Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Triennale in

Statistica per l'Economia e l'Impresa

Sulla distribuzione del numero di scambi negli incontri di tennis nel circuito professionistico

Relatore: Prof.Francesco Lisi
Dipartimento di Scienze Statistiche

Laureando: Mirko Gabriel Briglia
Matricola n.1222903

Anno Accademico 2021/2022

A Pier Paolo Briglia

Indice

Elenco delle figure

Elenco delle tabelle

Capitolo 1

Introduzione

1.1 Cenni Generali

Il gioco del tennis segue una logica semplice: due giocatori si affrontano, colpendo una palla cercando di mandarla nell'area di campo avversario. Chi fa rimbalzare la palla due volte nel proprio campo, o la manda a rete, perde il punto.

Nella sua struttura è molto più complesso di così: esistono infatti colpi di diritto, di rovescio, le volée, lo smash, il servizio e infine l'ace, il servizio che tocca terra nell'apposita area senza aver contatto né con la rete né con la racchetta dell'avversario.

Il gioco si è evoluto molto, soprattutto negli ultimi anni grazie all'utilizzo di nuovi materiali e nuove tecnologie.

L'evoluzione dei materiali ha portato a un'evoluzione dei giocatori, a cui oggi sono chieste differenti caratteristiche tecniche e fisiche.

In risposta a ciò, le superfici di gioco sono state modificate per salvaguardare le sfumature di uno sport che rischiava di essere sommerso da partite monotematiche, dominate da punti diretti ottenuti con il servizio.

1.2 Tema affrontato

Il tema trattato in questa tesi è quello del numero di scambi *(rally)* durante una partita di tennis. Questo tema è sempre stato citato in vari articoli e blog, come ad esempio (1) (3) (5), ma mai trattato approfonditamente nel dettaglio.

Le informazioni sul numero di scambi vengono spesso racchiuse in variabili categoriali (1-4 scambi, 5-8 scambi, 8-12 scambi e +12 scambi ad esempio) anche dagli enti ufficiali come ATP e WTA, ma esse non danno un quadro completo riguardante l'andamento dei *rally* di una partita.

In questa tesi verranno analizzati i *rally* singolarmente, come anche già fatto da Kovalchik & Ingram (2018) e Carboach (2019), ponendo rilevanza anche ai punti che hanno prodotto *rally* di lunghezza superiore a 12.

Un'ipotesi fondamentale è quella presente nel libro *Analyzing Wimbledon* (9) sull'indipendenza di ogni punto dall'altro; come anche dimostrato da Kovalchik (2016) (3) è possibile asserire che vincere un punto al servizio è un processo *iid*, ovvero "indipendente e identicamente distribuito".

L'obiettivo è quello di ottenere una distribuzione che si avvicini il più possibile a quella reale osservata, tenendo in considerazione anche i casi in cui ci sia un numero di scambi elevato per singolo punto.

Nel circuito professionistico di tennis, noi spettatori assistiamo a scambi sempre di alta intensità, ci esaltiamo quando uno di questi dura molto tempo e magari termina con un colpo spettacolare, ma la domanda è: "Ogni scambio è davvero così?".

Una partita di tennis ha una durata variabile, che mediamente va dalle due alle tre ore, quindi è lecito pensare, che al suo interno, venga assegnata una discreta quantità di punti.

Ogni punto può invece variare di durata: è possibile sia terminare un punto con un ace (vincente al servizio) oppure addirittura anche attendere 28 scambi. Analizzando nel dettaglio una partita professionistica, si nota che la maggior parte dei punti termina dopo pochi scambi.

Studiare la "durata del punto" può essere utile per eseguire un'analisi sullo stile di gioco più ideale, ovvero portare il giocatore a prediligere uno stile di gioco più aggressivo, finalizzato a terminare il punto il prima possibile, oppure uno stile di gioco che miri a far durare lo scambio più colpi possibili.

Competere nel tennis professionistico, richiede al giocatore di vincere le partite non solo con la tecnica, ma creando efficaci strategie e tattiche in base ai diversi tornei, round, set e alle diverse qualità degli avversari. Con l'applicazione delle tecniche di profilazione delle performance, i giocatori possono essere valutati e comparati. (Fitzpatrick A., 2021)

Lo stile di gioco è anche correlato significativamente con le caratteristiche fisiche del giocatore (Kovalchik & Ingram, 2018), come il peso e l'altezza.

Le diverse superfici di gioco si tramutano in differenti velocità di gioco. Ogni superficie è caratterizzata dai suoi coefficienti di frizione e restituzione, che influenzano l'interazione fra la superficie e la palla al suo rimbalzo.

In una superficie con un basso coefficiente di frizione, come ad esempio l'erba, la palla perde meno velocità orizzontale rispetto a una superficie con alto coefficiente di frizione, mentre in una superfice con un basso coefficiente di restituzione, la palla rimbalza più bassa di una superficie con un alto coefficiente di restituzione. (Yixiong C., 2019).

La lunghezza media dei *rally* in erba quindi tenderebbe a essere più bassa rispetto alle altre superfici, comportando di conseguenza una maggioranza di ace e servizi vincenti. Questo grazie a un rimbalzo più veloce e più basso della palla.

Prima di proseguire con una prima analisi descrittiva dei dati raccolti, è necessario indicare come sono stati considerati, per questa ricerca, i *rally* all'interno di un singolo punto, prendendo ad esempio, per semplicità, 2 giocatori che chiameremo "A" e "B".

- Con 0 scambi includiamo tutti i punti dove il giocatore A, al servizio, fa doppio fallo, ovvero non riesce a servire nel rettangolo avversario diagonale rispetto a dove batte;
- Con 1 scambio includiamo tutti i punti dove il giocatore A al servizio fa ace, un servizio vincente oppure il giocatore B riesce a rispondere, ma commette un errore e il punto finisce;
- Con 2 scambi includiamo tutti i punti dove, al servizio del giocatore A, B riesce a rispondere nel campo di A, il quale commette un errore e termina il punto.

Per tutti gli altri valori dei *rally*, si procede con il sistema sopra indicato.

Capitolo 2

Processo di raccolta dati

2.1 Match Charting Project e dataset ottenuto

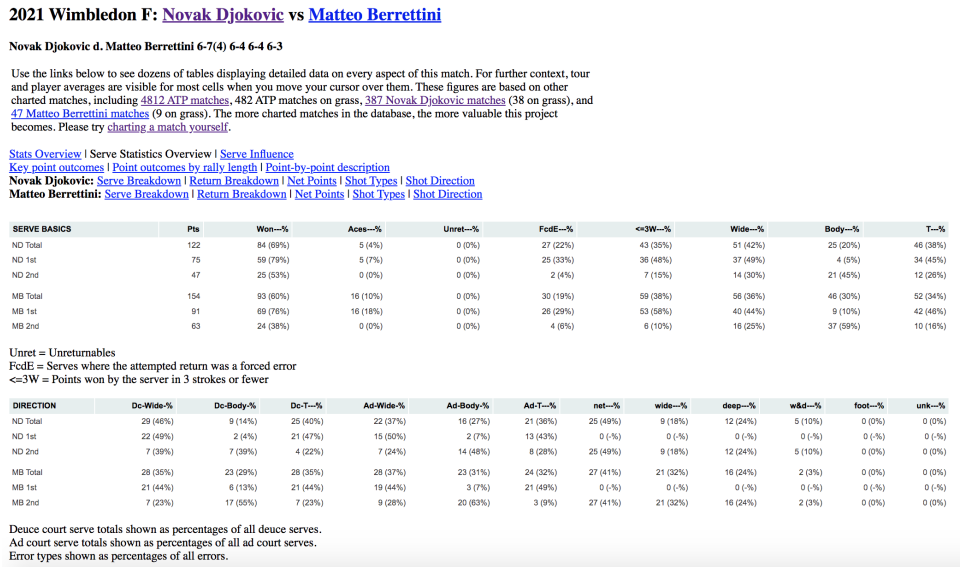
Il Match Charting Project è il maggiore sforzo indipendente di *crowdsourcing* per raccogliere dati, punto per punto, delle partite di tennis professionistico.

Creato da Jeff Sackmann (5), è un sistema di codifica di informazioni che chiunque, con un po' di esercizio, può imparare a utilizzare per registrare ogni colpo di una partita: la direzione del servizio, la direzione e profondità della risposta, la direzione dei colpi di scambio, la tipologia e la direzione degli errori, e altro ancora. Una singola partita genera un'incredibile quantità d'informazioni.

Il vero potenziale del progetto, risiede nella ricchezza di spunti di ricerca e analisi che un'aggregazione di dati di queste dimensioni è in grado di offrire. Ogni partita aggiunta al database, che si accresce giornalmente, contribuisce a una maggiore consapevolezza del tennis professionistico nella sua interezza.

Organizzazioni come gli ATP, WTA, ITF e i tornei del Grande Slam, registrano qualche dato, ma in modalità differente e raramente lo rendono pubblico.

Ogni match salvato, è stato suddiviso in varie sezioni contenenti statistiche differenti, come quelle sul servizio, sull'influenza del servizio, sui tiri e quella dalla quale abbiamo raccolto i dati per questa tesi, il "*point-by-point description*" ovvero la descrizione di ciò che accade punto per punto.



2021 Wimbledon F: Novak Djokovic vs Matteo Berrettini

Novak Djokovic d. Matteo Berrettini 6-7(4) 6-4 6-4 6-3

Use the links below to see dozens of tables displaying detailed data on every aspect of this match. For further context, tour and player averages are visible for most cells when you move your cursor over them. These figures are based on other charted matches, including 4812 ATP matches, 482 ATP matches on grass, 387 Novak Djokovic matches (38 on grass), and 47 Matteo Berrettini matches (9 on grass). The more charted matches in the database, the more valuable this project becomes. Please try charting a match yourself.

