Lecture 13. DTMC Applications

DTMC for Tennis Scoring

DTMC for Baseball Scoring

Sim, Min Kyu, Ph.D.
mksim@seoultech.ac.kr

서울과학기술대학교 데이터사이언스학과
데이터사이언스와 비즈니스포텐셜

DTMC for Tennis Scoring

About

RESEARCH QUARTERLY FOR EXERCISE AND SPORT
2020, VOL. 91, NO. 3, 361–372
https://doi.org/10.1080/02701367.2019.1666203

Routledge
Taylor & Francis Group

Check for updates

The Winning Probability of a Game and the Importance of Points in Tennis Matches

Min Kyu Sim [a] and Dong Gu Choi [b]

[a]Seoul National University of Science and Technology; [b]Pohang University of Science and Technology

ABSTRACT

Purpose: This study builds a stochastic model of a discrete-time Markov chain (DTMC) that fits well with a dataset of professional playing records. **Methods**: The point-by-point dataset of Men's single matches played in the Association of Tennis Professionals (ATP) tour from 2011 to 2015 is analyzed. A long-debated assumption on the *iid*-ness in the point winning probability of the server is statistically tested. A DTMC model is then developed to analyze the dataset further. **Results**: The statistical test results indicate that the identicality of point winning probabilities is not a valid assumption. For example, the server's point winning probability from scores 40:0, 30:15, 15:30, and 0:40 are significantly different. On the other hand, the independence is a generally valid assumption except for 40:15 where who won the previous point influences the point winning probability. Game winning probabilities and the importance of each point in winning a game are analyzed using the DTMC model by court surfaces and player groups of the different levels of serve effectiveness. **Conclusion**: Extensive empirical validation concludes unsealed debates over the stochastic models for tennis. The presented results reveal interesting properties in professional tennis matches.

ARTICLE HISTORY
Received 5 September 2018
Accepted 5 September 2019

KEYWORDS
Sports analytics; probability model; big data; ATP tour

Highlights

- Tennis scoring for a regular game (where the score transits $0 \to 15 \to 30 \to 40 \to 45$) is modeled as a DTMC.
- Used ATP database, men's singles point-by-point records from 2011 to 2015, containing 10,912 matches.
- The matches include 28,245 sets, 271,856 games, and 1,672,696 points.
- Is winning a point *iid(independent and identically distributed)*?
- What is the importance of moment for each score?

DTMC - first version

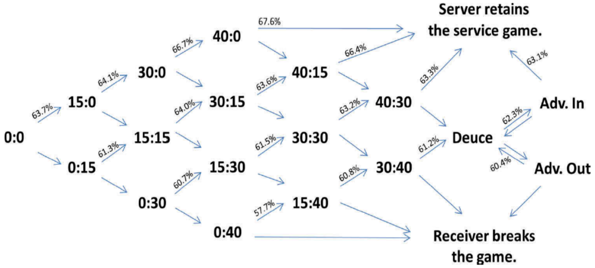

Figure 1. Markov chain diagram for a regular game—The point winning probabilities for servers are marked.

- Does the Markov Property hold?

Identicalty - Checking iid(1)

- The point winning probability is statistically not identical.
- The right side table indicates the point winning probability from each score (state) along with statistical difference.
- For example, the point winning probability for a server is 61.29% at `0:15` and 60.66% at `15:0`, and this difference is 14.55 in z-statistics.
- The long debate on the identicality is resolved.

	All	
0:0	63.73	
0:15	61.29	14.55
15:0	64.09	
0:30	60.66	11.65
15:15	63.96	13.97
30:0	66.71	
0:40	57.75	8.48
15:30	61.49	8.83
30:15	63.56	17.96
40:0	67.59	
15:40	60.79	7.91
30:30	63.24	13.95
40:15	66.37	
30:40	61.24	7.82
40:30	63.34	
Deuce	62.27	
AdvOut	60.43	9.03
AdvIn	63.08	
All	63.54	

Independence - Checking iid(2)

- The Markov Property holds for all cases, except for `40:15`.
- If independence does not hold for all states, it is not a proper DTMC.
- Discussion point
  - What's special about `40:15`?
  - How to resolve this issue?

Table 3. Test of path dependency for server's point winning probabilities in regular games.

