	90%信用上限	オッズ比	β_0	打率	π_1
大崎 雄太朗	-1.30	-2.53	-0.24	0.27	0.27	0.22	0.23
鬼崎 裕司	-0.75	-1.36	-0.17	0.47	0.31	0.26	0.27
バルディリス	-0.46	-0.89	-0.05	0.63	0.32	0.29	0.29
松田 宣浩	-0.45	-0.87	-0.04	0.64	0.29	0.26	0.27
ヘルマン	0.42	0.06	0.78	1.52	0.30	0.32	0.33
大引 啓次	0.52	0.06	0.97	1.68	0.24	0.26	0.27
西川 遥輝	0.56	0.02	1.09	1.75	0.24	0.26	0.27

表 9. 阿部慎之助選手についての分割表.

	0	1
0	158	76
1	76	22

して作成する. I は指標関数であり, 等式を満たすとき 1, 他の場合 0 の値を取る. 具体例として阿部慎之助選手についての分割表を, 表 9 に示した.

この表に対して独立性についてのカイ二乗検定を行った. 検定の結果, 帰無仮説が棄却された選手は, 直前の打席での成功・失敗が, 次の打席結果に影響を与えると解釈できる. 10%水準で有意になった選手を「調子の波がある」選手とする. 「調子の波がある」選手を表 10 にまとめた. 理解を容易にするために, クラメールの連関係数も併記した. カイ二乗検定では有意とされない選手についても, 表 7-8 に上げた「調子の波がある」選手については, 並べて記載した. 有意となった選手には, p 値の右に傍点を振った.

連関係数が正の値をとるときは, 直前の打席結果がヒットだった場合, 次の打席でのヒットが出やすくなると解釈でき, 負の値をとるとき, 直前の打席結果がアウトだった場合, 次の打席でのヒットが出やすくなると解釈できる. 表 7-8 と表 10 を比較すると, カイ二乗検定で「調子の波がある」と判断された選手は, 提案手法で「調子の波がある」と判断された選手にすべて含まれている. このことは提案手法が従来手法と矛盾しないことを示している. 逆に提案手法で「調子の波がある」と判断された選手が, カイ二乗検定で「調子の波がある」と判断されるとは限らない. これは提案手法がヒット・アウトの系列の情報や相手投手の情報を利用して, より感度の高い検出を行っているためである. また仮説検定では, 打席結果の情報を, 「有意」か「有意でない」という二値の解釈に落としている. 提案手法は各パラメータから調子の波の強さ

表 10. カイ二乗検定の結果 10% 水準で有意になった打者一覧(表 7-8 に上げた選手もあわせて記載).

		p 値	連関係数	打席数
セ・リーグ	梶谷 隆幸	0.05	. 0.14	219
	鳥谷 敬	0.09	. -0.09	444
	阿部 慎之助	0.09	. -0.10	333
	堂林 翔太	0.10	. -0.11	295
	山田 哲人	0.10	. 0.10	295
	福留 孝介	0.11	. -0.12	212
	クラーク	0.16	. -0.08	346
	中村 悠平	0.22	. -0.10	201
パ・リーグ	ヘルマン	0.06	. 0.10	434
	大引 啓次	0.07	. 0.11	327
	鬼崎 裕司	0.08	. -0.12	258
	西川 遥輝	0.09	. 0.12	232
	バルディリス	0.10	. -0.09	420
	松田 宣浩	0.11	. -0.08	484
	大崎 雄太朗	0.14	. -0.15	133

や、その信用区間といった、より多くの情報を与えることができる。

5.3 ヒットの予測

提案モデルでは打者 j とその直前の打席結果、および投手 k が決まれば、ヒットを打つ確率を予測できる。直前の打席結果がアウトだった場合、打者 j と投手 k の対戦でヒットが出る確率は、

$$(5.1) \quad p_{0,j,k} = \text{logit}^{-1}(\beta_{0,j} + b_k).$$

直前の打席結果がヒットだった場合、

$$(5.2) \quad p_{1,j,k} = \text{logit}^{-1}(\beta_{0,j} + \beta_{1,j} + b_k)$$

と表す。直前の打席結果がアウトだった場合の、打者 j と投手 k の対戦回数を $n_{0,j,k}$ 、そのヒットの回数を $h_{0,j,k}$ と表す。直前の打席結果がヒットだった場合の、打者 j と投手 k の対戦回数を $n_{1,j,k}$ 、そのヒットの回数を $h_{1,j,k}$ と表す。実際のヒットの割合 $h_{0,j,k}/n_{0,j,k}$ 、 $h_{1,j,k}/n_{1,j,k}$ と、提案モデルによる確率の推定値 $p_{0,j,k}$ 、 $p_{1,j,k}$ を比較する。以降、単に実際の割合といった場合は $h_{0,j,k}/n_{0,j,k}$ または $h_{1,j,k}/n_{1,j,k}$ 、推定確率といった場合は $p_{0,j,k}$ または $p_{1,j,k}$ を指すものとする。推定確率と実際の割合の差を、図 3(a)および図 3(b)に示す。 $n_{0,j,k}$ または $n_{1,j,k}$ が大きくなるにつれ、実際の割合と予測確率の差は小さくなる傾向が見られた。

$n_{0,j,k}$ または $n_{1,j,k}$ が一定以上大きいにも関わらず、予測確率と実際の割合が大きく食い違う場合、その打者と投手の組み合わせに固有の特殊な事情がある可能性を考える。 $n_{0,j,k}$ または $n_{1,j,k}$ が 10 を超える組み合わせについて、実際の割合と予測確率の差が 0.2 を超えたものを表 11、12にまとめた。

阿部慎之助選手はベースラインとなる能力パラメータ β_0 が大きく推定された(表 5)優れた選手だが、スタンリッジ投手とは 13 回対戦して 1 回しかヒットを打っていない。いわばスタンリッジ投手は阿部慎之助選手の天敵といえるだろう。本モデルは打者の能力と投手の能力を測る基準となるため、このような特異な関係性を見出すことに利用できる。

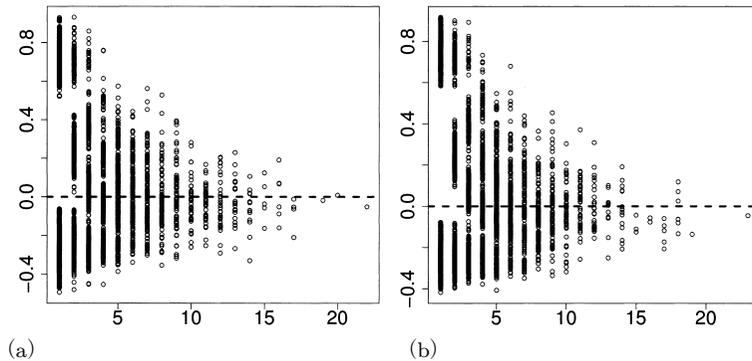


図 3. 推定確率と実際の割合の差. 横軸に n , 縦軸に差を取った((a)セ・リーグ, (b)パ・リーグ).

表 11. 予測確率と実際の割合の比較(セ・リーグ).

打者	投手	一打席前	実際の割合	予測確率	n
鳥谷 敬	大竹 寛	0	0.00	0.33	14
新井 良太	藤井 秀悟	0	0.00	0.29	12
阿部 慎之助	スタンリッジ	0	0.08	0.34	13
ルイス	菅野 智之	0	0.00	0.25	11
鳥谷 敬	三浦 大輔	0	0.09	0.34	11
鳥谷 敬	八木 亮祐	0	0.08	0.34	12
長野 久義	大野 雄大	0	0.54	0.31	13
梵 英心	メッセンジャー	0	0.09	0.31	11
バレンティン	メッセンジャー	0	0.12	0.33	17
菊池 涼介	メッセンジャー	0	0.00	0.21	14
マートン	澤村 拓一	0	0.50	0.30	12
新井 貴浩	三浦 大輔	0	0.50	0.30	12

表 12. 予測確率と実際の割合の比較(パ・リーグ).

打者	投手	一打席前	実際の割合	予測確率	n
大引 啓次	田中 将大	0	0.06	0.26	17
松井 稼頭央	十亀 剣	0	0.07	0.28	14
内川 聖一	西 勇輝	0	0.08	0.33	13
本多 雄一	牧田 和久	0	0.50	0.19	12
中村 晃	牧田 和久	0	0.64	0.26	11
大谷 翔平	田中 将大	0	0.00	0.26	11

5.4 パラメータ推定の安定性

4 節ではシミュレーションデータからパラメータ推定の安定性を議論したが, ここでは実データを用いて改めて検討する. 一般に項目反応理論では項目の困難度パラメータを正確に推定するためには, 多くの被験者が必要であるとされている. 被験者の数を少なくともった場合でも, 被験者の数を多くした場合と同様の推定結果が得られるならば, モデルとそのパラメータはより信頼できるものとなる. そこでパラメータ推定の対象となる打者が約半数となるように

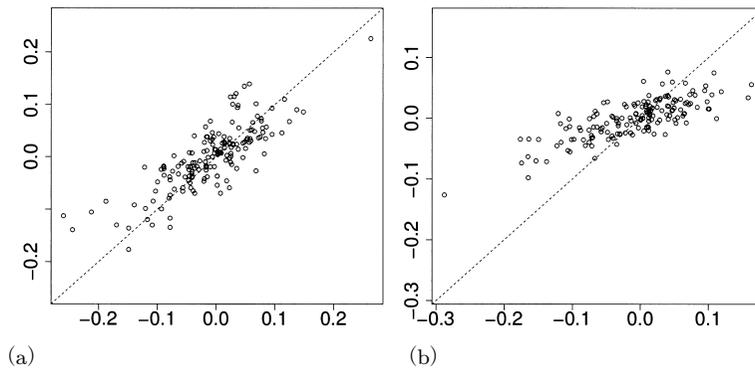


図 4. 投手難易度パラメータの推定結果の比較((a)セ・リーグ, (b)パ・リーグ).

ランダムサンプリングで打者を選び直した新たなデータセットに対してモデルのフィッティングを行い、5.1 節で推定した投手難易度パラメータと、新たなデータセットから推定した投手難易度パラメータを比較した。セ・リーグについての比較が図 4(a), パ・リーグについての比較が図 4(b)である。これらの図は縦軸にランダムサンプリングしたデータセットから推定した投手難易度パラメータ、横軸に 5.1 節で推定した投手難易度パラメータをとりプロットした。点はおおむね直線状に分布しており、相関係数はセ・リーグで 0.81, パ・リーグで 0.77 と高い値になった。決定係数はそれぞれ 0.77, 0.48 であった。このことはサンプルサイズが小さくても結果に矛盾が生じないことを意味しており、安定したパラメータの推定が行えた見込みが高いといえる。ただし、図 4(b)においては、線形回帰した場合に傾きが 1 より小さくなるような系統的な差が生まれている。ベイズモデルではサンプルサイズが小さい場合、事前分布の影響で推定値がシュリンクするためである。順序の関係は保たれているため、選手間の相対的比較をする上では大きな問題とならないが、提案モデルを利用する際は、この性質を認識する必要がある。

6. まとめ

本研究では打席結果を 0 または 1 にコード化したデータから、投手の能力と打者の能力を推定するモデルを提案した。ハミルトニアン・モンテカルロ法を用いたパラメータ推定の安定性はシミュレーションと事例研究で確かめた。事例研究では、従来用いられる被安打率では低く評価されるおそれのある優れた選手を見つけることができた。また 0 または 1 のプロセスにマルコフ性を仮定することで、打者の「調子の波」を捉えることができた。

謝 辞

本研究の貸与データは情報・システム研究機構の新領域融合研究プロジェクト『社会コミュニケーション』データ中心科学リサーチコモンズ事業『人間・社会データ』の支援を受けたものです。また、本研究の一部は科学科研究費(若手研究(B)No.15K21379)の助成を受けたものです。心より感謝申し上げます。

参 考 文 献

- Albert, J. (1992). Bayesian estimation of normal ogive item response curve using Gibbs sampling, *Journal of Educational Statistics*, **17**(3), 251–269.
- Albert, J. (2008). Streaky hitting in baseball, *Journal of Quantitative Analysis in Sports*, **4**(1), DOI: 10.2202/1559-0410.1085.
- Albert, J. and Bennett, J. (2003). *Curve Ball*, Springer-Verlag, New York. (加藤貴昭, 後藤寿彦 訳 (2004). 『メジャーリーグの数理科学〈上〉』, シュプリンガーフェアラーク, 東京.)
- Baker, F. B. (1992). *Item Response Theory: Parameter Estimation Techniques*, Marcel Dekker, New York.
- Bilog-MG (2005). Scientific Software International, <http://www.ssicentral.com/irt/index.html>.
- Bishop, C. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*, Springer, Cambridge. (元田浩, 栗田多喜夫, 樋口知之, 松本裕治, 村田昇 訳 (2008). 『パターン認識と機械学習 下(ベイズ理論による統計的予測)』, 丸善出版. 東京.)
- Bock, R. D. and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm, *Psychometrika*, **46**(4), 443–459.
- Chalmers. R. (2012). Mirt: A multidimensional item response theory package for the r environment, *Journal of Statistical Software*, **48**(1), 1–29.
- De Ayala, R. J. (2008). *The Theory and Practice of Item Response Theory*, The Guilford Press, New York.
- Fox, J. P. (2010). *Bayesian Item Response Modeling Theory and Applications*, Springer-Verlag, New York.
- Gelman, A. (1996). *Markov Chain Monte Carlo in Practice*, Chapman & Hall/CRC Interdisciplinary Statistics, London.
- Hambleton R. K., Swaminathan, H. and Rogers, H. J. (1991). *Fundamentals of Item Response Theory*, Vol. 2, Sage Publications, New York.
- Jong, M. G. D., Steenkamp, J. B. E. M., Fox, J. P. and Baumgartner, Hans (2008). Using item response theory to measure extreme response style in marketing research: A global investigation, *Journal of Marketing Research*, **45**(1), 104–115.
- Karlin, S. (1969). *A First Course in Stochastic Processes*, Academic Press, Cambridge. (佐藤健一, 佐藤由身子 訳 (1974). 『数理解析とその周辺 3 確率過程講義』, 産業図書, 東京.)
- Lord, F. M. (1952). A theory of test scores, *Psychometric Monographs*, No. 7, Psychometric Corporation, Richmond.
- Patz, R. J. and Junker, B. W. (1999a). A straightforward approach to Markov chain Monte Carlo methods for item response models, *Journal of Educational and Behavioral Statistics*, **24**(2), 146–178.
- Patz, R. J. and Junker, B. W. (1999b). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses, *Journal of Educational and Behavioral Statistics*, **24**(4), 342–366.
- Raykov, T. and Calantone, R. J. (2014). The utility of item response modeling in marketing research, *Journal of the Academy of Marketing Science*, **42**(4), 337–360.
- Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response analysis, *Journal of Statistical Software*, **17**(1), 1–25.
- 豊田秀樹 (2005). 『項目反応理論[理論編]』, 朝倉書店, 東京.
- Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data, *Journal of the Royal Statistical Society, Series B (Methodological)*, **38**, 290–295.

Measurements of Baseball Players' Batting Abilities

Ko Abe¹, Takenori Sakumura² and Toshinari Kamakura²¹Graduate School of Science and Engineering, Chuo University²Department of Industrial and Systems Engineering, Chuo University

Statistics of player performance is an important part of baseball. Many stats have been proposed to measure a batter's performance, including batting average, on-base percentage, and slugging percentage. In the field of baseball analytics, the "streakiness" of batter's ability is often discussed using a binary sequence of hitting outcomes for a player during a season.

Unlike previous studies, which use data from the batter, we take a different approach. To analyze a batter's performance, we simultaneously model the pitcher and batter's ability. To model a batter's streakiness, we employ an extension of a one-parameter logistic item response model. Item response theory (IRT) estimates both the subject's ability and item difficulty. In this study, the ability parameter and item difficulty parameter correspond to the batter's ability and pitcher's ability, respectively. Although simplicity is thought to make the one-parameter logistic model easy to interpret, our model incorporates numerous parameters. However, using the odds ratio allows athletes to be compared.

We express streakiness by the interactions of previous at bats and imposing the Markov property on batting data. Specifically, we use MCMC in the Hamiltonian Monte Carlo method (also called the hybrid Monte Carlo method). The computation of Gibbs sampling is complex and time consuming, but the Hamiltonian Monte Carlo method is easily computed once the prior distribution and the likelihood function are defined. Our simulation study shows that the true and estimated values agree well. Additionally, the calculated proportion of times that the credible interval contains the true value is close to the nominal value.

To demonstrate the usefulness of our proposed method, we applied it to analyze actual data from Japanese professional baseball. Two-way tables can measure the dependence of the previous success and the current success by the Pearson chi-square statistic and the corresponding p-value of the test of independence. The results provide more information and are consistent with the results of chi-square test. Because comparing streakiness in the hypothesis test is difficult, we ranked streaky players from the credible intervals and the posterior means. IRT requires many subjects to estimate item difficulty parameters. Although we estimated the parameters using fewer batters, the results from our method are similar to those from IRT.

バレーボール各国代表チームのレーティング手法 の提案および結果予測・大会形式評価への応用

小中 英嗣[†]

(受付 2016 年 12 月 20 日；改訂 2017 年 5 月 2 日；採択 8 月 28 日)

要 旨

バレーボールの各国代表チームのランキングである FIVB ランキングは、ランキングポイントの設計に統計的な根拠が無く、さらに、主要な国際大会を独占的に主催する特定の国が優遇されていることなどが原因となり、その値を実力の定量的評価には活用できない。そこで本研究では、過去の試合結果(各セットでの得点)から各チームの実力(レーティング)を算出する手法を提案する。具体的には、レーティング差がロジスティック回帰モデルを通し各セットでの得点率を説明するモデルを仮定する。また、試合結果からレーティングを算出するアルゴリズムをあわせて提案する。

リオデジャネイロオリンピックを含む複数の国際大会での実際の得点率と、提案手法により算出したレーティング差から予測される得点率との間に中程度から強い相関(相関係数がおおよそ 0.65 から 0.80)があり、その相関はおおむね FIVB ランキング(ポイント)差に基づくもの(相関係数がおおよそ 0.45 から 0.70)より強いことが分かった。さらに、提案手法に基づくレーティングにより各大陸間の実力差を定量的に明らかにし、FIVB ランキングが実力を正しく反映していないことの根拠を明示する。この事実に基づき、具体的な事例としてリオデジャネイロオリンピックの予選および本選の大会形式の不備を指摘する。具体的には、敗退行為を誘引しかねない最終予選出場国の決定方法、および実力を反映していない FIVB ランキングにより引き起こされたオリンピック本選での不均衡な組み分けについて指摘する。

キーワード：スポーツ，バレーボール，レーティング，ロジスティックモデル。

1. 研究背景

本稿ではバレーボール各国代表チームのレーティング手法の提案とその検証を目的とする。

スポーツにおいて、試合結果に基づき選手・チーム間の序列を定めるランキング制度は、大会への参加基準やシード対象選手(チーム)の選定基準となるなど、その適切な設計が要求される事項である。

全ての選手(チーム)が等しい回数総当たりする(できる)場合は、勝利数や勝率がランキングに最も適した指標であることは論を待たない。しかし、多くの選手(チーム)が参加し、総当たりが不可能な場合に適切に順序を定めることには困難を伴う。

多くのスポーツでは、過去の一定期間の各大会での成績(順位)を数値化し、その総計(ランキングポイント)に基づいたランキングを採用している(Ray, 2011)。ここではこれを加算方式

[†]名城大学 理工学部：〒468-8502 愛知県名古屋市天白区塩釜口 1-501

と呼ぶ。バレーボールの各国代表チームのランキングも加算方式により算出されている(FIVB, 2016)。この手法では、各チームが自分自身の結果のみと得点算出基準を参照することでランキングポイントを算出でき、この簡便さは利点である。しかし、各選手・チーム間の合計ポイントの差や比が実力差を定量的に表すためには、各大会・各順位に対する付与ポイント、大会参加条件、大会形式(シード権の設定など)を含めた総合的な設計が不可欠である。

一部のスポーツでは、ポイント交換(point exchange)と呼ばれる、各対戦結果に基づき対戦チーム間でランキングポイントを総和が変わらないように交換する手法に基づきランキングポイントを算出している。最も有名な例はチェスのイロレーティング(Elo rating)(Elo, 1979)であり、球技ではラグビーがイロレーティングを若干修正した手法を採用している(World Rugby, 2014)。この手法では、各チームの実力がある期間一定であり、その期間内に十分な対戦数が行われれば、算出されたランキングポイントが実力に相当する値に近づく特徴を持っている。ただし、この算出には対戦時点での両チームのランキングポイントが既知である必要があり、原理的には期間開始時点での全チームのランキングポイントおよび期間内の全対戦の結果を必要とする。また、得点差などによるポイント交換量の調整などを含んでいるスポーツも多く、その量の調整によりランキングが適切に実力を反映できているかどうかが明確でなくなる点は加算方式と同様である。

ここで、ランキング(ranking)とレーティング(rating)を定義し、区別することとする。ランキングは各チーム間の順位を定めることであり、それに対しレーティングは各チームの実力に相当する値を算出することと定義する。レーティングの結果に基づいてランキングを作成することが可能である。

2つのチームが対戦した場合、大きく二つに分けて、

- (1)それぞれのチームが数年程度のある一定の期間にわたり安定的に持っている技能・実力
- (2)当日の調子・運、対戦チーム間の相性

の要因が得点や勝敗に影響すると仮定する。本稿では、バレーボールなどの2チームが互いの得点の多少を競うスポーツに対する狭義のレーティングとして

- ・上記の(1)に対応する、過去の試合結果から推定した、あるチームが得点する確率を説明するパラメータ

と定義する。加算方式のランキングポイントは上記の(1)および(2)の双方を区別せずに反映させる方法であり、これが(1)のみを反映させるようにするためには大会の参加条件、トーナメント形式、および順位ごとに付与されるポイントなど統一的な設計が必要である。それに対してイロレーティングに代表されるポイント交換手法は(2)の影響を排除して(1)のみを推定する、つまり狭義のレーティングを統計的に算出することを意図した手法である。以降本稿で「レーティング」は「狭義のレーティング」を指すこととする。

バレーボールに関する統計的な研究として、各技能(サーブ、レセプション、スパイク、ディグ、ブロック、およびセットなど)や戦略が得点にどのように寄与しているかを調査したものは多数ある(たとえば、Eom and Schutz, 1992; Zetou et al., 2007; Florence et al., 2008; Araújo et al., 2010; Ferrante and Fonseca, 2014; Burton and Powers, 2015など)。しかし、これらの研究には試合動画を利用した各プレーの詳細な分析が必要であり、その調査コストは無視できない。また、状況を細かく分析しすぎることは説明変数を増やしてその次元を上げてしまうが、活用できるデータ数には上限がある。このようにして「次元の呪い」(Indyk and Motwani, 1998)を引き起こしかねないため、推定するパラメータ数が試合数に対して十分少なくなるようなモデルの構築が肝要である。

これら技能の評価に関する研究が非常に多岐にわたるのに対し、バレーボールのチームのレーティングに関する研究は(著者の調査の範囲では)非常に少ない。Masseyの方法(Massey, 1997)に基づき、大学女子バレーボールの勝率を予測したもの(Knapper and McIlwain, 2015)や、ビーチバレーに適用したもの(Glasson et al., 2001)、公式ランキング差をもとにロジスティック重回帰分析を適用したもの(Dziedzic and Hunter, 2015)などが散見されるのみである。

バレーボールはサービスから始まる一連のプレイが必ずどちらかの得点で終わるという特徴を有する競技である。また、サービスから始まる一連のプレイを1試合中に150回から200回繰り返す。得点は試合進行の最小単位、かつ(総得点が多いほうが勝利するとは限らないが)勝敗に対する主要な原因であり、さらに公式記録に必ず記載されて調査が容易である。この観点から、本研究では各試合、各セットの得点を基本データとし、これに基づく統計的な分析を行う。

本研究では、各チームが得点する確率が、それぞれのチームのレーティングの差により説明できると仮定し、国際試合の結果からその統計的な算出を試みる。得点率の説明モデルにはロジスティック回帰モデルを利用する。対象となる全試合に対する最適化を行うことで、最適なレーティングを算出する。これは試合ごとの逐次更新に基づくポイント交換(イロレーティングなど)の欠点を補う明確な利点となる。

提案手法により算出したレーティングに基づき、いくつかの国際大会の結果を予測する。具体的には、ロンドンオリンピック、リオデジャネイロオリンピック世界最終予選1、およびリオデジャネイロオリンピックの各試合の結果および得点率を、その直近1年の主要国際大会の結果から算出したレーティングに基づき予測する。

その結果、

- レーティングから予測された得点率と各セットの実際の得点率には中程度の相関があり(相関係数が0.5強)、
- レーティングから予測された得点率と各試合の実際の得点率には中程度から強い相関がある(相関係数が0.7付近)

ことを示す。

これらの相関、および得点率の大小の予測結果をFIVBランキングおよびランキングポイントに基づく手法と比較し、提案するレーティングが優位であることを示す。さらに、定量的なレーティングに基づく大会形式の評価の応用として

- 世界最終予選が敗退行為を助長しかねない不適切な設計となっていたこと
- FIVBランキングが実力を反映していないことが原因でリオデジャネイロオリンピックでのプールわけに不均衡が生じ、実力の高いチームが早期敗退を強いられたこと

の2点を指摘する。

2. レーティングの定義と算出方法

本節ではまず対象とするバレーボールの世界ランキングの特性について議論し、その後レーティング算出アルゴリズムを提案する。

2.1 FIVB ランキング

国際バレーボール連盟(FIVB)はオリンピック、世界選手権などの主要国際大会の順位をランキングポイントに換算し、一定期間の和をそのチームのランキングポイントとし、その値で

表 1. FIVB Ranking Point System.

Standing	Competition name				
	Olympic	World Cup	World Championship		Continental Championship
			Men	Women	
1	100	100	100	100	30
2	90	90	90	90	26
3	80	80	80	80	22
4	70	70	70	70	18
5	50	50	62	58	10
6	—	40	56	—	14
7	—	30	50	50	10
8	—	25	—	—	5
9	30	5	45	45	—
10	—	5	—	—	3
11	20	5	40	40	—
12	—	5	—	—	—
13 Tie			36	36	2
15 Tie			33	33	—
17 Tie			30	30	—
21 Tie			25	25	—

世界ランキングを定めている (FIVB, 2016)。

順位に対しどの程度のランキングポイントを付与するか設計により、どのようなチームがランキング上位となるかが異なる。表 1 に、主要国際大会で得られるランキングポイントを示す。

順位決定にトーナメント形式を採用している大会(オリンピック、世界選手権)では順位決定戦を行わない場合があり(たとえば、オリンピックでは準々決勝敗退の 4 チーム間での順位決定戦を行わない)、その場合は同順位として同じランキングポイントが付与される。

この表から、世界大会での 1 位から 4 位までの順位ひとつの差は 10 点で統一されているが、5 位以下でのランキングポイントの設計法には一貫性が無いことが分かる。具体的には、オリンピック、ワールドカップでは 4 位と 5 位の差が 20 ポイントであるのに対し、世界選手権では 8(男子)、12(女子)ポイントである。それ以下の順位となると順位間でのポイント差はまた縮まる傾向にある。大陸選手権は 5 大陸(アジア、アフリカ、ヨーロッパ、北中アメリカ、南アメリカ)それぞれで 2 年ごとに開催されるが、大陸間の実力差はポイントには反映されていない。また、特定の大会を継続して同一国で開催し、その開催国が常にランキングポイントを獲得している、などの問題点も指摘できる(第 4 章で詳述する)。

これに対し、順位による加算方式を採用している中でも設計に一貫性のあるテニス(ATP ランキングポイント)(ATP World Tour, 2016)と比較する。一例としてグランドスラム(主要四大大会)でのランキングポイントを示す。

優勝：2000、準優勝：1200、ベスト 4：720、ベスト 8：360、ベスト 16：180、
 ベスト 32：90、ベスト 64：45、ベスト 128：10、予選通過：25、予選決勝：16、
 予選 2 回戦：8

ベスト 64 からベスト 4 までは 1 回勝ち抜くごとに付与ポイントが 2 倍、それ以降は 5/3 倍となっている。グランドスラム以下の大会(マスターズ 1000、500 シリーズ、250 シリーズ)では優勝者に付与されるランキングポイントは大会名に冠されている値(1000、500、250)であり

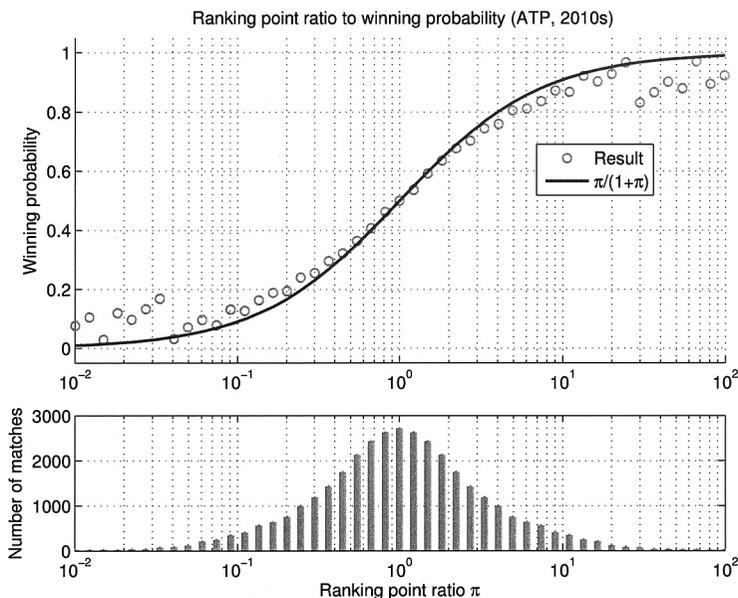


図 1. ATP ranking point ratio and winning probability.

それらは均等な比率を保っている。また優勝と準優勝以降の比率も保持されている。テニスのこのランキングポイントシステムは、どこまで勝ち抜いたかを二進数で表現していることとほぼ等価である。

このような一貫した付与ポイントの設計に加え、上位選手に対して主要大会への出場義務があり、ランキングによりシードを決定するトーナメント方式が安定して運営されている。したがって、その選手がどの規模の大会でどの程度まで勝ち抜いているかと合計ランキングポイントには強い関係ができ上がる。その結果、試合直前の対戦選手間のランキングポイント比が予測される勝率を説明することになる。図 1 に、試合直前の ATP ランキングポイント比と試合結果(勝率)の関係を示す。試合結果は 2010 年 1 月から 2016 年 1 月までの 18197 試合のもの (Sackmann, 2017) を利用した。ランキングポイント比を π としたとき、勝率が $\pi/(1+\pi)$ で近似できることが示されている。

2.2 レーティング算出方法

前述の通り、FIVB ランキングポイントの設計には数理的根拠が無く、明確な意図を持って設計されている ATP ランキングポイントと比較してもそれは明らかである。したがって、FIVB ランキングポイントによりそのチームの実力を推定し、さらに今後の対戦の勝敗予測などに利用するのは困難である(第 3 章および第 4 章で具体例を示す)。

そこで本稿では、試合結果として必ず公式に記載される各セット毎の得点に基づき、各チームの得点率を説明するパラメータを推定する手法を提案する。具体的には、各チーム i の実力を示すパラメータ(今後、これをチーム i のレーティングと呼ぶ)を r_i で示し、チーム i と j が対戦した場合、ひとつのプレイで i が得点する確率 $p_{i,j}$ が次式に従うと仮定する。

$$(2.1) \quad p_{i,j} = \frac{1}{1 + e^{-(r_i - r_j)}}.$$

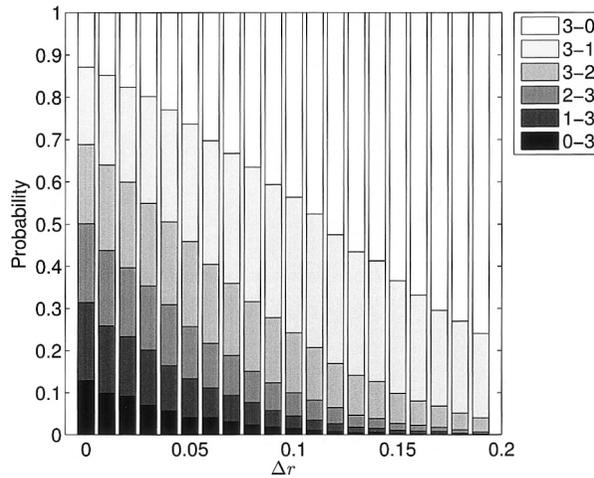


図 2. Won-lost sets probability.

この仮定はロジスティック回帰モデルと呼ばれ、チェスのレーティングであるイロレーティング (Elo, 1979) での勝率の仮定や、教育学における項目応答理論 (Hambleton, 1991; de Ayala, 2008) での問題の正答率の仮定など、幅広い分野で利用されているものと同じ形式である。大まかな目安として、パラメータ差 0.1 は、パラメータが大きいチームの 1 プレイでの得点率が 0.525 であることに対応する。この場合、パラメータの大きいチームが 1 セット取る (25 点得点する) までに小さいチームは平均で 22.6 点得点していることになる。

2 チームのパラメータ差 $\Delta r = r_i - r_j$ を 0, 0.01, 0.02, ..., 0.20 と選び、各プレイ間での得点確率が一定で独立であると仮定し、それぞれ 10^4 回の対戦をシミュレーションした結果を図 2 に示す。実力差が小さいところではセット数 3-2, 2-3 が多く、 $\Delta r = 0.05$ では実力が高いチームが 2-3 で敗れる割合が約 12.4% である。 Δr が 0.1 を超えると番狂わせ (レーティングが低いチームが勝利する) の確率は 5% 程度となり、0.2 を超えるとほぼ勝敗は明確であることが示唆されている。

2.2.1 ポイント交換法の特徴

イロレーティング (Elo, 1979) は、式 (2.1) の線形変換である

$$(2.2) \quad p_{i,j} = \frac{1}{1 + 10^{\frac{-(r_i - r_j)}{400}}}$$

により選手 i, j 間の勝率が説明できると仮定している。したがって、レーティングが正しく計算されていれば、その差を利用して試合結果を予測することが可能な方法である。ラグビーでは上式を折れ線で近似したものを利用している (World Rugby, 2014)。このようにしてランキングポイント差を実力差と関連付けているため、「日本初戦が番狂わせランキングトップ」 (World Rugby, 2015) のように、ある試合の結果がどの程度の番狂わせであったのかを定量的に評価することが可能となっている。

また、選手 (チーム) i, j 間の試合結果が

$$(2.3) \quad s_{i,j} = \begin{cases} 1 & (i \text{ wins}), \\ 0 & (i \text{ loses}). \end{cases}$$

で表されるとしたとき、試合後のレーティングが

$$(2.4) \quad r'_i = r_i + K(s_{i,j} - p_{i,j})$$

と更新される。 r_j の更新規則も同様である。式(2.4)は以下のように解釈できる：勝利したときはレーティングが増加する。これは、「勝利したのだから試合前のレーティングは過小評価であり、レーティングはより高く修正されるべきである」ということに他ならない。増加量は予測される勝率が高いほど小さい。これは「試合前のレーティングから予想される結果が起きているので、レーティングを大きく修正する必要は無い」ということである。敗北したときはこの逆である。

このような方式であるため、レーティングの計算には試合前の両チームのレーティングが必要であり、1チームのレーティングを計算するために、関与する全員(全チーム)の対戦結果が必要である。

式(2.4)の定数 K はレーティングの更新量を定める定数であり、チェスのプロレベルでは16が利用されている。 K が大きければ直近の対戦の影響が大きく、逆に小さければ過去の影響が大きい。この値の設定により、レーティングの性質が異なる。

r_i の真値が既知かつ一定であり、対戦が十分な回数実施されれば、式(2.4)の更新規則により r_i は真値に近づく。しかし、同程度の実力である二者が対戦した場合でも、対戦後に必ずレーティングが変動してしまうため、収束は保障されない。

2.2.2 提案手法

前述の、試合ごとのポイント交換手法の欠点は、1試合ごとの結果に対するレーティングの更新量(式(2.4)内のパラメータ K により定まる)の適切な設定が困難であること、およびレーティングの更新が試合ごとであり、レーティングが収束に至るまでには更新回数が少ないこと、の二点にまとめられる。本稿ではこれらの欠点を克服する手法を提案する。

提案手法の概要を以下に述べる。

- (1) 対象とする期間の全試合のデータをランダムにソートした後、順に選択し、その得点率に対して式(2.4)でレーティングを更新する。
- (2) 試合ごとの結果に対するレーティングの更新量(式(2.4)内の K)を動的に調整する。
レーティングの更新量が直前よりも大きくなることは K が大きすぎることを意味するので、その値を小さくする。
- (3) 上記の更新を、全チームのレーティングが収束するまで十分な回数繰り返す。

以下にアルゴリズムを示す。表2に記号の意味を示す。

レーティング算出アルゴリズム。

- (1) $\mathbf{r}^{(0)} = \mathbf{0}$ とし、収束判定のための $\epsilon_{th} > 0$ を小さな値に設定する。 $K > 0$ を定める。繰り返し回数を $k = 0$ とする。 N_S セットの結果を保持したデータベースを用意する。各セットの結果を $\langle i, j, s_i, s_j \rangle$ で示す。
- (2) データベースの各セットをランダムにソートする。
- (3) ソートされたデータベースからセットを順に取り出し、これを $\langle i, j, s_i, s_j \rangle$ とする。
- (4) レーティング r_i および r_j を以下の規則で更新する。

$$(2.5) \quad p_{i,j} = \frac{1}{1 + e^{-(r_i^{(k)} - r_j^{(k)})}}, \quad s_{i,j} = \frac{s_i}{s_i + s_j},$$

$$(2.6) \quad r_i^{(k+1)} = r_i^{(k)} + K(s_{i,j} - p_{i,j}), \quad r_j^{(k+1)} = r_j^{(k)} + K((1 - s_{i,j}) - (1 - p_{i,j})).$$

表 2. Notations.