Stats Overview | Serve Statistics Overview | Serve Influence
Key point outcomes | Point outcomes by rally length | Point-by-point description
Novak Djokovic: Serve Breakdown | Return Breakdown | Net Points | Shot Types | Shot Direction
Matteo Berrettini: Serve Breakdown | Return Breakdown | Net Points | Shot Types | Shot Direction

SERVE BASICS	Pts	Won--%	Aces--%	Unret--%	FcdE--%	<=3W--%	Wide--%	Body--%	T--%
ND Total	122	84 (69%)	5 (4%)	0 (0%)	27 (22%)	43 (35%)	51 (42%)	25 (20%)	46 (38%)
ND 1st	75	59 (79%)	5 (7%)	0 (0%)	25 (33%)	36 (48%)	37 (49%)	4 (5%)	34 (45%)
ND 2nd	47	25 (53%)	0 (0%)	0 (0%)	2 (4%)	7 (15%)	14 (30%)	21 (45%)	12 (26%)
MB Total	154	93 (60%)	16 (10%)	0 (0%)	30 (19%)	59 (38%)	56 (36%)	46 (30%)	52 (34%)
MB 1st	91	69 (76%)	16 (18%)	0 (0%)	26 (29%)	53 (58%)	40 (44%)	9 (10%)	42 (46%)
MB 2nd	63	24 (38%)	0 (0%)	0 (0%)	4 (6%)	6 (10%)	16 (25%)	37 (59%)	10 (16%)

Unret = Unreturnables
FcdE = Serves where the attempted return was a forced error
<=3W = Points won by the server in 3 strokes or fewer

DIRECTION	Dc-Wide-%	Dc-Body-%	Dc-T--%	Ad-Wide-%	Ad-Body-%	Ad-T--%	net--%	wide--%	deep--%	w&d--%	foot--%	unk--%
ND Total	29 (46%)	9 (14%)	25 (40%)	22 (37%)	16 (27%)	21 (36%)	25 (49%)	9 (18%)	12 (24%)	5 (10%)	0 (0%)	0 (0%)
ND 1st	22 (49%)	2 (4%)	21 (47%)	15 (50%)	2 (7%)	13 (43%)	0 (-%)	0 (-%)	0 (-%)	0 (-%)	0 (-%)	0 (-%)
ND 2nd	7 (39%)	7 (39%)	4 (22%)	7 (24%)	14 (48%)	8 (28%)	25 (49%)	9 (18%)	12 (24%)	5 (10%)	0 (0%)	0 (0%)
MB Total	28 (35%)	23 (29%)	28 (35%)	28 (37%)	23 (31%)	24 (32%)	27 (41%)	21 (32%)	16 (24%)	2 (3%)	0 (0%)	0 (0%)
MB 1st	21 (44%)	6 (13%)	21 (44%)	19 (44%)	3 (7%)	21 (49%)	0 (-%)	0 (-%)	0 (-%)	0 (-%)	0 (-%)	0 (-%)
MB 2nd	7 (23%)	17 (55%)	7 (23%)	9 (28%)	20 (63%)	3 (9%)	27 (41%)	21 (32%)	16 (24%)	2 (3%)	0 (0%)	0 (0%)

Deuce court serve totals shown as percentages of all deuce serves.
Ad court serve totals shown as percentages of all ad court serves.
Error types shown as percentages of all errors.

Attraverso funzioni di *web scraping* create appositamente per l'occasione, è stato possibile estrapolare il codice HTML del sito web dal quale, poi, ricavare il numero di scambi punto per punto.

Ciò che si ottiene sono le distribuzioni di frequenza relative e assolute per ogni *rally* per le varie categorie (superficie e sesso).

Il MCP, essendo un progetto *open source*, non contiene tutti i match effettivamente disputati nell'arco temporale tra il 2000 e oggi. Le analisi che sono state fatte precedentemente su questa tematica, avevano preso in esame solo 12 match maschili (Carboch, 2019) o 2448 match fra i due sessi (Kovalchik & Ingram, 2018).

Attraverso lo scraping dei dati dal MCP, abbiamo ottenuto un set di valori discretamente ampio, contente 5751 match maschili e 3413 match femminili, entrambi a partire dagli anni 2000 a oggi.

Per i match maschili sono stati analizzati 503946 punti, mentre per i match femminili 247392.

2.2 Analisi descrittive dei dati

2.2.1 Uomini

Superficie	Match Analizzati	Punti Analizzati
Erba	710	70333
Cemento	3332	290340
Terra	1709	142253
Totale	5751	503946

Tabella 2.1: Dimensione del campione dei dati per gli uomini diviso per superficie

Superficie	Mean	Mode	Median	$q_{2,5}$	q_5	q_{10}	q_{25}	q_{75}	q_{90}	q_{95}	$q_{97,5}$
Erba	3,2	1	2	0	1	1	1	4	7	10	12
Cemento	3,8	1	2	0	1	1	1	5	9	12	15
Terra	4,3	1	3	0	1	1	1	6	10	13	15

Tabella 2.2: Indicatori statistici maschili univariati divisi per superficie

Dalla tabella 2.2 è possibile notare come la moda sia 1: questo indica che, in tutte le superfici, il *rally* di lunghezza 1 è quello più frequente. Inoltre si nota che 3 quantili, q_5, q_{10}, q_{25} sono pari a 1, il che ci porta a dire che almeno il 25% dei *rally* sicuramente si è concluso entro il primo scambio.

Dagli altri quantili è possibile evidenziare le sostanziali differenze fra le tre superfici: sulla terra si tendono a produrre scambi più lunghi, mentre sull'erba, terreno veloce con basso coefficiente di frizione e alto coefficiente di restituzione (Fitzpatrick A., 2021), il 90% dei punti termina entro il settimo scambio.

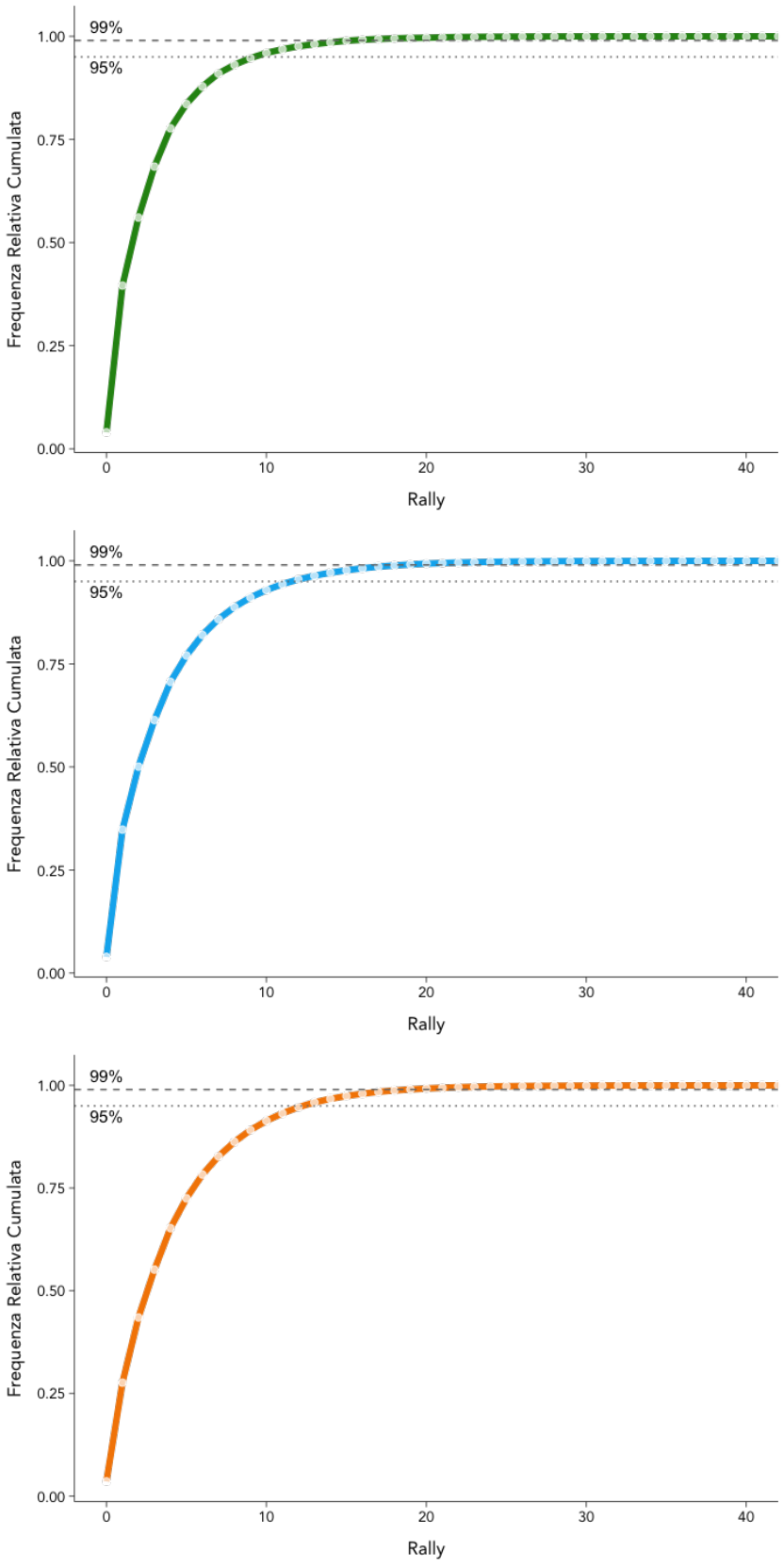


Figura 2.1: Funzioni di ripartizione empiriche

Non è un caso infatti che, come mostrato nella tabella A.1, le partite sulla terra abbiano prodotto delle osservazioni di *rally* più alte di tutte: si ha un punto finito dopo 71 o addirittura dopo 83 scambi. Nell'erba il massimo numero di *rally* registrato è stato invece 48.

L'analisi della tabella nell'appendice A.3, evidenzia come la frequenza relativa del *rally* pari a 0 (numero di doppi falli) sia pressoché costante, informazione che è alla base delle analisi che verranno eseguite successivamente.

