

# I contributi della metodologia statistica all'atletica leggera

Antonio Mussino

Università degli Studi di Roma "La Sapienza"

## 1. Introduzione

La Statistica è tra le discipline più funzionali alla descrizione e alla comprensione di tutte le sfaccettature del fenomeno sportivo e alla individuazione dei fattori che ne determinano le trasformazioni. Essa è alla base dello sport moderno: l'approccio ai criteri di misurazione, di costruzione dei punteggi e delle graduatorie, all'organizzazione di gare e tornei rientra di fatto nella metodologia statistica, sia che si usino semplici medie e percentuali, sia che si faccia ricorso a complessi modelli probabilistici.

È oggi crescente la domanda di tecniche statistiche avanzate per la soluzione dei problemi legati all'attività agonistica, con particolare riferimento a quella di alta prestazione. Nell'Atletica leggera questo interesse è enfatizzato dalla natura delle discipline praticate, nelle quali la misurazione è una componente fondamentale.

L'applicazione di modelli e tecniche statistiche avanzate è, infatti, condizionata dalla natura delle informazioni (variabili) che vengono rilevate. Se si tratta di variabili quantitative (tempi, lunghezze e così via) aumenta considerevolmente la possibilità di applicare anche i più complessi strumenti che la Statistica propone.

Sono così innumerevoli le applicazioni di modelli e tecniche statistiche che si possono incontrare nelle riviste di Scienza dello Sport, o sportive in senso lato; ma anche nelle riviste statistiche specializzate, di impostazione sia metodologica che

applicativa, vi è stata una proliferazione di contributi aventi per oggetto problemi relativi alla "regina degli sport".

Va infatti enfatizzato che, se è vero che il contributo statistico è di grande aiuto per risolvere problemi scientifici nell'ambito sportivo, è altrettanto vero che alcuni di questi problemi sono stati e saranno di stimolo per approfondimenti metodologici nei vari ambiti disciplinari della Statistica. In questo articolo sono stati presi in esame alcuni contributi che presentavano come obiettivo peculiare la soluzione di un problema connesso con l'atletica leggera: come detto, la bibliografia in questo settore è vastissima, per cui si è scelto di analizzare lavori tratti esclusivamente da pubblicazioni appartenenti alla letteratura statistica.

## 2. Studio del record e della prestazione eccezionale

Nell'Atletica leggera, come in genere negli sport individuali, l'interesse degli statistici si è concentrato prevalentemente sulla prestazione eccezionale, studiata come evoluzione dei record da chiunque stabiliti nelle diverse specialità e non come sequenza di performance personali di un singolo atleta.

Iniziamo con un lavoro di Chatterjee S. e Chatterjee S. (1982): si tratta di un'analisi dei tempi dei vincitori nelle gare di corsa sui 100, 200, 400 e 800 metri maschili alle Olimpiadi, dal 1900 al

1976. L'obiettivo degli Autori è quello di trovare se esista un modello che descrive l'evoluzione dei tempi osservati, considerando le quattro specialità nel contesto di due diversi modi di classificazione dei dati: per anni e per distanza. Inoltre viene esaminato e tenuto sotto controllo l'effetto dell'altitudine, che viene utilizzata come covariata.

Nell'analisi delle singole specialità il primo passo è quello di costruire un modello regressivo per i tempi osservati. Un primo esame dei dati mostra che per tutte le specialità: i) i tempi migliorano, ii) il tasso di miglioramento è decrescente, iii) è prevedibile un tempo *limite* più basso.

Gli Autori prendono il tempo in una qualsiasi prova ( $t_j$ ) come una funzione dell'anno ( $j$ ) e considerano il modello lineare  $t_j = \alpha + \beta_j$  e quello esponenziale  $t_j = \theta \exp(j\theta_j)$ . Come risultato propongono un semplice modello regressivo non lineare, che soddisfa le considerazioni di cui sopra e che consente di prevedere i tempi delle diverse specialità:

$$t_j = \theta_1 + \theta_2 \exp(j\theta_3) + \varepsilon_j, \quad \theta_2 > 0, \theta_3 < 0,$$

dove  $\varepsilon_j$  è un errore casuale.