N_T	Number of teams
$\mathbf{r} = (r_1, \dots, r_{N_T})^T$	Rating vector
N_S	Number of sets
$\langle i, j, s_i, s_j \rangle$	Result of one set. Team i and j scored s_i and s_j points in a set. N_S tuples are stored in database.
ϵ_{th}	Threshold value
K	Parameter used in rating update
k	Iteration index
$\mathbf{0}, \mathbf{1}$	Column vector composed of zeros and ones with suitable dimensions
$\ \mathbf{x}\ $	Euclidean norm of vector x

データベース内の全セットに対して上記の更新を行う。

- (5) $\|\mathbf{r}^{(k+1)} - \mathbf{r}^{(k)}\| > \|\mathbf{r}^{(k)} - \mathbf{r}^{(k-1)}\|$ の場合, $K \leftarrow 0.5K$ とする。
 (6) $\|\mathbf{r}^{(k+1)} - \mathbf{r}^{(k)}\| < \epsilon_{th}$ が成り立つ場合, $\mathbf{r}^{(k+1)}$ を出力としてアルゴリズムを終了する。そうでない場合 $k \leftarrow k + 1$ としてステップ(2)へ戻る。

このレーティングは 2 チーム間の差のみに意味がある(間隔尺度である)ため, 分かりやすさのために全体に定数を加減算してもかまわない。提案手法ではレーティングの和が保存されるので, $\mathbf{r}^{(0)} = \mathbf{0}$ から計算を開始すると $r_i = 0$ は全チーム内で中央の実力を持つことを意味する。また,

$$(2.7) \quad \mathbf{r} \leftarrow \mathbf{r} - (\max \mathbf{r}) \cdot \mathbf{1}$$

とすれば, レーティングの最上位を必ず 0 とし, そこからの差を表すことができる。

2.3 他の手法との関係および提案手法の拡張性

対の比較結果から全体の順位付けを行う他の手法として, Bradley-Terry モデル(BT モデル)(Rao and Kupper, 1967; Davidson, 1970)がよく知られており, スポーツのレーティングに関する研究も多い。BT モデルでは, チーム i とチーム j が対戦した場合, i が j に勝利する確率 $p_{i,j}$ が各チームのパラメータ $\pi_i, \pi_j > 0$ を用いて

$$(2.8) \quad p_{i,j} = \frac{\pi_i}{\pi_i + \pi_j}$$

で表すことができると仮定する。この仮定は,

$$(2.9) \quad p_{i,j} = \frac{\pi_i}{\pi_i + \pi_j} = \frac{1}{1 + \frac{\pi_j}{\pi_i}} = \frac{1}{1 + \exp(\log \frac{\pi_j}{\pi_i})} = \frac{1}{1 + \exp(\log \pi_j - \log \pi_i)}$$

と変形できるため,

$$(2.10) \quad r_i = \log \pi_i$$

の非線形変換により提案手法(式(2.1))と等価である。

これら二つの手法は等価であるが, 文献(Ray, 2011)の調査では, BT モデルに基づくレーティング手法を採用している競技は無く, それに対し提案手法の線形変換であるイロレーティングはいくつかの競技団体で採用されている。

また, 項目応答理論(Hambleton, 1991; de Ayala, 2008)では, 項目弁別度 a_i , 項目難易度 b_i , 個人能力パラメータ θ_j を用いて, 個人 j が項目 i に正解する確率 $p_{i,j}$ を

$$(2.11) \quad p_{i,j} = \frac{1}{1 + \exp(-a_i(\theta_j - b_i))}$$

により説明するモデルが採用されることがあり、そのパラメータ推定手法なども広く研究されている。

これをスポーツに拡張すると、大会 k でチーム i とチーム j が対戦した場合、 i が j に勝利する確率 $p_{i,j}$ が

$$(2.12) \quad p_{i,j} = \frac{1}{1 + \exp(-a_k(r_i - r_j))}$$

で説明されると仮定するモデルも構築することができる。 $a_k > 0$ は大会ごとの番狂わせの起こりやすさ・起こりにくさを意味するパラメータと解釈できる。

これらの、「既存競技での採用状況」、および「モデルの拡張性とその解釈」の観点から、本研究では(提案手法と等価ではあるが)BTモデルを採用せず、提案手法(式(2.1))を採用する。

3. レーティング算出および試合結果予測への適用

本章では、前章で提案したレーティング算出手法を実装し、各国代表チームのレーティングを算出する。

3.1 使用したデータ

以下のデータを使用した。オリンピック本選2大会(ロンドン, リオデジャネイロ)およびオリンピック世界最終予選, 男女計6大会を予測対象とし、それぞれの直近1年以内を目安とし、FIVB主催大会および各大陸のオリンピック予選の結果に基づきレーティングを算出した。このことにより、各チームが1年程度の期間を通じて安定的に保持している技能・実力をレーティングとして抽出できる。

データセット W1

- 2011年ワールドカップバレーボール(女子)[2011年11月]
- ロンドンオリンピック ヨーロッパ予選(女子)[2012年5月], 南アメリカ予選(女子)[2012年5月], 北アメリカ予選(女子)[2012年5月], アフリカ予選(女子)[2012年2月], 世界最終予選(女子)[2012年5月]
- 2012年ワールドグランプリ[2012年6月-7月]

予測対象：ロンドンオリンピック(女子)[2012年7月-8月]

データセット W2

- 2015年ワールドカップバレーボール(女子)[2015年8月-9月]
- リオデジャネイロオリンピック ヨーロッパ予選(女子)[2016年1月], 南アメリカ予選(女子)[2016年1月], 北アメリカ予選(女子)[2016年1月], アフリカ予選(女子)[2016年2月]

予測対象：リオデジャネイロオリンピック 世界最終予選1(女子), 2016年5月

データセット W3

W2に

- リオデジャネイロオリンピック 世界最終予選1(女子)[2016年5月]
- 2016年ワールドグランプリ[2016年6月-7月]

を追加したもの。

予測対象：リオデジャネイロオリンピック(女子)[2016年8月]

表 3. Root mean squares (RMSs).

	by Set		by Game	
Dataset W1	0.0697	0.0429	Dataset M1	0.0594 0.0381
Dataset W2	0.0701	0.0369	Dataset M2	0.0612 0.0305
Dataset W3	0.0710	0.0390	Dataset M3	0.0611 0.0321

データセット M1

- 2011 年ワールドカップバレーボール(男子) [2011 年 11 月–12 月]
- ロンドンオリンピック ヨーロッパ予選(男子) [2012 年 5 月], 南アメリカ予選(男子) [2012 年 5 月], 北アメリカ予選(男子) [2012 年 5 月], アフリカ予選(男子) [2012 年 1 月], 世界最終予選(男子) [2012 年 6 月]
- 2012 年ワールドリーグ [2012 年 5 月–7 月]

予測対象：ロンドンオリンピック(男子) [2012 年 7 月–8 月]

データセット M2

- 2015 年ワールドリーグ [2015 年 5 月–7 月]
- 2015 年ワールドカップバレーボール(男子) [2015 年 9 月]
- リオデジャネイロオリンピック ヨーロッパ予選(男子) [2016 年 1 月], 南アメリカ予選(男子) [2015 年 10 月], 北アメリカ予選(男子) [2016 年 1 月], アフリカ予選(男子) [2016 年 1 月]

予測対象：リオデジャネイロオリンピック 世界最終予選 1(男子) [2016 年 5 月]

データセット M3

M2 に

- リオデジャネイロオリンピック 世界最終予選 1(男子) [2016 年 5 月]
- 2016 年ワールドリーグ [2016 年 6 月–7 月]

を追加したもの。

予測対象：リオデジャネイロオリンピック(男子) [2016 年 8 月]

3.2 結果

3.2.1 数値的な安定性

アルゴリズムの数値的な安定性を確認するため、以下の数値実験を行った。データセット W3 に対し、初期レーティングベクトルの要素を正規乱数で与えた後和が 0 となるよう定数を加え、提案アルゴリズムを 25 回実行した。パラメータは $\epsilon_{th} = 10^{-5}$, $K = 0.1$ とした。得られた各国のレーティングの最大値と最小値の差の上限は 2.2×10^{-5} 、標準偏差の上限は 4.7×10^{-6} であり、以降の分析に対し十分な精度の値が安定して得られた。

表 3 に、それぞれのデータセットのレーティング算出対象の試合について、レーティングから予測される得点率 p_k と実際の得点率 s_k (k はセットまたは試合番号) との残差の二乗平均平方根 RMS (次式)を示す。

$$(3.1) \quad RMS = \sqrt{\frac{1}{N_S} \sum_{k=1}^{N_S} (p_k - s_k)^2}$$

データセット W3 を例とすると、RMS はセットごとで 0.0710、試合ごとで 0.0390 であった。セットごとの残差 $p_k - s_k$ を調べたところ、95% のセットが $|p_k - s_k| < 0.139$ を満たしていた。残差 0.139 は平均で 1 セットあたり約 6 点程度の予測誤差に相当する。試合ごとでは 95% の試

表 4. Correlation coefficients based on proposed rating method.

	Correlation coefficient		Correlation coefficient	
	by Set	by Game	by Set	by Game
Dataset W1	0.5110	0.7909	Dataset M1	0.5313 0.6775
Dataset W2	0.5549	0.7244	Dataset M2	0.4764 0.6413
Dataset W3	0.5605	0.7384	Dataset M3	0.5607 0.7381

合が $|p_k - s_k| < 0.076$ を満たしている. 3, 4, 5 セットの試合ではそれぞれ約 9.8, 13.1, 15.1 点の予測誤差に相当する.

3.2.2 主結果

表 4 に説明変数をレーティング差から予測される得点率(式(2.1)), 決定変数を予測対象の試合での実際の得点率(セットごと, および試合ごと)としたときの相関係数を示す. 図 3 および図 4 に, W3 および M3 で算出したレーティング差とリオデジャネイロオリンピックで行われた実際の試合ごとの得点率を示す. これらの図中の実線は式(2.1)で予測される得点率を示している.

3.3 考察

表 4 の結果より,

- レーティングから予測された得点率と各セットの実際の得点率には中程度の相関があり(相関係数が 0.5 強),
- レーティングから予測された得点率と各試合の実際の得点率には中程度から強い相関がある(相関係数が 0.7 付近)

ことが確認できた.

また図 3 および図 4 より, 番狂わせはレーティング差が小さい対戦で若干起こっており, 一定以上のレーティング差では番狂わせが起きていないことが示されている.

これらの結果より, 提案手法に基づくレーティングは各チームの実力を適切に評価し, その後の試合結果をある程度予測できることが示された.

3.4 FIVB ランキングとの比較

本節では提案手法に基づくレーティングと FIVB ランキングのどちらが試合結果をより正確に予測できるかについて議論する.

全てのデータセットに対し, 予測対象とした大会での各セットおよび各試合の得点率と, 直前の FIVB ランキング差および FIVB ランキングポイント差との相関係数を求めたものを表 5 に示す. ロンドンオリンピック(データセット W1 および M1)については, 開催国枠で出場したイギリスの世界ランキングが極端に低く(女子 69 位, 男子 92 位), 外れ値となる恐れがあったため, イギリスの試合結果を除外した結果もあわせて示す.

ランキング差の算出にはランキングの順位をそのまま使用しているため, 相関係数はマイナスになる(ランキング差がマイナスになるほど自分のランキングが高い). 一部強い相関を示している大会もあるが, おおむね提案手法よりも弱い相関を示している. データセット M1 とデータセット M1*(イギリスを除外したもの)を比較すると, 極端にランキング差およびランキングポイント差の絶対値が大きいデータにより擬似的な相関が導入されていることが分かる.

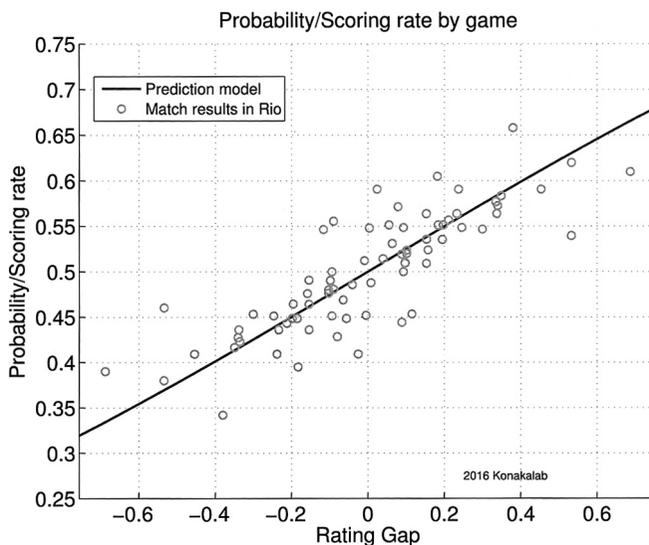


図 3. Rate difference and scoring rate in each game (Rio Olympic Games 2016, Women).

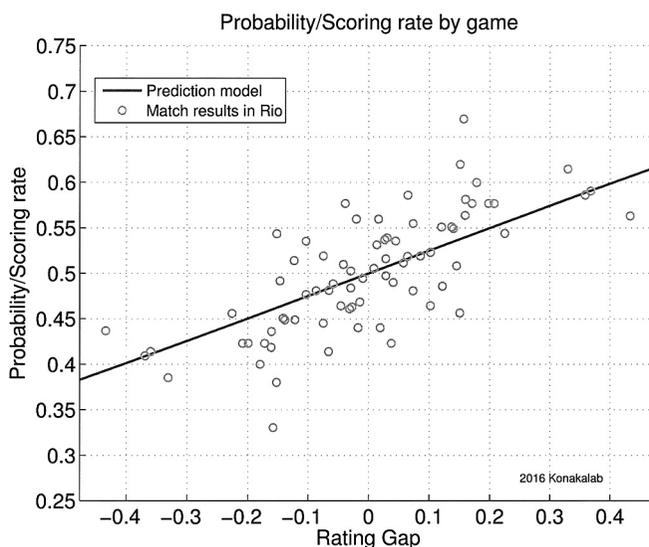


図 4. Rate difference and scoring rate in each game (Rio Olympic Games 2016, Men).

相関が低くなる要因は、ランキングおよびランキングポイントが間隔尺度となっていないことである。ランキングまたはランキングポイントに何らかの非線形な変換を施すことにより相関が見かけ上高くなることも考えられるが、相関を高くするように選択された変換が定性的な意味を持つとは限らない。この点において、レーティング差は得点確率を説明している、つまりレーティングは間隔尺度である、という統計的意味が明確である提案手法が優れている。

図 5 に W3 に基づき算出したレーティングとその時点での FIVB ランキングポイントとの関係を示す。

表 5. Correlation coefficients: FIVB ranking and scoring rate.

	Correlation coefficient			
	FIVB ranking gap		FIVB ranking point gap	
	by Set	by Game	by Set	by Game
Dataset W1	-0.3074	-0.5491	0.4179	0.5983
Dataset W1*	-0.3768	-0.5417	0.3763	0.5267
Dataset W2	-0.5078	-0.6578	0.3658	0.4901
Dataset W3	-0.4948	-0.6629	0.4713	0.6446
Dataset M1	-0.4789	-0.6105	0.5362	0.7016
Dataset M1*	-0.4587	-0.6331	0.4731	0.6512
Dataset M2	-0.3624	-0.5380	0.3171	0.4641
Dataset M3	-0.5992	-0.7899	0.5342	0.7166

*: except GBR.

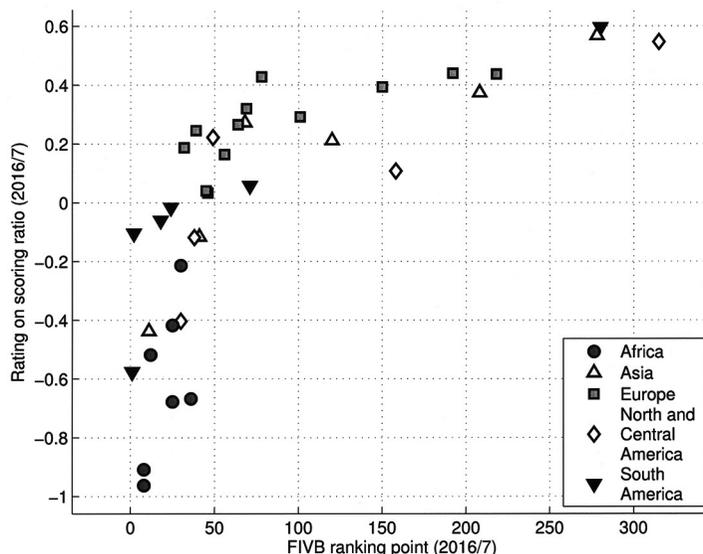


図 5. FIVB ranking point and proposed rating (2016/7, Women).

図 5 より、レーティングが 0.4 前後のヨーロッパ各国間でのランキングポイントの差が大きいことや、ランキングポイントが 40 前後の各国間でのレーティング差が大きいことが分かる。この結果からも、FIVB ランキングポイントがチームの実力を正しく反映していないことが分かる。この原因については次節で議論する。

4. 大会形式の評価への応用

本章では提案手法に基づくレーティングに基づく大会形式の評価(主に不適切な設計の指摘)について述べる。

4.1 世界最終予選の不適切な設計

リオデジャネイロオリンピック世界最終予選の参加国を以下に示す。

- 世界最終予選 1: アジア (FIVB ランキングアジア内上位 3 カ国および開催国(日本)), ヨーロッパ(予選 2・3 位), 北アメリカ(予選 2 位), 南アメリカ(予選 2 位). このうち 4 カ国(アジア最上位国とそれを除く上位 3 カ国)がオリンピック出場.
- 世界最終予選 2: アフリカ(予選 2・3 位), 北アメリカ(予選 3 位), 南アメリカ(予選 3 位). このうち最上位 1 カ国がオリンピック出場.

いずれの最終予選も 1 回戦総当りで, 順位は勝利数が多い順で決められた. この方式では北アメリカ予選で世界最終予選 2 を「選択」することでオリンピックへの出場権を獲得しやすくなる可能性がある. この時点で敗退行為を誘発しうる不適切な制度設計であり, 実際に最終予選 1 の出場国が疑問を呈している(奥村, 2016).

表 6 に世界最終予選直前のレーティング, 世界最終予選前最新の FIVB ランキングポイント(2015 年 10 月発表), 最終的なオリンピック出場権, および各予選での順位を示す. 国名は IOC の 3 文字コード(International Olympic Committee, 2016)で示す. BRA は開催国, CHN および SRB は 2015 年ワールドカップでオリンピック出場権を得ているため, 各大陸予選には参加していない.

前節での考察と同様に, 表 6 は, FIVB ランキングポイントが実力を適切に評価できていないことを示している. 各大陸選手権も含めて, 大会の順位に基づいてポイントが付与されている(表 1)ので, チーム数の多いヨーロッパには高レーティングにも関わらずポイントが少ないチームがあり(特に NED, GER), それに対してそれ以外の大陸の上位チーム(JPN, KOR(アジア), ARG(南アメリカ), KEN(アフリカ)など)はレーティングに対して付与ポイントが多い. FIVB ランキングが「世界」ランキングとして実力を正しく表せていないことが分かる.

レーティング以外にも, 大規模な国際大会での得点率を利用することでも FIVB ランキングの不適切な設計を指摘することができる. バレーボールでは各大陸から代表チームが出場する国際大会(ワールドグランプリなど)が毎年開催され, 大陸間の実力差は定性的には明確である. 表 7 に 2016 年ワールドグランプリでの大陸間対戦での得点率を示す. アフリカを除き, 各大陸で最上位の成績であったチーム(BRA, CHN, RUS, USA)を除外し, 中堅チーム間対戦のみの得点率もあわせて示す.

この結果からもアフリカ各国チームの実力が劣っていることは明確であり, また北アメリカは大陸 1 位(USA)とそれ以下との実力差が大きいことも読み取れる.

これらの統計に基づく, 北アメリカ予選に参加する USA は明確に実力が高く, 1 位での通過がほぼ確実である. 北アメリカの 2 位は世界最終予選 1 に回ることとなるが, ここにはヨーロッパの 2 位および 3 位, およびアジアの上位国(JPN, THA, KOR)が参加する. つまり, レーティングが 0.40 から 0.25 程度の 4 チームの出場が確定しており, レーティングが 0.2 付近の北アメリカ 2, 3 番手から見ると勝利が難しい相手ばかりである. したがって, この中で 4 位以上となることも困難が予想される.

それに対し, 世界最終予選 2 の対戦国は南アメリカ, アフリカからの 3 カ国であり, いずれもレーティングは 0 以下である. 北アメリカの 3 位になり世界最終予選 2 に出場すれば, そこで 1 位を目指す方がハードルが低いように思われる.

北アメリカ大陸予選は 4 チームの 1 回戦総当りであり, かつ USA が 3 勝することがほぼ確実視されている. したがって, 2 位と 3 位を争う 2 チームの順位は直接対決 1 試合の結果のみで決まる可能性は低くは無い.

この状況で行われた北アメリカ予選の初戦で DOM-PUR の対戦が行われ, セット数 3-1 で DOM が勝利した. USA は実力どおり 3 連勝し, DOM が北アメリカ 2 位で最終予選 1 へ, PUR が北アメリカ 3 位で最終予選 2 へ進出した. DOM は最終予選 1 でヨーロッパ, アジアの強豪

表 6. Rating just before Rio WOQT (2016/5) and qualifying rankings.

IOC Code	FIVB Ranking		Qualifiers(**)						
	Rating	Points (*)	OL	AF	EU	NA	SA	WOQT1	WOQT2
'BRA'	0.5951	302	✓						
'CHN'	0.5583	310	✓						
'USA'	0.4839	350	✓			1			
'SRB'	0.4308	212	✓						
'RUS'	0.4014	253	✓		1				
'NED'	0.3913	82	✓		2			2	
'GER'	0.3713	93			5				
'JPN'	0.3440	225	✓					3	
'ITA'	0.3330	163	✓		3			1	
'THA'	0.2770	83						5	
'TUR'	0.2472	117			4				
'BEL'	0.2115	76			6				
'CZE'	0.1962	38							
'KOR'	0.1927	146	✓					4	
'POL'	0.1875	44			7				
'PUR'	0.1613	70	✓			3			1
'BUL'	0.0951	57							
'DOM'	0.0685	186				2		6	
'CAN'	0.0494	64				4			
'ARG'	0.0386	87	✓				1		
'CRO'	-0.0427	51			8				
'CUB'	-0.0627	52							
'PER'	-0.0659	51					2	7	
'COL'	-0.1147	41					3		3
'KAZ'	-0.1430	42						8	
'VEN'	-0.1467	16					4		
'KEN'	-0.2257	58			3				2
'CMR'	-0.4291	51	✓	1					
'EGY'	-0.5280	34		2					
'CHL'	-0.6191	11					5		
'ALG'	-0.6747	51		4					4
'TUN'	-0.6906	25		6					
'BOT'	-0.9197	22		5					
'UGA'	-0.9715	8		7					

(*) FIVB Senior World Ranking - Women. As per 6th October 2015 (FIVB, 2015). It is the latest FIVB ranking before WOQTs.

(**) OL: Olympic games, AF: Africa, EU: Europe, NA: North and Central America, SA: South America, WOQT: World Olympic Qualifying Tournament

国と対戦し、8チーム中6位でオリンピック出場を逃している。それに対しPURは実力の劣る南アメリカ、アフリカの3チームと対戦し、1セットも落とすことなく3戦全勝の1位でオリンピック出場権を得ている。

この状況下での世界最終予選の設計は上記の理由から明らかに不適切であり、該当するチームに対し不要な疑いが向けられかねない。なおかつ、前述の通り、各大陸の上位チームには実力(実際の試合で得点を取る能力)に見合わないランキングポイントが多く付与される傾向にあ

表 7. Scoring ratio in inter-continental matches in World Grand Prix 2016.

(a) All teams							(b) Excluding the strongest team in four continents						
(*)	Teams	Opponents					(*)	Teams	Opponents				
		AF	AS	EU	NA	SA			AF	AS	EU	NA	SA
AF	2	—	0.318	0.377	0.422	0.381	AF	2	—	0.318	0.377	0.421	0.381
AS	5	0.682	—	0.494	0.494	0.444	AS	4	0.682	—	0.487	0.516	0.427
EU	11	0.623	0.506	—	0.494	0.479	EU	10	0.623	0.513	—	0.523	0.529
NA	6	0.579	0.506	0.506	—	0.482	NA	5	0.579	0.482	0.477	—	0.482
SA	4	0.619	0.556	0.525	0.518	—	SA	3	0.619	0.571	0.471	0.518	—

(*) continent names; AF: Africa, AS: Asia, EU: Europe, NA: North and central America, SA: South America

り、公式に発表されるランキングポイントを信用すると却って大陸間の実力差が見えにくくなってしまふこともあわせて問題である。

敗退行為・談合試合が起こるかどうかは大会方式を設計する上で非常に重要な点である。サッカーで談合試合の疑惑をもたれている有名な事例が1982年ワールドカップスペイン大会で起こっている(Smyth, 2014)。一次リーグ最終戦で対戦した2チームが、お互いが一次リーグを突破できるように試合を操作したのではないかという疑惑である。原因は一次リーグ最終節の試合時間が異なり、当該の2チームがともに一次リーグを勝ち抜く条件が試合前に確定していたことである。この事例を受け、1986年大会から一次リーグ最終節の2試合は同時刻に開始するよう変更されている。

他のスポーツでも、ロンドンオリンピックのバドミントンでは準々決勝の組み合わせを有利にするために4組のペアが故意の失点を繰り返したことが問題となった(Gilmour, 2012)。これは故意に負けることがその先で明らかに有利になるよう、不適切に一次リーグ戦を設計したことが大きな要因である(これ以前の大会のようにトーナメント戦であれば敗退行為は起こりえない)。これを受け、続くリオデジャネイロオリンピックでは大会形式が変更されている。

Jリーグで2シーズンのみ実施された2ステージ+チャンピオンシップ制についても、敗退行為の可能性を排除するよう制度設計が行われたことが公式に発表されている(J League, 2013; 泉・小中, 2016)。

上記の事例に鑑み、バレーボールでも最終予選制度の適切な再設計が望ましい。

4.2 リオデジャネイロオリンピックの予選プールの不均衡

表8にオリンピック本選における予選プールを示す。各プールは直前のFIVBランキングに基づいて決定されたが、ランキングとレーティングによる実力の評価には大きな隔たりがあり、Pool Bに実力国が集まっていることが分かる。ITAはPool Bで6チーム中5位となり準々決勝に進出できなかったものの、レーティングでは全体の7番目であり、Pool Bの上位4カ国が全て準々決勝を通過していることから、適切に組み分けが設計されていれば準々決勝進出が十分期待できる実力であったと思われる。

Pool Aではアジアおよび南アメリカの3カ国が実力以上のランキングを得ていることが分かる。特にJPNは主要国際大会であるワールドカップバレーボールの1977年以降全ての大会に開催国として参加し、ランキングポイントを獲得している影響が大きい。ワールドカップの出場国数は12カ国のみ、かつ大陸ごとの参加枠があるため、JPNは必ず実力の劣る数チームと対戦でき、ランキングポイントを獲得できる。それに対しヨーロッパからワールドカップへの出場枠は2つのみであり、出場権を獲得できないチーム(2015年大会ではITA, NEDなど)は

表 8. Rio 2016 women's volleyball pools and results.

Pool A					Pool B				
IOC code	FIVB ranking	Rating	Result (Pool)	Result (Final)	IOC code	FIVB ranking	Rating	Result (Pool)	Result (Final)
BRA	3 (Host)	0.5963	1	5T	USA	1	0.5464	1	3
RUS	4	0.4365	2	5T	CHN	2	0.5678	4	1
JPN	5	0.3737	4	5T	SRB	6	0.4396	3	2
KOR	9	0.2115	3	5T	ITA	8	0.3934	5	9T
ARG	12	0.0574	5	9T	NED	14	0.4276	2	4
CMR	21	-0.4174	6	11T	PUR	16	0.2222	6	11T

当然ランキングポイントも 0 であり、これが世界ランキングでの不当な低評価の原因である。

ランキングに基づいた大会制度設計の大前提は実力を適切に反映したランキングであり、その実現のためには、大会の参加権の与え方、および各大会・各順位におけるランキングポイントの抜本的な修正が必要である。

5. 結論

本論文では、バレーボールの各国代表チームの実力の定量的な評価方法として、各セットでの得点率を説明するモデルを構築し、そのパラメータをレーティングとして定義した。またレーティングの具体的な計算アルゴリズムを示し、実際の国際大会の試合結果に基づき値を算出した。

算出したレーティングは公式のランキングである FIVB ランキングよりも試合結果を適切に予測できることを、オリンピックなど複数の大会の結果により確認した。レーティング差が世界ランキング(ポイント)差よりも適切に実力差を説明できることに基づいて、世界最終予選やオリンピックでの大会形式の不適切な設計、およびその原因である FIVB ランキングポイントの不適切な設計を具体的な事例に基づき指摘した。

参 考 文 献

- Araújo, R.M., Castro, J., Marcelino, R. and Mesquita, I.R. (2010). Relationship between the opponent block and the hitter in elite male volleyball, *Journal of Quantitative Analysis in Sports*, **6**(4), 1–12.
- ATP World Tour (2016). Rankings FAQ, <http://www.atpworldtour.com/en/rankings/rankings-faq>, accessed 2016/6/14.
- Burton, T. and Powers, S. (2015). A linear model for estimating optimal service error fraction in volleyball, *Journal of Quantitative Analysis in Sports*, **11**(2), 117–129.
- Davidson, R.R. (1970). On extending the Bradley-Terry model to accommodate ties in paired comparison experiments, *Journal of the American Statistical Association*, **65**(329), 317–328, <http://www.jstor.org/stable/2283595>.
- de Ayala, R.J. (2008). *The Theory and Practice of Item Response Theory (Methodology in the Social Sciences)*, Guilford Press, New York.
- Dziedzic, E. and Hunter, G. (2015). Predicting the results of tennis and volleyball matches using regression models, and applications to gambling strategies, *Proceedings of the 5th International Conference on Mathematics in Sport*, 32–40.
- Elo, A.E. (1979). *Ratings of Chess Players Past and Present*, Arco Publishing, New York.

- Eom, H.J. and Schutz, R.W. (1992). Statistical analyses of volleyball team performance, *Research Quarterly for Exercise and Sport*, **63**(1), 11–18.
- Ferrante, M. and Fonseca, G. (2014). On the winning probabilities and mean durations of volleyball, *Journal of Quantitative Analysis in Sports*, **10**(2), 91–98.
- FIVB (2015). FIVB Senior World Ranking - Women. As per 6th October 2015, http://www.fivb.org/en/volleyball/VB_Ranking_W_2015-10.asp, accessed 2017/2/10.
- FIVB (2016). FIVB volleyball world rankings, <http://www.fivb.org/en/volleyball/Rankings.asp>, accessed 2016/6/14.
- Florence, L.W., Fellingham, G.W., Vehrs, P.R. and Mortensen, N.P. (2008). Skill evaluation in women's volleyball, *Journal of Quantitative Analysis in Sports*, **4**(2).
- Gilmour, R. (2012). Badminton match-fixing scandal: How and why the four pairs were disqualified from the London 2012 Olympics, <http://www.telegraph.co.uk/sport/olympics/badminton/9444025/Badminton-match-fixing-scandal-how-and-why-the-four-pairs-were-disqualified-from-the-London-2012-Olympics.html>, accessed 2015/3/2.
- Glasson, S., Jeremiejczyk, B. and Clarke, S.R. (2001). Simulation of women's beach volleyball tournaments, *Australian Society for Operations Research*, **20**(2), 2–7.
- Hambleton, R. (1991). *Fundamentals of Item Response Theory (Measurement Methods for the Social Science)*, new edition, Sage Publications, California.
- Indyk, P. and Motwani, R. (1998). Approximate nearest neighbors: Towards removing the curse of dimensionality, *Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing*, 604–613.
- International Olympic Committee (2016). List of all National Olympic Committees in IOC protocol order, <https://www.olympic.org/national-olympic-committees>, accessed 2016/6/15.
- 泉武志, 小中英嗣 (2016). J1 リーグ 2 ステージ + ポストシーズン制度の統計的分析, *日本オペレーションズ・リサーチ学会論文誌*, **59**, 21–37.
- J League (2013). 10 月 30 日開催 大会方式変更に関する意見交換会についてのチェアマンコメント, <http://www.jleague.jp/release/article-00005433/>, accessed 2017/2/14.
- Knapper, E. and McIlwain, H. (2015). Predicting wins and losses: A volleyball case study, *The College Mathematics Journal*, **46**(5), 352–358, <http://www.jstor.org/stable/10.4169/college.math.j.46.5.352>.
- Massey, K. (1997). Massey Rating, <http://www.masseyratings.com/>, accessed 2016/6/14.
- 奥村信哉 (2016). 不公平感漂う「もう一つ」の最終予選! 日本大会出場の監督ら相次ぎ疑問の声 知られざるプエルトリコ大会, <http://www.sankei.com/sports/news/160524/spo1605240002-n1.html>, accessed 2016/5/25.
- Rao, P.V. and Kupper, L.L. (1967). Ties in paired-comparison experiments: A generalization of the Bradley-Terry model, *Journal of the American Statistical Association*, **62**(317), 194–204.
- Ray, S. (2011). The methodology of officially recognized international sports rating systems, *Journal of Quantitative Analysis in Sports*, **7**(4).
- Sackmann, J. (2017). <https://github.com/JeffSackmann>, accessed 2017/4/4.
- Smyth, R. (2014). World Cup: 25 stunning moments ... No3: West Germany 1-0 Austria in 1982, <http://www.theguardian.com/football/blog/2014/feb/25/world-cup-25-stunning-moments-no3-germany-austria-1982-rob-smyth>, accessed 2015/3/3.
- World Rugby (2014). Rankings Explanation, <http://www.worldrugby.org/rankings/explanation>, accessed 2016/6/14.
- World Rugby (2015). World rankings confirm Japan's victory as biggest shock, <http://www.rugbyworldcup.com/news/111746>, accessed 2016/6/14.
- Zetou, E., Moustakidis, A., Tsigilis, N. and Komninakidou, A. (2007). Does effectiveness of skill in complex i predict win in men's Olympic volleyball games?, *Journal of Quantitative Analysis in Sports*, **3**(4).

Statistical Rating Method of Volleyball National Teams to Predict Results and Determine Competition Format Design

Eiji Konaka

Faculty of Science and Technology, Meijo University

The Fédération Internationale de Volleyball (FIVB), the world's governing body for volleyball, regularly ranks its member nations' teams. The FIVB Board of Administration has designed a point system for select FIVB world and other official competitions. However, the point system does not have a clear mathematical or statistical background. Consequently, this system cannot be used as a quantitative measure of a team's skill.

This paper proposes a novel mathematics-based rating and ranking system of national volleyball teams. The rating, which is a parameter reflecting the skill of the team, is calculated based on the scoring ratio of teams in each major international competition. A logistic regression model is employed to explain the scoring ratio with respect to the rating difference between two teams. Additionally, an iterative rating calculation method is proposed. Numerical experiments demonstrate the stability of the proposed method. The correlation coefficient between the proposed rating difference and the results of several major international competitions (e.g., Rio Olympic Games) is about 0.7. This value shows a strong correlation that is higher than that of the FIVB ranking (point) difference. The proposed rating is used to highlight the improper design of the ranking point attribution system and the competition format of the World Olympic Qualifying Tournament and Rio Olympic Games.