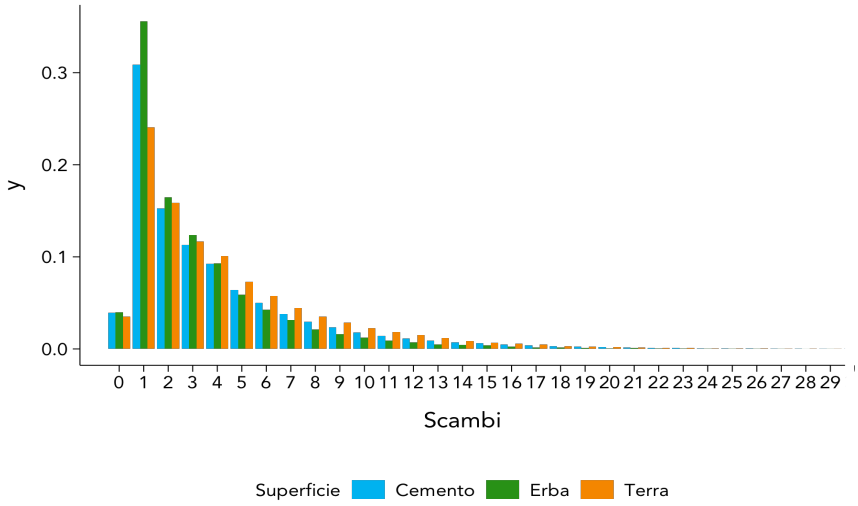


Figura 2.2: Distribuzione aggregata degli scambi maschili nelle tre superfici

Nell'analisi eseguita da Carboch (2019), si sono esaminati i *rally* fino a 13, includendo nella categoria *"13+"* tutti i *rally* di lunghezza superiore. In questa tesi vengono considerati tutti i *rally* singolarmente, senza escludere quelli con lunghezza elevata.

2.2.2 Donne

Superficie	Match Analizzati	Punti Analizzati
Erba	912	65705
Cemento	1833	130235
Terra	686	51449
Totale	3431	247389

Tabella 2.3: Dimensione del campione dei dati per le donne diviso per superficie

Come era lecito aspettarsi, le donne ottengono punti con scambi più lunghi degli uomini e, sempre considerando la differenza fra le superfici, l'erba risulta essere ancora il terreno dove vengono giocati gli scambi più corti.

Superficie	Mean	Mode	Median	$q_{2,5}$	q_5	q_{10}	q_{25}	q_{75}	q_{90}	q_{95}	$q_{97,5}$
Erba	3,5	1	3	0	1	1	1	5	8	10	12
Cemento	3,9	1	3	0	0	1	1	5	9	11	13
Terra	4,2	1	3	0	1	1	1	6	9	12	14

Tabella 2.4: Indicatori statistici femminili univariati divisi per superficie

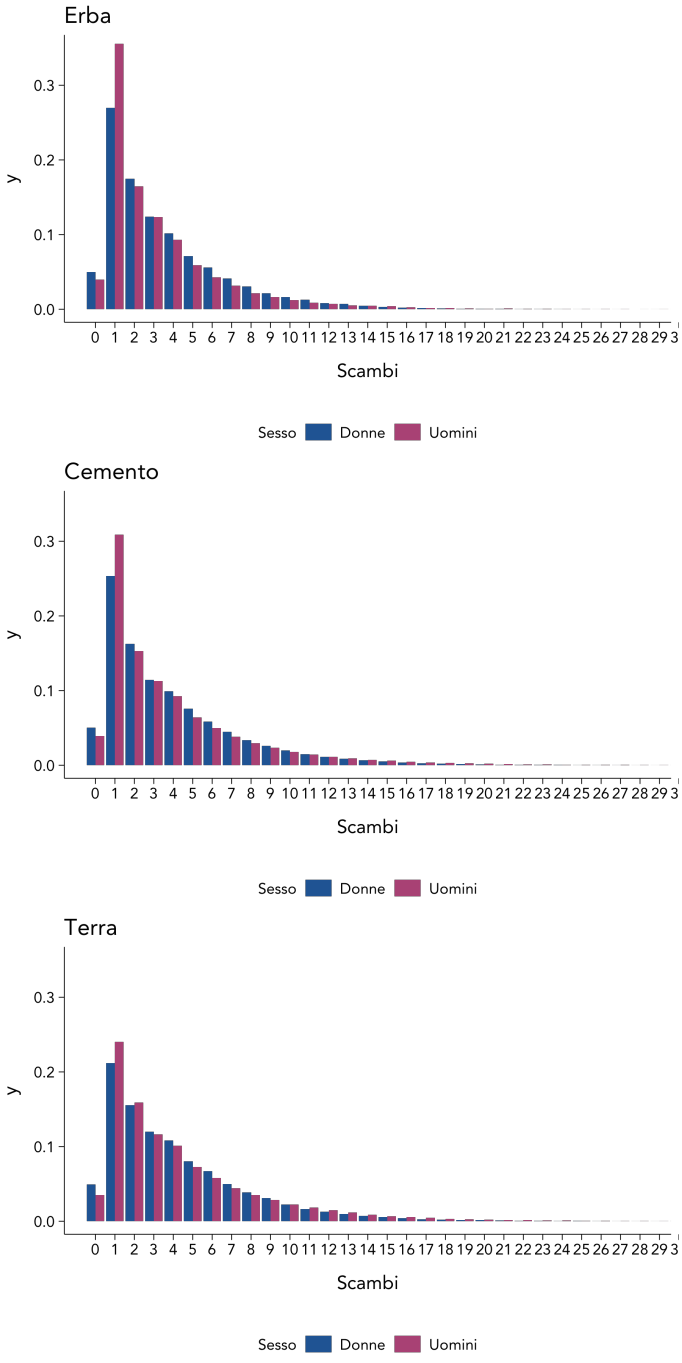


Figura 2.3: Differenze fra maschi e femmine nella distribuzione degli scambi, divise per superficie

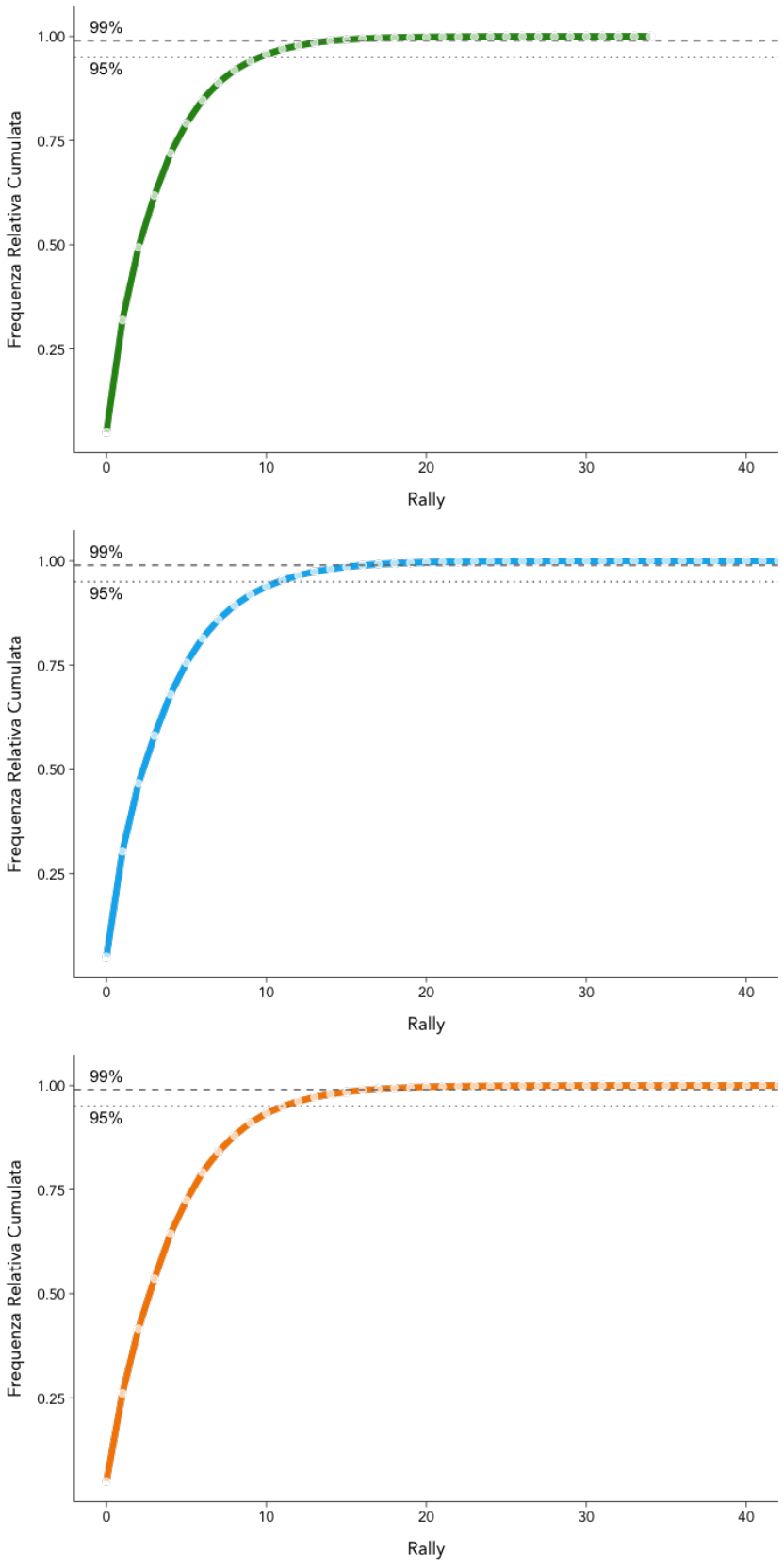


Figura 2.4: Funzioni di ripartizione empiriche

Capitolo 3

Descrizione dei modelli

3.1 Modelli preesistenti

L'obiettivo di questa ricerca è quello di costruire un modello probabilistico che abbia il compito di descrivere la frequenza dei *rally* negli incontri di tennis professionistici.

Kovalchik & Ingram (2018) avevano proposto un modello di $QuasiPoisson$ a questo scopo, nell'ambito della stima della durata dei incontri di tennis professionistici.

In questo articolo per il *Journal of Quantitative Analysis in Sports*, i ricercatori hanno creato un modello per la durata di un match di tennis prendendo in esame l'influenza dei fattori determinanti per la durata di un punto.

A tale scopo hanno utilizzato il metodo di simulazione Monte Carlo, dove hanno inserito in input due parametri usati anche in questa analisi, ovvero il bonus al servizio δ (somma delle probabilità che hanno i due giocatori di fare punto quando sono al servizio, che noi chiameremo ν) e il malus al servizio Δ (differenza delle probabilità che hanno i due giocatori di fare punto al servizio che noi chiameremo ω).

Kovalchik fa una divisione fondamentale fra tempo "in gioco" e tempo "fuori dal gioco".

Usando la terminologia di Klaassen e Magnus (2014), come probabilità vengono usate le proporzioni osservate dai dati.

Per stimare la durata del tempo del gioco, prima vengono considerati il numero di colpi giocati per punto $\mathcal{S}$, poi il tempo atteso per giocare uno scambio, condizionatamente al numero di colpi effettuati, e infine la somma di quest'ultimo fornisce la stima della durata del match.