Current state	Last point won by	Server's winning prob. in the next point (%)	Number of observations	z-statistics
15:15	Server	63.97	60,429	0.12
	Receiver	63.94	62,222	
30:15	Server	63.63	78,443	0.74
	Receiver	63.4	36,960	
15:30	Server	61.86	23,157	1.46
	Receiver	61.29	44,208	
40:15	Server	66.84	73,346	5.33***
	Receiver	64.96	24,010	
30:30	Server	63.15	41,420	−0.53
	Receiver	63.32	42,057	
15:40	Server	61.38	8,672	1.31
	Receiver	60.59	25,945	
40:30	Server	63.35	52,787	0.06
	Receiver	63.33	32,736	
30:40	Server	61.38	21,043	0.55
	Receiver	61.14	30,690	
Deuce	Server	62.37	58,193	0.76
	Receiver	62.16	58,084	

*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

Discussion - What's special about `40:15`?

- Whether the score at `40:15` progressed from `30:15` or `40:0` makes a significant statistical difference for every surface.
- Our interpretation is as follows:
  (1) `40:0` is the score where a receiver has been dominated by the opponent's serve. From this score, the receiver's scoring a point to arrive at `40:15` implies that the receiver has familiarized himself with his opponent's serve. In other words, the receiver has adapted somewhat and is in better shape.
  (2) On the other hand, `40:15` coming from `30:15` does not have such an advantageous status change for the receiver.

Discussion - How to resolve this issue?

- `40:15` is the score at which the past score affects the future transition
- An appropriate correction for the model is to split the `40:15` state into `40:15|40:0` and `40:15|30:15`, where the previous score is denoted next to a vertical bar, "|".
- Figure 2 presents the accordingly modified diagram.

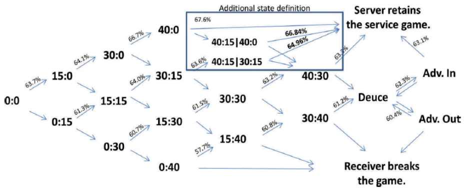

Figure 2. Remedy for the dependence at 40:15 (All courts).

- The score `40:15` is now divided into the two states depending on the previous score.
- This added state ensures the independence assumption of the model.

Limiting probability

- The limiting behavior of the Markov chain reveals the probability of winning a game from each score.
- At the beginning of a regular game, a server is likely to win the game with a probability of 78.9% and a receiver's probability of breaking the game is 21.1%.
- When the game reaches the state of 30:0, the likelihood of the server winning the game is higher than 95%.

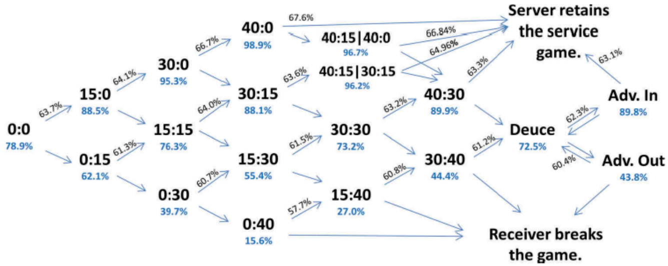

Figure 3. Probability of a server winning for a regular game—The probability is marked below the state.

Importance of a point in winning a game

- It is often said that "All plays count in any sports," but not all points are equally important for winning a match.
- **Definition 1**. The importance of a point in winning a game is defined as the difference between the probabilities of winning a game in the two subsequent states.
  - Importance of a point from state s = P[The prob. for Player A winning a game | Player A wins a point at the state s]- P[The prob. for Player A winning a game | Player A loses a point at the state s]
- **Example**
  - $importance(30:15) = winning\_prob(40:15) - winning\_prob(30:30)$
- The importance of points is relatively increased in the later part of the game and the importance is also higher in a fluctuating game.

Further analysis

- Surface
  - Different court surface leads to different probabilities?
- Serving skills
  - Given the high server's advantage in Men's Singles, what differences are made by players' service ability?

blank

DTMC for Baseball Scoring

About

JOURNAL OF SPORTS SCIENCES
https://doi.org/10.1080/02640414.2024.2355423

Routledge
Taylor & Francis Group

Check for updates

SPORTS PERFORMANCE

A measure of the importance of moment for ball-strike counts in a baseball plate appearance

Dohyun Lee[a], Jeongeon Lee[a], Tonghoon Suk[b] and Min Kyu Sim[a]

[a]Department of Data Science, Seoul National University of Science and Technology, Seoul, South Korea; [b]Department of Operation Research, OTIS Elevator Company, Farmington, CT, USA