重力モデルを用いたサッカー選手の動きの定量化

土田 潤¹・宿久 洋²

(受付 2016 年 12 月 30 日；改訂 2017 年 3 月 10 日；採択 5 月 15 日)

要 旨

サッカー選手の動きに関して、ボールを保持している選手に着目した研究は多数存在する。また、ボールを持っていない選手を含めた研究は戦術などの、選手全体の動きに関する評価が多く、個人への評価は少ないのが現状である。本稿では、全体の動きを考慮しつつ、選手個人の動きを評価する指標として、選手質量を用いる。選手質量の定義には重力モデルを用いる。重力モデルにおいて選手質量はパラメータとなるが、重力モデルの式が対数線形モデルと同様のモデル式として定義されることから、選手質量の推定が可能となり、選手質量について、対数線形モデルにおける主効果と同様の意味付けができることを報告する。また、本稿ではトラッキングデータを用いて、選手間密度、および選手間距離を算出し、選手質量を推定した結果を報告する。加えて、推定された選手質量に対して、階層ベイズモデルを用いてどのようなプレーが選手質量に影響を与えているかをモデリングした結果を報告する。

キーワード：階層ベイズモデル、スポーツデータ解析、対数線形モデル。

1. はじめに

サッカーは世界的に非常に人気のスポーツであり、多くのプロチームが存在する。プロチームは勝利のためにデータを活用することが多く、データ活用の論文も多く発表されている。例えば、Rue and Salvesen (2000)では、ゲームのスコアを予測するモデルをベイジアン Dynamic 一般化線形モデルを用いて構築している。Dobson and Goddard (2008)では、ホーム・アウェイなどの情報を用いることで、ゲームスコアの推定を行っている。これらの研究に代表されるように、多くの研究は各試合の集計結果を用いたデータを活用することが主流である。また、サッカー選手の動きやサッカーのシーンをどのように観測するかも研究対象となっている。Ekin et al. (2003)や Ren and Jose (2005)では、サッカー中継から、サッカー中継のどのようなシーンなのかを判別するモデルを提案しており、選手の動きを画像検出の方法を用いて類型化し、ズーム、リプレイなどのシーンを判別している。

サッカーのチーム戦術の分野では、ロボットサッカーにおいて研究が盛んである。その理由としてロボットサッカーは、4人チームであり、選手がロボットであることから個々の能力の差が少ないため、人対人に比べ、戦術によって勝敗が決定しやすいことが挙げられる。また、フィールドの大きさが人対人のサッカーに比べ小さく、カメラ一台でフィールド全体を写すことができ、データの観測が比較的容易であることも戦術研究が盛んになる要因であると考えられる。Nadarajah and Sundaraj (2013)では、ロボットとボール間の距離やそれぞれの速度を用

¹ 同志社大学大学院 文化情報学研究科：〒610-0394 京都府京田辺市多々羅都谷 1-3

² 同志社大学 文化情報学部：〒610-0394 京都府京田辺市多々羅都谷 1-3

いて、ロボットのポジショニングに関する戦術や、ゴールをどのように守ればよいのかを総合的に報告している。

しかし、近年の観測技術の発達により、人対人のサッカーにおいても、ボールを持った選手だけでなく、ボールを持っていない選手に関するデータが得られるようになり、ある時点でのフィールドにいる選手全員のプレーを分析することが可能になった。全員を同時に、特に、ボールを持っていない選手を分析することは重要である。加藤(1999)では、サッカーにおいて1人の選手がボールをプレーするのは多くても2分から3分であり、選手がボールを持っていない状態での動きの解析が求められていることが述べられている。

実際に、ボールを持っていない選手を含めたサッカーの分析は近年多く行われている。神谷他(2016)では、選手のトラッキングデータを用いて、サッカーの戦況変化の自動検出を選手の守備位置や選手間の距離を用いて行なわれている。また、成塚 他(2016)では、選手間の距離や角度の特徴について考察が行われている。可視化についてもボールの動きだけでなく、船山他(2016)では、ホットゾーンの利用人数の動的表示やカーネル密度推定を用いてフィールドのどこに人が集中しているかを可視化している。これらの研究の多くは、フィールド全体ないし、選手全体について考察していることが多く、個人の能力について言及しているものは少ない。

個人の能力や戦術の評価については、トラッキングデータではなく、ビデオデータを用いて行われている研究がある。瀧・長谷川(1998)では、チームという集団の動きの中で個人が作る領域について考察し、ボロノイ領域を拡張することで個人の特徴量を算出し、得られた特徴量から個人の動きを評価している。特に、サッカーという限られたエリアでの集団の密度は、各個人の周りの密度の集合として還元されると考えられることを述べている。つまり、選手個人の動きによって作られる密度を評価をすることで、全体の密度が評価できるとしている。また、Moura et al. (2015)では2012年ヨーロッパチャンピオンシップのデータに対して、個人の動きに対してPCAを適用し、選手のポジションやその範囲がチームごとによって異なることを示している。このように、選手全員を用いた、選手個人の動きに対する定量的評価はまだ初期段階であると考えられる。

本稿では、瀧・長谷川(1998)の考え方にに基づき、個人の周りの密度を評価することで、全体を考慮しながら、選手個人の動きを評価することを考える。まず、選手間の範囲で作られる密度はその選手個人の質量によって表現されるものとして捉え、選手個人の質量を選手間で構成される範囲の密度を表すパラメータとして重力モデルを用いて定式化する。このとき、選手間の範囲で作られる密度はデータとして与えられるが、個人の質量はパラメータである。そのため、データから個人質量を推定する必要があるが、個人質量の推定が、対数線形モデルの主効果の推定と同様であることを報告する。

本稿では、実際のトラッキングデータを用いて、個人質量を推定し、さらに、推定された個人の質量を高める行動とはどのような行動であるのかを階層ベイズモデルによって表現した結果を報告する。2章では、重力モデルを用いるためのデータの作成法および、個人質量の推定法について述べる。3章では、推定された個人の質量について階層ベイズモデルを適用し、どのような行動が質量に関連しているかを明らかにする。4章では、まとめとして、得られた結果に対する考察および、今後の課題について述べる。

2. 重力モデルを用いた選手の質量推定

この章では重力モデルを用いて選手の質量を定式化する。そして、選手質量の推定が、対数線形モデルの推定法と同一となることから、選手質量の意味について考察する。

2.1 重力モデルを用いた選手質量の定式化

重力モデルは Tinbergen (1962) に代表されるように経済貿易のモデルとして用いられている。簡略化された重力モデルは対象 i, j 間で観測される量を F_{ij} とすれば,

$$(2.1) \quad F_{ij} = \frac{M_i M_j}{D_{ij}}$$

で定義される。ここで、 D_{ij} は対象 i, j 間の距離であり、 M_i は対象 i の質量である。両辺の対数を取れば,

$$(2.2) \quad \log F_{ij} = \log M_i + \log M_j - \log D_{ij}$$

と表される。経済貿易のモデルでは、 F_{ij} を国 i, j 間の貿易量、 D_{ij} を国 i, j 間の距離、 M_i を国 i の総貿易量としている。

本稿では F_{ij} を選手 i, j 間の密度とする。 F_{ij}, D_{ij} が与えられたもとの、 $\log F_{ij}^* = \log F_{ij} + \log D_{ij}$ と置けば、式 (2.2) は

$$(2.3) \quad \log F_{ij}^* = \log M_i + \log M_j$$

与えられる。サッカーのフィールドの大きさは状況によらず一定であり、多くの場合、フィールドに計 22 人の選手でいることから、質量の総和は一定であるとする。

重力モデルの良い点として、選手間密度および、選手間距離が与えられれば、選手質量を定義することができる点が挙げられる。また、 i, j を選手だけでなく、フィールドを分割した区分とすれば、フィールドの質量として解釈も可能である。質量を解釈すれば、質量が大きい選手や区分の周辺に選手が集まっていることを意味する。

本稿では、図 1 のような選手間領域と選手領域を用いて、選手間密度 F_{ij} を定義する。選手 i, j 間の距離 D_{ij} を半径、対象 i を中心とする円を選手 i, j 間の領域 R_{ij} とし、 S_k を i の対戦チームの選手 k の領域とする。 S_k は i の対戦チームの選手 k を中心とする半径 1.7m の円とした。領域の半径 1.7m としたのは成人の平均身長であり、選手が一步、領域方向に進み、手を伸ばせば領域 R_{ij} に侵入できるためである。このとき、 R_{ij} と S_k が交わる k の個数を R_{ij} の面積で割ったものを F_{ij} とする。つまり、 F_{ij} は以下の式で与えられる。

$$(2.4) \quad F_{ij} = \frac{\sum_{k \in O_i} I(R_{ij} \cap S_k \neq \phi)}{D_{ij}^2 \pi}$$

ここで、

$$R_{ij} = \{x \in \mathbb{R}^2 \mid \|x - y_i\| \leq D_{ij}\}$$

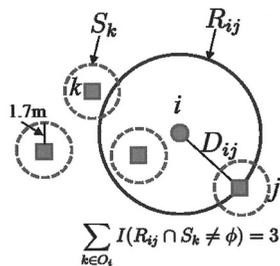


図 1. カウントする領域と選手の持つ領域。

$$S_k = \{x \in \mathbb{R}^2 \mid \|x - y_k\| \leq 1.7\}$$

であり、 O_i は対象 i の対戦チームの選手集合、 y_i は対象 i の座標ベクトルである。また、 $I(\cdot)$ は定義関数であり、 π は円周率である。式(2.4)を F_{ij}^* で書き直せば、

$$(2.5) \quad F_{ij}^* = \frac{\sum_{k \in O_i} I(R_{ij} \cap S_k \neq \phi)}{D_{ij} \pi}$$

となる。経済で用いられている重力モデルでは F_{ij} は対称であることが多いが、式(2.4)および、式(2.5)は非対称、つまり、 $F_{ij} \neq F_{ji}$, $F_{ij}^* \neq F_{ji}^*$ となることに注意されたい。

これは、中心が異なる円を用いている点、および、対戦相手の数を数えているためであり、領域 R_{ij} で数え上げる対象選手をフィールドにいる選手全員としても、必ずしも対称となるわけではない。また、対象 i の座標ベクトルを2次元としたのは、選手座標が X, Y 座標によって取られていることを想定しているためであり、一般に2次元である必要はない。

式(2.5)を用いて選手間密度を定義した場合、選手質量が大きい選手には、近くに多くの選手が存在しやすいことを表す。つまり、分母である選手間距離が小さく、分子である近くにいる選手の数が多い傾向があると考えられる。攻撃側の視点に立てば、質量の大きい選手は相手に厳しくマークされていることが読み取れ、質量が小さい選手はフリーであることが読み取れる。相手のゴールに近い位置で質量が大きい選手は囷として、フリーになる選手を作る役割を持つことができる。また、相手のゴールに近い位置で質量が小さい選手は、フリーに近い状態であると考えられるため、その選手がボールを得ることができれば、得点を期待することができる。守備側の視点に立てば、質量の大きい選手は近くに相手選手がおり、相手がフリーになるのを防いでいると考えることができる。一方で、質量が小さい選手は近くに相手選手が少なく、その選手自体がフリーの状態であり、フリーという意味で、守備に積極的に貢献しているわけではないと考えられる。

2.2 選手質量の推定

式(2.3)は和について制約を入れれば、対数線形モデルと同様となる。ここでは、対数線形モデルと同様、 $\log F_{ij}^*$ を i, j についての総和を0とし、それぞれの選手質量の総和を0とする。 $\log F_{ij}^*$ の総和が一定というのは、フィールドの大きさおよび、選手の数が一定であることから、自然な仮定である。以上のことからモデル式は以下ようになる。

$$(2.6) \quad \log F_{ij}^* = \log \lambda_i + \log \kappa_j + \mu$$

ただし、

$$\sum_{i \in O} \sum_{j \in O} \log F_{ij}^* = 0, \quad \sum_{i \in O} \log \lambda_i = 0, \quad \sum_{j \in O} \log \kappa_j = 0$$

である。ここで、 O はフィールドにいる選手の集合である。パラメータの推定量は対数線形モデルと同様、

$$\mu = \frac{1}{|O|^2} \sum_{i \in O} \sum_{j \in O} F_{ij}^*, \quad \log \lambda_i = \frac{1}{|O|} \sum_{j \in O} F_{ij}^* - \mu, \quad \log \kappa_j = \frac{1}{|O|} \sum_{i \in O} F_{ij}^* - \mu$$

で与えられる。 μ は総和を一定にするパラメータであり、どれだけフィールドにフリーである領域が残っているかを表すパラメータと解釈することができる。 $\log F_{ij}^*$ を i, j についての総和を0というデータの制約のため、 μ は0として推定される。実際には、与えられたデータである $\log F_{ij}^*$ の平均を求め、 $\log F_{ij}^* - \mu$ を新しいデータとし、その行平均、列平均を用いて選手質

量を求めればよい。

式(2.6)を対数線形モデルとして捉えれば、選手質量の対数は対数線形モデルの行主効果、列主効果としてとらえることができる。また、対数線形モデルであることから、対象 i に対して行主効果、列主効果の2つが推定される。対象 i の行主効果および列主効果の値は F_{ij}^* が非対称であることから、一般には異なる。つまり、 $\log \lambda_i \neq \log \kappa_i$ である。そのため、2種類の選手質量の対数 $\log \lambda_i, \log \kappa_i$ はそれぞれ解釈が異なる。行主効果の推定値 $\log \hat{\lambda}_i$ は選手 i を中心とする円の平均的な密度を表している。したがって、行の質量が大きい選手は周りに相手チームの選手が多く、その選手の周りにスペースが少ないことを表している。したがって、選手 i がどれだけ相手チームの選手にマークされているかがわかる。選手質量として解釈すれば、行の質量はその選手自身が相手選手をどれだけ引き付けているかを示す量であり、大きい選手の周りに他の選手が集まりやすいことを示す。

列主効果の推定値 $\log \hat{\kappa}_i$ は相手チームの選手を中心とし、選手 i との距離を半径とする円の平均的な密度について表している。そのため、 $\log \hat{\kappa}_i$ が大きい選手は、相手チームを中心とした円の半径が小さくなりやすい位置にいる傾向がある。このことから、列の質量が大きい選手は相手チームの選手との重心にいることを示していると考えられる。つまり、守備の意味では、相手チームの重心にいることによりボール回しを止められる位置にいると考えられ、攻撃の意味では、動くことで、どちらにスペースを作るかを決められるポジションにいると考えられる。選手質量として解釈すれば、質量が大きい選手は行側の選手質量の釣り合いを保つための質量が大きいことを意味している。

3. 階層ベイズモデルを用いた選手質量と選手の動きのモデル

本章では、2016年スポーツデータ解析コンペティションで提供されたデータスタジアム提供のトラッキングデータ及びボールタッチデータから、2章で定義した重力モデルを用いて選手質量を推定する。そして、ボールタッチと推定された選手質量との関連を階層ベイズモデルによってモデリングした結果を報告する。

3.1 データ概要およびデータの整形

分析対象として、2016年スポーツデータ解析コンペティションで提供されたデータスタジアム提供のトラッキングデータ及びボールタッチデータを用いる。それぞれ、2016年2月27日に行われた磐田対名古屋、広島対川崎、鳥栖対福岡、柏対浦和、湘南対新潟、神戸対甲府、横浜FM対仙台、FC東京対大宮、および2016年2月28日に行われたG大阪対鹿島の計9試合のデータであり、トラッキングデータでは、1/25秒ごとに各選手の X, Y 座標を計測している。ボールタッチデータでは、ボールタッチが起こった際の、ボールタッチを行った選手の X, Y 座標および、アクションが1/30秒のフレーム単位によって記述されている。ボールタッチデータは、あくまで、ボールタッチが起きたときのみ記録される。

ボールタッチデータとトラッキングデータは観測されているフレームレートが異なるため、ボールタッチデータを一度秒に直し、再度1/25秒のフレームに直した。端数は四捨五入し、トラッキングデータとフレーム数が同じ場合のみ結合を行った。また、 X, Y 座標はトラッキングデータに合わせ、ボールタッチが行われていない場合のボールの位置 X_t, Y_t を以下のように線形補完した。つまり、

$$X_{t+1} = X_t + a_x, Y_{t+1} = Y_t + a_y \quad (t = 0, 1, \dots, T-1)$$

とした。ここで、 a_x, a_y は変化量であり、

表 1. 階層ベイズモデルに用いた説明変数.

変数名	説明
選手位置	選手位置に関するダミー変数 (図 2)
ボール位置	ボール位置に関するダミー変数 (図 2)
累積パス回数	その時点までに行ったパスの総回数
累積シュート回数	その時点までに行ったシュートの総回数
ボールとの距離	ボールとのユークリッド距離. 単位は図 2 での座標の単位
時速	選手の移動速度

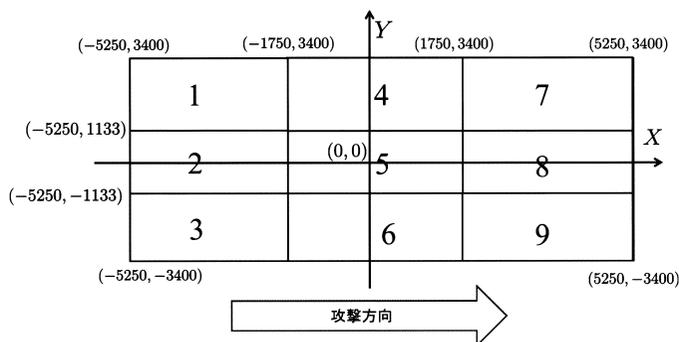


図 2. 計測したフィールドのスケール及びボール分割の位置について.

$$a_x = (X_0 + X_T)/T, \quad a_y = (Y_0 + Y_T)/T$$

である. X_0, Y_0 は観測されているボールの X, Y 座標であり, X_T, Y_T は X_0, Y_0 の次に観測されているボールの X, Y 座標である. T は X_0, Y_0 と X_T, Y_T が観測される間の総フレーム数である.

どのような行動が選手質量と関わるかを調べるため, 説明変数として, 行動に関わる変数を選択した. 用いた変数は表 1 のとおりである. 選手位置, ボール位置は図 2 のようにそれぞれフィールドを 9 分割したものに, その位置に選手, またはボールがいた場合 1, そうでない場合は 0 とするダミー変数を用いた.

これらの説明変数は各フレームで観測されている. 例えば, ボールとの距離では, 各時点で, 選手の位置と線形補完されたボールの位置で算出されている. また, ボール位置および選手位置は, 攻撃方向によって自陣か敵陣かが異なるため, 攻撃方向をそろえ, 算出している.

目的変数は, 各時点での目的変数としては, 対数線形モデルの行主効果, つまり選手質量の対数を用いる. しかし, i, j 間に相手チームの選手がいない場合, $\log F_{ij}^*$ の値が $-\infty$ となることから, 今回は, F_{ij}^* の数え上げを行う際に, チームごとにデータを分割した. つまり, 各チームごとに自チームの選手を行, 相手チームの選手を列にした密度行列を各フレームごとに作成し, 選手質量の推定を行った. 対象となる選手はゴールキーパーを含むフィールドにいる選手全員である. したがって, 退場などの特別な場合を除き, 密度行列の大きさは 11×11 である.

今回の分析では, 行主効果について着目した. 目的変数の値の例として, 図 3 に示す. 縦軸は, 推定された行主効果, つまり, 選手質量の対数を表している. 行主効果として解釈すれば, 0 以上では相対的に質量が大きく, 相手選手を引き付けているということを表す. また, 横軸は, フレーム数であり, 25 フレームで 1 秒を表す. 図 3(a) は川崎の FW である大久保選手の推定された行主効果, つまり選手質量の対数であり, 図 3(b) は広島 MF の茶島選手の選手

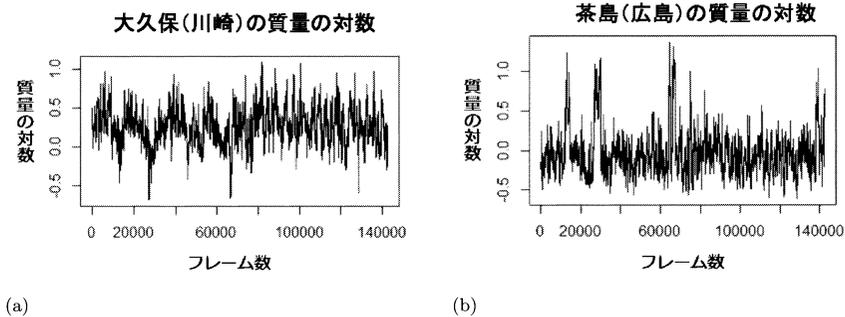


図 3. 推定された選手質量の対数の例.

質量の対数である。大久保は、全体的に 0 以上の値を取っており、前線の選手として、相手の DF にマークされていることがわかる。大久保は前年の得点王であり、また、この試合が、第 1 節であることからマークが厳しいことは、直観的である。また、時折、0 未満の値を取っているが、このとき、川崎のゴール前での攻防が行われていたことから、相手チームの選手が大久保を無視して、攻撃参加していたと考えられる。

一方、茶島は全体的に 0 未満の値を取っている。MF というポジションであることから、相手選手に張り付くことより、どの相手にボールがわたってもすぐに行動できるようなポジションを取っているため、行主効果が低く推定されたと考えられる。また、一時的に大きな値を取っているが、このときは相手ゴール前での攻防であり、ゴール前へのパスに反応するために茶島がゴールまで待機している場面であった。

3.2 階層ベイズモデル

前節で見た通り、選手質量の値が大きくなる場面と小さくなる場面がポジションによって異なると考えられる。FW のような攻撃的なポジションでは攻撃の際に質量は大きくなりやすく、一方で、DF や GK のような守備的なポジションでは守備の場面において質量が大きくなりやすいと考えられる。同様に、チームによっても異なると考えられる。これは、各チームにおいて、戦術が異なることや、チームが選手にどのような役割を与えているかによって、質量を大きくすることが目的となる選手も存在すれば、質量を小さくすることが目的の選手も存在するからである。さらに、前半後半によっても行動の質が異なると考えられる。特に後半は交代が行われやすく、選手は前半と異なる行動をとると考えられる。以上より、チーム、選手、前後半によって回帰係数の値が異なると考え、階層モデルを適用する。また、途中出場の選手は相対的にサンプル数が少ないため、ベイズモデルを用いる。

適用したモデルは Rossi et al. (2005) のモデルであり、以下で表現される。

$$\begin{aligned}
 (3.1) \quad & \begin{cases} y_{it} = \mathbf{x}'_{it}\beta_i + \varepsilon_{it} \\ \beta_i = \Theta \mathbf{z}_i + \mathbf{v}_i \end{cases} \\
 & \varepsilon_{it} \sim N(0, \tau_i) \\
 & \tau_i \sim n_e s q_i / \chi^2(n_e) \\
 & \mathbf{v}_i \sim N(0, \mathbf{V}_\beta) \\
 & \text{vec}(\Theta) | \mathbf{V}_\beta \sim N(\text{vec}(\hat{\Theta}), \mathbf{V}_\beta \otimes \mathbf{A}^{-1})
 \end{aligned}$$

表 2. 階層の説明変数.

説明変数名	説明
チームフラグ	各チーム ID を表すダミー変数
前後半フラグ	前半後半を表すダミー変数

$$V_{\beta} \sim IW(n_{\beta}, V)$$

ここで, $n_e, n_{\beta}, sq_i, V, A, \Theta$ はそれぞれハイパーパラメータである. ハイパーパラメータの設定には, Rossi et al. (2005) に則し, $n_e = 3, n_{\beta} = k + 3, sq_i = s_i^2, V = n_{\beta}I, A = 0.01I, \text{vec}(\Theta) = 0$ とした. ここで, $y_{it} \in \mathbb{R}$ ($i = 1, 2, \dots, n, t = 1, 2, \dots, T_i$) は目的変数であり, $x_{it} \in \mathbb{R}^p$ は説明変数, $z_i \in \mathbb{R}^k$ は個人の属性を表す階層の説明変数であり, $\beta_i \in \mathbb{R}^p$ は対象 i に対する回帰係数ベクトル, $\Theta \in \mathbb{R}^{p \times k}$ の階層に対する回帰係数行列である. n は対象数であり, T_i は対象 i の総フレーム数である. s_i^2 は対象 i の目的変数の標本分散であり, I は単位行列である. また, $\text{vec}(\cdot)$ は vec オペレータ, \otimes はクロネッカー積を表す.

説明変数の一覧は表 1 である. 目的変数は各時点での対象 i の推定された行主効果を用いた. 階層に用いた変数は表 2 の通りである. したがって, 対象 i はそれぞれの選手と前半後半の組み合わせとなる. ただし, 累積パス回数および累積シュート回数はその試合全体で算出した. また, 各対象でキックオフの時点から 5 秒まではデータから除外した. また, 総フレーム数が 1250 フレーム以下の対象も除外したデータ行列を作成した. このとき, 対象数は 444 となった.

事後分布の推定には, Gibbs sampler を用いた. 総サンプル数は 200000 回とし, 焼き入れ数を 100000 回とした. また, 初期値はランダムに発生したものを 3 回行った. 収束判定には目視による確認と, Gelman and Rubin (1992) の方法を用いて, \hat{R} 値が 1.1 未満であるかを基準として確認した.

β_i の推定値には事後平均値を用いた. 各 i の事後平均値のヒストグラムを図 4 に示す. 図 4 の縦軸は度数であり, 横軸は事後平均値である. 選手位置_1 および, ボール位置の 1 がないのは, ダミー変数の基準として用いたためである.

図 4 から, 選手位置_8 の係数が大きいものが多いことがわかる. つまり, 相手ゴール前の付近にいる選手の質量は大きくなりやすいことがわかる. また, 自陣ゴール前の付近である選手位置_2 も同様に大きくなりやすい傾向がみられる. これは, DF や GK などのゴールに関わる位置にいる選手は相手選手を引き付ける, もしくは引っ張られやすいということが読み取れる. また, 累計パス回数の回帰係数のヒストグラムを見れば, 回数を増やせば質量が大きくなる選手とそうでない選手, つまり, 回帰係数が正である選手と負である選手が半分に分かれている. このような結果となったのは, ポジションによる影響ではないかと推察される. ボールとの距離の回帰係数の値が小さいが, これは, 単位が他の変数と異なり, 説明変数に用いた値が大きいためである. また, 回帰係数の値が負である選手が多いことから, ボールとの距離の値が大きければ大きいほど, 選手質量が小さくなる傾向があることが読み取れる. これは, ボールを持っている選手をフリーにすることは危険であるという認識が推定結果として現れたのではないかと考えられる. 時速の推定値のヒストグラムを見れば, 素早い動きをする選手の質量が大きくなりやすい傾向が見て取れる. このことから, 速い動きにつられ, 相手選手が移動していると考えられるが, そもそも, 素早い動きをしなければならぬ場面が, 質量が大きくなりやすい場面であるとも考えられる.

以上から, 相手選手を引き付ける方法として, 相手ゴール正面のエリアで素早い動きを行えば, 一部 DF はつられ, 囷として機能するのではないかと解釈できる. フィールドが一定である

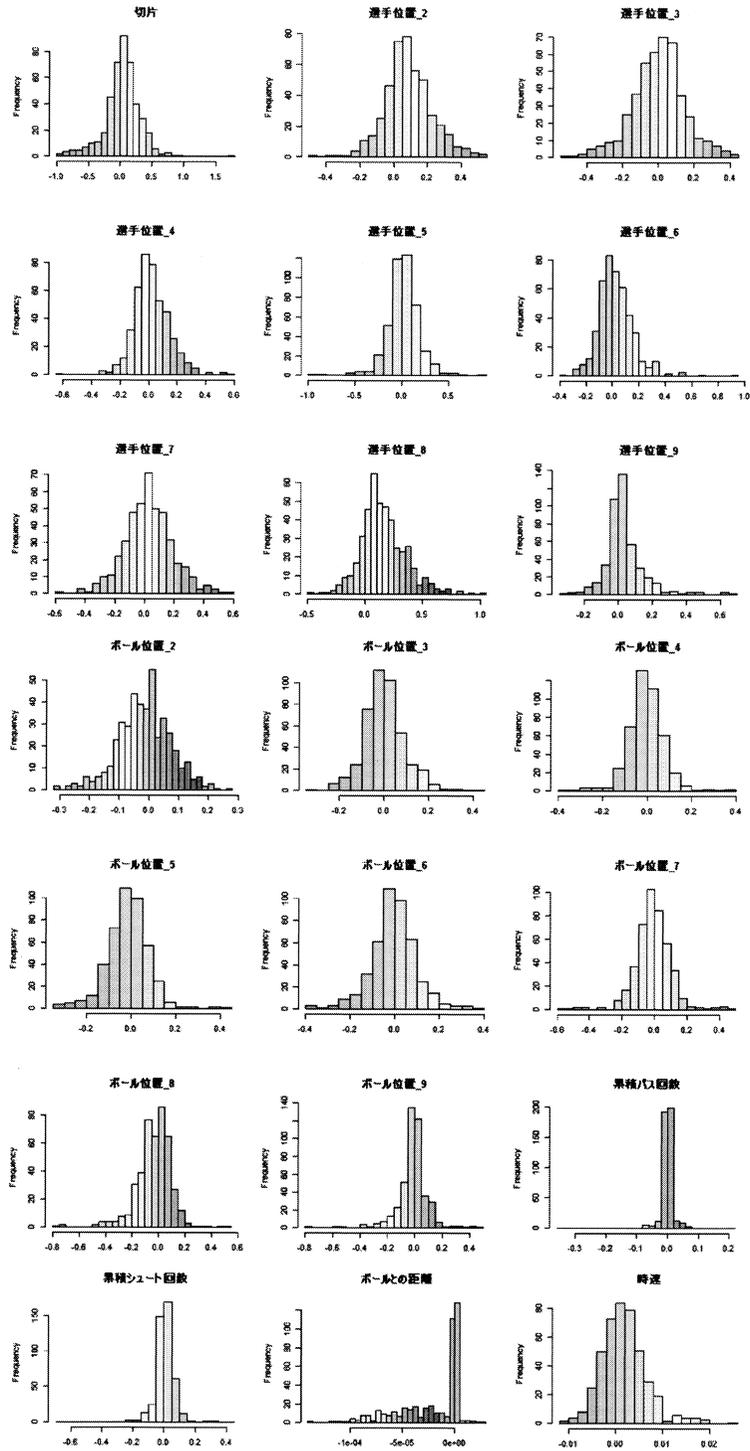


図 4. 回帰係数の事後平均値のヒストグラム.

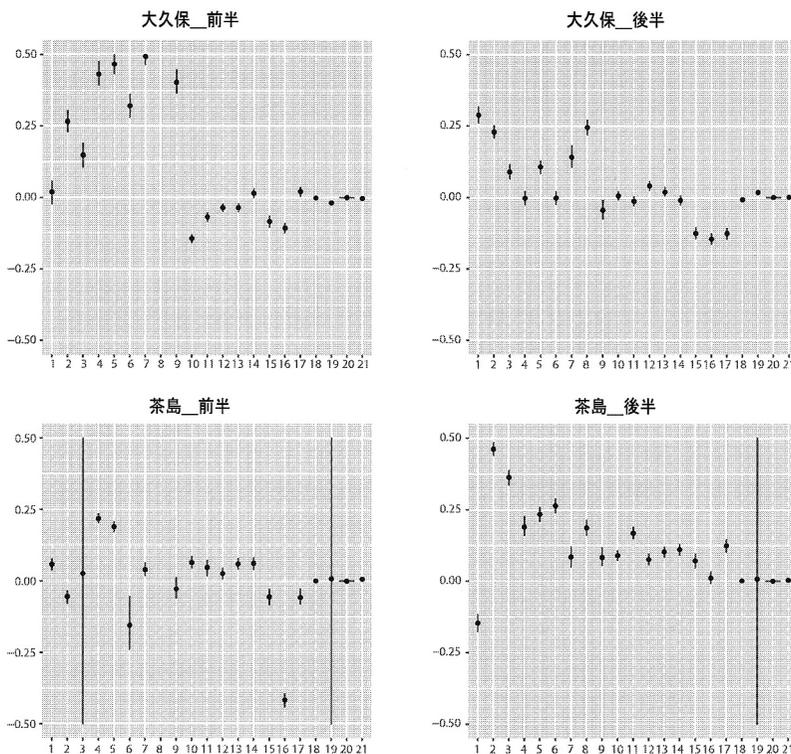


図 5. 大久保および茶島の回帰係数の事後分布。

ことから、一部の選手の付近に他の選手が集まれば、その他の選手がフリーになりやすい。そのため、ゴール付近に、一人は素早い動きで相手を引き付け、もう一人は相対的にゆっくりと動けば、相対的にフリーを作れるのではないかと考えられる。

次に、個人の回帰係数の事後分布について考察を行う。図 5 のバイオリンプロットは大久保および、茶島の回帰係数の事後分布である。縦軸は回帰係数の値であり、横軸の番号と説明変数は表 3 のように対応している。バイオリンプロットの黒点は事後平均値の値である。表 4 はそれぞれの事後平均を表している。

大久保、茶島それぞれ、前半と後半で異なる傾向を表している。大久保の前半では、どの選手位置にいても質量が上昇する傾向にあるが、後半になれば、一部の位置で 0 に近い値、または負の値を取るようになっていく。しかしながら、切片の値が大きく、依然として、相手を引き付ける力が高いことがわかる。累積シュート回数の係数に着目すれば、前半では、シュート回数が増えても、質量が増加する傾向はみられないが、後半はシュート回数が増えれば、質量が上昇する傾向があることがわかる。

茶島は、前半は選手位置によって質量が上昇する傾向や、低い傾向が散見されるが、後半は、どの位置にいても、どの位置にボールがあったとしても、ボール位置_1 に比べ、選手質量が上昇する傾向がある。しかしながら、後半の切片が負の値を取るから、平均的な質量は大きくなりにくいことがわかる。

茶島の累計シュート回数および、選手位置_2 の値はほとんどが 0 であるため茶島の累計シュート回数および、選手位置_2 の回帰係数の事後分布の分散が大きい。

表 3. 変数番号と説明変数の対応表.

番号	説明
1	切片
2~9	選手位置 2~9
10~17	ボール位置 2~9
18	累積パス回数
19	累積シュート回数
20	ボールとの距離
21	時速

表 4. 大久保および茶島の回帰係数の事後平均.

	大久保 前半	大久保 後半	茶島 前半	茶島 後半
切片	0.019	0.288	0.058	-0.146
選手位置_2	0.266	0.229	-0.055	0.462
選手位置_3	0.148	0.090	0.036	0.364
選手位置_4	0.431	-0.002	0.219	0.190
選手位置_5	0.466	0.107	0.189	0.235
選手位置_6	0.321	-0.002	-0.155	0.264
選手位置_7	0.503	0.141	0.039	0.086
選手位置_8	0.620	0.245	0.527	0.187
選手位置_9	0.403	-0.044	-0.028	0.083
ボール位置_2	-0.144	0.005	0.065	0.090
ボール位置_3	-0.069	-0.013	0.048	0.168
ボール位置_4	-0.036	0.041	0.026	0.076
ボール位置_5	-0.035	0.019	0.059	0.103
ボール位置_6	0.015	-0.010	0.061	0.111
ボール位置_7	-0.084	-0.126	-0.056	0.071
ボール位置_8	-0.107	-0.145	-0.417	0.011
ボール位置_9	0.021	-0.126	-0.058	0.124
累積パス回数	-0.002	-0.007	0.001	0.002
累積シュート回数	-0.019	0.018	0.014	0.017
ボールとの距離	0.000	0.000	0.000	0.000
時速	-0.003	0.001	0.006	0.003

次に、階層の説明変数 Z の回帰係数行列である Θ の事後分布について確認する。図 6 は各説明変数に対する回帰係数 Θ の事後分布を表している。縦軸は回帰係数の値であり、横軸の番号と説明変数は表 3 のように対応している。バイオリンプロットの黒点は事後平均値を表している。また、図 6 に、鹿島の係数がないのは、鹿島を基準にしているためである。表 5 は Θ の事後平均を表している。図 6 と同様の理由で鹿島は記載されていない。図 6 の前後半フラグを確認すれば、切片 0 より大きい点に事後平均があり、0 の部分が分布の裾であることから、前半の方が平均的な密度は大きいことが示唆される。これは、前半は後半に比べ選手が比較的元気であり、厳しいマークを付けることが可能であるためと考える。一方で、チームごとでは、切片を除く変数に大きな違いがみられない。そのため、ボールの位置や選手の位置、時速やパスなどの行動によって、選手質量の差がチームごとでは変わらないと考えられる。この結果は、位置や行動による質量の変化はチーム要因でなく、選手個人の影響が大きいためであると考えられる。他の変数に比べ、切片についてはチームごとで異なる傾向がある。これは、

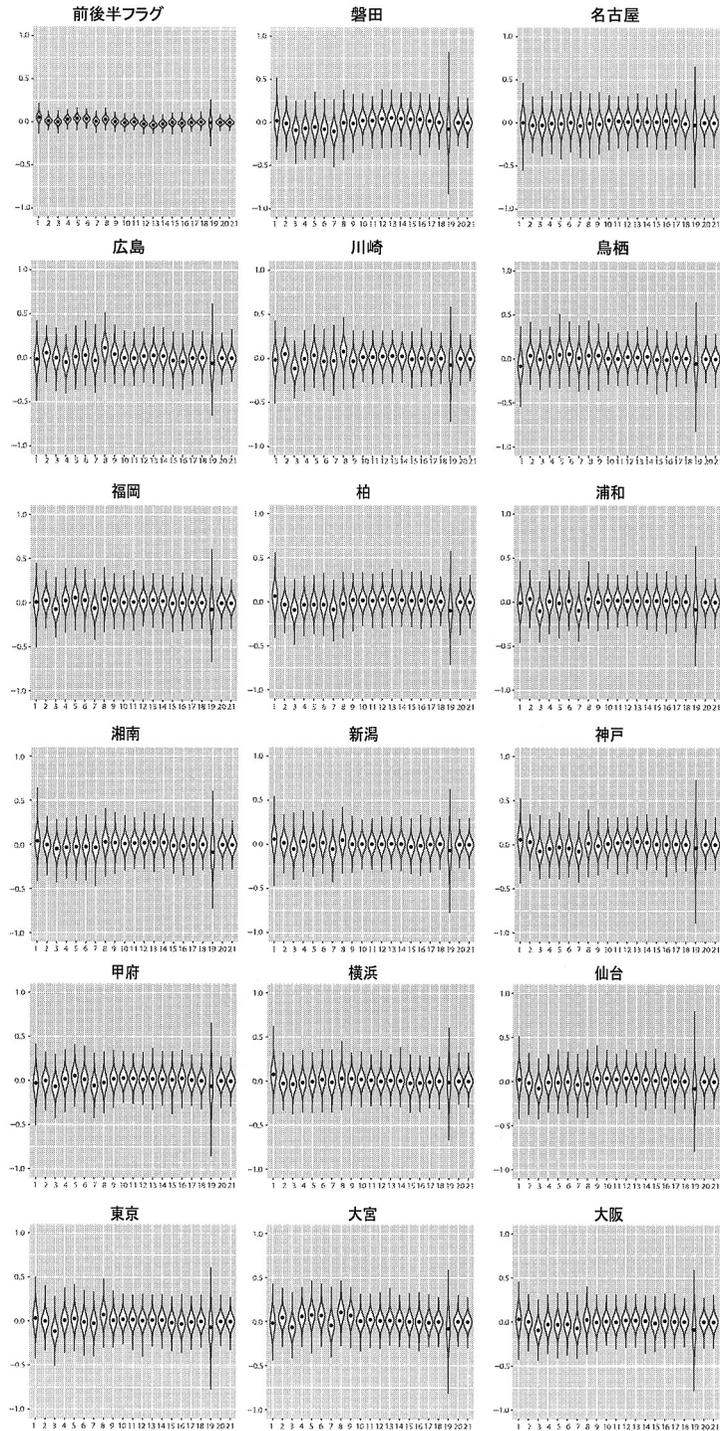


図 6. 階層の説明変数の回帰係数の事後分布.