Il tempo "fuori dal gioco" è l'insieme del tempo impiegato da un giocatore nella preparazione al servizio, della durata dei cambio-campo e dei cambio-set.

Come accennato in precedenza, la Kovalchik utilizza un modello di $Quasi - Poisson$ definito per gli uomini con $\lambda = 2.89 - 1\delta, \Phi = 3.3$ e per le donne con $\lambda = 2.33 - 0.7\delta, \Phi = 2.7$.

Una considerazione da fare su questo modello, che prende comunque i dati dal MCP, è che esclude completamente gli scambi che noi consideriamo aventi lunghezza 0, includendoli negli scambi aventi lunghezza 1.

3.2 Modelli non condizionati

Un modello non condizionato è un modello probabilistico dove i parametri della distribuzione non sono condizionati dai dati osservati.

Dall'analisi empirica dei dati, è possibile proporre alcuni modelli utili per descrivere al meglio le informazioni:

- *Poisson Distribution*
- *QuasiPoisson Distribution*
- *Zero-Modified Poisson Distribution*
- *Zero-One-Modified Poisson Distribution*
- *Zero-One-Modified Geometric Distribution*

Il primo modello viene scartato senza ulteriori analisi, perchè cade l'ipotesi sulla quale si basa la distribuzione di Poisson:

$$E[X] = Var[X] \tag{3.1}$$

Infatti se prendiamo ad esempio i valori di media e varianza per i nostri dati sugli incontri degli uomini sul cemento, si ottiene una $\mu = 3.8$ e una varianza di $\sigma = 14.561$.

Per ovviare a questa problematica, si prende come riferimento un modello di *QuasiPoisson*, dove viene inserito un parametro aggiuntivo, Φ, tale che

$$Var(X) = \Phi\mu \tag{3.2}$$

misura la sovra/sottodispersione dei dati (Ver Hoef, J.M., Boveng, P.L., 2007).

Le altre fanno parte della famiglia delle distribuzioni *"modificate"*, dove si calcola separatamente una parte del supporto della variabile, o perchè assente nel fenomeno (-*Truncated*) o perchè presente in eccesso (-*Inflated*).

Nei modelli di probabilità discreti, spesso il valore del supporto della variabile eliminato è lo 0, da qui il nome *Zero-Truncated*, *Zero-Inflated* (Hussain, T., 2020).

Per ovviare all'esclusione in questi modelli della probabilità in 0, viene alterato il suo valore, e si parla di *Zero-Modified Distribution*.

La funzione di densità di una *Poisson* è definita come

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} \quad k \geq 0 \tag{3.3}$$

quindi è possibile derivarne la sua versione *Zero-Modified*

$$P(X = k) = \frac{\lambda^k}{k!}\frac{1}{(e^{-\lambda} - 1)} \quad k \geq 1 \tag{3.4}$$

Lo stesso procedimento si applica anche alla *Geometric Distribution* e a valori del supporto diversi da 0, come 1 ad esempio:

$$P(X = k) \begin{cases} p(1-p)^k & k = 0 \\ p(1-p)^{k-1} & k = 1 \\ p(1-p)^{k-2} & x \geq 2 \end{cases} \tag{3.5}$$

3.3 Applicazione dei modelli non condizionati ai dati reali

Un altro modello che verrà preso in considerazione, oltre a quelli precedentemente elencati, è un modello generato da un campione di numeri che vanno da 0 a 29 (viene preso 29 come riferimento perchè, oltre questo valore, i *rally* perdono di significatività statistica).

L'obiettivo di questi modelli è quello di descrivere in modo più veritiero possbile, la reale distribuzione di frequenza dei *rally*, perciò, per ognuna delle distribuzioni citate sopra, verrà eseguito uno studio di simulazione con le singole distribuzioni di probabilità.

La variabile che si vuole analizzare in questa ricerca è il *rally* R che ha un supporto discreto variabile a seconda della superficie o sesso analizzato.

Superficie	Sesso	S_R
Erba	U	$\{0, 1, \ldots, 47, 48\}$
Cemento	U	$\{0, 1, \ldots, 59, 60\}$
Terra	U	$\{0, 1, \ldots, 82, 83\}$
Erba	D	$\{0, 1, \ldots, 33, 34\}$
Cemento	D	$\{0, 1, \ldots, 47, 48\}$
Terra	D	$\{0, 1, \ldots, 47, 48\}$

Si inizia ad esaminare il modello di *QuasiPoisson*, con il λ e il Φ calcolati con l'applicazione di un GLM (*Generalized Linear Model*) ai *rally* osservati sulle varie superfici e nei due sessi.

Superficie	Sesso	λ	Φ
Erba	U	3.1705	3.0281
Cemento	U	3.8299	3.8018
Terra	U	4.3353	3.8186
Erba	D	3.5101	2.6898
Cemento	D	3.8725	3.1639
Terra	D	4.1976	3.2043

Successivamente si passa ad esaminare il modello *Zero-One-Modified Poisson Distribution*, dove si sono stimati i parametri p_0 e p_1 con le frequenze relative osservate rispettivamente in $R = 0$ e $R = 1$ e il parametro λ tramite la minimizzazione della differenza fra le frequenze relative osservate e quelle che il modello simulerà:

$$\lambda = \min \sum_{i=0}^{\max S_R} |p_i^{oss} - p_i^{sim}| \tag{3.6}$$

dove p_i^{sim} è la frequenza relativa della simulazione generata da un campione avente vettore di probabilità pari alla distribuzione della *Zero-One-Modified Poisson Distribution*.

Analogamente a quanto visto per la *Zero-One-Modified Poisson Distribution*, si procede con l'esaminare il modello generato dalla *Zero-One-Modified Geometric Distribution*, con la differenza che al posto di stimare λ, si stima p.

Superficie	Sesso	p_0	p_1	λ
Erba	U	0.0399	0.3558	3.37
Cemento	U	0.0392	0.3088	3.48
Terra	U	0.0353	0.2405	3.43
Erba	D	0.0498	0.2695	3.41
Cemento	D	0.0051	0.2534	3.62
Terra	D	0.0494	0.2116	3.64

Tabella 3.1: Tabella dei parametri per la *Zero-One-Modified Poisson Distribution* con λ costante

Superficie	Sesso	p_0	p_1	p
Erba	U	0.0399	0.3558	0.285
Cemento	U	0.0392	0.3088	0.236
Terra	U	0.0353	0.2405	0.236
Erba	D	0.0498	0.2695	0.272
Cemento	D	0.0051	0.2534	0.236
Terra	D	0.0494	0.2116	0.203

Tabella 3.2: Tabella dei parametri per la *Zero-One-Modified Geometric Distribution* con p costante

3.4 Confronto fra i modelli con indicatori di ottimalità

L'obiettivo è quello di confrontare i dati ottenuti dalle varie simulazioni con i nostri dati reali, verificando così l'ottimalità del modello.

Sono stati creati degli indici che si basano sugli scarti fra le frequenze osservate relative e le frequenze simulate relative. Per semplicità, li chiameremo Δ divisi in 3 casistiche differenti:

- Δ_{10} : Scarti fra le frequenze relative fino a 10 *rally*
- Δ_{20} : Scarti fra le frequenze relative fino a 20 *rally*
- Δ : Scarti fra le frequenze relative considerando tutti i *rally*

Modello	Δ_{10}	Δ_{20}	Δ
QuasiPoisson	0.36656	0.37312	0.37529
Zero-One-Modified Poisson	0.33631	0.37269	0.37528
Zero-One-Modified Geometric	0.02538	0.03370	0.03583
Simulato dal campione probabilistico	0.00709	0.00812	0.00846

Tabella 3.3: Tabella di confronto di ottimalità fra i modelli non condizionati per la superficie *erba* negli incontri maschili

Modello	Δ_{10}	Δ_{20}	Δ
QuasiPoisson	0.32085	0.32832	0.33121
Zero-One-Modified Poisson	0.35912	0.42225	0.42835
Zero-One-Modified Geometric	0.01919	0.02969	0.03222
Simulato dal campione probabilistico	0.00420	0.00500	0.00522

Tabella 3.4: Tabella di confronto di ottimalità fra i modelli non condizionati per la superficie *cemento* negli incontri maschili

Modello	Δ_{10}	Δ_{20}	Δ
QuasiPoisson	0.24779	0.32832	0.37529
Zero-One-Modified Poisson	0.35753	0.43548	0.44220
Zero-One-Modified Geometric	0.03894	0.05870	0.06163
Simulato dal campione probabilistico	0.00566	0.00703	0.00737

Tabella 3.5: Tabella di confronto di ottimalità fra i modelli non condizionati per la superficie *terra* negli incontri maschili

Modello	Δ_{10}	Δ_{20}	Δ
QuasiPoisson	0.23681	0.24566	0.24642
Zero-One-Modified Poisson	0.31871	0.35966	0.36101
Zero-One-Modified Geometric	0.03605	0.04096	0.04148
Simulato dal campione probabilistico	0.00745	0.00854	0.00879

Tabella 3.6: Tabella di confronto di ottimalità fra i modelli non condizionati per la superficie *erba* negli incontri femminili

Modello	Δ_{10}	Δ_{20}	Δ
QuasiPoisson	0.23197	0.23782	0.23893
Zero-One-Modified Poisson	0.34038	0.39727	0.40013
Zero-One-Modified Geometric	0.02907	0.03104	0.03204
Simulato dal campione probabilistico	0.00520	0.00649	0.00669

Tabella 3.7: Tabella di confronto di ottimalità fra i modelli non condizionati per la superficie *cemento* negli incontri femminili

Cio che questi indicatori evidenziano è che, come ci si poteva aspettare, il modello generato dal campione avente probabilità pari alle frequenze relative osservate, è il migliore per descrivere i dati presentati.

Però, in fase di previsione di un possibile *outcome* di uno scambio, non è presente questa informazione, quindi ci si può basare sul modello della *Zero-One-Modified Geometric*.

Un'altra informazione utile da evidenziare è che i due modelli legati alla distribuzione di *Poisson*, abbiano numeri pressoché simili.