Il miglior tempo ottenibile per una specialità è:

$$\theta_1 \text{ (per } j \rightarrow \infty)$$

Analizzando singolarmente le quattro specialità con questo modello gli Autori trovano che l'interpolazione è buona e quindi è possibile prevedere gli andamenti cercati.

Il secondo problema affrontato è se l'altitudine abbia avuto effetto sui tempi riportati nelle quattro specialità. Utilizzando la rappresentazione grafica dei residui (che sono gli scarti fra i tempi osservati e quelli stimati dal modello regressivo) rispetto all'altitudine, gli Autori osservano che i due punti rappresentativi delle sedi di Città del Messico e di Monaco, le uniche ad avere un'altitudine apprezzabilmente differente dal livello del mare, presentano una collocazione negativa (cioè

i risultati sono migliori), ma non distante in modo significativo dai valori stimati dal modello. Pertanto gli Autori concludono che per le altitudini registrate e per le quattro specialità la prestazione non è influenzata significativamente dall'altitudine, sia perché i partecipanti ai Giochi Olimpici sono atleti di altissimo livello, sia perché gli effetti possono essere stati eliminati dall'allenamento e dall'ambientamento in quota.

Gli Autori propongono ulteriori interessanti analisi sulle relazioni fra i tempi delle diverse prove. La stima del miglior tempo ottenibile per i 200 metri è meno del doppio del miglior tempo stimato per i 100 metri, indicando che il tempo di avvio ha un peso considerevole per le gare brevi e di massima velocità; una volta che la velocità massima sia stata raggiunta, questa viene mantenuta per tutta la prova. Diverse sono, infatti, le relazioni fra e con i tempi nei 400 e negli 800 metri: ad esempio il miglior tempo per i 400 metri è più grande del doppio del miglior tempo per i 200 metri.

Il lavoro di Chatterjee e Chatterjee è stato lo spunto per molte applicazioni su questo tema. Tryfos P. e Blackmore R. (1985) criticano l'impostazione precedente in quanto non idonea a spiegare il processo che genera il risultato, ma anche perché non può accettarsi l'ipotesi di base di indipendenza tra le osservazioni. Essi considerano i record come la realizzazione di un processo aleatorio e, con una semplice estensione della teoria delle statistiche d'ordine, ottengono la densità congiunta e la covarianza dei record in due periodi qualsiasi.

Gli Autori considerano una specialità atletica e una serie che mostra il record per quella specialità lungo un periodo di tempo. La serie fornisce alcune informazioni circa i record probabili nel futuro, nell'ottica di prevedere possibili limiti delle prestazioni atletiche.

È chiaro che un semplice trend estrapolato dalla serie dei tempi dei record non è soddisfacente per due ragioni: la prima è che un approccio di questo tipo non fa nessun tentativo di spiegare che

cosa abbia generato le osservazioni e la seconda è che gli stimatori dei parametri e le previsioni possono non avere le proprietà desiderate (indipendenza).

Gli Autori come applicazione calcolano le previsioni dei futuri record mondiali nelle sei maggiori specialità della corsa su lunghe distanze (1 km, 1 miglio, 5 km, 10 km, 20 km e la maratona) per il periodo dal 1983 al 1997.