表 5. 階層の回帰係数の事後平均値.

	前後半フラグ	磐田	名古屋	広島	川崎	鳥栖	福岡	柏	浦和	湘南	新潟	神戸	甲府	横浜	仙台	東京	大宮	大阪
切片	0.053	0.019	0.003	-0.015	-0.017	-0.083	0.007	0.065	-0.011	0.040	0.060	0.057	-0.029	0.076	0.022	0.033	-0.018	0.031
選手位置_2	0.012	-0.011	-0.029	0.059	0.051	0.037	0.027	-0.031	0.037	-0.005	0.018	0.030	-0.001	-0.026	-0.017	0.000	0.045	0.003
選手位置_3	0.001	-0.090	-0.028	0.001	-0.114	-0.006	-0.074	-0.093	-0.102	-0.049	-0.052	-0.075	-0.067	-0.034	-0.079	-0.116	-0.066	-0.093
選手位置_4	0.029	-0.068	-0.007	-0.048	-0.003	0.021	0.025	-0.033	0.007	-0.034	0.033	-0.046	0.018	-0.017	-0.009	0.009	0.058	-0.028
選手位置_5	0.041	-0.052	-0.011	0.015	0.037	0.048	0.058	-0.031	-0.012	-0.028	-0.011	-0.030	0.055	-0.005	-0.011	0.029	0.075	-0.031
選手位置_6	0.036	-0.078	0.004	0.033	-0.032	0.054	0.029	-0.032	0.013	-0.025	0.020	-0.041	0.014	-0.004	-0.004	-0.010	0.069	-0.026
選手位置_7	0.008	-0.103	-0.033	-0.029	-0.026	0.009	-0.061	-0.087	-0.095	-0.034	-0.053	-0.075	-0.054	-0.013	-0.036	-0.024	-0.041	-0.071
選手位置_8	0.031	-0.002	-0.006	0.116	0.081	0.039	0.044	-0.022	0.035	0.031	0.049	0.015	-0.021	0.030	-0.025	0.075	0.106	0.024
選手位置_9	0.004	-0.009	-0.017	0.044	-0.030	0.039	0.023	0.022	-0.001	0.022	0.003	-0.015	0.021	0.028	0.036	0.010	0.071	-0.003
ボール位置_2	-0.008	0.022	0.030	0.001	0.016	0.005	0.004	0.016	0.021	0.010	0.004	0.011	0.030	0.020	0.037	0.022	0.007	0.007
ボール位置_3	0.002	0.023	0.020	-0.001	0.017	0.001	0.010	0.015	0.019	0.016	0.004	0.020	0.025	0.011	0.025	0.017	0.023	0.000
ボール位置_4	-0.021	0.043	0.013	0.023	0.026	0.023	0.024	0.026	0.016	0.020	0.003	0.027	0.016	-0.002	0.039	0.007	0.009	0.020
ボール位置_5	-0.033	0.056	0.025	0.030	0.030	0.019	0.031	0.028	0.015	0.025	0.008	0.037	0.021	0.008	0.037	0.014	0.014	0.019
ボール位置_6	-0.020	0.047	0.012	0.023	0.026	0.027	0.022	0.021	0.012	0.024	0.006	0.026	0.016	0.003	0.023	0.014	0.011	0.012
ボール位置_7	-0.002	0.038	0.010	-0.026	-0.007	-0.008	-0.004	0.016	0.017	-0.011	-0.025	0.004	0.013	-0.023	0.011	-0.015	-0.001	-0.013
ボール位置_8	-0.011	0.039	0.024	-0.040	0.004	-0.009	0.001	0.017	0.018	-0.014	-0.013	0.001	0.029	-0.019	0.027	-0.029	-0.001	0.011
ボール位置_9	-0.003	0.020	0.026	0.003	-0.005	0.013	0.007	0.006	0.004	0.002	0.001	0.006	0.010	-0.008	0.004	-0.007	-0.011	0.004
累積パス回数	0.002	0.005	-0.010	0.007	0.003	0.005	0.005	0.006	0.006	0.002	0.006	0.001	0.001	0.000	0.002	0.000	-0.001	0.003
累積シュート回数	-0.005	-0.073	-0.026	-0.055	-0.072	-0.053	-0.072	-0.097	-0.083	-0.085	-0.065	-0.039	-0.059	-0.013	-0.081	-0.066	-0.078	-0.085
ボールとの距離	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.001	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.001
時速	-0.002	0.000	-0.001	0.000	-0.001	-0.003	-0.002	-0.002	-0.001	-0.001	-0.001	-0.001	-0.003	-0.001	-0.003	-0.001	-0.003	-0.001

チームの戦術により、積極的な攻撃をしているかどうかや、前線にいる選手の数に違いがあり、平均的なチームの質量が変改してゐるのではないかと考えられる。

4. おわりに

本稿では、重力モデルを用いて、選手質量を定義し、選手の動きを定量的に評価する方法を提案した。また、定義した選手質量について、トラッキングデータを用いて、どのような行動と関連があるかを階層ベイズモデルによってモデリングを行った。

モデリングの結果から、選手の位置、選手の速度及び、ボールとの距離をうまく利用すれば、フリーの選手を作ることが可能となること等が示唆された。しかしながら、モデリングの際には行側のみに着目し、列側を考慮していない。このことは、選手付近のフリーについてのみモデリングしており、ゴール近く of 攻防を見るのには適しているが、中盤の攻防において、特にパスを回す場面においては適切でない場合がある。

一方で、列側の相手の重心にどうかは、フィールドの中盤の攻防における役割をモデリングする際に必要であると考えられる。また、重力モデルを利用するための密度として今回は単純に相手との距離を利用した円を利用したが、実際には、選手の向きや相手選手の向き、ゴールへの方向など、選手の動きやすい方向によって重みが異なると考えられる。移動方向や向きがわかるのであれば、その向きや速度を用いて重みづけを行い、円形ではなく楕円形によって密度を測ることが考えられる。また、本分析ではゴールキーパーもフィールドプレイヤーとして扱ったが、その性質上、分析対象から除外することも考えられる。さらには、一部のシーンを切り取り、ボールに関わるであろう一部プレイヤーのみを対象とする方法も考えられる。

今回用いたデータでは、1フレームが1/25秒とデータ観測期間が短いため、各フレーム単位で密度行列を作成した今回の分析では、選手質量の自己相関が非常に高い。そのため、推定された選手質量の対数を直接用いるのではなく、一階差分を取った選手質量の対数を用いる方法なども考えられる。また、フレーム単位で密度行列を作るのではなく、シーンごとの場面に区切ることや、秒単位や分単位の代表値を用いる方法なども考えられる。加えて、本分析では、ボールの位置や、ボールとの距離は選手の動きに関連しないことを暗に仮定していたが、実際には、ボールの位置はプレイした選手や近くにいる選手に依存すると考えられる。そのため、ボール位置および、ボール距離をフィールド上の選手位置によって説明するモデルを階層ベイズに組み込むことも考えられる。つまり、ボール位置のダミー変数行列を \mathbf{W} とし、選手位置行列を \mathbf{Z} とし、選手 i の質量の対数を y_i 、選手 i に対する説明変数行列を \mathbf{X}_i とすれば、

$$\begin{aligned} y_i &= \mathbf{W}\boldsymbol{\theta}_i + \mathbf{X}_i\boldsymbol{\beta}_i + \varepsilon_i (i = 1, \dots, n) \\ \mathbf{w}_t &\sim \text{Multi}(1, \boldsymbol{\pi}_t) (t = 1, \dots, T); \\ \pi_{kt} &= \frac{\exp\{z'_i\gamma_k\}}{\sum_{\ell=1}^K \exp\{z'_i\gamma_\ell\}} (k = 1, \dots, K) \end{aligned}$$

と、モデルを構築することが考えられる。ここで、 γ_ℓ, β_i はそれぞれボール位置を目的変数としたときの説明変数に対する回帰係数ベクトル、選手質量の対数を目的変数としたときの回帰係数ベクトルであり、 $\boldsymbol{\theta}_i$ は選手質量の対数を目的変数としたときのボール位置に対する回帰係数であり、 ε はそれぞれ誤差の分布である。また、 $\text{Multi}(1, \boldsymbol{\pi})$ は試行回数1、確率ベクトル $\boldsymbol{\pi}$ を持つ多項分布を表す。このようにモデル化することにより、選手位置によってボール位置が決定され、そのボール位置が選手質量に影響を与えているというモデルを構築することができる。これらのモデルと本稿で用いたモデルとの比較を当てはまりや予測の意味で比較する必要

がある。これらは今後の課題である。

また、本稿では、敵と味方との非対称性について、考察が十分にできていない。この非対称性についてデータからの意味付けを行うことが今後の課題である。

謝 辞

本研究で用いたデータはデータスタジアム株式会社より提供していただいたものである。ここに感謝申し上げます。また、査読者の方々より貴重なコメントとアドバイスをいただきました。心より感謝申し上げます。

参 考 文 献

- Dobson, S. and Goddard, J. (2008). Forecasting scores and result and testing the efficiency of the fixed-odds betting market in Scottish League Football, *Statistical Thinking in Sports* (eds. J. Albert and R. H. Koning), 91–119, CRC Press, New York, U.S.A.
- Ekin, A., Tekalp, A. M. and Mehrotra, R. (2003). Automatic soccer video analysis and summarization, *IEEE Transactions on Image Processing*, **12**, 796–807.
- 船山貴光, 山田実俊, 山本義郎, 内山秀一 (2016). 選手の動きによるチーム戦略と試合の流れの可視化, 「スポーツデータ解析における理論と事例に関する研究集会」, 統計数理研究所共同研究レポート, No. 363, 91–94.
- Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences, *Statistical Science*, **7**, 457–511.
- 神谷哲太, 中西航, 泉裕一郎 (2016). トラッキングデータを用いたサッカーの試合における戦況変化の抽出, 「スポーツデータ解析における理論と事例に関する研究集会」, 統計数理研究所共同研究レポート, No. 363, 77–82.
- 加藤久 (1999). サッカーの戦術とコンピュータ分析の現状と課題, *オペレーションズ・リサーチ: 経営の科学*, **44**, 125–131.
- Moura, F. A., Santana, E. J., Vieira, N. A., Santiago, P. R. P. and Cunha, A. S. (2015). Analysis of soccer players' positional variability during the 2012 UEFA European Championship: A case study, *Journal of Human Kinetics*, **47**, 225–236.
- Nadarajah, S. and Sundaraj, K. (2013). A survey on team strategies in robot soccer: Team strategies and role description, *Artificial Intelligence Review*, **40**, 271–304.
- 成塚拓真, 卯田純平, 山崎義弘 (2016). サッカーの対戦的特徴に現れる普遍的な統計性の探求, 「スポーツデータ解析における理論と事例に関する研究集会」, 統計数理研究所共同研究レポート, No. 363, 83–90.
- Ren, R. and Jose, J. M. (2005). Football video segmentation based on video production strategy, *Lecture Notes in Computer Science*, **3408**, 433–446.
- Rossi, P. E., Allenby, M. G. and McCulloch, R. (2005). *Bayesian Statistics and Marketing*, Wiley, Chichester, England.
- Rue, H. and Salvesen, O. (2000). Prediction and retrospective analysis of soccer matches in league, *Journal of the Royal Statistical Society: Series D*, **49**, 399–418.
- 瀧剛志, 長谷川純一 (1998). チームスポーツにおける集団行動解析のための特徴量とその応用, *電子情報通信学会論文誌*, **81**, 1802–1811.
- Tinbergen, J. (1962). *Shaping the World Economy: Suggestions for an International Economic Policy*, Twentieth Century Fund, New York, U.S.A.

Quantitative Evaluation of Soccer Players' Movements

Jun Tsuchida¹ and Hiroshi Yadohisa²

¹Graduate School of Culture and Information Science, Doshisha University

²Department of Culture and Information Science, Doshisha University

Many studies have examined the movement of soccer players with the ball. Some studies have even investigated the movement of soccer players without the ball, but they tend to focus on evaluating the overall movements when executing gameplay strategies. Only a few studies have evaluated the soccer players themselves.

In this paper, we use player mass as an indicator to evaluate soccer players based on overall movements by a gravity model. In the gravity model, player mass is a parameter reflecting the movements of all players. Because the estimated player mass is equivalent to the parameters of a log-linear model, it corresponds to the main effect of the log-linear model. We calculated the density and distance among players to estimate the player mass for actual tracking data. Applying the estimated player mass and data from recorded gameplays to a Bayesian hierarchical model reveals the relationships between player mass and player movement.

トラッキングデータを用いたサッカーの 試合における戦況変化の抽出

神谷 啓太¹・中西 航²・泉 裕一朗¹

(受付 2016 年 12 月 31 日；改訂 2017 年 3 月 1 日；採択 4 月 14 日)

要 旨

サッカーの試合において「戦況」とは両チームの攻守のやり取りの中で徐々に変化していくものである。戦況変化を自動的に抽出できれば試合を有利に進めるための戦略が立てられるだけでなく、観戦者に対する情報提供などの面で有用である。本研究では、戦況変化を選手やボール位置の時系列的な振る舞いの変化と考え、統計的变化検知手法である ChangeFinder を使用することでサッカーの試合における戦況変化の抽出を試みる。ChangeFinder とは 2 段階の VAR モデルのオンライン学習を行うことで非定常かつ多ノイズな時系列データへの変化検知が可能な手法である。トラッキングデータ等からボール位置、前線位置、コンパクトネス、守備脆弱度、攻撃率の 5 種類の指標の入力変数を作成した。実験の結果、検出された変化点に対応するように、その直前で VAR モデルパラメータの時系列的振る舞いに大きな変動が存在することが確認された。そのパラメータ変動から想定される戦況変化の内容は実際のプレー内容と概ね合致していたとともに、想定される戦況変化が検出できたことが確認された。

キーワード：サッカー、戦況、変化点検知、時系列分析、ChangeFinder。

1. はじめに

数多くのスポーツで、センサやデータ計測員を駆使して多量のデータを取得し、統計的手法によってそれらを分析し、得られた示唆や知見をチーム戦略や選手評価に活用することが一般的となっている。サッカーにおいても、ボール支配率やパス・シュート本数などの従来のスタッツを始め、ここ数年では選手の走った経路を逐次記録したトラッキングデータなどが計測されている。サッカーでは両チーム合わせ 22 人の選手とボールが互いに影響し合いながら 45 分ハーフを連続的にプレーしており、取得されるデータにはそのような複雑な動きがすべて記録されている。サッカーの試合の大局を俯瞰的に把握するためには、この複雑なデータから選手や観戦者にとって解釈可能な形で示唆や知見を発掘する研究や分析が望まれる。

関連する研究の一つとして、Kijima et al. (2014) はサッカーの試合展開を支配する単純な法則性に言及している。一見複雑に見える選手とボールの位置変化はフラクタル構造を持つというもので、サッカーの試合はある普遍的な法則に従うことを示唆している。しかし一方で、観戦者や選手が感じる「戦況」は試合中に大きく変化することがある。例えば、攻勢であったのにいつの間にか守勢に転じている、停滞していた状況が一気に動き出す、などである。このよう

¹ 東京大学大学院 工学系研究科：〒113-8656 東京都文京区本郷 7-3-1

² 東京工業大学 環境・社会理工学院：〒152-8552 東京都目黒区大岡山 2-12-1

な戦況変化は上記の不変性の崩れとも考えられ、戦況変化を適切に把握することができれば、試合を有利に進めることに繋がる。しかし現状、実際に試合を観戦もしくはプレーすること以外に戦況変化を把握する方法が存在しない。戦況変化を自動的に抽出できれば試合を有利に進めるための戦略が立てられるだけでなく、観戦者に対する情報提供などの面で有用であり、その研究意義は大きい。

これまで、取得されたデータは各選手のプレー特徴やパフォーマンスの定量的な評価・分析に活用されるなど、主として個別的な分析に活用されてきた (Football LAB, 2017; Gudmundsson and Wolle, 2014; Sarmiento et al., 2014)。しかしながら、22 人の選手とボールとの相互作用を考慮した大局的な分析はいまだ少ない。数少ない例として、Moura et al. (2012) や高橋 他 (2015) は、選手の相対的な位置関係から算出される指標を用いて、攻撃時と守備時との選手配置の相違や、局面での優勢・劣勢の判定を試みている。

これに対し本研究では、攻守の切り替えのような明らかに観測される事実よりも大局的な視点にたち、上述の「戦況」の変化を統計学的に扱うことを考える。そこで、戦況変化の自動抽出に向け、まずは戦況変化とは何かを検討する必要があるが、上述した例はいずれも選手やボールの時系列的な振る舞いの変化した結果であると解釈できる。そのため、選手やボールに関するデータの時系列的振る舞いをモニタリングし、その振る舞いの変化した点を検出することで戦況の変化を抽出することが可能となると考えられる。以上を踏まえると、統計的变化点検出 (山西, 2009) の枠組みを使用し、データの時系列的な振る舞いの変化点を検出することでサッカーの試合における戦況変化の抽出を行うことができると考えられる。

なお、サッカーを始めとする集団競技の戦況は、選手個々の意思決定に依る動きだけでなく、チーム全体としての動き、もしくは選手相互間やボールとの関係から発生するものである。したがって戦況が代表されるようなある一つの変数を見つけることは難しく、複数の変数および変数間の関係に着目して分析する必要がある。また、実用を考える上で、リアルタイムに戦況変化が抽出されることが望ましい。すると、多次元時系列データを対象に、確率モデルとして時系列モデルを仮定して、リアルタイムに変化点度合いのスコアを計算していく手法である ChangeFinder (Takeuchi and Yamanishi, 2006) が適用可能であると考えられる。

本研究の目的は、トラッキングデータを用いてサッカーの試合における戦況変化の抽出を行うことである。第 2 章で ChangeFinder について説明した後、第 3 章で分析に用いる変数を選定する。第 4 章で実データに ChangeFinder を適用し、検出された戦況変化に関する解釈を行う。最後に、本研究のまとめと今後の課題を第 5 章でまとめる。

2. ChangeFinder を用いた戦況変化の抽出手法

本章では ChangeFinder を用いた戦況変化の抽出手法について説明する。まず、ChangeFinder を適用するにあたり必要となる VAR モデル (Vector Autoregressive model; 多変量自己回帰モデル) について確認した後、ChangeFinder の基本原理および VAR モデルのオンライン学習方法について説明する。

2.1 VAR モデル

まず、初期値の平均が μ である d 次元時系列変数 $\{x_t : t = 1, 2, \dots\}$ を K 次の VAR 過程によってモデル化すると、以下のように定式化される。

$$(2.1) \quad x_t = \sum_{i=1}^K w_i(x_{t-i} - \mu) + \mu + \varepsilon$$

ここで、 $\omega_i \in \mathbf{R}^{d \times d} (i = 1, \dots, K)$ は d 次元パラメータ行列であり、 ε は平均 0, 分散共分散行列 Σ のガウス分布 $\mathcal{N}(0, \Sigma)$ に従う確率変数である。

すると、上記の VAR モデルによって定式化された x_t に関する確率密度関数は以下のように表すことができる。

$$(2.2) \quad p(x_t|\theta) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2} (x_t - \omega)^T \Sigma^{-1} (x_t - \omega) \right\}$$

なお、VAR モデルのパラメータをまとめて $\theta = \{\omega_1, \dots, \omega_K, \mu, \Sigma\}$ と表記し、 $\omega = \sum_{i=1}^K \omega_i (x_{t-i} - \mu) + \mu$ である。また、 T は転置を表す。

2.2 ChangeFinder

ChangeFinder では観測データに対し 2 段階の VAR モデルのオンライン学習を行う。1 段階目の VAR モデルの学習において観測値に対する外れ値スコアを計算し、その後、平滑化した外れ値スコアを入力とした 2 段階目の VAR モデルを学習することによって、変化点スコアの計算を行う機構となっている。すなわち、外れ値スコアが一定期間連続して高水準である場合に、観測データにおける変化を検出する手法と換言できる。本研究においては、選手やボールの時系列的な振る舞いを VAR モデルで表現し、その振る舞いが変化した点を検出することが目的である。選手やボールの振る舞いが一定期間連続してこれまでの振る舞いと異なっている場合に戦況が変化したと考えられるため、ChangeFinder の枠組みにおいて外れ値ではなく変化点を検知することが戦況変化を抽出することに相当する。

いま、時刻 $t-1$ までの観測値 x_1, \dots, x_{t-1} が得られているとする。すると、後述する SDAR (Sequentially Discounting AR model learning) アルゴリズムと呼ばれるオンライン忘却型学習アルゴリズムを用いることで、時刻 $t-1$ における確率密度関数 $p_{t-1}(x)$ が推定される。ひとたび観測値 x_t が観測されると、時刻 t での外れ値スコアが対数損失によって以下のように算出される。

$$(2.3) \quad \text{Score}(x_t) = -\log(p_{t-1}(x_t))$$

次に、上記の方法によって算出した外れ値スコアに対して、以下の T 次移動平均を計算する。

$$(2.4) \quad y_t = \frac{1}{T} \sum_{i=t-T+1}^t \text{Score}(x_i)$$

新たに得られた時系列データ $\{y_t : t = 1, 2, \dots\}$ を VAR モデルで 2 段階目のモデル化を行い、再び SDAR アルゴリズムを用いて学習を行う。 y_t が得られた際に学習された確率密度関数を q_t とすると、 y_t の対数損失 $-\log(q_{t-1}(y_t))$ も 1 段階目と同様に算出される。

最後に、上記の対数損失に対し T' 次移動平均を計算した結果を時刻 t における変化点スコア $\text{Score}(t)$ とする。

$$(2.5) \quad \text{Score}(t) = \frac{1}{T'} \sum_{i=t-T'+1}^t \{-\log(q_{i-1}(y_i))\}$$

この変化点スコア $\text{Score}(t)$ が大きいほど時刻 t における状態変化の度合いが大きいことを意味する。

2.3 SDAR アルゴリズム

SDAR アルゴリズムでは、観測値 x_t が観測される度に VAR モデルのパラメータである

$\theta = \{\omega_1, \dots, \omega_K, \mu, \Sigma\}$ を学習する. この際, 忘却効果を取り入れることによって過去の観測で得られた情報の影響を徐々に低減していく. これにより, もともと定常過程の取り扱いを旨としている VAR モデルを, 形式的に非定常なモデルの学習へ拡張したものとなっている.

まず, 各種パラメータおよび統計量の初期値 $\hat{\mu}, \hat{\Sigma}, C_i (i = 1, \dots, K)$ を定める. ここで, $\{C_i : i = 1, \dots, K\}$ は自己共分散関数である.

x_t を観測する度に, 以下の更新式を計算する.

$$(2.6) \quad \hat{\mu} \leftarrow (1-r)\hat{\mu} + rx_t$$

$$(2.7) \quad C_j \leftarrow (1-r)C_j + r(x_t - \hat{\mu})(x_{t-j} - \hat{\mu})^T$$

上式における $r (0 < r < 1)$ が忘却パラメータであり, 新たなデータから計算された統計量と過去のデータから計算されている統計量の更新比を制御する. r が大きいほど忘却の度合いが大きいこととなる.

次に, 以下の自己共分散関数とパラメータ行列に関する YuleWalker 方程式を解く (北川, 2005).

$$(2.8) \quad \sum_{i=1}^K \omega_i C_{j-i} = C_j \quad (j = 1, \dots, K)$$

最後に, 上記の解を $\hat{\omega}_1, \dots, \hat{\omega}_K$ とおき, 以下を計算する.

$$(2.9) \quad \hat{x}_t \leftarrow \sum_{i=1}^K \hat{\omega}_i (x_{t-i} - \hat{\mu}) + \hat{\mu}$$

$$(2.10) \quad \hat{\Sigma} \leftarrow (1-r)\hat{\Sigma} + r(x_t - \hat{x}_t)(x_t - \hat{x}_t)^T$$

観測値 x_t が観測される度に, 式(2.6)~(2.10)を繰り返す.

3. 入力変数の検討

3.1 使用したデータ

本研究で使用したデータは, 2015 明治安田生命 J1 リーグ 1st ステージ第 2 節鹿島アントラーズ対湘南ベルマーレ戦および第 17 節松本山雅 FC 対湘南ベルマーレ戦の計 2 試合に関して, 1/30 秒毎にパスやタックルなどボール周辺で発生したイベントおよびその発生時刻と位置を取得したボールタッチデータと, 1/25 秒毎に選手及び審判のピッチ上での位置を取得したトラッキングデータの 2 種類である. なお, これらのデータはデータスタジアム株式会社から提供を受けたものである.

3.2 入力変数の選定

VAR モデルの入力変数として用いる指標について検討を行う. まず, 戦況を表すのに十分な入力変数を用意できるよう, 試合全体の流れを表現すると思われる指標や, 各チームの攻勢・守勢を表す指標を複数作成した. 具体的には, 以下のような検討により (a) ボールの動き, (b) 選手の動き, (c) ボールポゼッションの変化, (d) 選手位置の変化というカテゴリを設け, 変数候補を選定した.

まず, 1 章で述べたとおり, 戦況は直接的にはボールや選手位置の時系列的振る舞いで表現されると考えられる. そこで, ボールの動きを表す変数として (a-1)「ボール位置」を候補とする. また, 選手の動きを表す変数として, (b-1)「各選手位置」(22 人)と, その集計量である (b-2)「各チーム選手位置の重心」, (b-3)「両チーム選手位置の重心」および既往研究 (Kijima et

al., 2014)で提案された(b-4)「前線位置」という指標を候補とする。

次に、観戦者は、両チームが攻撃あるいは守備を行っている度合いやそのプレー内容によって戦況を感じ取っていると考えられる。そこでの戦況の変化とは、例えば1章で述べたような、攻勢だった試合展開が守勢に変化したというものである。そこで、ボールポジションの変化を表す変数として(c)「攻撃率(詳細後述)」を候補とする。

さらに、サッカーの戦術についての書籍(松田 他, 2015; 西部, 2013; 西部・北, 2011)およびプレーヤーを含む複数の専門家へのヒアリングを通して、一般に「守備はコンパクトに、攻撃はワイドに」行うことが良いとされていることに着目した。「コンパクト」とはピッチ上で同チームの選手が近接していることで、特にピッチ前後方向の位置関係を指す。「ワイド」とはピッチ上で同チームの選手が散らばっていることで、特にピッチ左右方向の位置関係を指す。このことから、選手位置の変化を表す変数として、両チームそれぞれの(d-1)「オフサイドラインの座標」(d-2)「最前線の選手の座標」を候補とする。また、各チーム最前線の選手から最後方の選手までの距離を表す(d-3)「コンパクトネス(詳細後述)」、ピッチ左右方向の選手の幅を表す(d-4)「ワイドネス」も候補とする。さらに、コンパクトとワイドという相反する目標に対する攻守のせめぎ合いに注目した(d-5)「守備脆弱度(詳細後述)」を候補とする。

なお、戦況を生起するプレーは、実際には選手個人により行われているという観点からは、各選手のパスやシュートなどの「アクション」やボールを保持していないときの攻撃の一環としてのダッシュ(「スプリント」)も変数候補となり得る。ただし、今回は以下の理由から、これらの変数は候補としていない。まず、同一のアクションであっても、各時点で他に取り得たアクションの選択肢によってその意味合いは異なるため、選択肢集合の情報を有さずにアクションを変数とすると、結果の解釈上困難が生じる。また、スプリントを抽出するためには、目の前の相手に対する守備のような必要に迫られた走行との区別が必要であるが、各時刻の位置データのみからこれを区別することは容易ではないためである。

続いて、以上の変数候補から、実際に利用する変数の選定を行う。まず、VARのパラメータの推定可能性から、なるべく少ない変数でVARを構築することが望ましい。そこで、多大な次元増加をもたらす選手個人を示す変数は採用せず、それを代表する集計的な変数を用いることを前提とする。すなわち、(a)「ボール位置」、(b)選手の動きを代表する変数としての「各チーム選手位置の重心」「両チーム選手位置の重心」「前線位置」、(c)攻撃率、(d)選手配置を代表する変数としての「オフサイドラインの座標」「最前線の選手の座標」「コンパクトネス」「ワイドネス」「守備脆弱度」が変数候補として残る。

さらに、複数の変数を統合することで別の変数と同様の意味を表す場合や、同じカテゴリの変数間に相関がある場合が考えられる。次元削減の観点および共線性のない必要最低限な変数組という観点から、「オフサイドラインの座標」と「最前線の選手の座標」は同様の意味を表す「コンパクトネス」で代表させる。また、「ワイドネス」と「守備脆弱度」はどちらも選手の散らばり度合いを表す指標であり、実際にも相関を持っていることが確認されたため、攻守のせめぎ合いの状況をより明示的に表すであろう「守備脆弱度」で代表させる。同様に相関を持つ「各チーム選手位置の重心」、「両チーム選手位置の重心」および「前線位置」は、外れ値に強いと考えられる「前線位置」で代表させる。

以上の検討により、最終的に、ボール位置、前線位置、コンパクトネス、守備脆弱度、攻撃率の5種類の指標を分析に用いる変数として選定した。

3.3 入力変数の説明

以下、それらの指標について詳細な説明を行う。特に、前線位置、コンパクトネス、守備脆弱度については図1に概念図を示している。なお、提供を受けたデータは1/25秒や1/30秒間

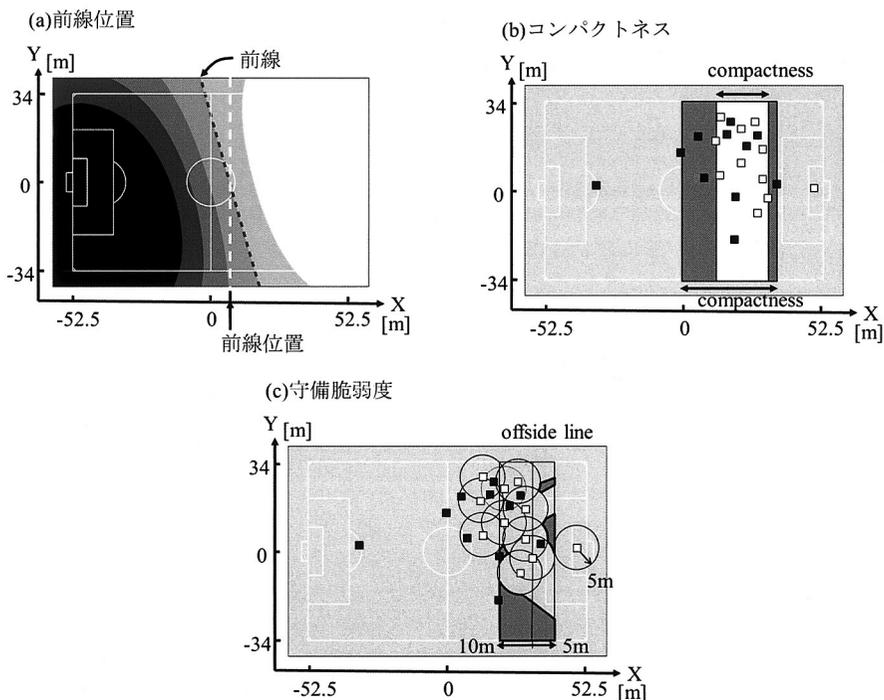


図 1. 入力変数の概念図。(a) 前線位置. 両チームの選手がなす支配領域が均衡する前線の X 座標に関する平均値. 図中の白色破線に対応。(b) コンパクトネス. 一番前方の選手と後方 2 番目の選手の X 座標上の距離. 図中の灰色と白色の四角形の幅にそれぞれ対応。(c) 守備脆弱度. 自軍のオフサイドラインより前方 10m, 後方 5m の長方形のうち, 最寄りの味方選手から 5m 以上離れており, 最近傍選手が相手選手であるような地点の合計面積の割合. 図中の灰色で囲まれている図形の合計面積が占める割合に相当.

隔であったが, VAR モデルへの適用を踏まえ, 分析に使うデータはすべて 1 秒間隔となるよう加工した. また, ピッチ中央を原点とし, コートの長辺方向を X 方向, 短辺方向を Y 方向と定義する.

まず, 試合全体の流れを表現する指標としてボール位置を選定した. 攻守の切り替わりによるボールの上下動のみを考慮するとし, X 座標のみを採用した. 時刻 t におけるボール位置を $ball(t)$ とする.

前線位置とは, 両チームの選手がなす支配領域が均衡する X 座標として定義した. この変数は全選手の動きを代表するため, 各チームの攻勢と守勢に合わせ値が増減すると予測される. 既往研究 (Kijima et al., 2014) に従い, 以下のとおり作成した.

- i) チームで正負の異なるガウスクERNEL (標準偏差 3m) を各選手が時刻 t に位置する座標に設定する.
- ii) 全てのKERNELを足し合わせ, ピッチ上で値が 0 となる線分を前線とする.
- iii) 前線の X 座標に対する平均値を算出し, これを時刻 t における前線位置とする.
- iv) i)~iii) を全時刻について算出し, 時刻 t における前線位置を $frontLine(t)$ とする.