Modello	Δ_{10}	Δ_{20}	Δ
QuasiPoisson	0.16949	0.17508	0.17566
Zero-One-Modified Poisson	0.32224	0.38430	0.38735
Zero-One-Modified Geometric	0.03341	0.05634	0.06162
Simulato dal campione probabilistico	0.00767	0.00921	0.00959

Tabella 3.8: Tabella di confronto di ottimalità fra i modelli non condizionati per la superficie *terra* negli incontri femminili

3.5 Modelli condizionati

Adesso l'obiettivo è quello di far dipendere i parametri dei nostri modelli da delle variabili, che rappresentano alcuni fattori determinanti per la durata del *rally* negli incontri professionistici:

- **Bonus** ν, ovvero la somma delle probabilità che hanno i due giocatori di vincere il punto quando si trovano al servizio $(p_A + p_B)$ (Newton, P. K., 2005),
- **Malus** ω, ovvero la differenza delle probabilità che hanno i due giocatori di vincere il punto quando si trovano al servizio $(p_A - p_B)$,
- **Differenza di altezza** h^-, ovvero la differenza assoluta fra le stature dei due giocatori in cm,
- **Somma di altezza** h^+, ovvero la somma fra le stature dei due giocatori in cm.

La prima variabile esplicativa dà un'indicazione sulla bravura dei due giocatori ad aggiudicarsi il punto quando hanno loro il servizio, il che ci porta ad affermare che quando essi battono il servizio si mettono in condizioni tali da essere in vantaggio sulla chiusura del punto, ergo con valori alti di ν si avranno *rally* più brevi.

Discorso simile vale per la seconda variabile esplicativa: se c'è molta differenza di livello allora ci sarà uno dei due giocatori che tenderà a fare molti punti quando è al servizio e conseguentemente si avranno *rally* più brevi.

Si è deciso di considerare la differenza di altezze dei giocatore perchè, come anche dimostrato da Kovalchik (2020), essere più alti porta a una percentuale di ace superiore, che comporta *rally* più brevi.

La somma delle altezze è indirettamente legata a quest'ultimo concetto: in una partita giocata fra giocatori alti, quindi con una h^+ elevata, è lecito aspettarsi *rally* più brevi. Inoltre considerando la somma delle altezze, si vanno a distinguere i casi in cui h^- pari a 0, sia dovuta ad esempio a 2 giocatori alti entrambi 190cm oppure 170cm.

Il modello proposto da Kovalchik & Ingram (2018), come spiegato nella sezione 3.1, fa parte di un modello più grande che mira a spiegare la durata del tempo di gioco nelle partite di tennis professionistico.

Ciò che è utile allo scopo di questa tesi, è il modello che propongono basato sulla *QuasiPoisson* e condizionato al parametro ν, da noi definito.

3.5.1 *Zero-One-Modified Geometric Distribution* con parametro p variabile

Come si è visto dai dati precedentemente analizzati, il numero di doppi falli ($R = 0$) rimane pressoché costante per le 3 superfici, quindi p_0 si ipotizza non alterarsi.

Poniamo una seconda ipotesi di fondo, ovvero che anche quando $R = 1$ la p_1 non è influenzata dalle caratteristiche delle variabili esplicative.

A questo punto si vuole rendere variabile il parametro p del modello della *Zero-One-Modified Geometric Distribution* definito precedentemente, facendolo dipendere dalle nostre variabili esplicative $(x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5})$ e ponendo quindi

$$p_i = \frac{exp(\sum_{j=1}^{k} \beta_j x_{ij})}{1 + exp(\sum_{j=1}^{k} \beta_j x_{ij})} \tag{3.7}$$

Data la (3.7) la funzione di verosimiglianza associata è:

$$\begin{aligned} l(\beta) = \sum_{i=1}^{n} log \left(\frac{exp(\sum_{j=1}^{k} \beta_j x_{ij})}{1 + exp(\sum_{j=1}^{k} \beta_j x_{ij})} \right) + \\ + \sum_{i=1}^{n} (R_i - 2) \left[1 - \sum_{i=1}^{n} \frac{exp(\sum_{j=1}^{k} \beta_j x_{ij})}{1 + exp(\sum_{j=1}^{k} \beta_j x_{ij})} \right] \end{aligned} \tag{3.8}$$

Si tratta quindi di massimizzare (3.8) rispetto a $\hat{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5)$.

Una volta stimati i $\hat{\beta}$, è possibile calcolare le probabilità che all'X-esimo punto dell'Y-esimo match, con le quali si ottiene un determinato numero di scambi, ovvero calcolare le nostre $\hat{p_i}$ con la (3.7).

Fatto ciò, si otterrà un vettore di lunghezza pari al numero di punti analizzati, dove i valori di $\hat{p_i}$ variano a seconda del match di riferimento e della lunghezza relativa del *rally*.

Ogni $\hat{p_i}$ genererà una distribuzione diversa, attraverso la $zomGeom(p_0, p_1, \hat{p_i})$ in base ai regressori, quindi si passerà a simulare, per ogni punto, un solo valore di *rally* e si analizzerà la distribuzione marginale.

Dalla massimizzazione di (3.8), è risultato che le variabili *Bonus* e *Malus* non sono significative, quindi si è provato anche un modello a 3 esplicative in cui queste variabili non vengono considerate, il che ha prodotto risultati quasi uguali al modello con 5 esplicative.

3.6 Applicazione dei modelli condizionati ai dati reali

3.6.1 Confronto dei dati dei due modelli analizzati

Dai grafici della figura 3.2 si evidenzia come il modello proposto da Kovalchik & Ingram (2018) sia più efficace per la superficie terra rispetto al cemento, e soprattutto rispetto all'erba, oltre ad essere meno efficace di quello prodotto in questa ricerca.

Si procede ora con l'analisi degli indici di ottimalità:

Le tabelle 3.13, 3.14, 3.15 confermano quanto si poteva evincere dal grafico, cioè che il modello di Kovalchik & Ingram (2018) funziona al suo meglio sulla terra sugli incontri maschili. Sugli incontri femminili invece presenta un'efficacia decisamente inferiore

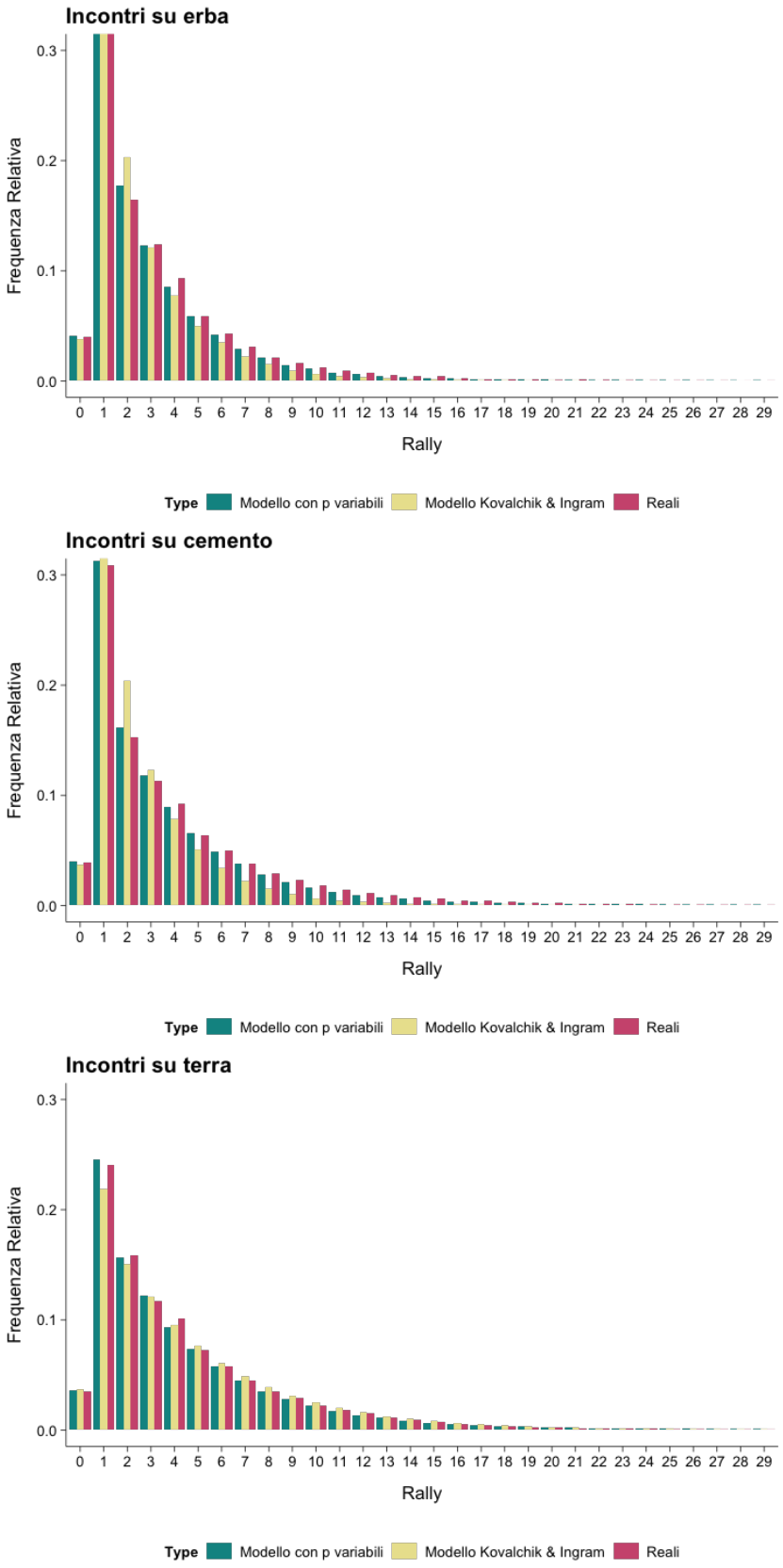


Figura 3.1: Confronto tra i modelli presentati in 3.5.1 e i dati reali osservati per gli incontri maschili

Superficie	Sesso	β_1	*Std.Error*	β_2	*Std.Error*
Erba	U	-13.5251	1.4380	-0.1230	0.0543
Cemento	U	-8.1527	0.5925	0.0028	0.0227
Terra	U	-15.6345	0.7915	1.1366	0.0294
Erba	D	-31.9095	1.7296	0.0908	0.0434
Cemento	D	-2.5367	1.1809	0.1588	0.0277
Terra	D	-28.2791	1.8008	-0.2141	0.0464