In un periodo di tempo – in genere si considera un anno – una specialità è corsa più volte. I tempi dei vincitori in ognuna delle prove sono influenzati da un largo numero di fattori: qualità dei corridori, luogo delle gare, condizioni meteorologiche e altro. Gli Autori prendono  $X_t$  come la variabile aleatoria che rappresenta il tempo di corsa più veloce nel periodo  $t$  e la sua funzione di densità. Prendono come il record nel periodo  $t$ , chiaramente  $Y_t = \min. \{X_t + X_2, \dots, X_t\}$ . Supponendo disponibili le  $T$  osservazioni,  $Y_1, Y_2, \dots, Y_T$ , che si considerano realizzazioni di questo processo aleatorio, il problema è di determinare la densità congiunta e la covarianza di  $Y_t$  e  $Y_{t+k}$  ( $k = 1, 2, \dots$ ). Un'altra considerazione critica nei confronti degli studi precedenti è l'impossibilità di studiare i record nel lungo periodo di tempo, perché possono riflettere importanti cambiamenti nei fattori nutrizionali, ereditari, tecnologici, tecnici e di allenamento.

Prima di fare qualsiasi applicazione per prevedere i record mondiali futuri, gli Autori suggeriscono di stabilire un periodo di tempo rispetto al quale si può assumere che le distribuzioni dei tempi di corsa siano approssimativamente le stesse. I fattori nutrizionali, tecnologici, di allenamento e così via, causano profonde modifiche nell'andamento delle distribuzioni nel tempo, rendendole incomparabili: il periodo base massimo consentito per gli Autori è di 15 anni.

Anche Robinson M.E. e Tawn J.A. (1995) sviluppano metodi statistici, basati sulle tecniche del valore estremo, per stimare la migliore prestazione possibile di una popolazione di atleti in com-

petizione. Questi metodi sono applicati all'analisi dei dati dei 3000 m femminili.

Un *cliché* in atletica è che i record esistano per essere battuti. Con lo sviluppo delle tecniche di allenamento e dell'attrezzatura, migliorata la scienza dietetica e aumentata la partecipazione, i record sono battuti con maggiore frequenza e con miglioramenti più consistenti nel caso in cui le prestazioni provengano da una popolazione fissata. Nel contesto di una popolazione che migliora pertanto, è molto probabile che un record venga battuto, ma se il margine con cui il record è battuto è sorprendentemente ampio si può sospettare che la prestazione sia stata «aiutata» con l'uso di sostanze farmacologiche. Nonostante i miglioramenti considerevoli nei controlli sul doping, vi sono stati infatti molti casi sospetti di uso di sostanze non consentite non diagnosticati. In questi casi, ossia quando un atleta risulta negativo al test per doping, ma la sua prova desta qualche sospetto è utile studiare statisticamente le sue prestazioni passate e confrontare il suo trend con quello degli altri atleti a livello internazionale.

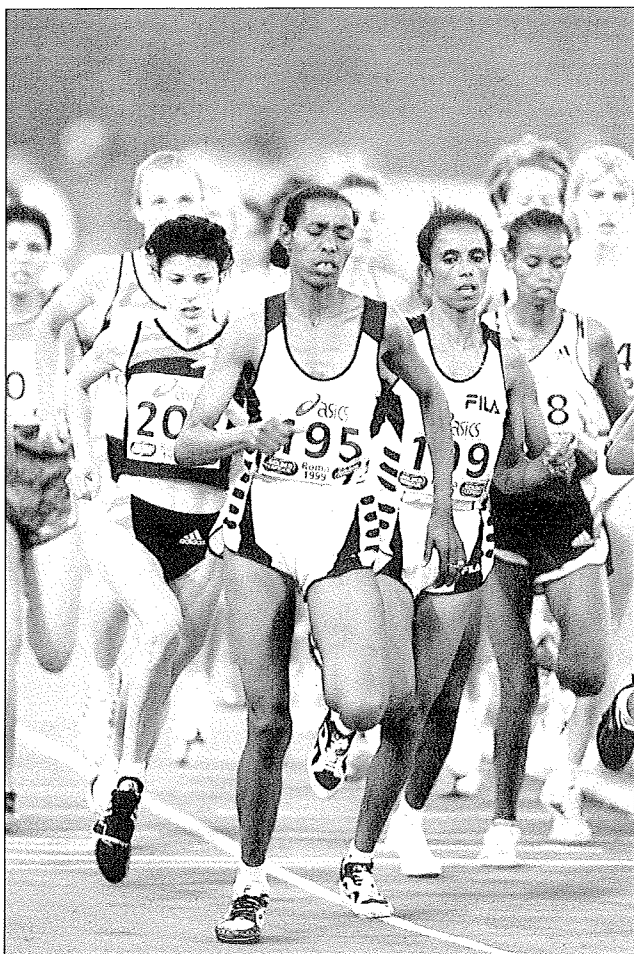
La metodologia sviluppata dagli Autori può essere applicata a molteplici discipline, ma qui si concentra l'attenzione sull'atleta cinese Wang Junxia. L'atleta, nel campionato nazionale a Pechino il 13 settembre 1993, migliorò il record mondiale dei 3000 m femminili di 10.43 secondi; il giorno successivo, corse in un tempo di 8 minuti e 6.11 secondi, battendo di nuovo il record di 6.08, e in entrambe le occasioni risultò negativa al controllo antidoping. La spiegazione dei cinesi era che queste prestazioni erano influenzate da una combinazione di vita, abitudine all'altitudine e una rara dieta, ma molti atleti si lamentarono sostenendo che sulle prestazioni dell'atleta cinese vi fosse stato l'effetto di una o più sostanze dopanti. Una terza possibilità è che l'atleta provenisse da una popolazione differente da quella degli atleti partecipanti alle competizioni a livello mondiale, essendo il 1993 il primo anno in cui gli atleti cinesi erano seriamente impegnati nelle competizioni internazio-