コンパクトネスに関しては、ホーム・アウェイチーム毎に「一番前方に存在する選手の X 座標」と「後方より 2 番目に存在する選手の X 座標」の距離として定義し、時刻 t において $compact(HOME)(t)$ および $compact(AWAY)(t)$ とする。各チームの選手がその時刻において、どれだけピッチ上に広く存在しているかを示す指標である。攻撃時はピッチを広く使うためコンパクトネスは大きくなり、守備時は相手のプレーエリアを狭めるべくコンパクトネスは小さくなると予想される。

チームの守備力の脆弱性を表す指標として、自陣の最終ライン付近において相手選手が侵入している程度を守備脆弱度と定義し、時刻 t において $defense(HOME)(t)$ および $defense(AWAY)(t)$ とする。具体的には図 1(c) に示すように、「自軍のオフサイドラインより前方 10m, 後方 5m の長方形のうち、最寄りの味方選手から 5m 以上離れており、最近傍選手が相手選手であるような地点の合計面積の割合」である。攻撃時にボールを失って守備へ転換した直後には守備脆弱度は上昇し、守備陣営が整うと守備脆弱度は減少すると予想される。一方、カウンターや人数をかけた攻撃時には守備脆弱度は上昇することが予想される。

最後に、ホームチームの直近 10 分間における攻撃頻度を表す指標として攻撃率 $Attack(t)$ を定義する。ボールタッチデータ中に離散的に取得されている各チームの攻撃アクションを内挿し、1 秒毎の連続的な値として算出した。

$$(3.1) \quad Attack(t) = \frac{1}{6000} \int_{t-6000}^t f(t') dt'$$

$$(3.2) \quad f(t) = \begin{cases} 1 & \text{if } t \geq 0 \text{ and } A(t) = HOME \\ 0 & \text{if } t \geq 0 \text{ and } A(t) = AWAY \\ 0.5 & \text{if } t < 0 \end{cases}$$

ただし、 $A(t)$ は $t \geq 0$ において時刻 t 以前の最新の攻撃を行ったチーム (HOME または AWAY) を示す。

4. 戦況変化の抽出実験

4.1 適用条件

ChangeFinder の入力変数として、ボール位置、前線位置、コンパクトネス(ホーム・アウェイチームごと)、守備脆弱度(ホーム・アウェイチームごと)、攻撃率の 5 種類・計 7 つのデータを 7 次元変数として設定した。また、VAR モデルの次数 K は、オフラインでの VAR 推定を行った場合の最小 AIC 値を参考に、 $K = 5$ と決定した。忘却率 r や平滑化パラメータ T, T' についてはそれぞれ $r = 0.01$, $T = 50$, $T' = 5$ と設定した。鹿島・湘南戦および松本・湘南戦の前後半毎に ChangeFinder を適用し、戦況変化の抽出を試みる。

なお、SDAR アルゴリズムで用いる VAR モデルの各種パラメータおよび統計量の初期値 $\hat{\mu}, \hat{\Sigma}, C_i (i = 1, \dots, K)$ は、前後半それぞれの開始 60 秒間で得られたデータから計算することとし、この間の変化点スコアは算出しない。

4.2 適用結果および結果の考察

松本・湘南戦後半を例として、図 2 の上段に入力時系列データ x_t を、下段に ChangeFinder による変化点スコアの出力値 $Score(t)$ を示す。図 2 の上段において、1 行目にボール位置と前線位置、2 行目に湘南ベルマーレと松本山雅 FC のコンパクトネス、3 行目に湘南ベルマーレと松本山雅 FC の守備脆弱度がそれぞれ灰色の太線および黒色の細線で、そして 4 行目に松本山雅 FC の攻撃率が示されている。ボール位置と前線位置に関しては、正方向に値が上昇するに

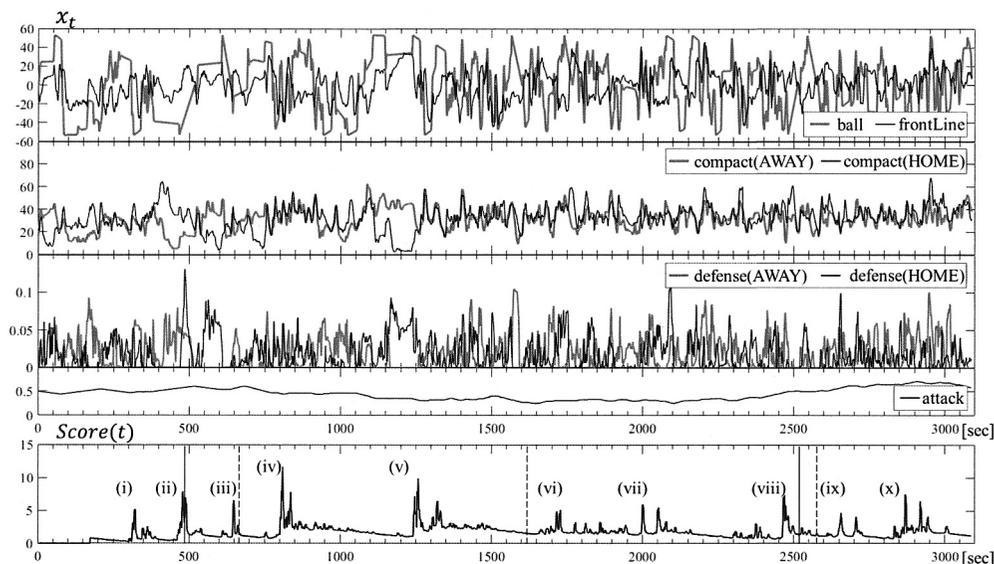


図 2. 松本・湘南戦後半における入力データおよび変化点スコア出力結果. 上段 1 行目にボール位置 (ball) と前線位置 (frontLine), 2 行目に湘南ベルマーレと松本山雅 FC のコンパクトネス (compact), 3 行目に湘南ベルマーレと松本山雅 FC の守備脆弱度 (defense) をそれぞれ灰色の太線および黒色の細線で, 4 行目に松本山雅 FC の攻撃率 (attack) を示す. AWAY および HOME はそれぞれ湘南ベルマーレと松本山雅 FC に対応する. 下段は ChangeFinder による変化点スコアの出力値 $Score(t)$ である. また, 下段グラフ中の実線および破線は, 松本および湘南による得点が生じた時刻である.

つれ, 湘南ベルマーレの攻撃方向に進出していることを意味する. 下段は各時刻について算出された変化点スコア $Score(t)$ であり, 値が大きいほど戦況変化の度合いが大きいことを表している.

出力された変化点スコアを見ると, 図中に (i)~(x) で示す合計 10 箇所です比較的大きな戦況変化が発生したと推定されたことが分かる. 紙面には掲載していないが鹿島・湘南戦前後半および松本・湘南戦前半の結果と比較すると, 松本・湘南戦後半ではより高頻度で変化点が検出されていた. 松本・湘南戦後半では, 53 分・87 分に松本が 2 点, 56 分・72 分・88 分に湘南が 3 点を獲得している. 片方のチームの得点の直後にもう片方のチームが得点するなど, 戦況の移り変わりが比較的大きかった試合であったことが, 本結果において比較的高頻度に変化点が検出された理由として考えられる. なお, 図 2 下段には松本による得点に対応する時刻を実線で, 湘南によるものを破線で示した. ゴールおよびその周辺の時刻にて変化点スコアが上昇していることが分かる.

次に, 検出された変化点が具体的にどういった戦況の変化を表しているのか, 入力データの振る舞いと実際のプレーを比較することで結果の解釈および検証を行う. ただし, 図 2 から分かるように入力データは複雑な動きをしているため, 入力データから変化点検出の原因を直接分析するのは困難である. そこで, パラメータ μ の各時刻における推定値 $\hat{\mu}$ に着目する. このパラメータは VAR モデルの式 (2.1) より, 自己回帰分を除いた平均的な値と解釈できる. この推定値 $\hat{\mu}$ の振る舞いに着目し, 検出された変化点との関係を分析することによって, どういった戦況の変化を表しているのか解釈を進めることが可能になると考える.

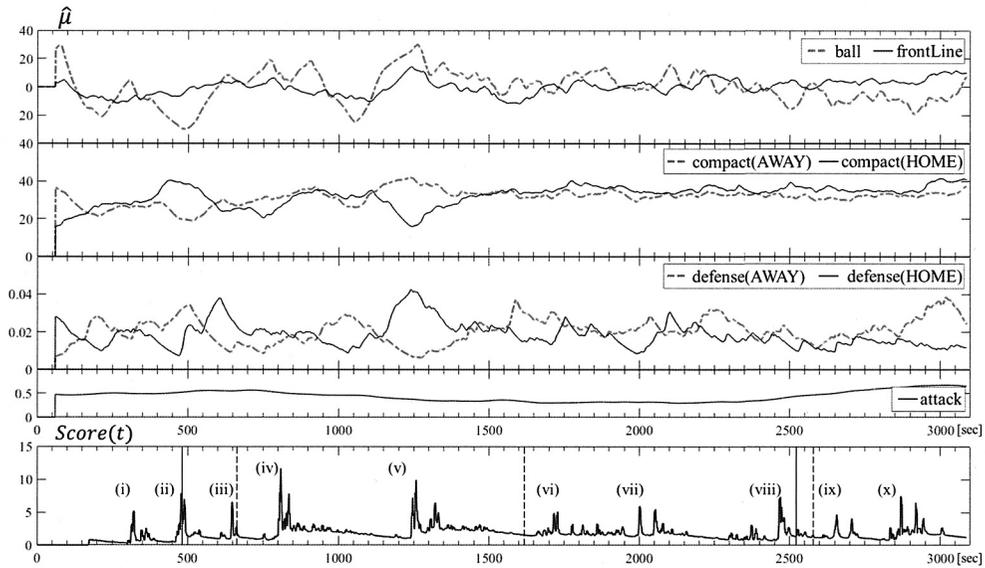


図 3. 松本・湘南戦後半におけるパラメータ μ の推定値 $\hat{\mu}$ の推移および変化点スコア出力結果。図中のグラフと変数の対応は図 2 に同じ。ただし、図 2 における灰色の太線が、本図では灰色の破線へ置き換わっている。また、試合開始 60 秒間のパラメータは推定されていない。

そこで、図 3 に松本・湘南戦後半における推定値 $\hat{\mu}$ の推移と変化点の検出結果を示す。図中のグラフと変数の対応は図 2 と同様である。ただし視認性の向上のため、図 2 における灰色の太線が、図 3 では灰色の破線へと置き換わっている。図 2 で複雑な振る舞いをしてきた入力値が、図 3 の推定値 $\hat{\mu}$ では比較的滑らかな振る舞いをしていることが分かる。また、推定値の時系列的振る舞いに変動があると、それに対応するように直後 1 分後あたりで変化点スコアが上昇していることが分かる。なお、ChangeFinder では第 1 段階および第 2 段階で移動平均処理を行っているため、観測データの入力から変化点の検出までに遅延が生じることに注意が必要である。

表 1 に、各戦況変化点 (i) ~ (x) について代表的な $\hat{\mu}$ の変動の様子およびそれから想定される戦況変化の内容、ならびに実際にボールタッチデータから確認されたプレーの内容をまとめた。例えば戦況変化点 (i) の直前では、ボールが湘南攻撃方向に移動し、松本の守備脆弱度が上昇するという変動がパラメータ推定値 $\hat{\mu}$ の中で確認された。この結果より、湘南が攻勢へと転換したことが戦況変化の内容として想定される。実際にボールタッチデータより、湘南が複数のパスを繋ぎながら攻撃を展開している様子が確認できた。また、戦況変化点 (vii) では松本の守備脆弱度の低下およびその直後の上昇が確認されており、松本が陣営を立て直したうえで攻勢へと展開したことが想定される。実際のプレーを確認すると、1 点ビハインドである松本が飯尾 (MF) から石原 (MF) へ、喜山 (MF) から阿部 (FW) への選手交代を行っており、攻撃的な戦術へと変更した際の影響が検知されたものと考えられる。他の変化点についても、表中にあるように、変化点スコア上昇の原因となるようなパラメータ推定値 $\hat{\mu}$ の変動から想定される戦況変化の内容が、実際のプレー内容と概ね合致していることが確認できた。

以上の結果を踏まえると、本分析方法により、攻勢・守勢への転換やセットプレーによる連

表 1. 検出された変化点の解釈および実際のプレーとの関係.

#	代表的な $\hat{\mu}$ の変動	想定される戦況変化	実際に確認されたプレー
i	ボールが湘南攻撃方向に移動 松本の守備脆弱度が上昇	湘南の攻勢への転換	湘南がパスを繋ぎ攻撃を展開
ii	ボールが松本攻撃方向に移動 湘南のコンパクトネスが低下 松本のコンパクトネスが上昇 湘南の守備脆弱度が上昇 松本の守備脆弱度が低下	松本の攻勢への転換	松本が PK を獲得
iii	湘南のコンパクトネスが上げ止まり 松本のコンパクトネスが下げ止まり 松本の守備脆弱度が回復	松本の守勢の実現	湘南が FK を獲得し松本が守備体勢に
iv	ボールが松本攻撃方向に回復 松本のコンパクトネスが上昇	松本の攻勢への転換	松本が自陣で FK を獲得しピンチを脱する
v	松本のコンパクトネスが低下 松本の守備脆弱度が上昇 湘南の守備脆弱度が低下	湘南の攻勢への転換 松本の守勢の実現	湘南の CK/FK による連続的な攻撃
vi	湘南の守備脆弱度が回復	湘南の守備の実現	湘南得点後の松本キックオフ
vii	松本の守備脆弱度が低下および上昇	松本の攻勢への転換	松本の複数選手の交代による 攻撃的な戦術への変換か
viii	ボール位置が短時間に前後	試合の活性化	松本得点直後に湘南の得点
ix	ボールが松本攻撃方向に移動 湘南の守備脆弱度が上昇	松本の攻勢への転換	松本のシュートを含む試合展開
x	ボールが松本攻撃方向に移動 湘南の守備脆弱度が上昇	松本の攻勢への転換	同点弾を狙った松本の攻撃

統攻撃, 連続得点による試合の活性化や選手交代による戦術変更など, 想定される戦況変化は概ね検出できていると考えられる. ただし, 実際の試合映像などを確認し, 検出された戦況変化が妥当なものであるのか, また, 未検出となった戦況変化等が存在しないかを専門的知見と共に検討を重ねたい. また, 現在は検出された変化点と戦況変化の原因と考えられる実際のプレー内容との関係をパラメータ μ のみで見ているが, 戦況のより高度な解釈に向け, μ 以外のパラメータについても精査を行う必要がある. さらに, 検出された戦況変化に対応するプレーにはセットプレーや選手交代などアウトオブプレーが多く含まれている. 実際, セットプレーを機に戦況が大きく変わることは多々あるが, 変化点(i)や(x)などインプレー中の戦況変化の抽出も重要であろう. インプレー中の戦況変化の抽出のためには, ChangeFinder に適用する時系列データからアウトオブプレー中のデータを排除する, 一つのインプレーデータ毎に ChangeFinder を適用するなど, 分析方法を修正することが解決策として考えられる.

5. おわりに

本研究では, 統計的変化点検出手法である ChangeFinder をトラッキングデータに適用することで, サッカーの試合における戦況変化の抽出を行った. ChangeFinder の入力変数の検討にあたっては, 変数間の相関分析を通して共線性のない必要最低限な変数組となるよう配慮し, ボール位置, 前線位置, コンパクトネス, 守備脆弱度, 攻撃率の 5 種類の指標を選定した. 松本山雅 FC 対湘南ベルマーレ戦後半のトラッキングデータに ChangeFinder を適用した結果では, 検出された合計約 10 か所の変化点に対応するように, その直前でパラメータ μ の推定値 $\hat{\mu}$ の時系列的振る舞いに変動が確認された. また, 変化点スコア上昇の原因となるパラメータ推定値 $\hat{\mu}$ の変動から想定される戦況変化の内容と, 実際のプレー内容がおおむね合致している

ことが確認できた。本研究は、選手交代のような戦術変更が実戦況にどのような効果を与えたかを測定・分析するツールとしての活用可能性も考えられる。

今度の課題として、実際の試合動画を専門的知見をもとに観察し、検出された戦況変化が妥当かつ十分なものであるか検証を進めたい。また、より高度な戦術分析の実現に向け、VARモデルの μ 以外のパラメータについても精査し、変化点が検出された原因について分析を重ねる必要がある。検出された変化点の前後でどのような戦況であったのか記述を行うことも今後の課題として挙げられる。そして、データの背後に潜む戦況に関する潜在構造を推定し、戦況を変化させる因子の検出が可能となれば、より高度な戦術分析の実現へ向け前進するであろう。

謝 辞

本研究で使用したデータはデータスタジアム株式会社から提供を受けたものである。また、本研究の貸与データは情報・システム研究機構の新領域融合研究プロジェクト『社会コミュニケーション』データ中心科学リサーチコモンズ事業『人間・社会データ』の支援を受けている。なお、本分析を進める中で、矢野慎一氏(東京大学・当時)にサッカーにおける専門的知見を提供いただいた。

参 考 文 献

- Football LAB (2017). Football LAB (フットボールラボ)とは, <http://www.football-lab.jp/pages/about/>, 2017年2月28日閲覧.
- Gudmundsson, J. and Wolle, T. (2014). Football analysis using spatio-temporal tools, *Computers, Environment and Urban Systems*, **47**, 16–27.
- Kijima, A., Yokoyama, K., Shima, H. and Yamamoto, Y. (2014). Emergence of self-similarity in football dynamics, *The European Physical Journal B*, **87**(2), 1–6.
- 北川源四郎 (2005). 『時系列解析入門』, 岩波書店, 東京.
- 松田浩, 鈴木康浩 (2015). 『サッカー守備戦術の教科書 超ゾーンディフェンス論』, カンゼン, 東京.
- Moura, F. A., Martins, L. E. B., Anido, R. D. O., De Barros, R. M. L. and Cunha, S. A. (2012). Quantitative analysis of Brazilian football players' organisation on the pitch, *Sports Biomechanics*, **11**(1), 85–96.
- 西部謙司 (2013). 『サッカーで大事なことは、すべてゲームの中にある：クローズアップ 世界トップクラスの技術とアイデア』, 出版芸術社, 東京.
- 西部謙司, 北健一郎 (2011). 『サッカー戦術とは何か? が誰でも簡単に分かるようになる本』, 毎日コミュニケーションズ, 東京.
- Sarmiento, H., Marcelino, R., Anguera, M. T., Campaniço, J., Matos, N. and Leitão, J. C. (2014). Match analysis in football: A systematic review, *Journal of Sports Sciences*, **32**(20), 1831–1843.
- 高橋翔, 長谷山美紀 (2015). サッカー映像におけるグループ戦術解析のための優先度推定に関する一考察, 映像情報メディア学会技術報告, **39**(49), 7–12.
- Takeuchi, J. and Yamanishi, K. (2006). A unifying framework for detecting outliers and change points from time series, *IEEE Transactions on Knowledge and Data Engineering*, **18**(4), 482–492.
- 山西健司 (2009). 『データマイニングによる異常検知』, 共立出版, 東京.

Tracking Data to Extract Changes in Football Game Situation

Keita Kamiya¹, Wataru Nakanishi² and Yuichiro Izumi¹

¹Department of Civil Engineering, The University of Tokyo

²Department of Civil and Environmental Engineering, Tokyo Institute of Technology

In football, a “game situation” gradually changes the interaction of teams’ attacks and defenses. The ability to automatically extract a change in a football game will facilitate the development of advanced strategies as well as provide richer information to the spectators. In this research, we regard the change in a football game situation as a change in time series behavior of players and a ball.

We attempt to extract the change using ChangeFinder, which is a statistical change detection method. ChangeFinder can detect changes in nonstationary and multi-noise time-series data via online learning of the two-step VAR model. Input variables are created from the tracked data to create five types of indicators: ball position, front line position, compactness, defense vulnerability degree, and attack rate. The experiments confirm that a large fluctuation in the time-series behavior of VAR model parameters occurs just prior to a detected change point. The contents of the change in a game situation assumed from the parameter variation roughly agree with the actual play contents, suggesting that the model can detect changes in football game situations.

ドロネー分割と階層的クラスタリングを用いた 集団スポーツにおけるフォーメーション 解析手法の提案

成塚 拓真¹・山崎 義弘²

(受付 2016 年 12 月 28 日；改訂 2017 年 5 月 5 日；採択 5 月 9 日)

要 旨

集団スポーツにおいて、選手同士の相対的な位置関係(フォーメーション)が戦術上重要な意味を持つ。しかし、現在のところ、フォーメーションを定量的に解析する確立された手法は存在しない。そこで、本稿では、ドロネー分割を用いた新たなフォーメーションの解析手法を提案する。本手法では各選手を母点とするドロネー分割をネットワークと見なし、その隣接行列をフォーメーションのパターンと定義する。これにより、フォーメーションの時間変化の解析や異なる時刻間での定量的な比較が可能となる。本稿では、サッカートラッキングデータを用い、本手法と階層的クラスタリングを用いたフォーメーションの分類を試みる。

キーワード：フォーメーション，ドロネー分割，階層的クラスタリング。

1. はじめに

対戦型スポーツは、選手の動きを制限する“ルール”，および、敵や味方との相互作用に起因する“不確定性”という二つの要素が、数秒から試合全体までのスケールに渡って階層的に競合した系であると捉えることができる。対戦型スポーツが有するこうした性質は、マーキング、パス回し、フォーメーション形成といった多様な動きを生み出す要因となっている。

対戦型スポーツの中でも、集団で行われる球技などでは、選手同士の連携が戦術上重要な意味を持つ。例えば、サッカーでは、各選手が味方との距離を一定に保つことにより、全体として様々なフォーメーションを形成する。フォーメーションの形成は各選手の周りに一定の守備範囲あるいは勢力圏を作り出し、これにより、相手選手へのマーキングやパス回しを効率的に行うことができる。Taki et al. (1996)は各選手の持つこうした領域を優勢領域と呼び、「ある選手が他のどの選手よりも速く到達可能なフィールド上の領域」と定義した。優勢領域の典型的な例は、ボロノイ領域であろう(杉原, 2009)。これは、フィールド上の各点をどの選手に近いかによって分割したもので、最も単純な定義といえる。サッカーにおけるボロノイ領域の基本的な性質は、Kim (2004)や Fonseca et al. (2012)によって調べられている。優勢領域のより現実的な定義は、各選手の位置に加えて速度と加速度を考慮する方法で、運動モデルと呼ばれている。これまでに、運動モデルを用いた試合分析として、プレッシャー度合いの提案やパ

¹ 中央大学 理工学部：〒112-8551 東京都文京区春日 1-13-27

² 早稲田大学 先進理工学研究科：〒169-0072 東京都新宿区大久保 3-4-1

スコアの同定などの研究が行われている (Taki and Hasegawa, 2000; Fujimura and Sugihara, 2005; Nakanishi et al., 2010; Gudmundsson and Wolle, 2014). 優勢領域を用いたゲーム分析については Gudmundsson and Horton (2017) に詳しい解説がある.

一方, 試合中に激しく変動する選手のポジションを捉える試みが最近行われている (Bialkowski et al., 2014a; Bialkowski et al., 2014b; Lucey et al., 2013; Lucey et al., 2014; Wei et al., 2013). これらの研究では, 各選手を背番号などの固有の ID ではなく各時刻で逐次割り当てられたポジションによって識別することにより, ポジションの入れ替わりなどの動的な性質を捉えることに成功している. ポジションの割り当て方には様々な手法が存在するが, 特に, Bialkowski et al. (2014a) は, ヒートマップを用いた方法を提案している. この手法は, フォーメーションを定量的に捉える上でも有用な手法であり, 実際, チームの重心位置を中心とする座標系において 4-4-2 などの特徴的なフォーメーションを検出できることが示されている. しかし, ある時間間隔での平均位置を基にした手法であるため, フォーメーションの数秒単位での時間変化の解析や異なる時刻間のフォーメーションを比較するのが難しいという難点もある.

以上を踏まえた上で, 本研究ノートでは, フォーメーションを定量的に特徴付けるための新たな手法を提案する. 提案手法は, ボロノイ領域から自動的に求まるドロネー分割を基にしたもので, フォーメーションの時間変化の解析や異なるフォーメーションの定量的な比較を行う上で非常に有用である.

2. ドロネー分割を用いたフォーメーションの特徴付け

本研究では, データスタジアム社から提供された 2016 年度サッカー J1 リーグ第 1 節のトラッキングデータのうち, 2016 年 2 月 27 日に行われた磐田対名古屋の試合データを用いる. トラッキングデータには 0.04 秒ごとに取得した全選手の位置座標が含まれている. 以下の解析では, キーパーを除く各チーム 10 人の選手に着目し, 名古屋のチームの解析結果を示す. なお, 解析にはプログラミング言語 “python” のパッケージ (numpy, scipy, pandas, networkx, scikit-learn, matplotlib) を用いた.

あるチームのフォーメーションを特徴付けるためには, 選手同士の相対的な位置関係を抽出する必要がある. 2次元平面においてこれを実行するには, 各選手に固有の領域を割り当てた上で, その領域の隣接関係によって相対位置を決めれば良い. ここでは, そのような領域の自然な定義として, 各選手を母点としたボロノイ領域を用いる. これにより, ボロノイ領域の隣接関係はドロネー分割によって与えられる. 特に, ドロネー分割は各選手をノードとするネットワークと見なすことができるので, 以下ではドロネーネットワークと呼ぶことにする. ドロネーネットワークは, 母点間に枝が存在する場合に 1, 存在しない場合に 0 を割り当てることにより, 隣接行列 A によって明確に定義できる. なお, フィールドには境界が存在するため, 各ボロノイ線と境界との交点を新たなボロノイ点とすることによって境界の影響を考慮する.

このようにフォーメーションを隣接行列によって定義することの利点は, フォーメーションの時間変化の解析や異なる時刻間の定量的な比較が可能になる点である. 具体的に, あるチームの異なる時刻 t, t' における隣接行列を $A(t), A(t')$ とすれば, それらの非類似度 $D_{tt'}$ は次のような量で測ることができる:

$$(2.1) \quad D_{tt'} = \sum_{i=1}^{10} \sum_{j=1}^{10} [A_{ij}(t) - A_{ij}(t')]^2.$$

ここで, 非類似度を二乗ユークリッド距離として定義するのは, 次節で Ward 法による階層的クラスタリングを行うためである.

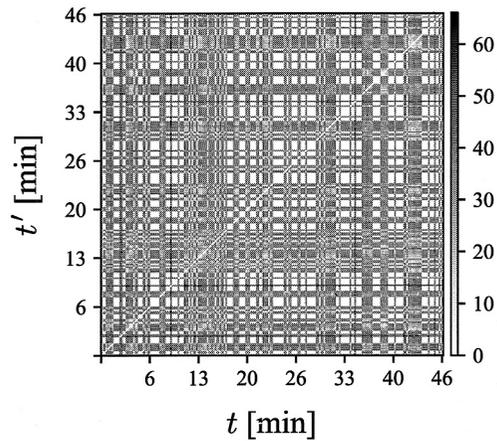


図 1. 非類似度行列 D の可視化. 非類似度 $D_{tt'}$ の値が濃淡で表されている.

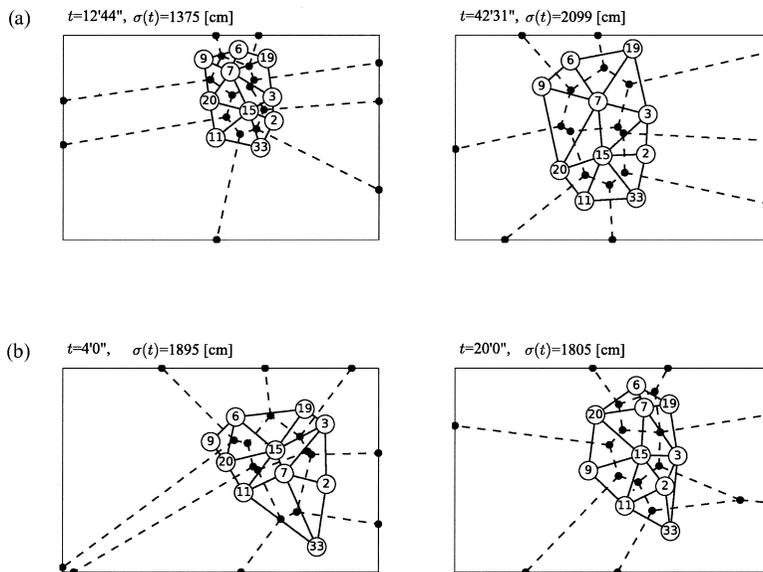


図 2. 異なる時刻のフォーメーションの比較例. (a) $D_{tt'} = 0.0$ の場合. (b) $D_{tt'} = 30.0$ の場合.

図 1 は $D_{tt'}$ を要素とする非類似度行列 D を可視化したものである. この図において, 特定の列または行に着目すると, 非類似度の小さい領域(色の薄い領域)が繰り返し現れることが分かる. これは, 1 試合の中で類似のフォーメーションが何度も間欠的に出現することを意味する. また, 図 2(a)に示したのは $D_{tt'} = 0$ となるような二つの時刻のフォーメーションであるが, $D_{tt'} = 0$ であることを反映して, 各選手の相対的な位置関係が全く同じであることが確認できる. なお, このときのチームの拡がりは, 次の式で定義される慣性半径 $\sigma(t)$ によって特徴づけられる:

$$(2.2) \quad \sigma(t) = \sqrt{\frac{1}{10} \sum_{j=1}^{10} |\mathbf{x}_c(t) - \mathbf{x}_j(t)|^2}.$$

ここで、 $\mathbf{x}_c(t)$ はチームの重心位置、 $\mathbf{x}_j(t)$ は選手 j の位置座標である。図 2(a) では二つの時刻の慣性半径が全く異なることから、本手法は各チームの慣性半径に依らず、選手同士の相対的な位置関係だけを抽出できる手法であることが分かる。一方、図 2(b) は $D_{tt'} = 30.0$ の場合の例である。この場合、7 番と 15 番、9 番と 20 番の選手が互いに位置を入れ替えている。このように、2 選手が位置を入れ替えると非類似度が大きくなる。

3. 階層的クラスタリングによるフォーメーションの分類

次に、前節の手法を応用し、階層的クラスタリングによるフォーメーションの分類を試みる。具体的な手順は以下の通りである：

- (1) 数秒ごとにドロネーネットワークを求める。
- (2) 非類似度行列 D を用いて階層的クラスタリングを行う。
- (3) デンドログラムを高さ h_c でカットし、 N_c 個のクラスターを抽出する。
- (4) クラスター内の全ドロネーネットワークをチームの重心位置を中心とする座標系に変換し、慣性半径で規格化する。各選手の平均位置からの拡がりをも楕円で表し、おおまかなポジションを可視化する。

まず、ステップ(1)で数秒ごとにドロネーネットワークを求める。ただし、本稿では選手交代の影響を取り除くため、前半だけのデータを用いた。次に、ステップ(2)では式(2.1)から得られる非類似度行列 D を用いて階層的クラスタリングを行う。ここでは、実用的な方法として知られる Ward 法を用いた。なお、Ward 法では、任意のクラスター C_1 と C_2 の距離は以下で与えられる (Pang-Ning et al., 2005)：

$$(3.1) \quad h(C_1, C_2) = V(C_1 \cup C_2) - [V(C_1) + V(C_2)].$$

ここで、 $V(C)$ はクラスター C 内の各点から重心までの二乗ユークリッド距離の和である。これにより、初期状態(各点が一つのクラスターを形成する状態)でのクラスター間距離は式(2.1)の二乗ユークリッド距離に等しくなる。

クラスタリングを行うと、図 3(a) のようなデンドログラムが得られる。ここで、デンドログラムの縦軸はクラスター併合時のクラスター間距離(クラスター併合距離)を表す。ステップ(3)では得られたデンドログラムをある高さ h_c でカットすることにより、 N_c 個のクラスターを抽出する。ここでは、クラスター数に対してクラスター併合距離が急激に増加し始める点として $h_c = 15$ を採用し[図 3(b)], $N_c = 12$ 個のクラスターが得られた。

このようにして得られた 12 個のクラスター内には様々な時刻のドロネーネットワークが含まれており、これらは互いに類似したフォーメーションを表す。そこで、ステップ(4)では次の手順で各クラスターに含まれる平均的なフォーメーションを可視化する。まず、クラスター内の全てのドロネーネットワークについて、選手 j の位置座標を

$$(3.2) \quad \tilde{\mathbf{x}}_j(t) = \frac{\mathbf{x}_j(t) - \mathbf{x}_c(t)}{\sigma(t)}$$

と変換する。この変換により、チームの拡がりとも重心の位置に依らずに各フォーメーションを比較することが可能となる。次に、各クラスターごとに、式(3.2)によって変換した後の各選手

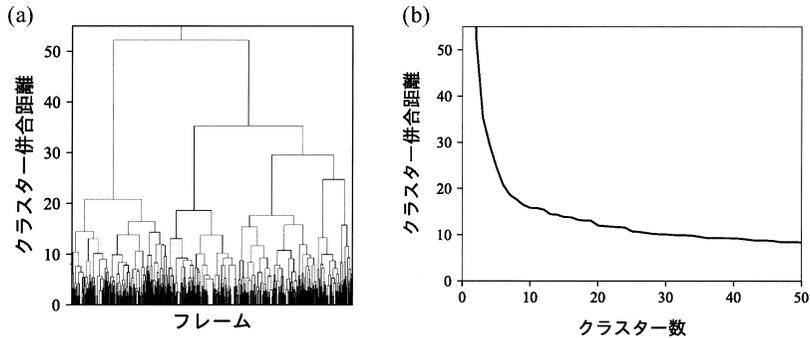


図 3. 階層的クラスタリングで得られた (a) デンドログラム, (b) クラスタ併合距離とクラスター数の関係. この場合, $h_c = 15$ を採用する.

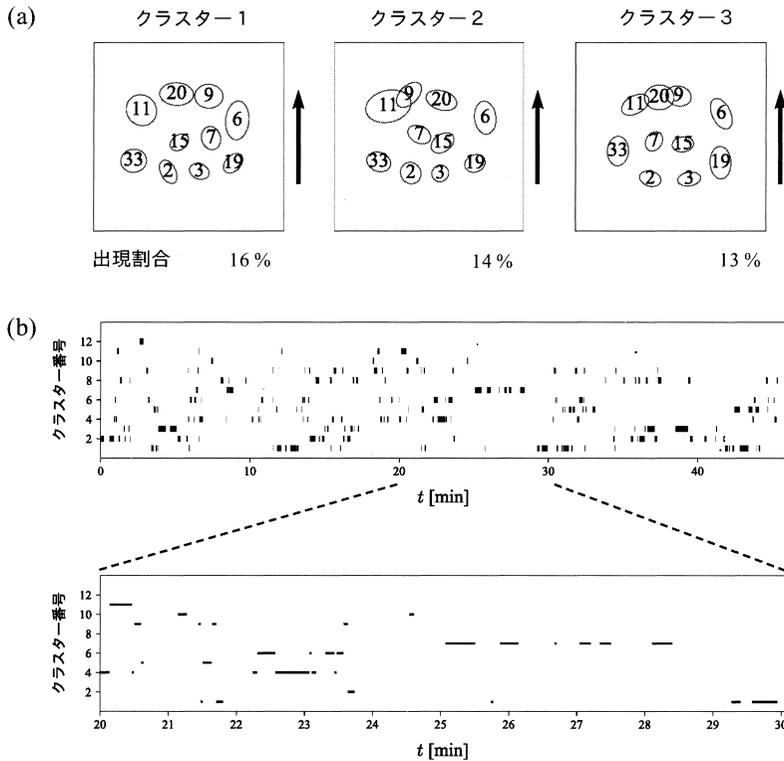


図 4. 階層的クラスタリングで得られた (a) サイズの大きい上位 3 クラスタ (矢印は攻撃方向, 番号は各選手の背番号), (b) クラスタの時間変化 (縦軸はクラスター番号).

の座標を全て重ね書きする. その上で, 選手ごとに分散共分散行列を求め, その固有ベクトル方向を長軸, 短軸とする楕円を描くと図 4(a) が得られる. ただし, 長軸, 短軸の長さは分散共分散行列の固有値のルートとして定めた. ここでは $N_c = 12$ 個のクラスターのうち, サイズの大きい上位 3 クラスタまでを示しており, 試合中に頻りに現れるフォーメーションを可視化

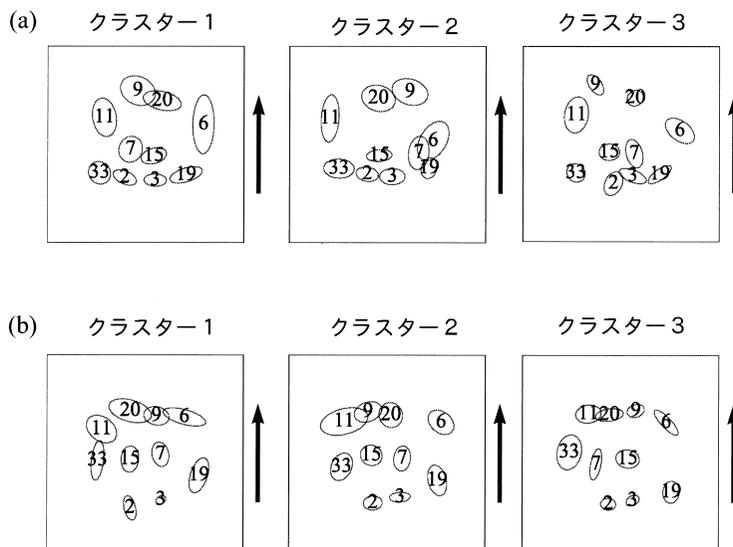


図 5. (a)守備時, (b)攻撃時に分けてクラスタリングした結果. サイズの大きい上位 3 クラスタを示している. 矢印は攻撃方向, 番号は各選手の背番号を表す.