β_3	*Std.Error*	β_4	*Std.Error*	β_5	*Std.Error*
0.1380	0.0619	-0.0025	0.0008	2.1385	0.2431
0.1945	0.0253	0.0014	0.0004	1.1603	0.0998
0.0941	0.0034	0.0032	0.0005	2.3936	0.1334
-0.0730	0.0494	0.0037	0.0011	5.2537	0.2958
0.0380	0.0328	-0.0015	0.0007	4.0982	0.2019
0.4692	0.0542	0.0049	0.0011	4.6460	0.3070

Tabella 3.9: Stima e Std.Error dei parametri del modello *Zero-One-Modified Geometric* con p variabili

Modello	Δ_{10}	Δ_{20}	Δ
Kovalchik & Ingram	0.15779	0.18129	0.18397
Zero-One-Modified Geometric con p variabili	0.34886	0.04126	0.44490

Tabella 3.10: Tabella di confronto di ottimalità fra i modelli condizionati per la superficie *erba* negli incontri maschili

Modello	Δ_{10}	Δ_{20}	Δ
Kovalchik & Ingram	0.25734	0.30722	0.31299
Zero-One-Modified Geometric con p variabili	0.02992	0.04126	0.04307

Tabella 3.11: Tabella di confronto di ottimalità fra i modelli condizionati per la superficie *cemento* negli incontri maschili

Modello	Δ_{10}	Δ_{20}	Δ
Kovalchik & Ingram	0.06058	0.06909	0.07128
Zero-One-Modified Geometric con p variabili	0.02387	0.02913	0.03279

Tabella 3.12: Tabella di confronto di ottimalità fra i modelli condizionati per la superficie *terra* negli incontri maschili

Modello	Δ_{10}	Δ_{20}	Δ
Kovalchik & Ingram	0.31899	0.35346	0.35453
Zero-One-Modified Geometric con p variabili	0.03048	0.03880	0.04460

Tabella 3.13: Tabella di confronto di ottimalità fra i modelli condizionati per la superficie *erba* negli incontri femminili

Modello	Δ_{10}	Δ_{20}	Δ
Kovalchik & Ingram	0.37566	0.42637	0.42868
Zero-One-Modified Geometric con p variabili	0.03679	0.04182	0.04793

Tabella 3.14: Tabella di confronto di ottimalità fra i modelli condizionati per la superficie *cemento* negli incontri femminili

Modello	Δ_{10}	Δ_{20}	Δ
Kovalchik & Ingram	0.45691	0.51321	0.51467
Zero-One-Modified Geometric con p variabili	0.05245	0.06587	0.07366

Tabella 3.15: Tabella di confronto di ottimalità fra i modelli condizionati per la superficie *terra* negli incontri femminili

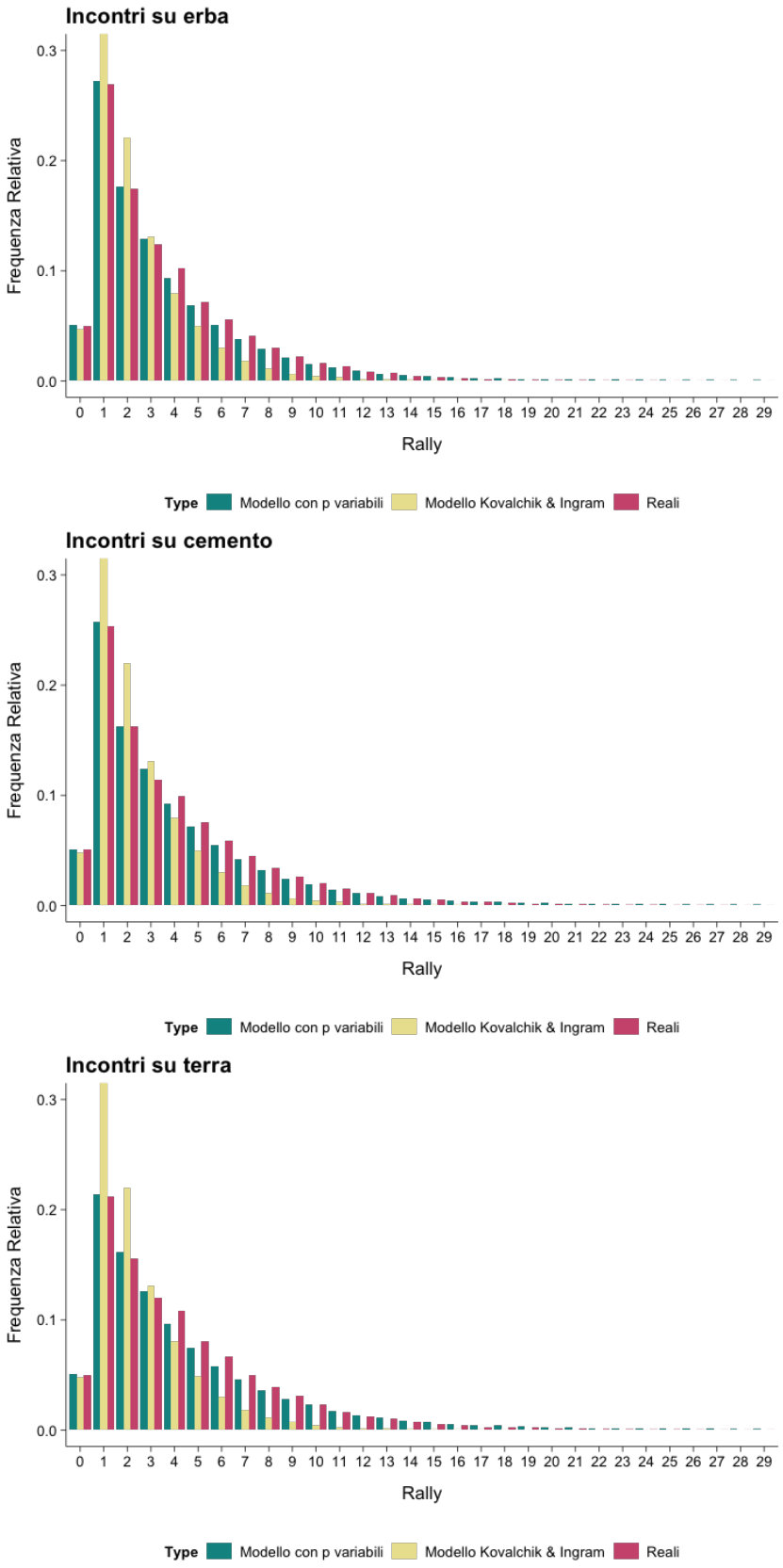


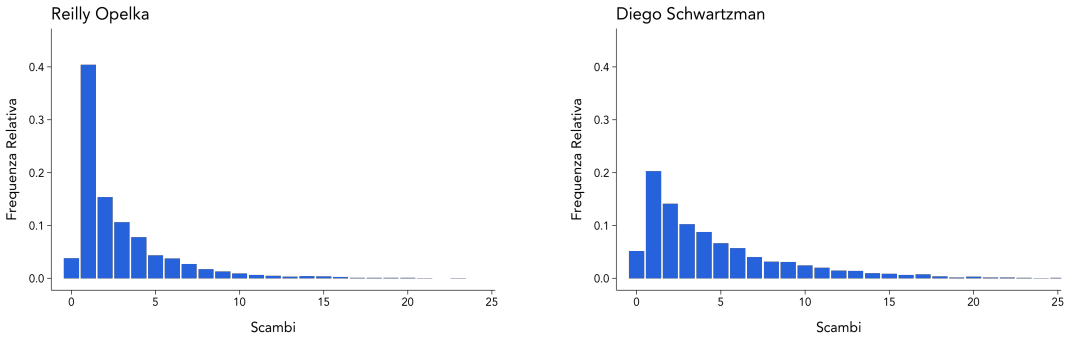
Figura 3.2: Confronto tra i modelli presentati in 3.5.1 e i dati reali osservati per gli incontri femminili

3.7 Confronto fra giocatori professionisti

Passiamo ora ad analizzare alcuni giocatori del circuito professionistico: Reilly Opelka (211cm), Ivo Karlovic (211cm), Diego Schwartzman (170cm) e i membri del *Big Three*.

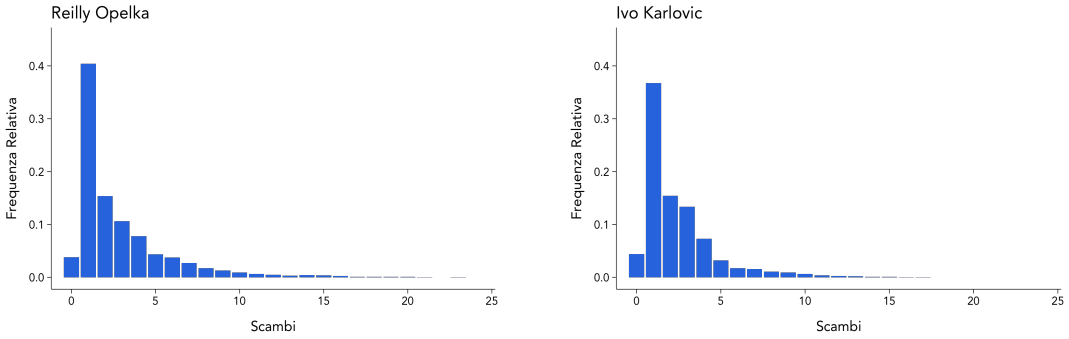
Il *Big Three* è un soprannome comune usato per indicare il trio formato d aRoger Federer, Rafael Nadal e Novak Djokovic, i 3 giocatori più vincenti negli ultimi 2 decenni.

Inoltre, sono considerati fra i migliori giocatori di tennis di sempre.