ABSTRACT

This study constructs a discrete-time Markov Chain (DTMC) model for a baseball plate appearance (PA) employing Major League Baseball's pitch-by-pitch dataset. Based on the DTMC model, we propose a novel measure for a baseball PA, termed the Importance of Moment (IOM). The IOM quantifies the criticality of each ball-strike count situation, by assessing the probabilistic difference between the pitcher's and hitter's favourable outcomes (out vs reaching base). If the favours significantly vary right after a particular ball-strike count, then the count is deemed critical and is assigned a high IOM value. We empirically verify that IOM explains pitchers' behaviour of fastball speed. We then further investigate whether the behaviour of ace pitchers differs significantly from the majority. Several interesting properties are found from the analysis. Firstly, the path independence assumption generally holds, with the exception of the ball-strike count of 2B1S. Second, pitchers tend to throw the faster fastball at counts with higher IOM values. Lastly, ace pitchers are capable of pitching even faster fastball in two-strike situations in which IOM is high. The DTMC effectively models the probabilistic structure of a baseball PA, and the proposed IOM measure serves as a useful tool for explaining player behaviour.

ARTICLE HISTORY
Received 6 November 2023
Accepted 8 May 2024

KEYWORDS
MLB; baseball; plate appearance; probability model; Markov chain

Highlights

- Discrete-time Markov Chain (DTMC) model leveraging the extensive PA data from the MLB database.
- Propose a metric, termed importance of moment, to quantify the significance of moments during a single baseball PA.
- Proposed metric effectively elucidates variations in pitchers' fastball speeds across different ball-strike counts.

Dataset

- Pitch-level dataset consisting of 2,867,154 pitches with 40 features
- PA level dataset comprising 727,031 PAs with 11 features
- Recorded for MLB matches between 2015 and 2018.

Table 2. Pre-processed dataset.

No.	ab_id	count_type	pitch_type	ball_speeds (mph)
1	2015000001	SSBBO	FaFaFaBrFa	[92.9,92.8,91,75.4,92.9]
2	2015000002	BO	FaFa	[93.3, 89.3]
⋮	⋮	⋮	⋮	⋮
727,031	2018190000	BSSO	FaFaFaFa	[97.7, 97.3, 95.9, 95.8]

*Note on count_type: B (ball), S (strike), O (out), R (reaching base).
*Note on pitch_type: Fa (fastball), Br (breaking ball).

Transition diagram

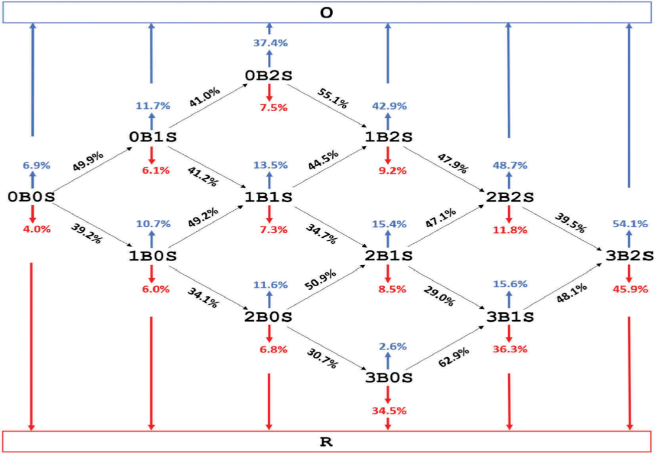

Figure 1. DTMC transition diagram for a PA. There are 12 transient states (notated in a form of xByS) and two absorbing states (O and R). The color-coded arrows represent different types of transitions: blue for an upward arrow indicating a transition to state O, red for a down arrow indicating a transition to state R, and black for either a right-upward or right-downward arrow indicating a transition to another transient state. Note that all arrows leading out of a single state sum up to 100%.