した図となっている. この図において, 各クラスター(フォーメーション)の違いは主に中央に位置する 2 選手の位置の入れ替わりに対応していることが確認できる. また, 図 4(b)はクラスターの時間変化を表した図であるが, 数分ごとに別のクラスターへの遷移, すなわちフォーメーションの組み替えが行われていることが分かる. 特に, 特定のフォーメーションが様々な時刻において繰り返し出現していることが大きな特徴である.

次に, より詳細な場合として, 攻撃時と守備時に分けてクラスタリングを行った. ここで, 2 チーム全体での重心位置が, 3 分割したフィールドの相手陣地側(アタッキングサード)にある場合を攻撃時, 自陣側にある場合を守備時とした. 各局面で得られたフォーメーションを図 5 に示す. まず, 守備時には 4 バックで最終ラインを形成し, おおよそ 4-4-2 というフォーメーションが現れているのが特徴である. 一方, 攻撃時には 4 トップでフロントラインを形成し, おおよそ 2-4-4 というフォーメーションが現れていることが分かる. このように, 本手法は局面ごとの特徴的なフォーメーションを抽出するのにも有用である.

4. おわりに

本稿では, ドロネー分割に基づく新たなフォーメーション解析手法を提案した. 提案手法の特徴は, 各時刻におけるフォーメーションを隣接行列によって明確に定義できる点である. これにより, これまでに難しかったフォーメーションの時間変化の解析や異なる時刻間での定量的な比較が可能となった. 本稿では, 階層的クラスタリングによるフォーメーションの分類を行うに留まったが, 今後, 以下のような発展・応用可能性が考えられる.

まず, 非類似度を定義する際に隣接行列に重み付けを行うことで, より正確なクラスタリングが実現できる可能性がある. 例えば, 式(2.1)による定義では, 枝を多く持つ選手ほど枝が張り替わったときの非類似度への寄与が大きい. この問題への対処としては, 隣接行列の要素 $A_{ij}(t)$ を次数 $k_i(t)$ (選手 i の持つ枝の本数) で規格化した

$$D_{tt'} = \sum_{i=1}^{10} \sum_{j=1}^{10} \left[\frac{A_{ij}(t)}{k_i(t)} - \frac{A_{ij}(t')}{k_i(t')} \right]^2$$

という定義への拡張が考えられる。また、選手間距離や共有するボロノイ線の長さによって重み付けすれば、慣性半径の違いを反映したクラスタリングも可能と考えられる。一方、本手法を用いた試合分析としては、各クラスタの出現規則やクラスタ間の遷移、相手チームのフォーメーションとの相関といったフォーメーションの相関構造の解析が考えられる。さらに、動的なネットワーク解析として、枝の生成・消滅・張替えの過程やネットワーク指標の時間変化などを調べることも重要な課題である。こうした解析を行うことによって、対戦型スポーツにおける集団運動の理解が深まるとともに、チームを特徴付ける新たな指標の提案などにもつながると期待している。

謝 辞

本研究で用いたデータはデータスタジアム株式会社様より提供して頂きました。深く感謝致します。本研究は情報・システム研究機構の新領域融合研究プロジェクト「社会コミュニケーション」データ中心科学リサーチコモンズ事業「人間・社会データ」の支援を受けたものです。

参 考 文 献

- Bialkowski, A., Lucey, P., Carr, P., Yue, Y., Matthews, I. and Ram, F. (2014a). Win at home and draw away: Automatic formation analysis highlighting the differences in home and away team behaviors, *Proceedings of 8th Annual MIT Sloan Sports Analytics Conference*, 1–7.
- Bialkowski, A., Lucey, P., Carr, P., Yue, Y., Sridharan, S. and Matthews, I. (2014b). Large-scale analysis of soccer matches using spatiotemporal tracking data, *2014 IEEE International Conference on Data Mining*, 725–730.
- Fonseca, S., Milho, J., Travassos, B. and Araújo, D. (2012). Spatial dynamics of team sports exposed by Voronoi diagrams, *Human Movement Science*, **31**, 1652–1659.
- Fujimura, A. and Sugihara, K. (2005). Geometric analysis and quantitative evaluation of sport teamwork, *Systems and Computers in Japan*, **36**, 49–58.
- Gudmundsson, J. and Horton, M. (2017). Spatio-temporal analysis of team sports, *ACM Computing Surveys*, **50**, 22.
- Gudmundsson, J. and Wolle, T. (2014). Football analysis using spatio-temporal tools, *Computers, Environment and Urban Systems*, **47**, 16–27.
- Kim, S. (2004). Voronoi analysis of a soccer game, *Nonlinear Analysis: Modelling and Control*, **9**, 233–240.
- Lucey, P., Bialkowski, A., Carr, P., Morgan, S., Matthews, I. and Sheikh, Y. (2013). Representing and discovering adversarial team behaviors using player roles, *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2706–2713.
- Lucey, P., Bialkowski, A., Carr, P., Yue, Y. and Matthews, I. (2014). How to get an open shot: Analyzing team movement in basketball using tracking data, *Proceedings of the 8th Annual MIT Sloan Sports Analytics Conference*, 1–10.
- Nakanishi, R., Maeno, J., Murakami, K. and Naruse, T. (2010). An approximate computation of the dominant region diagram for the real-time analysis of group behaviors, *Proceedings of the 13th Annual RoboCup International Symposium*, 228–239.

- Pang-Ning, T., Steinbach, M. and Kumar, V. (2005). *Introduction to Data Mining*, Addison Wesley, Boston, USA.
- 杉原厚吉 (2009). 『なわばりの数理モデル—ボロノイ図からの数理工学入門—』, 共立出版, 東京.
- Taki, T. and Hasegawa, J. (2000). Visualization of dominant region in team games and its application to teamwork analysis, *Proceedings of the Computer Graphics International 2000*, 227–235.
- Taki, T., Hasegawa, J. and Fukumura, T. (1996). Development of motion analysis system for quantitative evaluation of teamwork in soccer games, *Proceedings of 3rd IEEE International Conference on Image Processing*, **3**, 815–818.
- Wei, X., Sha, L., Lucey, P., Morgan, S. and Sridharan, S. (2013). Large-scale analysis of formations in soccer, *2013 International Conference on Digital Image Computing: Techniques and Applications, DICTA 2013*, 1–8.

Characterization of the Formation Structure in Team Sports

Takuma Narizuka¹ and Yoshihiro Yamazaki²

¹Department of Physics, Faculty of Science and Engineering, Chuo University

²Department of Physics, School of Advanced Science and Engineering, Waseda University

In team sports, whether to maintain or rearrange a team formation is an essential strategy, but there is not an established method to analyze the influence of different formations. We propose a method to identify the formation structure based on Delaunay triangulation. The adjacency matrix obtained from the Delaunay triangulation for each player is regarded as the formation pattern. Our method allows time-series analysis and a quantitative comparison of formations. A classification algorithm of formations is proposed by combining our method with hierarchical clustering.

サッカーの攻撃におけるプレーの 最適化アルゴリズムの開発

徐 広孝^{1,2}・大澤 啓亮³・見汐 翔太²・
安藤 梢⁴・鈴木 宏哉⁵・西嶋 尚彦⁶

(受付 2017 年 1 月 10 日；改訂 3 月 15 日；採択 3 月 30 日)

要 旨

近年、スポーツパフォーマンスデータの分析が盛んに行われているが、ビッグデータを用いた研究は少ない。本研究の目的は、サッカーのビッグデータを用いて、攻撃においてシュートまでたどり着くためのプレーを最適化するアルゴリズムを考案することであった。データスタジアム株式会社から提供された 2013 年の J1 全 306 試合の攻撃データを、先行研究の測定項目に従って達成データセットに変換した。測定項目間のオッズ比から連動確率行列を作成し、次の手順による最適化アルゴリズムを作成した。(1) 攻撃プレーから達成項目を保存する。(2) シュートに対する連動確率に基づいて達成項目を降順でソートする。(3) 未達成項目を達成項目間に挿入した場合の確率を計算する。(4) その確率が達成項目間の確率よりも高い場合は未達成項目を挿入する。(5) 二重ループで挿入を行う。このアルゴリズムで最適化を適用し、ハーフタイムなどの短期的な場面や、数カ月単位の長期的な場面で活用する方法が提案された。

キーワード：サッカー、Jリーグ、攻撃プレー、最適化アルゴリズム、ビッグデータ。

1. はじめに

1993 年に Jリーグが開幕して以来、日本のサッカーの競技水準は著しく向上し、1998 年の初出場から 5 大会連続で FIFA (Fédération Internationale de Football Association) ワールドカップへの出場を果たした。海外リーグでプレーする日本人選手も年々増加し、日本サッカーの競技水準は世界に近づいているといえる。しかし、FIFA ワールドカップの結果は 2010 年大会のベスト 16 が最高成績であり、2014 年大会では一勝もあげることなくグループリーグで敗退した。FIFA (2016) の男子世界ランクは 2016 年 12 月時点で 45 位であり、世界の上位には及んでいない。

日本サッカー協会技術委員会 (2010) は、世界をスタンダードとして世界大会を分析し、様々

¹ 筑波大学附属駒場中・高等学校：〒154-0001 東京都世田谷区池尻 4-7-1

² 筑波大学大学院 人間総合科学研究科：〒305-8574 茨城県つくば市天王台 1-1-1

³ 日本スポーツ振興センター：〒107-0061 東京都港区北青山 2-8-35

⁴ 新渡戸文化学園：〒164-0012 東京都中野区本町 6-38-1

⁵ 順天堂大学 スポーツ健康科学部：〒270-1695 千葉県印西市平賀学園台 1-1

⁶ 筑波大学 体育系：〒305-8574 茨城県つくば市天王台 1-1-1

な強化策を推進している。さらに、指導者の仕事としてゲーム分析・評価、計画立案等が挙げられている(日本サッカー協会技術委員会, 2001)。鈴木・西嶋(2002)は、ゲームパフォーマンス分析は、一般的にゲーム分析(game analysis)とゲーム統計(game statistics)に大別されるとしている。ゲーム分析では専門家の視認的方法によって、技術、戦術、技能、チーム力などが質的に評価、記述され、ゲーム統計ではゲームパフォーマンスの分析手法(Hughes, 1996)を用いて得点数、失点数、シュート数などの計数データを扱うとしている。鈴木・西嶋(2002)は、ゲーム分析では分析者の主観性および恣意性を排除することはできないと指摘しており、Frank and Miller(1986)もまた、コーチのゲーム観察の正確性は3割程度であると述べていることから、客観的な測定値に基づいたゲームパフォーマンス分析が重要であるといえる。

ゲームパフォーマンスの分析には、データの取得が必要である。サッカーにおけるデータの取得法の原点は、競技知識をもった人間が試合を見て判断したものを記録する方法であり(加藤, 2016)、1960年代は手作業が主流であった。太田 他(1969)はグラウンド上の線を目印にして選手の移動を視覚的にとらえ、5分間に1枚のペースで手記する方法を用い、選手の移動距離や移動パターンを分析して戦術面に活用した。しかし、この方法ではデータの取得に多くの人数が必要であった。1980年代には、ビデオに録画された試合映像からゲームパフォーマンスを中心とした統計を取って分析する方法が主流となり(田中, 1984; 難波・清, 1988)、ハードウェアやソフトウェアの進化に伴って、データの取得方法も変化した。内山 他(1989)はビデオ撮影者1名、ゲーム展開(場所、プレイヤー、プレイの結果)を口頭で言う者1名、コンピュータに入力する者1名の計3名で役割を分担し、試合中にリアルタイムでデータを取得し、ハーフタイムなどに活用する方法を考案した。Hughes(1993)やErdmann(1991)は、ビデオとコンピュータを使用した解析の重要性を指摘した。

移動距離などのCGS単位系や、シュートの回数や決定率などの頻度または割合で測定できるパフォーマンスは比較的主観に依存する程度が低く、サッカー経験のある者であれば、少ない誤差で測定することができると考えられる。しかし、これらの測定可能なパフォーマンスはサッカーの一部を表現しているに過ぎず、複合的な技能やプレー構造などの測定できない領域については、選手や指導者がパフォーマンスデータに基づいて推測するしかない。そこで、鈴木 他(2000)はシュート技能の因果構造を明らかにし、山田 他(2000)はディフェンスプレッシング技能の因果構造を明らかにした。これらの研究では、複数の専門家によるデルファイ法(Linstone and Turoff, 1975)を伴う特性要因分析を適用して技能の因果構造を定性的に分析し、内容的妥当性を確認した。さらに、構造方程式モデリングを適用して技能領域間の因果構造モデルを検証し、技能評価尺度を構成した。

近年では、ハードウェア、ソフトウェアのさらなる進化によってデータ収集が飛躍的に向上し、スポーツにおいてもビッグデータを活用した研究がなされるようになった(徐 他, 2014; 丸山 他, 2015)。ビッグデータとは、「その厳密な定義はないが、小規模では成し得ないことを大きな規模で実行し、新たな知の抽出や価値の創出によって市場、組織、さらには市民と政府の関係などを変えること」と説明されている(Viktor and Kenneth, 2013)。スポーツにおけるデータ分析の根源的なテーマは試合に勝つための情報収集であることから(加藤, 2016)、サッカーのビッグデータによって、従来にはない有益な情報を得ることが期待されている。しかしながら、サッカーのビッグデータを用いた研究はまだ歴史が浅く、研究の蓄積が必要である。

本研究は、徐 他(2014)が開発したJリーグの攻撃力の測定項目を利用し、ビッグデータ解析による攻撃プレーの最適化アルゴリズムを考案した。攻撃プレーの最適化とは、「シュートにたどり着く確率を最大化するためのプレーを導き出す」ことである。サッカーの試合の目的は相手に勝つことであり、そのためには得点の獲得が必要であることから、攻撃局面はボールを奪ってからゴールを決める(あるいはボールを失う)までとなる。しかし、攻撃プレーを最適

化することを考えると、分析の対象はボールを奪ってからシュートにたどり着くまでとしたほうがよい。なぜなら、仮に理想的なプレーでシュートにたどり着いたとしても、シュート技能が低ければゴールを決める可能性が低くなるからである。つまり、ゴールを決めるかどうかはシュート技能への依存度が高いため、シュートにたどり着くまでのプレーと、その先のシュートを分けたほうが望ましいといえる。シュートにたどり着くためには、サッカー特有の技術と戦術を駆使しなければならない。技術と戦術の関係について、Jan (1989)は、「技術は戦術の重要な基礎である」と述べ、Deniau (1977)は、「技術は、戦術に役立てるために存在している」と述べている。すなわち、攻撃力は、技術と戦術の双方の側面から評価されるべきである。

これらのことを踏まえ、続く第2章では徐 他 (2014)の攻撃力測定項目について述べ、3章で本研究のデータへの適用結果を述べる。第4章では攻撃プレーの最適化アルゴリズムを説明し、第5章で活用方法の例を提示する。

2. 攻撃力の測定項目

徐 他 (2014)は、データスタジアム株式会社が測定した2011年のJリーグ Division1 (J1)と Division2 (J2)の全686試合のデータを使用して、選手とチームの攻撃力を評価する指標を作成した。データの変数は200を超え、一試合で測定されるレコード数は2,000程度であり、一年間で測定されるデータ行列はおよそ200列×137万行のビッグデータであった。このデータのレコードは、ボールを保持したプレイヤーが、ドリブルをする、パスを出す、パスを受けるなどの何かしらのアクションを起こすごとに、位置情報と共に1行記録される。1行が1アクションで構成されるため、このデータセットを「アクションデータセット」とした。測定項目は、鈴木・西嶋 (2002)や山田 他 (2000)の手続きに準じたデルファイ法を伴う特性要因分析によって、7因子39項目が作成された(図1)。ボールを奪ってからシュートにたどり着く(またはボールを失う)までのプレーは複数のアクションの集合であるため、測定項目の規準に基づいてアクションデータセットをプレーデータセットに変換した。プレーデータセットは1行が1プレーとなっており、レコード数はおよそ189万行であった。

徐 他 (2014)は、サッカーの「攻撃力」は、「技術力」と「シュート生産力」の因子から成ると仮定した。シュートを放つためには、相手よりも優位な状態をつくる必要があり、これが戦術の

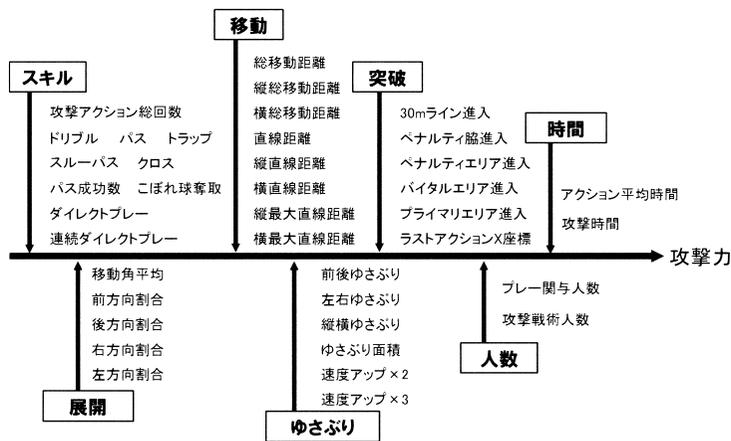


図1. 攻撃力の測定項目(徐 他, 2014).

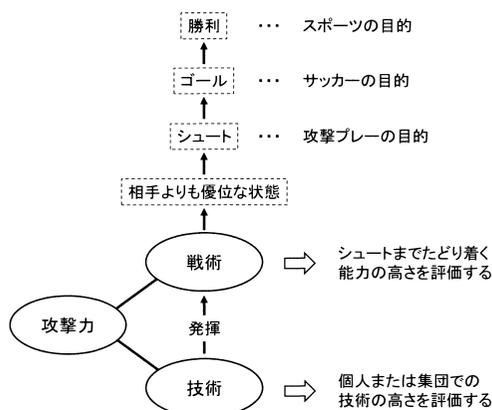


図 2. サッカーの目的と攻撃力の構造.

目的である。「シュート生産力」は、シュートにたどり着く力の大きさを表すことから、戦術技能の高さであると言い換えられる。ゆえに、技術、戦術の順に技能を発揮し、相手よりも優位に立った結果、シュートを放つことができるという因果関係が成り立つ(図 2)。これらのことから、測定項目は技術力と戦術力の両方を評価すると考えられる。

測定項目とシュートの関係を明らかにするために、CART アルゴリズム (Breiman et al., 1984) による決定木分析(分類二進木)を適用し、測定項目において、シュートした/しなかったの割合の差が最大になるような分割点を求め、これを測定項目の達成基準とした。具体的には、まず、シュートした/しなかったの 2 水準カテゴリカルデータ(以下、変数「シュート」とする)をルートノード(基準変数)とし、次式に従ってジニ係数を算出した。

$$(2.1) \quad I(S) = 1 - \sum_{j=1}^J p_j^2$$

ここで、 S は変数「シュート」、 J は水準数 2 である。 p はシュートした/しなかったの確率であり、ダミー変数であることから水準 j の分散は $p_j(1-p_j)$ となる。この分散が 0.25 であれば、シュートした/しなかったの割合が半々となり最も判別しにくい状態にあるが、サッカーではシュートにたどり着いたプレー数よりも、シュートにたどり着かなかったプレー数の方が少ないため、現実的にそうなることはない。

続いて、1つの予測変数(測定項目)を二つに分岐させる。すべての予測変数は量的尺度であるため、それぞれの分割点を探す必要がある。予測変数のレコードをソートして重複のない測定値をカウントし、その数を M とする。測定値の配列を V としたとき、分割点の候補は、 V_1 と V_2 の間から V_{M-1} と V_M の間までの $M-1$ 個である。分割点よりも大きい/小さいの二値をとるダミー変数を仮定して、次式に従い $M-1$ 個の分岐基準を計算した。

$$(2.2) \quad \Delta I = I(S) - \{P(S_L)I(S_L) + P(S_R)I(S_R)\}$$

ここで、 $P(\cdot)$ は分岐確率、 $I(\cdot)$ はジニ係数、 L は分割点よりも大きい、 R は分割点よりも小さいことを表す。 $M-1$ 個の分岐基準の中から最大のものを選び、その分割点を採用した。分割点で分岐した二つのノードにおいて、シュートした確率が高い方のノードに注目し、その割合をシュート貢献度とした。シュート貢献度は、ある測定項目の達成基準を満たすプレーをした場合に、シュートにたどり着ける確率がどの程度であるかを意味する。決定木分析は、本来で

あれば複数の予測変数で樹木を成長させるが、今回は分割点を見つけるという用途で用いた。

続いて、プレーデータセットのレコードから、シュートしたプレーのみを抽出し、さらにすべての変数を達成基準に従って達成した／しなかったの二値に変換し、これを達成データセット ($n=9,327$) とした。技術的に難しいプレーをすることは、スポーツにおいて重要であることから、2パラメータ・ロジスティック・モデル(2PLM)の項目反応理論 (Load, 1952) を適用し、項目困難度 (b) を算出した。なお、項目反応理論の前提条件となる一次元性は、テトラコリック相関行列から主因子解法の因子分析を行って、第一因子寄与率が 48.0% であり、第二因子以下に格段の差があることを確認した。

ここまで、すべての測定項目に対して、達成基準、シュート貢献度、困難度を算出した。これらのパラメータを使用して、プレーに対してシュート生産力、技術力を算出した。シュート生産力は、そのプレーにシュートまでたどり着く力がどの程度あったか、すなわち戦術技能の高さを意味する。技術力は、戦術の遂行に必要な技術技能の高さを意味する。シュート生産力 (P) は、

$$(2.3) \quad P = \frac{\sum_{i=1}^m (p_i a_i)}{A}$$

によって求められ、技術力 (B) は、

$$(2.4) \quad B = \frac{\sum_{i=1}^m (b_i a_i)}{A}$$

によって求められる。ここで、 A は達成基準を満たした測定項目数、 m は測定項目数、 p はシュート貢献度の配列、 b は困難度の配列、 a は達成した／しなかったの 1 または 0 をとる配列である。すなわち、プレーが始まってから終わるまでに達成基準を満たした項目のシュート貢献度と困難度のそれぞれの平均値を求めている。徐 他 (2014) は、 A で除する理由について、「仮に m で除した場合、無駄に多くのアクションをして達成基準を満たす項目を増やせばスコアを高めることが可能であるが、無駄なアクションをせずにシュートにたどり着いたプレーにも価値がある」と述べている。攻撃力は、因果構造に基づいてシュート生産力と技術力の積の平方根とし、因子間の相関関係を確認した。その結果、攻撃力とシュート生産力、攻撃力と技術力の相関関係が高値、シュート生産力と技術力の相関関係が低値であったことから、攻撃力を二次元で測定するモデルであることが確認された。この攻撃力指標によって、選手、チームそれぞれの技能を評価することが可能となった。しかし、どのようなプレーをすればシュートにたどり着けるかという課題は解決されていない。

3. 本研究で使用するデータの攻撃力指標への適用

第 2 章の測定項目と手続きに従い、データスタジアム株式会社から提供された 2013 年の J1 のパフォーマンスデータ (全 306 試合) を用いて達成基準を再計算し (表 1)、達成基準を満たした場合を 1、満たさなかった場合を 0 とした達成データセットを作成した。サッカーのプレーは動的であり、プレー内のある動きがその後の動きに影響すると考えられる。つまり、ある測定項目の達成は、別の項目の達成に影響するという関係が成り立つ。そこで、変数「シュート」を加えた 40 個の測定項目間の関係性をオッズ比によって求め、確率に変換した。

$$(3.1) \quad P = \frac{odds}{1 + odds} \quad odds = \frac{p(1-p)}{q(1-q)}$$

ここで、 p は片方の測定項目の達成率、 q はもう一方の測定項目の達成率である。

ある測定項目が別の測定項目に影響を及ぼすという仮定から、この確率を連動確率と命名し、すべての測定項目間の連動確率行列 (PE) を作成した (表 2)。

表 1. 測定項目の達成基準.

要因	測定項目	達成基準	要因	測定項目	達成基準	
スキル	攻撃アクション総回数	>=6	突破	30mライン進入	>=1	
	ドリブル	>=1		ペナルティ脇進入	>=1	
	パス	>=4		ペナルティエリア進入	>=1	
	トラップ	>=2		バイタルエリア進入	>=1	
	スルーパス	>=1		プライマリエリア進入	>=1	
	クロス	>=1		ラストアクションX座標	>=22	
	パス成功数	>=3		時間	アクション平均時間	>=1.05
	こぼれ球奪取	>=2			攻撃時間	>=9.5
	ダイレクトプレー	>=3		展開	移動角平均	>=31.3
	連続ダイレクトプレー	>=1			前方向移動割合	>=3.2
移動	総移動距離	>=59	後方向移動割合		>=3.3	
	縦総移動距離	>=53	右方向移動割合		>=4.8	
	横総移動距離	>=25	左方向移動割合	>=5.4		
	移動距離	>=50	ゆさぶり	前後ゆさぶり	>=2	
	縦移動距離	>=46		左右ゆさぶり	>=1	
	横移動距離	>=5		縦横ゆさぶり	>=2	
	縦最大移動距離	>=53		ゆさぶり面積	>=525	
	横最大移動距離	>=33		速度アップx2	>=2	
				速度アップx3	>=1	
			人数	プレー関与人数	>=4	
			攻撃戦術人数	>=3		

4. 攻撃プレーの最適化アルゴリズムの開発

4.1 ツリー方式

連動確率行列をもとに、攻撃プレーの最適化を試みる。最も簡素的なアルゴリズムは、変数「シュート」をルートに置き、すべての測定項目の連動確率を掛け合わせて総合連動確率を求めておき、実際に行われたプレーの中から、総合連動確率が最も高いものを探すことである。しかし、この方法の総合連動確率の組み合わせは 39 の階乗である。処理に要する時間の多さと、膨大な情報の中から人が有益な情報を探すことのコストの観点から、この方法は実用的ではなかった。

4.2 挿入方式

この方法は、実際のプレーにおいて達成された測定項目をソートし、連動確率が最も高くなるように未達成の測定項目を挿入していく方法である。その手順を以下に示す。

- ① すべての測定項目の配列 V の中から、実際に行われたプレーの達成項目を保存し、配列 A とする。
- ② A を「シュート」に対する連動確率について降順でソートし、「シュート」を要素の先頭とする。
- ③ A の要素間に、 V の要素を挿入した場合の確率 p を計算する。

$$(4.1) \quad p = PE(V_i, A_j)PE(V_i, A_{j+1})$$

ここで、 PE は連動確率行列である。次の条件を満たす場合に、 V の要素を A の要素間に挿入する。

$$(4.2) \quad p > PE(A_j, A_{j+1})$$

表 2. 運動確率行列.

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17	V18	V19	V20	V21	V22	V23	V24	V25	V26	V27	V28	V29	V30	V31	V32	V33	V34	V35	V36	V37	V38	V39	V40	
V1 シュート																																									
V2 こぼれ球ゲイン回数	81																																								
V3 トリプル	75	74																																							
V4 クロス	77	88	84																																						
V5 トラップ	62	74	62	82																																					
V6 攻撃アクション総回数	73	82	75	86	99																																				
V7 攻撃時間	73	86	80	88	97	99																																			
V8 アクション平均時間	68	89	81	91	89	93	99																																		
V9 スタック	92	65	85	94	70	74	75	62																																	
V10 プレー一回と人数	70	83	72	83	97	99	98	94	71																																
V11 攻撃戦術人数	60	77	67	76	90	91	89	85	63	94																															
V12 総移動距離	69	73	76	85	97	98	99	98	76	98	90																														
V13 縦総移動距離	68	72	73	83	94	96	97	96	96	85	99																														
V14 横総移動距離	71	75	75	86	97	98	98	94	76	98	91	99	96																												
V15 移動距離	65	64	75	80	80	83	89	98	75	84	75	92	92	85																											
V16 縦移動距離	71	63	76	81	77	87	97	84	81	73	92	95	83	99																											
V17 横移動距離	66	70	66	72	80	81	82	84	57	82	79	82	77	86	82	72																									
V18 縦最大移動距離	75	73	77	85	85	89	92	98	85	89	78	98	99	90	97	99	72																								
V19 横最大移動距離	65	69	70	81	88	90	89	76	76	90	78	91	88	92	76	67	78	77																							
V20 移動前後左右移動角平均	57	72	53	61	56	59	61	56	59	60	53	60	64	56	55	59	40	63	52																						
V21 前方向移動割合	53	67	67	75	90	91	88	78	60	90	83	90	90	86	70	71	70	80	79	68																					
V22 後方向移動割合	54	68	67	74	89	91	88	78	60	90	82	89	89	85	70	71	69	79	79	28	64																				
V23 右方向移動割合	68	75	70	80	93	94	92	80	69	93	85	92	88	94	70	69	78	78	86	80	77	77																			
V24 左方向移動割合	66	72	70	79	93	94	82	80	67	93	85	92	88	94	69	69	78	77	86	29	77	77	81																		
V25 ゆさぶ面積	72	70	75	84	96	97	88	99	76	97	88	99	98	98	88	80	93	91	64	89	88	93	93																		
V26 前後ゆさぶり回数	58	69	63	70	96	97	94	88	66	95	80	94	92	70	70	70	80	84	62	99	99	84	90																		
V27 左右ゆさぶり回数	62	70	66	74	94	94	92	84	68	92	81	92	87	95	69	68	71	76	87	68	78	79	99	89	81																
V28 縦横変換ゆさぶり回数	67	71	72	83	98	99	97	91	72	98	90	97	95	98	80	79	80	86	99	58	91	91	95	95	97	91	91														
V29 タイムアウトプレー回数	68	81	59	75	83	93	92	89	65	96	89	91	90	91	70	71	72	81	81	66	87	87	88	87	90	89	86	90													
V30 ダイレクトプレー最大連続回数	69	78	54	73	70	81	80	73	55	86	85	81	83	80	67	72	74	78	71	48	77	77	76	76	82	88	79	80	99												
V31 バス回数	58	81	67	83	97	99	98	94	69	99	90	97	96	97	81	78	80	87	89	62	91	91	94	94	96	96	93	88	97	89											
V32 スルーパス回数	69	71	67	78	82	83	83	82	82	80	72	81	78	80	76	79	67	80	66	57	72	72	75	75	80	70	80	74	71	81											
V33 パス成功回数	66	73	71	81	98	99	98	92	71	99	90	98	96	98	81	79	80	87	89	61	82	92	95	95	97	96	94	99	96	88	99	80									
V34 速度アップx2	62	81	64	79	96	98	98	96	88	98	82	96	94	96	75	75	74	84	89	70	93	93	94	93	90	89	96	92	93	99	72	98									
V35 速度アップx3	58	78	61	75	84	87	85	82	82	82	78	82	81	83	65	68	73	74	57	81	81	82	82	81	82	79	85	85	86	89	68	88	94								
V36 ペナルティエリア進入	96	80	78	91	71	78	79	73	97	77	72	79	78	88	74	79	77	83	78	55	61	62	78	75	79	66	69	75	73	79	71	82	74	71	85						
V37 ペナルティエリア進入	72	80	84	96	79	82	84	81	96	78	69	82	81	80	76	76	66	84	85	58	74	74	78	77	80	74	80	72	66	79	80	79	79	70	83						
V38 ライン30m進入	87	76	86	94	88	90	91	89	96	88	79	91	90	90	84	87	73	82	80	79	92	81	89	79	79	82	81	89	79	78	88	78	73	86	87	88	82	73	91	92	
V39 バイタルエリア進入	96	82	79	85	81	85	85	82	94	84	76	84	82	88	75	79	75	84	76	58	71	81	80	83	72	76	82	71	75	79	84	82	77	70	98	81	97				
V40 プライマリエリア進入	98	81	69	90	63	74	72	63	96	74	73	73	71	91	67	73	79	77	78	53	54	55	78	74	75	60	65	70	74	87	65	72	70	69	64	99	79	85	98		

処理①～③を、 V と A に対して二重ループさせ、 V から A に挿入された測定項目を「改善項目」、その連動確率を「改善確率」とする。

具体例を挙げて説明する。プレーの達成項目が「パス」、「縦移動距離」、「ペナ協進入」だとする(図3のA)。 V の要素数は39から3を引いた36となり、 A の要素数は3となる。 A の要素を「シュート」に対する連動確率について降順でソートすると、先頭の要素から「ペナ協進入」(0.72)、「縦移動距離」(0.71)、「パス」(0.58)となる(図3のB)。 A の先頭の要素に「シュート」を挿入すると、 A の要素間の連動確率は0.72, 0.62, 0.51となり(図3のC)、これよりも連動確率が高くなるように未達成の測定項目を挿入していく。まず、 V_1 「ドリブル」と A_1 「シュート」が0.88、 V_1 「ドリブル」と A_2 「ペナ協進入」が0.79なので、その積は0.70であり、 A_1 「シュート」と A_2 「ペナ協進入」の0.72よりも小さい。この場合、測定項目の挿入に該当しない(図3のD)。続いて、 V_2 「クロス」と A_1 「シュート」が0.92、 V_2 「クロス」と A_2 「ペナ協進入」が0.89なので、その積は0.82であり、 A_1 「シュート」と A_2 「ペナ協進入」の0.72よりも大きい(図3のE)。この場合、 A_1 「シュート」と A_2 「ペナ協進入」の間に V_2 「クロス」を挿入する(図3のF)。この処理を、 V と A のすべての要素に対して行って最適化した結果、「ペナエリ進入」、「クロス」、「ドリブル」が改善項目として挿入された(図3のG)。「シュート」を除く項目数が最適化の前後で3増加し、連動確率の積(最適化前が $0.72 \times 0.62 \times 0.51 = 0.23$ 、最適化後が $0.97 \times 0.95 \times 0.89 \times 0.84 \times 0.79 \times 0.51 = 0.28$)と項目間の連動確率(「ペナ協進入」と「シュート」は0.72から0.82、「縦移動距離」と「ペナ協進入」は0.62から0.66)が高まった。

4.3 アルゴリズムの比較

ツリー方式と挿入方式のアルゴリズムの処理速度を比較するために、両アルゴリズムのプログラムをVisual Basic for Applicationで作成し、Microsoft Office Excel 2010のアドインを実装した。一般的なノートパソコン(Sony VAIO, Intel® Core™ i7-2670QM CPU 2.20 GHz, 8 GB RAM, Windows 10 64bit)を使用して、1試合1チームの全プレーを最適化し、結果を表示するまでに要した時間を計測した。その結果、ツリー方式は76.9時間、挿入方式は78.2秒であった。

ツリー方式のアルゴリズムはシンプルであるが、すべてのパターンを計算するため処理時間が長くなってしまふ。情報量が多いことは利点のようであるが、膨大な情報の中に有益な情報が埋もれてしまい、選手やコーチが有益な情報を見落としてしまうリスクが生じる。一方、挿入方式は処理に要する時間が短く、情報量がツリー方式に比べて少ない。項目間に別の項目を挿入した場合にシュート確率が高くなるという前提に基づいたアルゴリズムであるため、関係性の低い項目が提案されることがない。そのため、プレイヤーやコーチが戦略を考えるうえで、有益な情報を見つけやすいと考えられる。しかし、挿入方式では、ある項目が達成されなかった場合にどうなるかという情報が含まれていない。これは欠点のように思えるが、この点を考慮したアルゴリズムを考えた場合、一つの問題が生じる。それは、項目を減らして計算していくと、最終的にはシュートとの連動確率が最も高い項目だけが残ってしまい、プレーとして成立しないということである。どの項目を除き、どの項目を残すかについては、プレイヤーやコーチが試合状況を総合的に判断して考えるべきであり、コンピュータに判断させることはできない。

5. チームの攻撃特性の把握とシュート率の改善

攻撃プレーの最適化を適用して、2013年のJ1上位3チーム(サンフレッチェ広島、横浜Fマリノス、川崎フロンターレ)の特性を比較した。全34節を対象にして測定項目の達成率と、改善項目として提案される確率である改善率を算出した(表3)。改善率が高い測定項目は、「エ