(a) Reilly Opelka (b) Diego Schwartzman

Figura 3.3: Confronto fra la distribuzione degli scambi fra Opelka e Schwartzman



(a) Reilly Opelka (b) Ivo Karlovic

Figura 3.4: Confronto fra la distribuzione degli scambi fra Opelka e Karlovic

Nella figura 3.5 sono mostrate le distribuzioni di frequenza dei *Big Three*,

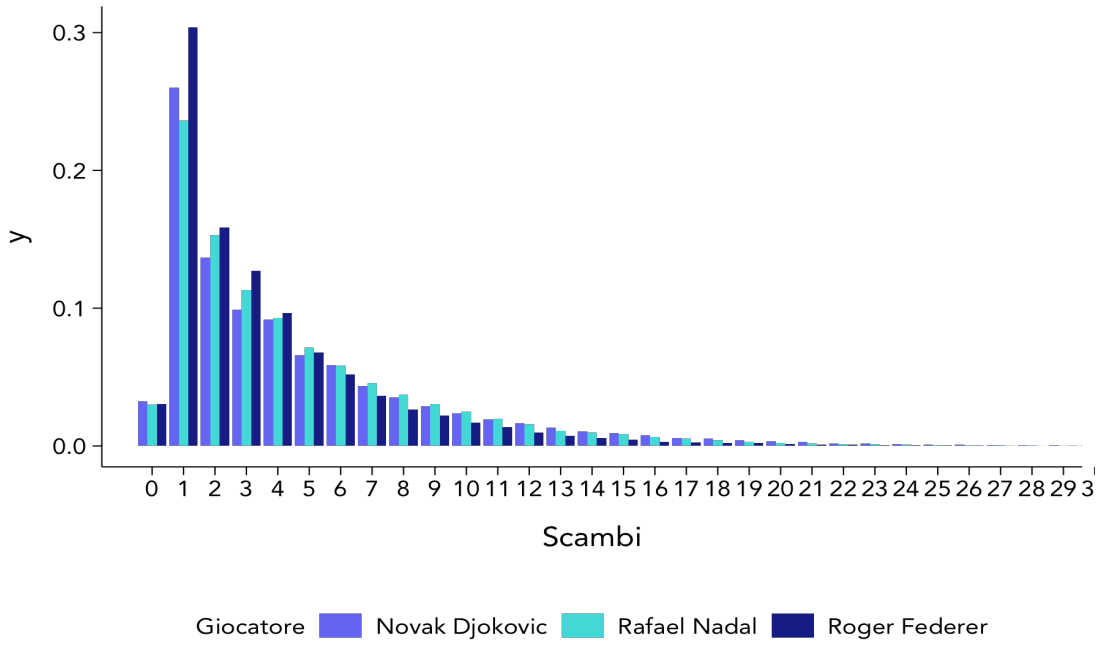


Figura 3.5: Distribuzione di frequenza relativa dei *Big Three*

Capitolo 4

Conclusione

In questa tesi si è provveduto a dimostrare l'esistenza, e valutare l'ottimalità, di più modelli probabilistici atti a descrivere la distribuzione di frequenza dei *rally* negli incontri professionistici di tennis.

Dalle analisi presentate, è risultato che il modello migliore per rappresentare i dati dei *rally* è il modello associato alla *Zero-One-Modified Geometric Distribution*.

Sono stati analizzati anche altri modelli, come ad esempio quello di *QuasiPoisson* o quello proposto da Kovalchik & Ingram (2018), che sono risultati meno ottimali.

E' possibile asserire che i *rally* R si distribuiscono come una *Zero-One-Modified Geometric Distribution*, avente parametri (p_0, p_1, p_i), dove p_i sono i parametri che dipendono dalle variabili indicate nella sezione 3.5.1, che variano da partita a partita.

Una prima difficoltà riscontrata in questa ricerca, è stata l'accesso ai dati e la loro "pulizia" per renderli leggibili e interpretabili. Il MCP (Match Charting Program) non prevedeva l'esportazione dei dati in formato CSV o formati simili, quindi si è provveduto a creare specifiche funzioni di *web scraping* mirate alla raccolta dei dati.

Successivamente c'è stato il problema di trovare possibili distribuzioni di probabilità che presentassero caratteristiche simili alla distribuzione osservata.

Un limite che ha presentato questa ricerca è stato quello che si sono dovuti stimare i parametri di alcune distribuzioni, solamente tramite le frequenze relative osservate disponibili, mentre invece sarebbe stato ottimale farle dipendere dai dati.

Conoscere le caratteristiche dell'avversario (nello specifico la probabilità di fare punto al servizio e la statura), permette, in fase di preparazione del match, di applicare il modello migliore per simulare la distribuzione degli scambi.

Grazie a questo si potrà aiutare l'atleta nella scelta della strategia migliore da applicare durante la partita, relativamente all'avversario e alla superficie del terreno di gioco.

Si potrebbe anche ampliare il modello descritto in questa tesi, aggiungendo altre variabili esplicative, come la velocità di gioco della palla, la percentuale di vincenti eseguiti, la zona di servizio preferita e tante altre variabili simili.

Con l'aggiunta di altri fattori quindi, potremmo rendere il parametro p_1 variabile, con interessanti conseguenze.

Appendice A

Appendice

Nella tabella A.5, da sinistra a destra, si hanno rispettivamente i dati relativi agli incontri maschili sull'erba, sulla terra e sul cemento, agli incontri femminili sull'erba, sulla terra e sul cemento.

Tabella A.1: Tabella Frequenze assolute osservate per gli incontri maschili

Rally	Erba	Cemento	Terra
0	2,814	11,399	5,021
1	25,080	89,874	34,250
2	11,617	44,456	22,623
3	8,708	32,887	16,604
4	6,543	26,975	14,371
5	4,156	18,627	10,352
6	2,997	14,562	8,221
7	2,209	11,078	6,323
8	1,497	8,554	5,032
9	1,140	6,788	4,093
10	877	5,244	3,202
11	629	4,161	2,604
12	519	3,282	2,124
13	350	2,664	1,654
14	309	2,106	1,252
15	278	1,767	978
16	176	1,376	790
17	128	1,108	666
18	97	876	457
19	80	730	376
20	42	557	318
21	67	438	243
22	48	331	191
23	29	280	159
24	20	197	121
25	17	160	69
26	16	130	67
27	12	92	41
28	4	86	41
29	8	59	25
30	5	41	23
31	2	24	28
32	4	32	19
33	2	25	15
34	3	14	10
35	0	16	7
36	1	8	6
37	0	15	3
38	0	9	1
39	1	6	3
40	0	4	2
41	1	4	4

Rally	Erba	Cemento	Terra
42	0	2	0
43	0	2	0
44	1	4	1
45	0	1	2
46	0	1	0
47	0	0	2
48	1	2	1
49	0	0	2
50	0	1	0
51	0	1	0
52	0	1	0
53	0	0	0
54	0	2	0
55	0	0	0
56	0	0	0
57	0	1	0
58	0	0	1
59	0	1	0
60	0	0	1
61	0	0	0
62	0	0	0
63	0	0	0
64	0	0	0
65	0	0	0
66	0	0	0
67	0	0	0
68	0	0	0
69	0	0	0
70	0	0	0
71	0	0	1
72	0	0	0
73	0	0	0
74	0	0	0
75	0	0	0
76	0	0	0
77	0	0	0
78	0	0	0
79	0	0	0
80	0	0	0
81	0	0	0
82	0	0	0
83	0	0	1

Tabella A.2: Tabella Frequenze assolute osservate per gli incontri femminili

Rally	Erba	Cemento	Terra
0	3,279	6,600	2,524
1	17,761	33,069	10,806
2	11,507	21,219	7,933
3	8,186	14,941	6,126
4	6,704	12,935	5,516
5	4,697	9,914	4,092
6	3,699	7,650	3,416
7	2,722	5,830	2,544
8	1,998	4,412	1,971
9	1,426	3,366	1,580
10	1,064	2,598	1,156
11	839	1,949	829
12	541	1,490	642
13	456	1,166	508
14	307	850	354
15	193	656	290
16	149	494	200
17	118	367	134
18	63	265	108
19	50	186	94
20	36	147	80
21	32	107	44
22	15	79	33
23	17	48	24
24	13	52	14

Rally	Erba	Cemento	Terra
25	6	24	18
26	4	25	10
27	3	20	5
28	0	11	5
29	3	7	4
30	2	7	6
31	1	7	0
32	1	5	4
33	0	3	0
34	1	0	1
35	0	2	2
36	0	1	0
37	0	0	1
38	0	0	0
39	0	4	0
40	0	0	0
41	0	0	0
42	0	0	0
43	0	0	0
44	0	1	0
45	0	0	0
46	0	0	0
47	0	0	0
48	0	1	1
49	0	0	0

Tabella A.3: Tabella Frequenze relative osservate per gli incontri maschili

Rally	Erba	Cemento	Terra	Rally	Erba	Cemento	Terra
0	0.0399	0.0392	0.0353	42	0	6.8714e-6	0
1	0.3558	0.30886	0.2405	43	0	6.8714e-6	0
2	0.1648	0.1527	0.1589	44	1.4187e-5	1.3743e-5	7.0226e-6
3	0.1235	0.1130	0.1166	45	0	3.4357e-6	1.4045e-5
4	0.0928	0.0927	0.1009	46	0	3.4357e-6	0
5	0.0590	0.0640	0.0727	47	0	0	1.4045e-5
6	0.0425	0.0500	0.0577	48	1.4187e-5	6.8714e-6	7.0225e-6
7	0.0313	0.0381	0.0444	49	0	0	1.404e-5
8	0.0212	0.0294	0.0353	50	0	3.4357e-6	0
9	0.0162	0.0233	0.0287	51	0	3.4357e-6	0
10	0.0124	0.0180	0.0224	52	0	3.4357e-6	0
11	0.0089	0.0143	0.0183	53	0	0	0
12	0.0074	0.0113	0.0149	54	0	6.8714e-6	0
13	0.0050	0.0092	0.0116	55	0	0	0
14	0.0044	0.0072	0.0088	56	0	0	0
15	0.0039	0.0061	0.0069	57	0	3.4357-6	0
16	0.0025	0.0047	0.0055	58	0	0	7.0225e-6
17	0.0018	0.0038	0.0047	59	0	3.4357e-6	0
18	0.0014	0.0030	0.0032	60	0	0	7.0225e-6
19	0.0011	0.0025	0.0026	61	0	0	0
20	5.9585e-4	0.0019	0.0022	62	0	0	0
21	9.5052	0.0015	0.0017	63	0	0	0
22	6.8097e-4	0.0011	0.0013	64	0	0	0
23	4.1142e-4	9.6200e-4	0.0011	65	0	0	0
24	2.8374e-4	6.7684e-4	8.4973e-4	66	0	0	0
25	2.4118e-4	5.5143e-4	4.8456e-4	67	0	0	0
26	2.2699e-4	4.4664e-4	4.7051e-4	68	0	0	0
27	1.7024e-4	3.1609e-4	2.8792e-4	69	0	0	0
28	5.6747e-5	2.9547e-4	2.8792e-4	70	0	0	0
29	1.1349e-4	2.0271e-4	1.7556e-4	71	0	0	7.0225e-6
30	7.0934e-5	1.4086e-4	1.6152e-4	72	0	0	0
31	2.8374e-5	8.2457e-5	1.9663e-4	73	0	0	0
32	5.6747e-5	1.0994e-4	1.3342e-4	74	0	0	0
33	2.8374e-5	8.5893e-5	1.0534e-4	75	0	0	0
34	4.2560e-5	4.8100e-5	7.0225e-5	76	0	0	0
35	0	5.4972e-5	4.9158e-5	77	0	0	0
36	1.4187e-5	2.7486e-5	4.2135e-5	78	0	0	0
37	0	5.1536e-5	2.1068e-5	79	0	0	0
38	0	3.0922e-5	7.0225e-6	80	0	0	0
39	1.4187e-5	2.0614e-5	2.1068e-5	81	0	0	0
40	0	1.3743e-5	1.4045e-5	82	0	0	0
41	1.4187e-5	1.3743e-5	2.8090e-5	83	0	0	7.0225e-6