Limiting probability

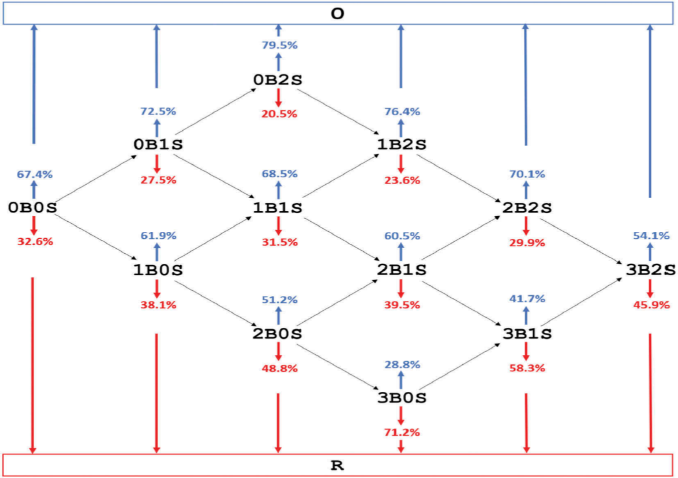

Figure 4. Transition diagram with limiting probability.

Importance of Moment

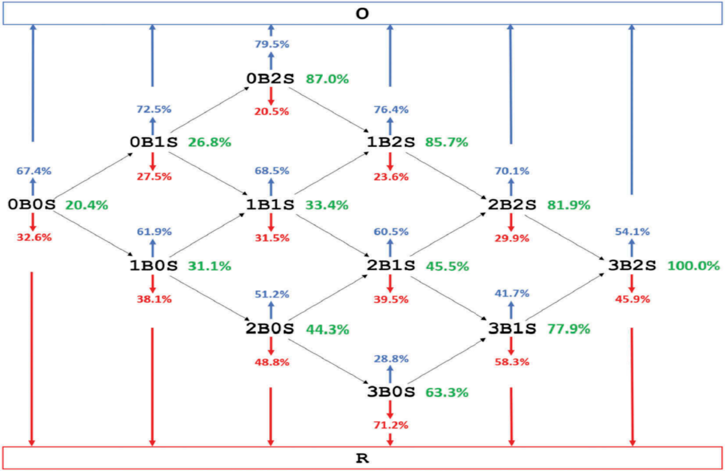

Figure 5. Importance of each count. The green values to the right of each count indicate the corresponding IOM value.

Empirical Analysis on fastball speed

Table 7. Importance and fastball speed for ball-strike counts.

	Importance		Fastball speed		
States	IOM (%)	Rank	Average (mph)	Rank	Number of obs.
3B2S	100.00	1	92.41	1	65,593
0B2S	87.02	2	92.28	3	77,079
1B2S	85.70	3	92.19	4	105,466
2B2S	81.90	4	92.29	2	92,476
3B1S	77.95	5	92.07	5	52,187
3B0S	63.33	6	91.86	8	28,549
2B1S	45.52	7	91.96	6	100,454
2B0S	44.28	8	91.89	7	77,247
1B1S	33.45	9	91.79	11	167,764
1B0S	31.15	10	91.70	12	189,501
0B1S	26.78	11	91.82	9	209,878
0B0S	20.39	12	91.79	10	486,062

Table 8. Fastball speed for ace pitchers.

	Importance		Fastball speed (all pitchers)			Fastball speed (ace pitchers)			
States	IOM (%)	Rank	Average	Rank	Number of obs.	Average	Rank	Number of obs.	Diff. of average speeds
3B2S	100.00	1	92.41	1	65,593	92.71	4	10,077	0.30
0B2S	87.02	2	92.28	3	77,079	92.87	1	13,293	0.60
1B2S	85.70	3	92.19	4	105,466	92.80	2	17,593	0.62
2B2S	81.90	4	92.29	2	92,476	92.75	3	14,976	0.46
3B1S	77.95	5	92.07	5	52,187	92.34	5	7,200	0.28
3B0S	63.33	6	91.86	8	28,549	92.06	11	3,787	0.20
2B1S	45.52	7	91.96	6	100,454	92.26	6	14,993	0.31
2B0S	44.28	8	91.89	7	77,247	92.18	9	10,808	0.30
1B1S	33.45	9	91.79	11	167,764	92.19	8	26,930	0.40
1B0S	31.15	10	91.70	12	189,501	92.01	12	28,012	0.31
0B1S	26.78	11	91.82	9	209,878	92.20	7	36,210	0.38
0B0S	20.39	12	91.79	10	486,062	92.07	10	79,222	0.28
Avg.	-	-	92.00	-	-	92.37	-	-	0.37

```
"Exceptional people, I have found, either start out being optimistic or learn to be
optimistic because they realize that they can't get what they want in life without
being optimistic.
- B. Rotella in How Champions Think: In Sports and in Life"
```