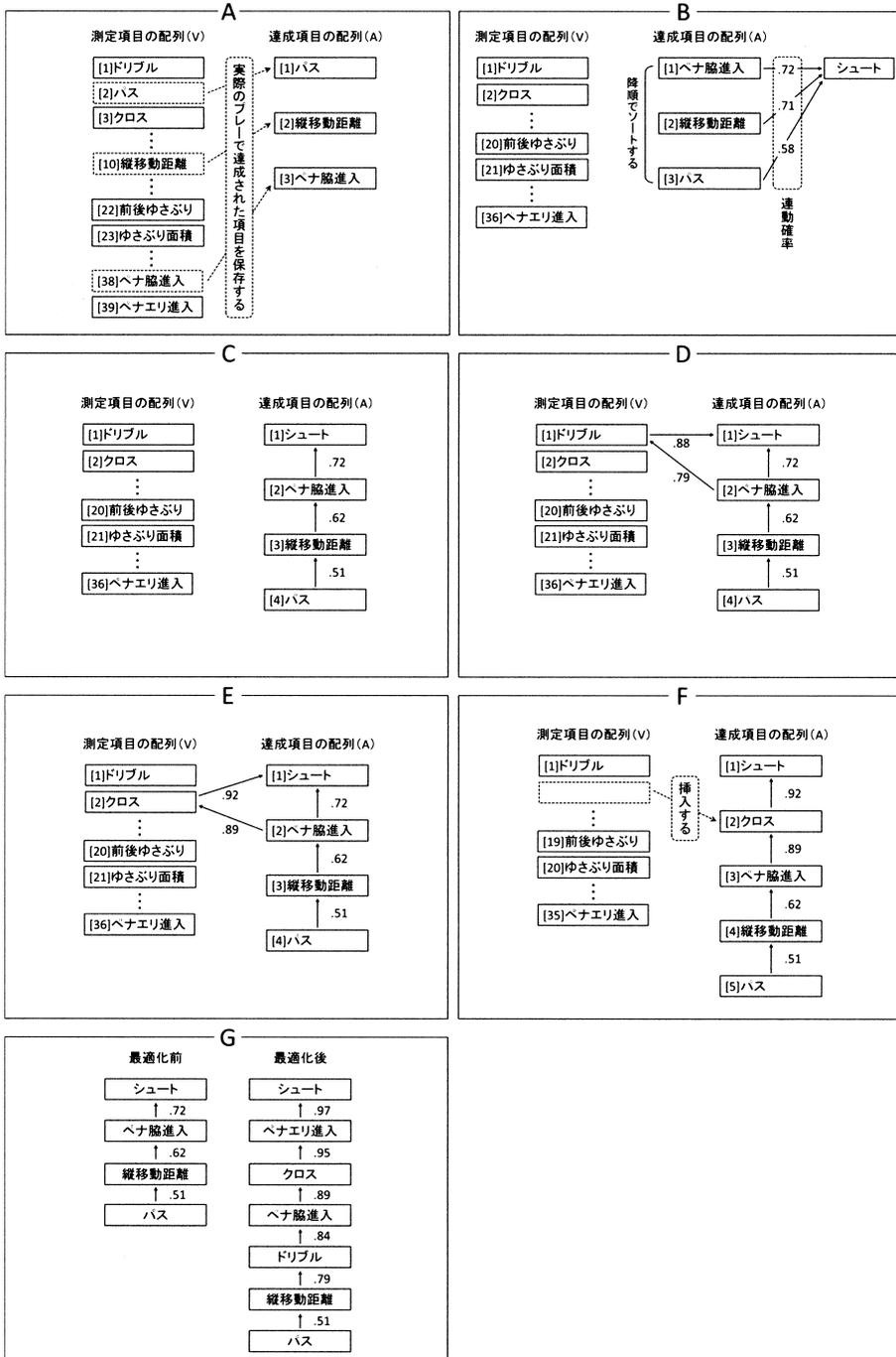


図 3. 攻撃プレーの最適化アルゴリズムの手順。

リア」要因に集中している。例えば「ペナルティエリア進入」は、ピッチにおけるペナルティエリアに進入できたかどうかを示す測定項目である。ペナルティエリアはゴールに近く守備が厚

表 3. 2013 年 J1 上位 3 チームの項目達成率と改善率、

要因	項目	広島				横浜F・マリノス				川崎					
		達成回数	改善回数	プレー回数	達成率	達成回数	改善回数	プレー回数	達成率	達成回数	改善回数	プレー回数	達成率		
スキル	攻撃アクション総回数	2,079	465	4,013	51.8	1,889	743	4,683	40.3	1,599	2,033	554	4,245	47.9	
	ドリブル	516	0	4,013	12.9	0	462	0	4,683	9.9	0	519	0	4,245	
	パス	1,817	1,038	4,013	45.3	1,635	1,478	4,683	34.9	31.6	1,693	1,237	4,245	39.9	
	スルーパス	417	160	4,013	10.4	429	111	4,683	9.2	2.4	548	142	4,245	12.9	
	トラップ	2,175	129	4,013	54.2	3.2	2,098	157	4,683	44.8	3.4	2,183	126	4,245	51.4
	クロス	579	1,446	4,013	14.4	36.0	445	1,380	4,683	9.5	29.5	352	1,513	4,245	8.3
	パス成功数	1,797	796	4,013	44.8	19.8	1,581	1,078	4,683	33.8	23.0	1,738	873	4,245	40.9
	こぼれ球奪取	87	1,406	4,013	2.2	35.0	92	1,082	4,683	2.0	23.1	91	1,142	4,245	2.1
	ダイルクトプレー	872	1,844	4,013	21.7	46.0	640	2,318	4,683	13.7	49.5	701	2,067	4,245	16.5
	連続ダイルクトプレー	2,695	82	4,013	67.2	2.0	2,893	85	4,683	63.9	1.8	2,731	105	4,245	64.3
移動	総移動距離	1,912	616	4,013	47.6	15.4	1,523	964	4,683	32.5	20.6	1,673	809	4,245	39.4
	縦総移動距離	1,553	751	4,013	38.7	18.7	1,202	940	4,683	25.7	20.1	1,318	858	4,245	31.0
	横総移動距離	2,214	311	4,013	55.2	7.7	1,902	506	4,683	40.6	10.8	2,078	383	4,245	49.0
	移動距離	1,095	837	4,013	27.3	20.9	988	648	4,683	21.1	13.8	1,134	609	4,245	26.7
	縦移動距離	835	1,014	4,013	20.8	25.3	658	838	4,683	14.1	17.9	826	815	4,245	19.5
	横移動距離	2,828	64	4,013	70.5	1.6	3,149	29	4,683	67.2	0.6	2,938	54	4,245	69.2
	縦最大移動距離	885	1,236	4,013	22.1	30.8	590	1,287	4,683	12.6	27.7	729	1,237	4,245	17.2
	横最大移動距離	1,808	259	4,013	45.1	6.5	1,740	259	4,683	37.2	5.5	1,640	282	4,245	38.6
	30mライン進入	1,150	1,559	4,013	28.7	38.8	1,170	1,724	4,683	25.0	36.8	1,213	1,523	4,245	28.6
	ペナルティ脱走	470	1,309	4,013	11.7	32.6	514	1,127	4,683	11.0	24.1	411	1,289	4,245	9.7
時間	ペナルティエリア進入	397	3,593	4,013	9.9	89.5	440	4,204	4,683	9.4	89.8	513	3,702	4,245	12.1
	ハイレベルエリア進入	920	1,606	4,013	22.9	40.0	903	1,914	4,683	19.3	40.9	1,088	1,535	4,245	25.6
	フレイムエリア進入	216	3,772	4,013	5.4	94.0	254	4,393	4,683	5.4	93.8	248	3,967	4,245	5.8
	ラストアクション/促進	1,425	818	4,013	35.5	20.4	1,604	1,101	4,683	34.3	23.5	1,578	947	4,245	37.2
	アクション平均時間	2,992	99	4,013	74.6	2.5	3,255	156	4,683	69.5	3.3	2,959	182	4,245	69.7
	攻撃時間	2,037	630	4,013	50.8	15.7	1,842	914	4,683	39.3	19.5	1,865	788	4,245	43.9
	移動角平均	3,375	0	4,013	84.1	0.0	3,837	0	4,683	81.9	0.0	3,496	0	4,245	82.4
	前方向移動割合	2,490	146	4,013	62.0	3.6	2,609	122	4,683	55.7	2.6	2,427	155	4,245	57.2
	後方向移動割合	2,429	166	4,013	60.5	4.1	2,559	139	4,683	54.6	3.0	2,444	168	4,245	57.6
	右方向移動割合	2,421	122	4,013	60.3	3.0	2,357	131	4,683	50.3	2.8	2,357	115	4,245	55.5
ゆさぶり	左方向移動割合	2,397	109	4,013	59.7	2.7	2,310	122	4,683	49.3	2.6	2,438	105	4,245	57.4
	前後ゆさぶり	506	1,354	4,013	12.6	33.7	429	1,151	4,683	9.2	24.6	472	1,243	4,245	11.1
	左右ゆさぶり	1,220	740	4,013	30.4	18.4	1,009	692	4,683	21.5	14.8	1,193	677	4,245	28.1
	縦横ゆさぶり	1,999	198	4,013	49.8	4.9	1,763	281	4,683	37.6	6.0	1,930	228	4,245	45.5
	ゆさぶり面積	1,641	698	4,013	40.9	17.4	1,134	1,048	4,683	24.2	22.4	1,325	907	4,245	31.2
	速度アップx2	533	1,985	4,013	13.3	49.5	399	2,149	4,683	8.5	45.9	418	2,069	4,245	9.8
	速度アップx3	733	960	4,013	18.3	23.9	659	767	4,683	14.1	16.4	634	849	4,245	14.9
	プレー量/人数	1,913	375	4,013	47.7	9.3	1,752	488	4,683	37.4	10.4	1,804	476	4,245	42.5
	攻撃戦術人数	2,169	299	4,013	54.0	7.5	2,317	211	4,683	49.5	4.5	2,104	302	4,245	48.6
	人数	最大値	3,375	3,772	4,013	84	94	3,837	4,393	4,683	82	94	3,496	3,967	4,245
中央値		1,641	698	4,013	41	17	1,523	767	4,683	33	16	1,578	788	4,245	37
最小値		87	0	4,013	2	0	92	0	4,683	2	0	91	0	4,245	2
平均値		1,528	846	4,013	38	21	1,465	942	4,683	31	20	1,483	873	4,245	35
標準偏差		868.7	872.3	0.0	21.6	21.7	951.0	1,002.6	0.0	20.3	21.4	876.1	897.4	0.0	20.6

くなる場所でもあるため、達成率が低くなり、改善率が高くなるのは必然的である。したがって、これらの項目よりも、選手が意思選択可能な項目に着目して分析することが望ましい。

「クロス」は、1プレーの中でクロスボールを上げられたかどうかを示す測定項目である。達成率は広島が14.4%、横浜が9.5%、川崎が8.3%であり、改善率は広島が36.0%、横浜が29.5%、川崎が35.6%であった。広島はクロスを多く上げられるチームであり、なおかつ改善率が高いことから、クロスをさらに活用することでシュート率を高めることができる。横浜は達成率と改善率が低いことから、クロス以外のプレーをいかした攻撃を組み立てるべきである。川崎は達成率が低い、改善率が高いことから、クロスをより多く上げることで、シュートチャンスが生まれると考えられる。このように、項目達成率によってチームの攻撃特性を把握することができ、改善率に基づくプレーの選択は、シュート率を高めることに貢献する。

6. おわりに

攻撃プレーの最適化を活用する場面はさまざまである。試合中にリアルタイムでデータを取得できれば、ハーフタイムに前半のプレーの最適化を行うことによって、後半にどのような攻撃プレーを選択すべきであるかをフィードバックすることが可能である。また、数カ月に渡って試合のデータを測定し、その間の全プレーを最適化すれば、項目の達成率はチームの攻撃プレースタイルを反映し、改善率はシュート率を高めるためのプレー選択における情報源となる。

加藤(2016)は、「ほとんどのクラブの担当者は、いわゆる統計的なバックグラウンドを持っていない」と述べている。サッカーに限らず、スポーツではデータを分析する専門職であるスポーツデータサイエンティストが定着しつつあるが、多くのチームが有能な人材を欲している段階であるがゆえに、分析の過程や結果は、理解が難であるものは避けるべきであろう。また、加藤はスポーツデータ解析の方向性として、「どのように攻撃を組み立てて得点するかというチームごとに異なるフィロソフィーの部分からスタートして、データの取得 → 活用 → フィードバック → 改善というプロセスを繰り返していくことが必要である」とも述べている。攻撃プレーの最適化はこれらの指摘の実現の一助となると考えられる。しかし、サッカーの技術や戦術、攻撃スタイルは変わっていくものであり、それに応じて測定項目の精査、改善が必要となる。また、連動確率がオッズ比から算出されているため、測定項目間の短方向の因果関係を考慮したアルゴリズムの開発等が次の課題である。

参 考 文 献

- Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). *Classification and Regression Trees*, Chapman & Hall, London.
- Deniau, G. (1977). *Tennistaktik*, Quickmedia, Gunzenhausen.
- Erdmann, W. S. (1991). Quantification of games — preliminary kinematic investigations in soccer, *Science and Football II* (eds. T. Reilly, J. Clarys and A. Stibbe), 174–179, E & FNSPON, London.
- FIFA (2016). FIFA Men's World Ranking, <http://www.fifa.com/fifa-world-ranking/>.
- Frank, I. M. and Miller, G. (1986). Eye witness testimony in sport, *Journal of Sports Behavior*, **9**, 38–45.
- Hughes, M. D. (1993). Notation analysis in football, *Science and Football II* (eds. T. Reilly, J. Clarys and A. Stibbe), 151–159, E & FNSPON, London.

- Hughes, M. D. (1996). Notational analysis, *Science and Soccer* (ed. T. Reilly), 343-361, E & FNSPON, London.
- Jan, K. (1989). *Taktik im Sport*, Verlag Karl Hofmann, Schorndorf.
- 徐広孝, 横尾智治, 安藤梢, 西嶋尚彦, 熊谷紗希, 猶本光, 鈴木宏哉, 山田庸, 中野貴博, 齋藤健治 (2014). Jリーグにおける選手とチームの攻撃力指標, スポーツデータ解析における理論と事例に関する研究集会, **1**, 21-26.
- 加藤健太 (2016). サッカーにおけるデータ分析とチーム強化, 通信ソサイエティマガジン, **37**, 29-34.
- Linstone, A. H. and Turoff, M. (1975). *The Delphi Method*, Addison-Wesley Publishing Company, Massachusetts.
- Load, F. M. (1952). A Theory of Test Scores, *Psychometric Monograph*, **7**, Psychometric Society, Richmond.
- 丸山智久, 鈴木惇一, 田畑耕司 (2015). ゾーンとパスから見える J1 チームの特徴, スポーツデータ解析における理論と事例に関する研究集会, **2**, 17-22.
- 難波邦雄, 清剛裕 (1988). 発達段階別に見たサッカーの基礎的技術の比較検討, サッカー医・科学研究, **8**, 125-148.
- 日本サッカー協会技術委員会 (2001). サッカー指導教本・2000 年度版, 日本サッカー協会, 東京.
- 日本サッカー協会技術委員会 (2010). U-12 指導指針, 日本サッカー協会, 東京.
- 太田哲夫, 浅見俊雄, 小宮喜久, 戸疋晴彦, 菊池武道, 森忠保 (1969). サッカーのゲーム分析, サッカー, **92**, 31-43.
- 鈴木宏哉, 西嶋尚彦 (2002). サッカーゲームにおける攻撃技能の因果構造, 体育学研究, **47**, 547-567.
- 鈴木宏哉, 山田庸, 大迫剛, 高橋信二, 西嶋尚彦 (2000). フォワード選手におけるゲームパフォーマンスからのシュート技能の計量, サッカー医・科学研究, **20**, 37-41.
- 田中和久 (1984). サッカー競技におけるスタイルの研究 最終ディフェンスラインの突破方法, サッカー医・科学研究, **5**, 49-56.
- 内山秀一, 今川正浩, 西野仁, 宇野勝 (1989). コンピュータを導入したサッカーのゲーム分析法, サッカー医・科学研究, **9**, 109-117.
- Viktor, M. S. and Kenneth, C. (2013). Letting the data speak, *Big Data: A Revolution That Will Transform How We Live, Work, and Think*, John Murray, London.
- 山田庸, 鈴木宏哉, 大迫剛, 高橋信二, 西嶋尚彦 (2000). ゲームパフォーマンスからのディフェンスプレッシャーの計量, サッカー医・科学研究, **20**, 32-36.

Development of Optimization Algorithm for Attack Play in Football

HirotaKa Jo^{1,2}, Keisuke Oosawa³, Syota Mishio²,
Kozue Ando⁴, Koya Suzuki⁵ and Takahiko Nishijima⁶

¹Junior and Senior High School at Komaba, University of Tsukuba

²Doctoral Program in Physical Education, Health and Sport Sciences, University of Tsukuba

³Japan Sports Council

⁴NitobeBunkagakuen

⁵Department of Health and Sports Science, Juntendo University

⁶Faculty of Health and Sport Sciences, University of Tsukuba

Although many analyses of sports performance data have been performed, few studies have worked with big data. The purpose of this study was to invent an optimization algorithm to increase player shot probability using big data. Using attacking data in all 306 matches in the J. League division 1 in 2013, supplied by DataStudiam Inc., we converted raw data to a binary dataset in accordance with the measurement items in a prior study. To create a cooperation probability matrix from the odds ratio between measurement items, we invented the “insertion algorithm,” which has the following procedure: (1) Store the ‘success’ items from attacking play; (2) sort the success items in descending order based on cooperation probability for “shoot”; (3) calculate probability in case inserting a ‘failure’ item between the success items; (4) if the probability was higher than the probability between the success items, insert the failure item; and (5) continue the insertion by double-loop. Team attack characteristics were compared by calculating the success rate and improvement rate adapted by the algorithm.

整数値自己回帰モデルの最近の発展

中嶋 雅彦¹・酒折 文武²・川崎 能典^{3,4}

(受付 2017 年 3 月 31 日；改訂 7 月 4 日；採択 7 月 4 日)

要 旨

整数値自己回帰モデル(Integer-valued Autoregressive Models, INAR モデル)は、自分自身の過去(整数値)に直接依存させて現時点での値(整数値)を説明するものであり、潜在過程に基づく動的な一般化線形モデルや一般化状態空間モデルとは、データへの接近法は大きく異なる。特に、自己回帰モデルの左辺と右辺において確率分布が整合的であるためには、周辺分布、自己回帰項、イノベーションのそれぞれの定式化が重要である。1980 年代後半から 90 年代初期にいくつかの研究が行われたあとやや停滞していたこの分野は、2000 年代後半から活性化してきている。本稿は、最近の整数値自己回帰モデルの発展を、ポアソン分布に基づく INAR モデルから出発して、定式化を変えた各種モデルを順に取り上げることで概観する。その過程で、必ずしも既存の文献で詳述ないし証明の与えられていない結果に関しては、著者なりに補った結果を付録に収めた。ポアソン分布の差から生成される分布に基づく INAR モデルについては、若干新しい提案と結果を加えることができた。その枠組みを利用した実データ解析例を、最後に紹介する。

キーワード：整数値時系列データ、間引き演算子、確率変数の分解、INAR(1) モデル、INAR(p) モデル、モーメント法。

1. はじめに

センサー技術など情報通信技術の発達に伴うデータ取得環境の変化の中で、様々な時系列データが収集・活用されるようになり、時系列データに対する統計的モデリングの重要性が高まっている。時系列解析においては、自己回帰移動平均(ARMA)モデルに代表される、イノベーションに正規分布を仮定した線形モデルが、現在でも最も頻繁に用いられるモデルであり、その対象は理学・工学のみならず経済学等の社会科学にも広がって久しい。

しかしながら、このような枠組みが、観測される時系列が整数値である場合でも妥当かどうかは注意を要する。例えば、マクロ経済時系列データの多くは、その取りうる値が非負で整数値とはいっても、概ね連続量と見なして差し支えないであろう。従って、イノベーションならびにデータに対する連続分布の仮定は近似として許される範囲として、あとは分布形とモデルの定式化の問題となる。

一方、観測データが計数時系列(count time series)で、とりわけ稀なイベントの計測から生じ

¹ 中央大学大学院 理工学研究科：〒112-8851 東京都文京区春日 1-13-27

² 中央大学 理工学部：〒112-8851 東京都文京区春日 1-13-27

³ 統計数理研究所：〒190-8562 東京都立川市緑町 10-3

⁴ 総合研究大学院大学 統計科学専攻：〒190-8562 東京都立川市緑町 10-3

ているような場合、すなわち 0, 1, 2 からただか 10 程度の計数の連なりである場合はもはや正規近似は妥当性を欠き、データ分布に何らかの離散分布を仮定することが必要である。計数時系列に対するモデリング法として多くの成功事例を集めながら拡張されてきたのは、一般化線形モデル (generalized linear model, GLM; Nelder and Wedderburn, 1972) の時系列版である動的一般化線形モデル (Dynamic GLM, 以下 DGLM と略) であった。このような問題設定では、一般化された状態空間表現に基づく更新アルゴリズムを実現するための新たな計算法の提案が必要とされた。例としては Kitagawa (1987) による数値積分フィルタが挙げられる。

一般化状態空間モデルないし階層構造モデルで少数計測データ (small count data) を分析することの大きなメリットは、背後にある生起度 (intensity) が仮に非定常に近い振る舞いをするような場合でも、生起度を表す潜在時系列モデルの側で柔軟性に富んだ適切なモデルを置いてやれば、少なくとも現象の記述には支障がない点であろう。DGLM に関しては West and Harrison (1997), Fahrmeir and Tutz (2001) を、一般状態空間モデルの推定法に関しては Kitagawa (2010) の第 14 章と第 15 章を参照されたい。

さて、本稿で取り上げる整数値自己回帰モデル (Integer-valued Autoregressive Models, INAR モデル) は、整数値時系列のモデリングのための別種のアプローチである。DGLM や一般化状態空間モデルでは背後の潜在時系列に系列の履歴情報が縮約されるわけであるが、INAR モデルは整数値の履歴自体にモデルを依存させる。そうすると、例えばイノベーションに連続分布を仮定した線形 AR(1) モデル $X_t = \alpha X_{t-1} + \epsilon_t$ からの類推で考えれば、自己回帰項 αX_{t-1} の部分が整数値となるよう定義を与え直し、かつ両辺の分布が整合する形でイノベーション ϵ_t の分布を定式化しなければならない。INAR モデルにおける自己回帰項の定義は、2.1 節で説明する通り「間引き (thinning)」で与えられる。間引きの構造は、自己回帰項が 0 から X_{t-1} の値を取りうる形で定式化されており、それが 0 に近いのか X_{t-1} に近いのか、換言すれば時系列として持続性 (persistence) が希薄なのか強いのかは、間引き項を生成する確率分布のパラメータ設定次第となっている。

以下では、INAR モデルとその拡張に関する文献をサーベイしておこう。著者らが文献上確認できた INAR に関する最も古い論文は McKenzie (1985) である。INAR(1) モデルを p 次に拡張した研究として、Al-Osh and Alzaid (1990), Du and Li (1991) がある。Du and Li (1991) では INAR(p) モデルのパラメータ $\alpha_1, \alpha_2, \dots, \alpha_p$ をユールウォーカー法により推定できることを示しており、Al-Osh and Alzaid (1990) では、INAR(p) モデルの周辺分布がポアソン分布のとき、 $\sum_{i=1}^p \alpha_i \leq 1$ ならば 1 次のときと同様に確率変数が退化せずに分解可能であることが述べられている。

こうして 80 年代後半から 90 年代初頭に散発的に行われていた研究は、2000 年代後半に入っようやく活性化する。Leonenko et al. (2007) は INAR(1) モデルの周辺分布をポアソン分布ではなく負の二項分布とした NBD INAR(1) モデルを提案し、過分散なデータに対する INAR モデルを提案している。Savani and Zhigljavsky (2007a, 2007b) では、NBD INAR(1) モデルのパラメータのモーメント法による推定について言及しており、3 種類の推定量を提案し、比較している。

Fokianos (2011) は周辺分布をポアソン分布とした PINAR(1) モデルを中心に解説を加えているが、ハンドブックの一章ということもありサーベイ的な側面も強いことから、現時点でこの分野の研究を始める際にまず目を通しておくべき文献と言える。Rajarshi (2012) も INAR を含む離散確率過程のモノグラフであり、間引き項を二項分布とし、周辺分布を一般的によく知られた離散型分布としたモデルについて広く述べている点に特徴がある。

間引き項を二項分布ではなく幾何分布や超幾何分布とした研究もある。Rinstić et al. (2009) は、間引き演算子を幾何分布、周辺分布を幾何分布とした NGINAR(1) モデルを導入し、パラ

表 1. INAR モデルの概要.

名称	間引き演算子	周辺分布	誤差分布	p 次の拡張	負の場合の拡張
PINAR	二項	ポアソン分布	ポアソン分布	あり	なし
NBD INAR	二項	負の二項分布	負の二項幾何分布	なし	なし
BINAR	二項・超幾何	二項分布	二項分布	あり	なし
SINAR	二項	スケラム分布	スケラム分布	あり	あり
NGINAR	負の二項	幾何分布	混合幾何分布	なし	なし
STINAR	負の二項	SDL 分布	混合 SDL 分布	なし	あり

メータ推定にも言及している. Weiß (2008) は, 間引き演算子を超幾何分布, 周辺分布を二項分布とした BINAR(1) モデルを考え, 確率変数の取りうる値が有限な場合の INAR モデルを提案した.

整数値全体をとる時系列データに対する INAR モデルとして Freeland (2010) は, ポアソン分布の差が従う分布としての Skellam 分布 (以下スケラム分布と記す) に着目し, 2つの独立な PINAR(1) モデルの差による SINAR(1) モデル (Skellam INAR(1) モデル) を提案した. なお, Freeland (2010) では, 同じパラメータを持つポアソン分布の差を取った研究しか述べられていない. また, Baretto-Souza and Bourguignon (2013) は, 2つの独立な NGINAR(1) 過程の差を考えた非対称離散ラプラス分布による STINAR(1) モデルを提案し, さらに負の相関をもつ場合への拡張を与えた. ここまで名前を挙げたモデル群を, その特徴と拡張性の現状という観点から整理したのが, 表 1 である.

以上を踏まえ本稿では, まず PINAR モデルから始めて, NBD INAR モデル, BINAR モデル, SINAR モデルの定義と性質について紹介する. 文献上必ずしも詳述ないし証明が与えられていない結果に関しては, 著者らなりに補った結果を付録に収めている. 本稿は基本的には研究詳解であるが, SINAR(1) モデルの拡張とパラメータ推定法に関しては, 若干新しい提案を加えることができたので 4.2 節に記し, 5 節で実データへの応用を試みた結果を報告する.

2. 平均と分散が等しい場合: PINAR モデル

非負整数値全体をサポートとする基本的な分布であるポアソン分布を X_t の周辺分布とする PINAR モデル (Rajarshi, 2012) は, INAR モデルの中で最も代表的なモデルである.

2.1 PINAR(1) モデル

PINAR(1) モデルは以下で表される:

$$(2.1) \quad X_t = \alpha \circ X_{t-1} + \varepsilon_t, \quad X_t \sim Po(\lambda).$$

いま, X_t を平均が λ であるポアソン分布 $Po(\lambda)$ に従う確率変数とする. また, $\alpha \in (0, 1]$, X を非負整数値とし, $\{Y_i\}$ を独立にベルヌーイ分布 $B(1, \alpha)$ に従う確率変数とする. このとき, $\alpha \circ X$ を確率変数として $\alpha \circ X = \sum_{i=1}^X Y_i$ と定義する. すなわち, $\alpha \circ X$ のとりうる値は 0 から X までの整数値であり, X を “間引く” という意味で \circ を間引き演算子 (thinning operator) と呼ぶ. このとき, $X_{t-1} \sim Po(\lambda), \alpha \circ X_{t-1} | X_{t-1} \sim B(X_{t-1}, \alpha)$ より, $\alpha \circ X_{t-1}$ が平均 $\lambda\alpha$ のポアソン分布に従うことが示せる. したがって, ポアソン分布の再生性より, $\{\varepsilon_t\} \sim Po(\lambda(1-\alpha))$ が成り立つこともわかり, PINAR(1) のパラメータは α と λ のみで表されることが確認できる.

PINAR(1) モデルに関する基本的な性質を見ておこう. 条件付き期待値, 条件付き分散は

$$E(X_t | X_{t-1}) = E(\alpha \circ X_{t-1} + \varepsilon_t | X_{t-1}) = \alpha X_{t-1} + (1-\alpha)\lambda,$$

$$V(X_t|X_{t-1}) = V(\alpha \circ X_{t-1} + \varepsilon_t|X_{t-1}) = \alpha(1 - \alpha)X_{t-1} + (1 - \alpha)\lambda$$

となる。また、 $E(X_t)$ は条件付き期待値の性質を用いると

$$E(X_t) = E(E(X_t|X_{t-1})) = E(\alpha X_{t-1} + (1 - \alpha)\lambda) = \lambda,$$

$V(X_t)$ は条件付き分散の性質を用いると

$$V(X_t) = V(E(X_t|X_{t-1})) + E(V(X_t|X_{t-1})) = \lambda$$

となり、周辺分布であるポアソン分布の期待値と分散に一致していることが確認できる。PINAR(1) モデルは、1 時点前の値に依存しているため 1 次のマルコフ性を持ち、状態 i から j に推移する確率は

$$p_{ij} = P(X_{t+1} = j|X_t = i) = \sum_{z=0}^{\min(i,j)} \binom{i}{z} \alpha^z (1 - \alpha)^{i-z} \frac{e^{-\lambda} \lambda^{j-z}}{(j-z)!}$$

となる。推移確率が時間 t に依存しないので、この時系列過程は斉時性が成り立つことが分かる。また、このマルコフ系列は任意の i, j に対して $p_{ij} > 0$ となることから、既約で非周期的であることが分かる。

2.2 PINAR(p) モデル

PINAR(1) モデルは p 次へ自然に拡張できる (Al-Osh and Alzaid, 1990 参照)。PINAR(p) モデルは

$$X_t = \alpha_1 \circ X_{t-1} + \alpha_2 \circ X_{t-2} + \cdots + \alpha_p \circ X_{t-p} + \varepsilon_t$$

で定義される。ここで、 $X_t \sim Po(\lambda)$, $\sum_{i=1}^p \alpha_i \leq 1$, $\alpha_i \circ X_{t-i} \sim B(X_{t-i}, \alpha_i)$ である。このとき、 $\alpha \circ X_{t-j} \sim Po(\alpha_j \lambda)$ となり、 $\varepsilon_t \sim Po((1 - \sum_{i=1}^p \alpha_i)\lambda)$ であることが分かる。

PINAR(p) モデルにおいて、 X_t の条件付き確率分布は

$$P(X_t = k|X_{t-1} = x_1, \dots, X_{t-p} = x_p) = \prod_{i=1}^p \sum_{z_i=0}^{\min(x_i, k)} \binom{x_i}{z_i} \alpha_i^{z_i} (1 - \alpha_i)^{x_i - z_i} \frac{\lambda^{k - \sum_{i=1}^p z_i} e^{-\lambda}}{(j - \sum_{i=1}^p z_i)!},$$

条件付き期待値、条件付き分散は

$$(2.2) \quad E(X_t|X_{t-1}, X_{t-2}, \dots, X_{t-p}) = \sum_{i=1}^p \alpha_i X_{t-i} + \left(1 - \sum_{i=1}^p \alpha_i\right) \lambda,$$

$$(2.3) \quad V(X_t|X_{t-1}, X_{t-2}, \dots, X_{t-p}) = \sum_{i=1}^p \alpha_i (1 - \alpha_i) X_{t-i} + \left(1 - \sum_{i=1}^p \alpha_i\right) \lambda$$

で与えられる。この証明は Al-Osh and Alzaid (1990) では示されていないので、付録 A.2 に記しておく。

2.3 PINAR モデルにおけるパラメータ推定

PINAR モデルにおけるパラメータ推定はモーメント法を用いるのが最も単純である。ここでは PINAR(1) モデルにおけるパラメータ α, λ のモーメント法での推定量について述べておく。まず、1 次の自己共分散 $\text{Cov}(X_t, X_{t-1})$ は

$$(2.4) \quad \text{Cov}(X_t, X_{t-1}) = E(X_{t-1}(\alpha \circ X_{t-1} + \varepsilon_t)) - E(X_t)E(X_{t-1}) = \alpha V(X_{t-1})$$

である。定常性より $V(X_t) = V(X_{t-1})$ であるので、1 次の自己相関 $\rho(1)$ は

$$\rho(1) = \frac{\text{Cov}(X_t, X_{t-1})}{\sqrt{V(X_t)V(X_{t-1})}} = \alpha$$

となることが分かる。よって、 α のモーメント法による推定量は、1 次の標本自己相関 $\hat{\rho}(1)$ を用いて $\hat{\alpha} = \hat{\rho}(1)$ とすればよい。ただし、 α は間引きの確率を表すため非負であるが、標本自己相関は負の値をとることもありうるため注意が必要である。さらに、 $X_t \sim Po(\lambda)$ なので、 $E(X_t) = \lambda$ であることから、 λ のモーメント法による推定量は標本平均 \bar{X} を用いて $\hat{\lambda} = \bar{X}$ とすればよい。なお、 $\hat{\lambda}$ は、 $E(\hat{\lambda}) = \lambda$ 、 $V(\hat{\lambda}) = \frac{1}{nT(\lambda)}$ であることから、不偏性と有効性をもつことがわかる。

もちろん最尤推定量などを考えることも可能であるが、推定量は陽関数としては得られず数値的に解くことになるため、最尤推定量のもつ漸近的な性質以上のことを言うことが困難となる。

3. その他の非負整数値自己回帰モデル

本節では、まず周辺分布を負の二項分布とした NBD INAR(1) モデルについて述べる。次に周辺分布を二項分布とした BINAR(1) モデルについて述べる。

3.1 過分散の場合：NBD INAR(1) モデル

これまで、INAR(1) の周辺分布をポアソン分布とした PINAR モデルについて述べた。ポアソン分布には平均と分散が等しいという特徴があるので、データの平均と分散がほぼ等しければ PINAR(1) は妥当なモデルであると考えられる。しかしながら、平均よりも分散が大きい、すなわち過分散の場合には PINAR(1) モデルは適切でない。そこで、過分散なモデルとして、INAR(1) モデルの周辺分布を負の二項分布 (NBD) とした NBD INAR(1) モデルについて説明する。なお、これらにおけるパラメータ推定も PINAR(1) モデルと同様にモーメント法などで行うことができる。詳しくは Leonenko et al. (2007), Savani and Zhigljavsky (2007a, 2007b) を参照されたい。

3.1.1 NBD INAR(1) モデル

NBD INAR(1) モデルの式を次のように定義する：

$$X_t = \alpha \circ X_{t-1} + \varepsilon_t, \quad \alpha \circ X_{t-1} | X_{t-1} \sim B(X_{t-1}, \alpha), \quad X_t \sim NBD(\gamma, \beta).$$

ここで $NBD(\gamma, \beta)$ はパラメータ γ, β の負の二項分布であり、その確率関数は

$$f(x) = \binom{\gamma + x - 1}{x} \left(\frac{\beta}{\beta + 1} \right)^\gamma \left(\frac{1}{\beta + 1} \right)^x \quad (x = 0, 1, 2, \dots)$$

である。このとき、 $\alpha \circ X_{t-1} \sim NBD(\gamma, \frac{\beta}{\alpha})$ である。証明は Leonenko et al. (2007) を参照されたい。

この NBD INAR(1) モデルにおいては、 X_t と $\alpha \circ X_{t-1}$ の従う負の二項分布の第 2 パラメータが異なるため、 ε_t の分布を再生性から求めることはできない。そこで、 ε_t の確率母関数を求めると、

$$G_{\varepsilon_t}(s) = \frac{G_{X_t}(s)}{G_{\alpha \circ X_{t-1}}(s)} = \alpha^\gamma \left(1 - (1 - \alpha) \frac{\beta}{\alpha + \beta} \left(1 - \frac{\alpha s}{\alpha + \beta} \right)^{-1} \right)^{-\gamma}$$

となる。これは、負の二項幾何分布 $NBG(\gamma, \frac{\beta}{\alpha + \beta}, \alpha)$ の確率母関数なので、 $\varepsilon_t \sim NBG(\gamma, \frac{\beta}{\alpha + \beta}, \alpha)$

であることが分かる. なお, パラメータ m, p, θ の負の二項幾何分布 $NBG(m, p, \theta)$ は, 以下の確率関数をもつ確率分布である:

$$f(x) = \sum_{k=0}^{\infty} \binom{k+x-1}{x} p^k (1-p)^x \binom{m+k-1}{k} \theta^m (1-\theta)^k \quad (x=0, 1, 2, \dots).$$

この確率母関数は

$$\begin{aligned} G(s) &= \theta^m (1 - (1-\theta)p(1-(1-p)s)^{-1})^{-m} \\ &= \left(\frac{\theta}{1 - (1-\theta)\frac{p}{1-(1-p)s}} \right)^m \end{aligned}$$

であり, 負の二項分布の確率母関数の中に幾何分布の確率母関数が含まれている. これが負の二項幾何分布とよばれる理由である.

3.1.2 Stochastic thinning による NBD INAR(1) モデル

先の NBD INAR(1) モデルでは, X_t の周辺分布および間引き項 $\alpha \circ X_{t-1}$ は負の二項分布であったが, 誤差項 ε_t は NBG 分布となった. 誤差分布も同じ負の二項分布とするために, Leonenko et al. (2007) では, NBD INAR(1) モデルにおけるパラメータ α を定数ではなくベータ分布に従って変動する Stochastic thinning (確率の間引き) とするモデルを提案している.