Tabella A.4: Tabella Frequenze relative osservate per gli incontri femminili

Rally	Erba	Cemento	Terra	Rally	Erba	Cemento	Terra
0	0.050	0.051	0.049	25	9.1057e-5	1.8390e-4	3.4267e-4
1	0.270	0.253	0.212	26	6.0704e-5	1.9156e-4	1.9581e-4
2	0.175	0.163	0.155	27	4.5528e-5	1.5325e-4	9.7907e-5
3	0.124	0.114	0.120	28	0	8.4286e-5	9.7907e-5
4	0.102	0.099	0.108	29	4.5528e-5	5.3637e-5	9.7907e-5
5	0.071	0.076	0.080	30	3.0352e.5	5.3637e-5	7.8325e-5
6	0.056	0.059	0.067	31	1.5176e-5	5.3637e-5	0
7	0.041	0.045	0.050	32	1.5176e-5	3.8312e-5	7.8325e-5
8	0.030	0.034	0.039	33	0	2.2987e-5	0
9	0.022	0.026	0.031	34	1.5176e-5	0	1.9581e-5
10	0.016	0.020	0.023	35	0	1.5325e-5	3.9163e-5
11	0.013	0.015	0.016	36	0	7.6624e-6	0
12	0.008	0.011	0.013	37	0	0	1.9581e-5
13	0.007	0.009	0.010	38	0	0	0
14	0.005	0.007	0.007	39	0	3.0649e-05	0
15	0.003	0.005	0.006	40	0	0	0
16	0.002	0.004	0.004	41	0	0	0
17	0.002	0.003	0.003	42	0	0	0
18	0.001	0.002	0.002	43	0	0	0
19	0.001	0.001	0.002	44	0	7.6624e-6	0
20	0.001	0.001	0.002	45	0	0	0
21	4.8564e-4	8.1987e-4	8.5179e-4	46	0	0	0
22	2.2764e-4	6.0533e-4	6.4618e-4	47	0	0	0
23	2.5799e-4	3.6779e-4	4.6017e-4	48	0	7.6624e-6	1.9581e-5
24	1.9729e-4	3.9844e-4	2.7414e-4	49	0	0	0

Tabella A.5: Frequenze reali e simulate per gli incontri analizzati

Rally	Frequenze reali	Frequenze simulate
0	5,628	5,804
1	50,160	50,892
2	23,234	24,976
3	17,416	17,241
4	13,086	11,992
5	8,312	8,301
6	5,994	5,925
7	4,418	4,081
8	2,994	2,956
9	2,280	2,056
10	1,754	1,509
11	1,258	1,072
12	1,038	840
⋮	⋮	⋮

Rally	Frequenze reali	Frequenze simulate
0	10,042	10,318
1	68,499	69,729
2	45,246	44,583
3	33,207	34,831
4	28,742	26,666
5	20,704	20,832
6	16,442	16,448
7	12,646	12,810
8	10,063	10,016
9	8,186	7,876
10	6,403	6,199
11	5,207	4,809
12	4,248	3,872
⋮	⋮	⋮

Rally	Frequenze reali	Frequenze simulate
0	22,798	22,927
1	179,748	181,857
2	88,911	93,644
3	65,774	68,637
4	53,950	51,685
5	37,254	38,255
6	29,124	28,450
7	22,155	21,923
8	17,107	16,445
9	13,576	12,441
10	10,488	9,500
11	8,322	7,291
12	6,564	5,528
⋮	⋮	⋮

Rally	Frequenze reali	Frequenze simulate
0	1,924	1,869
1	15,120	15,289
2	7,552	8,427
3	6,158	5,905
4	4,778	4,334
5	3,238	3,117
6	2,424	2,353
7	1,660	1,673
8	1,256	1,312
9	952	839
10	676	632
11	586	498
12	402	372
⋮	⋮	⋮

Rally	Frequenze reali	Frequenze simulate
0	2,608	2,676
1	15,350	15,613
2	10,726	10,777
3	8,200	8,590
4	7,068	6,614
5	5,476	5,270
6	4,352	4,137
7	3,446	3,403
8	2,596	2,646
9	2,222	2,060
10	1,754	1,715
11	1,346	1,389
12	1,154	1,035
⋮	⋮	⋮

Rally	Frequenze reali	Frequenze simulate
0	4,457	4,530
1	32,691	33,041
2	17,034	17,755
3	12,953	13,370
4	10,596	9,927
5	7,496	7,669
6	5,862	5,967
7	4,449	4,479
8	3,380	3,331
9	2,806	2,715
10	2,125	2,129
11	1,894	1,651
12	1,441	1,250
⋮	⋮	⋮

Bibliografia

[1] Carboch, J., Placha, K., Sklenarik, M., Rally pace and match characteristics of male and female tennis matches at the Australian Open 2017, *Journal of Human Sport and Exercise*, 13(4), 743-751

[2] O'Donoughe, P., *Statistics for Sport and Exercise Studies: An Introduction*, Routledge, 2001, 108-109

[3] O'Donoghue, P.,Ingram , B., A notational analysis of elite tennis strategy *Journal of Sports Sciences*, 2001, 19:2, 107-115

[4] Courel-Ibáñez, J., Sánchez-Alcaraz Martínez, B. J., Cañas, J., Game Performance and Length of Rally in Professional Padel Players, *Journal of human kinetics*, 2017, 55, 161–169s

[5] Lowrance, M., One More Shot: Predicting Wins in Women's Professional Tennis, *Southern Utah University Working Paper Series*, 2018

[6] Carboch J, Siman J, Sklenarik M, Blau M., Match Characteristics and Rally Pace of Male Tennis Matches in Three Grand Slam Tournaments, *Physical Activity Review 2019*, 7, 49-56

[7] Kovalchik, S.A., Ingram, M., Estimating the duration of professional tennis matches for varying formats *Journal of Quantitative Analysis in Sports*, vol. 14, no. 1, 2018, pp. 13-23

[8] Lisi, F., Grigoletto, M., Modeling and simulating durations of men's professional tennis matches by resampling match features, *Journal of Sports Analytics*, 7 (2021), 57–75

[9] Klaassen, F., Magnus, J.,R., *Analyzing Wimbledon: The Power of Statistics*. USA: Oxford University Press. 2014

[10] Fitzpatrick, A., Stone, J.A., Choppin, S., Kelley, J., Investigating the most important aspect of elite grass court tennis: Short points, *International Journal of Sports Science & Coaching*, 2021, 16(5):1178-1186

[11] Yixiong C., Haoyang L., Hongyou L., Miguel-Ángel G., Data-driven analysis of point-by-point performance for male tennis player in Grand Slams, *Motricidade, 2019*, vol. 15, n. 1, pp. 49-61

[12] Bijesh Y., Lakshmanan J., Visalakshi J., Jothilakshmi D., Sebastian G., K.G. Selvaraj, Shrikant I.B., Can Generalized Poisson Model replace any other count data models?, *Clinical Epidemiology and Global Health*, 11, 2021

[13] Liu, W., Tang, Y., Xu, A., Zero-and-one-inflated Poisson regression model, *Statistical Papers*, 62, 2021

[14] Dankmar B., Van Der Heijden P. G. M., The identity of the zero-truncated, one-inflated likelihood and the zero-one-truncated likelihood for general count densities with an application to drink-driving in Britain *The Annals of Applied Statistics*, Ann. Appl. Stat. 13(2), 1198-1211, (June 2019)

[15] Hussain T., A zero truncated discrete distribution: Theory and application to count data, *Pak.j.stat.oper.res.*, Vol.16, No.1, 2020, pp 167-190

[16] Norman L.J., Adrienne W.K., Kots S. *Univariate Discrete Distribution*, Wiley, Third Edition, 2005

[17] Klaassen, F., Magnus, J.,R., Are Points in Tennis Independent and Identically Distributed? Evidence From a Dynamic Binary Panel Data Model, *Journal of the American Statistical Association*, 2001

[18] Machar R., Stuart M., Whiteside D., Matchplay characteristics of Grand Slam tennis: implications for training and conditioning, *Journal of Sports Sciences*, 2016, 34:19, 1791-1798

[19] Ver Hoef, J.M. and Boveng, P.L., Quasi-Poisson vs. Negative Binomial Regression: How should we model overdispersed count data?, *Ecology*, 88: 2766-2772, 2007

[20] R-forge distributions Core Team *A guide on probability distributions* University Year 2008-2009

[21] Newton, P. K., Keller, J.B, Probability of winning at Tennis I. Theory and Data, *Studies in Applied Mathematics*,Massachusetts Institute of Technology, 114: 241-269, 2015

Sitografia

[1] Kovalchik, S., *Sizing up height advantage*, http://www.on-the-t.com/2020/10/25/,

[2] Kovalchik, S., http://on-the-t.com/2016/02/14/klaassen-magnus-hypothesis-1/

[3] Kuzdub, M., https://www.mattspoint.com/blog/rally-lengths-in-tennis-a-contrasting-perspective

[4] Vose Software, https://www.vosesoftware.com/riskwiki/Zero-modifiedcountingdistributions.php

[5] Sackmann, J., *The Match Charting Project*, http://www.tennisabstract.com/blog/2013/11/26

[6] Sackmann, J., *Searching For Meaning in Distance Run Stats*, http://www.tennisabstract.com/blog/2016/08/19,

[7] Orson, A., *Tennis? Yeah, Tennis*, Fourspaces, https://fourspacesplus.wordpress.com/2015/06/30

[8] Coppini, F., *Un punto di vista diverso sulla durata degli scambi*, https://www.tennisworlditalia.com/tennis/news/LezioniTennis