いま,

$$X_t = A_t \circ X_{t-1} + \varepsilon_t, \quad X_t \sim NBD^{(\gamma, \beta)}$$

とし, 間引きの確率を表すパラメータ A_t が定数ではなく, 独立にベータ分布に従うとする. すなわち, $A_t \circ X_{t-1} | X_{t-1} \sim B(X_{t-1}, A_t)$, $A_t \sim \text{Beta}(a, \gamma - a)$, $X_t \sim NBD^{(\gamma, \beta)}$ である. このとき, $A_t \circ X_{t-1} \sim NBD^{(a, \beta)}$ が成り立つことが示される. このことより, X_t と $A_t \circ X_{t-1}$ の第 2 パラメータがともに β のため, ε_t の分布が負の二項分布の再生性より $\varepsilon_t \sim NBD^{(\gamma-a, \beta)}$ となることがわかる. このモデルを, stochastic thinning による NBD INAR(1) モデルという.

3.1.3 Mixed thinning による NBD INAR(1) モデル

3.1.1 節の間引きパラメータが定数 α のモデルと, 3.1.2 節の確率変数 A_t のモデルを混合し, αA_t を間引きパラメータとする一般的なモデル

$$X_t = \alpha A_t \circ X_{t-1} + \varepsilon_t$$

も Leonenko et al. (2007) で提案されている. ここで, $\alpha A_t \circ X_{t-1} | X_{t-1} \sim B(X_{t-1}, \alpha A_t)$, $A_t \sim \text{Beta}(a, \gamma - a)$, $X_t \sim NBD^{(\gamma, \beta)}$ である. このとき, $\alpha A_t \circ X_{t-1} \sim NBD^{(a, \frac{\beta}{\alpha})}$, $\varepsilon_t \sim NBD^{(\gamma, \frac{\beta}{\alpha})} \cdot NBG(\gamma, \frac{\beta}{\alpha+\beta}, \alpha)$ となることが示される. このモデルを, mixed thinning (混合間引き) による NBD INAR(1) モデルという.

3.2 サポートが有限の場合: BINAR(1) モデル

ここまで, X_t のとりうる値が非負の整数全体のモデルについて考えてきた. しかし実際には, ある有限の値 n までしか値をとらない場合も考えられる. ここでは, 周辺分布を最も代表的な二項分布としたモデルについて説明する.

単純に INAR(1) モデルの周辺分布を二項分布とした場合, 誤差分布が一般的に知られた分布にならない. その改善策として 2 つの方法が提案されている. 誤差分布を用いず, 2 つの二項演算子を用いる方法と, 超幾何演算子を用いる方法である. 詳細は, Weiß (2008, 2009), Rajarshi (2012) を参照されたい. 原論文では両方とも BINAR(1) モデルとして言及されているが, ここでは区別のためにそれぞれ $\text{BINAR}_b(1)$ と $\text{BINAR}_h(1)$ と呼ぶこととする.

3.2.1 BINAR_b(1) : 2つの二項演算子を用いたモデル

まず, 2つの二項演算子を用いたモデルを説明する. $X_t \sim B(n, p)$, $\rho \in [\max(-\frac{1-p}{p}, -\frac{p}{1-p}), 1]$, $\alpha \circ X_{t-1} | X_{t-1} \sim B(X_{t-1}, \alpha)$ とする. また, $\beta = p(1 - \rho)$, $\alpha = \beta + \rho$ とする. このとき,

$$X_t = \alpha \circ X_{t-1} + \beta \circ (n - X_{t-1}) \quad (t \geq 1)$$

を BINAR_b(1) モデルという.

右辺により与えられる分布が X_t の分布, すなわち二項分布 $B(n, p)$ であることを確認しておこう. 右辺の確率母関数は

$$\begin{aligned} E(s^{\alpha \circ X_{t-1} + \beta \circ (n - X_{t-1})}) &= E((1 - \alpha + \alpha s)^{X_{t-1}}) E((1 - \beta + \beta s)^{n - X_{t-1}}) \\ &= (1 - \beta + \beta s)^n \left(1 - p + p \left(\frac{1 - \alpha + \alpha s}{1 - \beta + \beta s} \right) \right)^n \end{aligned}$$

となる. ここで, $\beta = p(1 - \rho)$, $\alpha = \beta + \rho$ を代入すると,

$$E(s^{\alpha \circ X_{t-1} + \beta \circ (n - X_{t-1})}) = (1 - p + ps)^n$$

となることが示される.

BINAR_b(1) モデルにおける X_t の 1 次の自己相関 $\rho(1)$ は ρ である. 証明は付録 A.3 に記した.

3.2.2 BINAR_h(1) : 超幾何演算子を用いたモデル

次に, 間引き演算子に超幾何演算子を用いたモデルについて説明する. いま, $X_t \sim B(N, p)$ とし, $n/N \diamond X_{t-1} | X_{t-1} \sim HG(N, X_{t-1}, n)$ とする. ここで $HG(N, m, n)$ は超幾何分布であり, その確率関数は

$$f(x) = \frac{\binom{m}{x} \binom{N-m}{n-x}}{\binom{N}{n}} \quad (x = 0, 1, 2, \dots, \min(m, n))$$

である. このとき, 次のモデルを BINAR_h(1) モデルという:

$$X_t = n/N \diamond X_{t-1} + \varepsilon_t \quad (t \geq 1).$$

このモデルにおいて, $n/N \diamond X_{t-1} \sim B(n, p)$ となることが示される(付録 A.4 参照). したがって, 二項分布の再生性より $\{\varepsilon_t\} \sim B(N - n, p)$ となることが分かる.

また, このモデルは PINAR(p) モデルと同様, 二項分布の再生性より p 次への拡張が可能であるが, Weiß(2008, 2009)では提案されていない.

4. \mathbb{Z} 上の整数値自己回帰モデル

これまででは, X_t が非負整数値をとる場合のモデルについて言及してきた. 本節では, 負の整数値も含む整数全体 \mathbb{Z} に拡張したモデルとして, 周辺分布をスケラム分布とするモデルについて述べる.

4.1 SINAR(1) モデル

ここでは, Freeland (2010)で提案されている 1 次の SINAR モデル(Skellam INAR モデル)について述べる.

まずはスケラム分布について確認しておく. X, Y が独立でそれぞれパラメータ λ_1, λ_2 のポアソン分布に従うとする. このとき, $Z = X - Y$ の従う分布を, パラメータ λ_1, λ_2 のスケラム分布という. ここでは, パラメータが λ_1, λ_2 のスケラム分布を $S(\lambda_1, \lambda_2)$ と表記する. スケ

ラム分布 $S(\lambda_1, \lambda_2)$ の確率関数は

$$f(z) = e^{-(\lambda_1 + \lambda_2)} \lambda_1^z \sum_{x=\max(0, -z)}^{\infty} \frac{(\lambda_1 \lambda_2)^x}{(x+z)! x!} \quad (x \in \mathbb{Z})$$

である。スケラム分布の性質については、例えば Alzaid and Omaid (2010) を参照されたい。

スケラム分布を用いて、SINAR(1) モデルは次のように定義される。 X_t, Y_t をそれぞれ独立でパラメータの等しい PINAR(1) モデルとする。つまり、

$$\begin{aligned} X_t &= \alpha \circ X_{t-1} + \delta_t, & X_t &\sim Po(\lambda), & \alpha \circ X_{t-1} | X_{t-1} &\sim B(X_{t-1}, \alpha), \\ Y_t &= \alpha \circ Y_{t-1} + \eta_t, & Y_t &\sim Po(\lambda), & \alpha \circ Y_{t-1} | Y_{t-1} &\sim B(Y_{t-1}, \alpha) \end{aligned}$$

とする。このとき、2つの PINAR(1) モデルの差を取り、 $Z_t = X_t - Y_t$ 、 $\alpha \star Z_{t-1} = \alpha \circ X_{t-1} - \alpha \circ Y_{t-1}$ 、 $\varepsilon_t = \delta_t - \eta_t$ とおくと

$$Z_t = \alpha \star Z_{t-1} + \varepsilon_t \quad \{Z_t\} \sim S(\lambda, \lambda)$$

である。このモデルを SINAR(1) モデルという。周辺分布がスケラム分布であり、 Z_t のとりうる値は整数全体となっていることがわかる。

Z_t が正負の値をとりうるとなると、その自己相関が正の場合のみならず負の場合についても積極的に検討が必要となる。Freeland (2010) では、奇数と偶数で X_t と Y_t を入れ替える方法が提案されている。すなわち、先ほどと同様に

$$X_t = \alpha \circ X_{t-1} + \delta_t, \quad Y_t = \alpha \circ Y_{t-1} + \eta_t, \quad X_t \sim Po(\lambda_1), \quad Y_t \sim Po(\lambda_2)$$

とし、

$$\begin{aligned} Z_t &= \begin{cases} X_t - Y_t & (t = 0, 2, 4, \dots) \\ Y_t - X_t & (t = 1, 3, 5, \dots) \end{cases} \\ \varepsilon_t &= \begin{cases} \delta_t - \eta_t & (t = 0, 2, 4, \dots) \\ \eta_t - \delta_t & (t = 1, 3, 5, \dots) \end{cases} \end{aligned}$$

のように偶数と奇数で X_t と Y_t の差の順序を入れ替えると、 $\rho(k) = (-\alpha)^k$ が成り立ち、1次の自己相関が負の場合を表現することができる。

4.2 SINAR(1) モデルの拡張

さて、Freeland (2010) では X_t, Y_t の2つの PINAR(1) モデルのパラメータが等しい場合のみを論じている。ここでは、 λ が異なるモデルへの拡張を提案する。 X_t, Y_t をそれぞれ独立な PINAR(1) モデルとし、そのポアソン分布の平均は異なるとする、すなわち

$$\begin{aligned} X_t &= \alpha \circ X_{t-1} + \delta_t \quad \{X_t\} \sim Po(\lambda_1), \quad \alpha \circ X_{t-1} | X_{t-1} \sim B(X_{t-1}, \alpha), \\ Y_t &= \alpha \circ Y_{t-1} + \eta_t \quad \{Y_t\} \sim Po(\lambda_2), \quad \alpha \circ Y_{t-1} | Y_{t-1} \sim B(Y_{t-1}, \alpha) \end{aligned}$$

とする。このとき、この2つの PINAR(1) モデルの差を取ったモデル

$$Z_t = \alpha \star Z_{t-1} + \varepsilon_t \quad \{Z_t\} \sim S(\lambda_1, \lambda_2)$$

を考える。ここで、 $Z_t = X_t - Y_t$ 、 $\alpha \star Z_{t-1} = \alpha \circ X_{t-1} - \alpha \circ Y_{t-1}$ 、 $\varepsilon_t = \delta_t - \eta_t$ である。 $\lambda_1 = \lambda_2$ のときは4.1節で述べた SINAR(1) モデルであり、このモデルはそれを拡張したものになっている。以降では、この拡張したモデルのことを改めて SINAR(1) モデルと呼ぶことにする。

この(拡張した)SINAR モデルの性質を述べる. ε_t のモーメント母関数, 期待値, 分散はそれぞれ

$$(4.1) \quad \begin{aligned} M_{\varepsilon_t}(s) &= \exp \left\{ -((1-\alpha)\lambda_1 + (1-\alpha)\lambda_2) + (1-\alpha)\lambda_1 e^s - (1-\alpha)\lambda_2 e^{-s} \right\} \\ E(\varepsilon_t) &= (1-\alpha)(\lambda_1 - \lambda_2) \\ V(\varepsilon_t) &= (1-\alpha)(\lambda_1 + \lambda_2) \end{aligned}$$

であり(付録 A.5 を参照), このモーメント母関数より $\varepsilon_t \sim S((1-\alpha)\lambda_1, (1-\alpha)\lambda_2)$ であることがわかる. そして,

$$M_{\alpha * Z_{t-1}}(s) = \exp \left\{ -(\alpha\lambda_1 + \alpha\lambda_2) + \alpha\lambda_1 e^s + \alpha\lambda_2 e^{-s} \right\}$$

であり, $\alpha * Z_{t-1} \sim S(\alpha\lambda_1, \alpha\lambda_2)$ であることもわかる. 条件付き期待値, 条件付きモーメント母関数はそれぞれ

$$(4.2) \quad \begin{aligned} E(Z_t | Z_{t-1} = z) &= \alpha z + (1-\alpha)(\lambda_1 - \lambda_2) \\ E(\exp \{s Z_t\} | Z_{t-1} = z) &= \frac{M_{\varepsilon_t}(s)}{P(Z_{t-1} = z)} \sum_{y=0}^{\infty} \frac{(\lambda_1(1-\alpha + \alpha e^s))^{z+y}}{(z+y)!} \frac{(\lambda_2(1-\alpha + \alpha e^{-s}))^y}{y!} \end{aligned}$$

である(付録 A.6 を参照).

パラメータ推定は, PINAR(1) モデルと同様にモーメント法などで行うことができ, 推定量

$$\begin{cases} \hat{\alpha} = \hat{\rho}(1) \\ \hat{\lambda}_1 = \frac{1}{2}(S_Z^2 + \bar{z}) \\ \hat{\lambda}_2 = \frac{1}{2}(S_Z^2 - \bar{z}) \end{cases}$$

が得られる. ここで, S_Z^2 は Z の不偏標本分散である. また, $\hat{\lambda}_1, \hat{\lambda}_2$ はそれぞれ, パラメータ λ_1, λ_2 の不偏推定量である.

5. 実例

整数値自己回帰モデルの実データへの適用として, サッカーにおける各試合の得失点差のデータを, 拡張した SINAR(1) モデルにあてはめた例を紹介する. データとしては, 2015 年サッカー J1 リーグ戦における, 鹿島アントラーズの各試合の得失点差を用いる. データは J リーグウェブサイトより取得した. 図 1 は全 34 試合(節)における得失点差の分布である. 標本平均は 0.47, 標本分散は 1.5 であった. また, 図 2 はその推移を表す折れ線グラフである. 正負の値をとりながら, 比較的上昇と下降を交互に繰り返している傾向があり, 負の相関がありそうに見える. この時系列データのコレログラムは図 3 のようになり, 1 次の自己相関係数は -0.16 と, 想定されたように負の値である. すなわち, -0.16 と値こそ大きくはないが相関があり, 各試合での得失点差は独立ではないといえる.

そこで, SINAR(1) モデル

$$Z_t = \alpha * Z_{t-1} + \varepsilon_t$$

によりモデル化することを考える. 1 次の自己相関が負であることから,

$$Z_t = \begin{cases} X_t - Y_t & (t = 0, 2, 4, \dots) \\ Y_t - X_t & (t = 1, 3, 5, \dots) \end{cases}$$

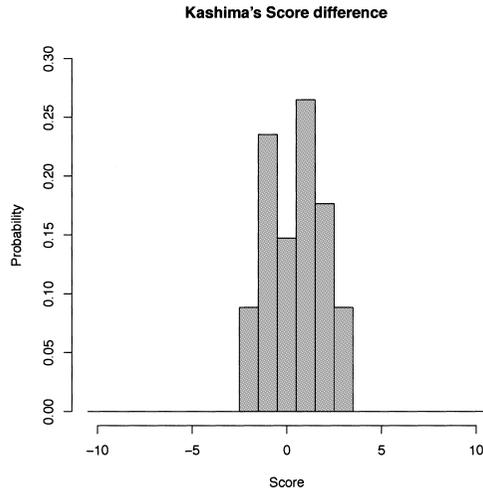


図 1. 2015 年鹿島アントラーズの得失点差のヒストグラム.

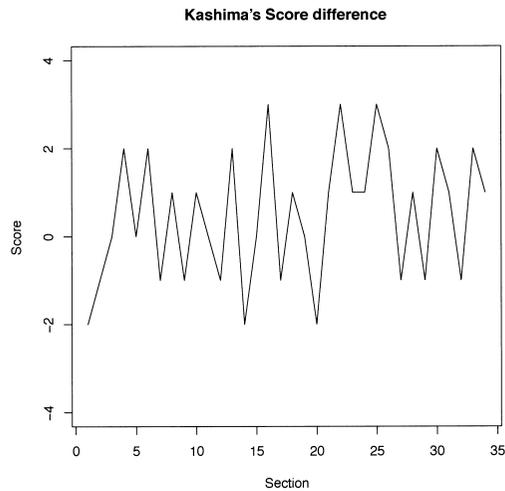


図 2. 2015 年鹿島アントラーズの得失点差のグラフ.

$$\varepsilon_t = \begin{cases} \delta_t - \eta_t & (t = 0, 2, 4, \dots) \\ \eta_t - \delta_t & (t = 1, 3, 5, \dots) \end{cases}$$

のように偶数と奇数で得点数 X_t と失点数 Y_t の差の順序を入れ替える, すなわち得失点差 Z_t の正負を交互に入れ替える. パラメータをモーメント法で推定すると

$$\begin{cases} \hat{\alpha} & = -\hat{\rho}(1) = 0.16 \\ \hat{\lambda}_1 & = 1.33 \\ \hat{\lambda}_2 & = 0.86 \end{cases}$$

が得られる. しかし, この結果自体の解釈は難しい.

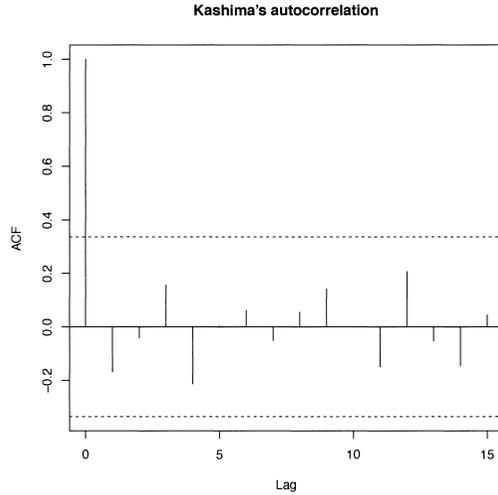


図 3. 2015 年鹿島アントラーズの得失点差のコレログラム.

1 次の自己相関が小さく、このモデルを用いての予測の精度はあまり高くはないが、1 期先予測などを行うことは可能である。例えば、第 34 節(最終節)の得失点差は 1 であり、これ以上の試合はなかったが、もし次の試合があったとした場合の得失点差を予測してみよう。例えば $Z_{35} = 0$ となる確率は、 $Z_{34} = 1$ である条件の下で、

$$\begin{aligned}
 P(Z_{35} = 0|Z_{34} = 1) &= \dots + P(\hat{\alpha} * Z_{34} = -3|Z_{34} = 1)P(\varepsilon_{35} = 3) \\
 &\quad + P(\hat{\alpha} * Z_{34} = -2|Z_{34} = 1)P(\varepsilon_{35} = 2) \\
 &\quad + \dots \\
 &\quad + P(\hat{\alpha} * Z_{34} = 2|Z_{34} = 1)P(\varepsilon_{35} = -2) \\
 &\quad + P(\hat{\alpha} * Z_{34} = 3|Z_{34} = 1)P(\varepsilon_{35} = -3) + \dots
 \end{aligned}$$

のように計算することができる。ここで、 $\hat{\alpha} = -0.16 < 0$ であり、 $t = 35$ が奇数であることから、

$$\begin{aligned}
 P(\alpha * Z_{34} = v|Z_{34} = z_{34}) &= \frac{P(\alpha \circ X_{34} - \alpha \circ Y_{34} = v, Z_{34} = z_{34})}{P(Z_{34} = z_{34})} \\
 &= \frac{1}{P(Z_{34} = z_{34})} \sum_x \sum_w P(\alpha \circ X_{34} = w|X_{34} = x)P(Y_{34} = x - z_{34}) \\
 &\quad \cdot P(X_{34} = x)P(\alpha \circ Y_{34} = w + v|Y_{34} = x - z_{34})
 \end{aligned}$$

により求まる。この結果をまとめたものが表 2 である。表 2 には、得失点差が独立なスケラム分布 $S(1.33, 0.86)$ に従うとした場合の確率も併記している。いずれのモデルを用いた場合においても得失点差が 0 の確率が最も大きく、続いて得失点差 1, 得失点差 -1 の確率が大きいことがわかる。しかし、その確率自体を比較してみると、相関を考慮した SINAR(1) モデルのほうが、全体的に得失点差が負の方向に分布が寄っていることがわかる。実際、得失点差の推移をみると、得失点差が 1 であった場合の次の節においては、ほとんどのケースで得失点差が負(とくに -1)になっており、SINAR(1) モデルではこの結果を反映していると解釈することができる。

表 2. SINAR モデルによる推定.

得失点差	-3	-2	-1	0	1	2	3
確率 (独立スケラム)	0.016	0.060	0.163	0.282	0.252	0.143	0.058
確率 (SINAR(1))	0.021	0.076	0.191	0.295	0.235	0.118	0.043
確率 (DGLM)	0.021	0.065	0.151	0.246	0.238	0.157	0.076

ところで、同じデータに対して DGLM で分析することは可能だろうか。少なくとも、得失点差系列自体に DGLM をあてはめるために、ポアソン過程の生起度(intensity)として単一の系列をどうモデル化して与えるかは自明ではないように思われる。この点に鑑みれば、正負双方を取り得る整数値時系列データに対しては、DGLM のようなパラメータ駆動型のモデルよりは、INAR モデルのような観測値駆動型のモデルのほうが構築しやすさからメリットがあると言えよう。

ここでは、得点系列と失点系列にそれぞれ別々の非定常ポアソンモデルをあてはめた結果とその含意を述べる。まずポアソンパラメータは時変(λ_t)とし、非負制約を課して $\log \lambda_t$ に対して $\log \lambda_t = \log \lambda_{t-1} + v_t$ というランダムウォーク型のシステム方程式を仮定する。システムノイズには正規分布を仮定し、 $v_t \sim N(0, \tau^2)$ とする。観測時系列 Y_t は、ポアソン分布 $P(Y_t | \lambda_t) = e^{-\lambda_t} (\lambda_t)^{Y_t} / Y_t!$ に従うものとする。ここでは状態変数の分布を短冊状に近似して Kitagawa (1987) の数値積分フィルタ・平滑化で、時変生起度 $\lambda_{t|T}$ ($t = 1, \dots, T$) を求める。状態の初期分布は、各系列の標本平均を $\bar{\lambda}$ と書くとする、 $N(\bar{\lambda}, \bar{\lambda}^2)$ で与えた。

システムノイズの分散(τ^2)は、得点系列に関しては 1.12×10^{-4} と、失点系列に関しては 4.01×10^{-3} と推定された。このとき最終の第 34 節でのフィルタ値はそれぞれ 1.657 と 1.067 と得られたので、ランダムウォーク予測で仮想的第 35 節もこの生起度で得失点が発生すると仮定する。10 万回のシミュレーションから得られた得失点差の相対頻度を、表 2 の最終行に示している。傾向としては、独立スケラム分布の結果に近いが、今回の DGLM では得失点差系列に見られる負の相関(高周波成分の強い平均回帰性)がモデルに取り込めていないことを考えれば自然である。2 本の対数生起度系列(潜在時系列)に関する多変量モデリングが現象説明能力の改善に資するかどうかは、今後の研究課題である。

独立であっても、得点と失点に DGLM をあてはめることの効用はある。時変生起度の平滑化分布を描画してみれば、これらの整数値時系列が定常という仮定が妥当であったかどうかのチェックを、少なくとも視覚的に行うことが可能である。図 4 が今回の推定結果である。ポアソンの生起度に非定常な振る舞いを許す枠組みで分析したわけだが、平滑化分布の中央値に多少の凹凸は見られるものの、得点、失点ともに定常ポアソン過程とみなして問題ないであろう。

6. おわりに

本研究では、整数値をとる時系列データに対するモデルとして近年研究が進んでいる PINAR モデル、NBD INAR モデル、BINAR モデル、SINAR モデルについて、その定義や性質、そしてパラメータ推定法を説明した。

さらに本論文では、SINAR(1) モデルに関する拡張を提案し、それに伴いいくつかの性質を示すことができた。これらの結果により、今までよりもさらに複雑な構造をもつ整数値時系列データに対するモデリングの可能性が広がったといえよう。しかしながら、拡張したモデルにおけるパラメータ構造の問題が残されており、これは今後の課題である。また、他のモデルに関する p 次への拡張への可能性、そしてそのパラメータ推定についての研究も残されている。最後に、パラメータ推定法が確立されていないモデルでの推定法の提案も今後の研究課題といえる。

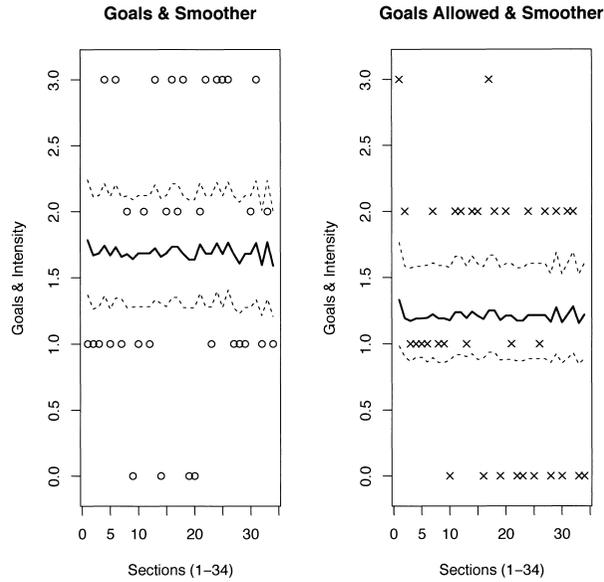


図 4. 得点(左パネル中 o)失点(右パネル中 x)の時系列と, DGLM で推定した時変生起度の平滑化系列. 平滑化系列は太実線が 50% 点, 細破線は下が 3% 点, 上が 97% 点. 横軸は試合の開催節.

謝 辞

改稿につながるコメントを寄せてくださった匿名の査読者にこの場を借りて感謝申し上げます。本研究を進めるにあたり、中嶋は統計数理研究所の特別共同利用研究員制度(平成 28 年度)を利用した。

付録. 証明

A.1 PINAR(p) モデルの誤差分布の導出

まず, $\alpha_i \circ X_{t-i}$ は

$$M_{\alpha_i \circ X_{t-i}}(s) = \exp \{ \lambda \alpha_i (s - 1) \}$$

である。ここで, $\sum_{i=1}^p \alpha_i \circ X_{t-i}$ の確率母関数は、ポアソン分布の再生性より

$$(A.1) \quad G_{\sum_{i=1}^p \alpha_i \circ X_{t-i}}(s) = \exp \left\{ \sum_{i=1}^p \alpha_i \lambda (s - 1) \right\}$$

となる。したがって, ε_t の確率母関数は,

$$G_{\varepsilon_t}(s) = \frac{G_{X_t}(s)}{G_{\sum_{i=1}^p \alpha_i \circ X_{t-i}}(s)} = \exp \left\{ \left(1 - \sum_{i=1}^p \alpha_i \right) \lambda (s - 1) \right\}$$

となる。□

A.2 (2.2), (2.3) の証明

まず, 条件付き期待値は,

$$\begin{aligned} E(X_t | X_{t-1}, X_{t-2}, \dots, X_{t-p}) &= E(\alpha_1 \circ X_{t-1} + \dots + \alpha_p \circ X_{t-p} + \varepsilon_t | X_{t-1}, X_{t-2}, \dots, X_{t-p}) \\ &= E(\alpha_1 \circ X_{t-1} | X_{t-1}) + \dots + E(\alpha_p \circ X_{t-p} | X_{t-p}) + E(\varepsilon_t) \\ &= \sum_{i=1}^p \alpha_i X_{t-i} + \left(1 - \sum_{i=1}^p \alpha_i\right) \lambda \end{aligned}$$

となる. 続いて条件付き分散は

$$\begin{aligned} V(X_t | X_{t-1}, X_{t-2}, \dots, X_{t-p}) &= V(\alpha_1 \circ X_{t-1} + \dots + \alpha_p \circ X_{t-p} + \varepsilon_t | X_{t-1}, X_{t-2}, \dots, X_{t-p}) \\ &= V(\alpha_1 \circ X_{t-1} | X_{t-1}) + \dots + V(\alpha_p \circ X_{t-p} | X_{t-p}) + V(\varepsilon_t) \\ &= \sum_{i=1}^p \alpha_i (1 - \alpha_i) X_{t-i} + \left(1 - \sum_{i=1}^p \alpha_i\right) \lambda \end{aligned}$$

となる. \square

A.3 BINAR_b(1) モデルにおける X_t の 1 次の自己相関 $\rho(1)$ の導出

1 次の自己共分散は

$$\text{Cov}(X_t, X_{t-1}) = \text{Cov}(\alpha \circ X_{t-1} + \beta \circ (n - X_{t-1})) = \rho V(X_{t-1})$$

となるので, 1 次の自己相関は $\rho(1) = \rho$ となることが分かる. \square

A.4 $n/N \diamond X_{t-1} \sim B(n, p)$ の証明

X と $n/N \diamond X$ の同時分布を求め, そのあと $n/N \diamond X$ の周辺分布を求める. まず, X と $n/N \diamond X$ の同時分布は

$$\begin{aligned} P(n/N \diamond X = k, X = x) &= P(n/N \diamond X = k | X = x) P(X = x) \\ &= \frac{\binom{x}{k} \binom{N-x}{n-k}}{\binom{N}{n}} \binom{N}{x} p^x (1-p)^{N-x} \\ &= \binom{n}{k} p^k (1-p)^{n-k} \binom{N-n}{x-k} p^{x-k} (1-p)^{N-n-(x-k)} \end{aligned}$$

である. ここで, x の範囲が $k \leq x \leq N - n + k$ であることに注意を払うと, $n/N \diamond X$ の確率関数は

$$\begin{aligned} P(n/N \diamond X = k) &= \sum_{x=k}^{N-n+k} \binom{n}{k} p^k (1-p)^{n-k} \binom{N-n}{x-k} p^{x-k} (1-p)^{N-n-(x-k)} \\ &= \binom{n}{k} p^k (1-p)^{n-k} \sum_{x'=0}^{N-n} \binom{N-n}{x'} p^{x'} (1-p)^{N-n-x'} \quad (x' = x - k \text{ とおく}) \\ &= \binom{n}{k} p^k (1-p)^{n-k} \end{aligned}$$

となるので, $n/N \diamond X \sim B(n, p)$ となることが分かる. \square

A.5 (4.1)の証明

まず,

$$(A.2) \quad \begin{aligned} M_{\alpha \star Z_{t-1}}(s) &= E(e^{s(\alpha \circ X_{t-1} - \alpha \circ Y_{t-1})}) \\ &= \exp \{ -(\alpha \lambda_1 + \alpha \lambda_2) + \alpha \lambda_1 e^s + \alpha \lambda_2 e^{-s} \} \end{aligned}$$

となる. ここで, (A.2)の結果を用いると, $M_{\varepsilon_t}(s)$ は

$$\begin{aligned} M_{\varepsilon_t}(s) &= \frac{M_{Z_t}(s)}{M_{\alpha \star Z_{t-1}}(s)} \\ &= \exp \{ -((1-\alpha)\lambda_1 + (1-\alpha)\lambda_2) + (1-\alpha)\lambda_1 e^s - (1-\alpha)\lambda_2 e^{-s} \} \end{aligned}$$

となる. これは $S((1-\alpha)\lambda_1, (1-\alpha)\lambda_2)$ のモーメント母関数であることから, $\varepsilon_t \sim S((1-\alpha)\lambda_1, (1-\alpha)\lambda_2)$ である. このことから, 期待値, 分散はそれぞれ,

$$\begin{cases} E(\varepsilon_t) = (1-\alpha)(\lambda_1 - \lambda_2) \\ V(\varepsilon_t) = (1-\alpha)(\lambda_1 + \lambda_2) \end{cases}$$

であることが分かる. □

A.6 (4.2)の証明

まず, 条件付き期待値は

$$\begin{aligned} E(Z_t | Z_{t-1} = z) &= E(E(\alpha \star Z_{t-1} | Z_{t-1} = z)) + E(\varepsilon_t) \\ &= \alpha z + (1-\alpha)(\lambda_1 - \lambda_2) \end{aligned}$$

である. 一方, 条件付きモーメント母関数は

$$\begin{aligned} E(\exp \{sZ_t\} | Z_{t-1} = z) &= E(\exp \{s(\alpha \star Z_{t-1} + \varepsilon_t)\} | Z_{t-1} = z) \\ &= \frac{M_{\varepsilon_t}(s)}{P(Z_{t-1} = z)} \sum_{y=0}^{\infty} P(X_{t-1} = z+y, Y_{t-1} = y) M_{\alpha \circ X_{t-1}}(s) M_{\alpha \circ Y_{t-1}}(-s) \\ &= \frac{M_{\varepsilon_t}(s)}{P(Z_{t-1} = z)} \sum_{y=0}^{\infty} \frac{(\lambda_1(1-\alpha + \alpha e^s))^{z+y}}{(z+y)!} \frac{(\lambda_2(1-\alpha + \alpha e^{-s}))^y}{y!} \end{aligned}$$

となる. これ以上は解析的に解くことができない. □

参 考 文 献

Al-Osh, M. A. and Alzaid, A. A. (1990). An integer-valued p th-order autoregressive structure (INAR(p)) process, *Journal of Applied Probability*, **27**(2), 314–324.

Alzaid, A. A. and Omair, M. A. (2010). On the poisson difference distribution inference and applications, *Bulletin of the Malaysian Mathematical Sciences Society*, **33**(1), 17–45.

Baretto-Souza, W. and Bourguignon, M. (2013). A skew true INAR(1) process with application, arXiv: 1306.0156 [stat.ME].

Du, J. G. and Li, Y. (1991). The integer-valued autoregressive (INAR(p)) model, *Journal of Time Series Analysis*, **12**, 129–142.

Fahrmeir, L. and Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear*

- Models*, 2nd ed., Springer-Verlag, New York.
- Fokianos, K. (2011). Count time series models, *Handbook of Statistics Vol.30* (eds. T. S. Rao, S. S. Rao and C. R. Rao), 315–347, North-Holland, Oxford.
- Freeland, R. K. (2010). True integer value time series, *ASTA Advances in Statistical Analysis*, **94**, 217–229.
- Kitagawa, G. (1987). Non-Gaussian state-space modeling of nonstationary time series (with discussion), *Journal of the American Statistical Association*, **76**, 1032–1064.
- Kitagawa, G. (2010). *Introduction to Time Series Modeling*, Chapman & Hall/CRC, Boca Raton.
- Leonenko, N., Savani, V. and Zhigljasky, A. (2007). Autoregressive negative binomial processes, *Annales de l'Institut de Statistique de l'Université de Paris*, **51**(1), 25–47.
- McKenzie, E. (1985). Some simple models for discrete variate time series, *Journal of the American Water Resources Association*, **21**(4), 645–650.
- Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models, *Journal of the Royal Statistical Society Series A*, **135**, 370–384.
- Rajarshi, M. B. (2012). *Statistical Inference for Discrete Time Stochastic Processes*, Springer, New Delhi.
- Rinstić, M. M., Bakouch, H. S. and Nastić, A. S. (2009). A new geometric first-order integer-valued autoregressive (NGINAR(1)) process, *Journal of Statistical Planning and Inference*, **139**, 2218–2226.
- Savani, V. and Zhigljavsky, A. (2007a). Efficient parameter estimation for independent and INAR(1) negative binomial samples, *Metrika*, **65**(2), 207–225.
- Savani, V. and Zhigljavsky, A. (2007b). Efficient estimation of parameters of the negative binomial distribution, *Communications in Statistics — Theory and Methods*, **35**(5), 1–17.
- Weiß, C. H. (2008). Thinning operations for modeling time series of counts — A survey, *ASTA Advances in Statistical Analysis*, **92**(3), 319–341.
- Weiß, C. H. (2009). A new class of autoregressive models for time series of binomial counts, *Communications in Statistics — Theory and Methods*, **38**(4), 447–460.
- West, M. and Harrison, J. (1997). *Bayesian Forecasting and Dynamic Models*, 2nd ed., Springer-Verlag, New York.

Recent Development of Integer-valued Autoregressive Models

Masahiko Nakajima¹, Fumitake Sakaori² and Yoshinori Kawasaki^{3,4}

¹Graduate School of Science and Engineering, Chuo University

²Faculty of Science and Engineering, Chuo University

³The Institute of Statistical Mathematics

⁴Department of Statistical Science, SOKENDAI

Integer-valued Autoregressive Models (INAR models) express current observations, which are integers that depend on past integer-value observations. INAR models completely differ from Dynamic Generalized Linear Models and General State-Space Models, which employ an unobservable/latent process to model integer-valued time series. In INAR models, the choice of the marginal and/or innovation distribution and the definition of the ‘autoregressive part’ are very important to ensure compatibility (in the distributional sense) of both sides of an autoregressive model. Although publications sporadically appeared from the mid 1980s to the early 1990s, notable contributions were not reported for more than a decade. However in the late 2000s, new results began appearing on a regular basis. In this paper, we introduce recent developments in INAR models beginning with Poisson INAR. For some theorems and propositions without proofs in existing literature, we compiled our notes in the appendix. Additionally, we include new results on an INAR model based on the difference of two Poisson distributions and use real data analysis as an illustration.