nali. Gli Autori soffermano la loro attenzione sul progressivo allargamento della base (degli atleti agonisti ad alto livello) e si interrogano sull'importanza che può avere, nella previsione delle prestazioni limite, l'ingresso di atleti provenienti da una popolazione differente da quella fino a quel momento analizzata. Dai dati non si può determinare se l'atleta abbia fatto ricorso al doping o se provenga da una popolazione differente. Conseguentemente gli Autori formulano alcuni modelli statistici per testare le ipotesi formulate dagli esperti di atletica leggera per valutare se il record sia compatibile con le tendenze in atto.

I modelli si distinguono in modelli di intradistanza e di interdistanza: i primi confrontano il record con le altre buone prestazioni nella stessa specialità e nello stesso anno. I secondi, invece, compa-

rano i risultati in altre specialità simili della corsa con distanze diverse: ad esempio la performance nei 3000 m può essere studiata in relazione a quelle sugli 800 m, sui 1500 m e così via. I 1500 m sono considerati la prova più simile ai 3000 m, infatti sono corsi con lo stesso stile: i 3000 m, in pratica, non possono essere corsi più velocemente di due volte l'ultimo tempo per i 1500 m. Gli Autori tengono conto di questa considerazione nella loro analisi e, attraverso un semplice modello, sviluppano le relazioni tra i parametri delle distribuzioni minime annuali per i 3000 m e 1500 m.

La conclusione cui gli Autori pervengono è che, con una probabilità di errore piuttosto bassa, il risultato sorprendente dell'atleta cinese sia da attribuire alla sua appartenenza ad una popolazione fino ad allora non indagata e perciò con differenti caratteristiche potenziali.



### 3. La maratona

Alla maratona sono connesse problematiche particolari, legate alla maggiore influenza di fattori ambientali e alla variabilità dei percorsi.

Smith R. L. e Corbett M. (1987), in riferimento al metodo di misurazione della lunghezza della maratona o di altre corse su strada, vogliono valutare quanto siano accurate le procedure utilizzate. La lunghezza ufficiale della maratona è 42 km e 195 m. Con l'accresciuta popolarità di questa prova l'attenzione sull'accuratezza della misurazione della sua lunghezza è stata crescente. Il metodo di misurazione più diffuso prevede l'utilizzo di una ruota calibrata di bicicletta. Il processo di misurazione avviene con una calibrazione iniziale su una distanza standard, seguita dalla misurazione vera e propria. Si tratta del cosiddetto metodo della bicicletta.

Una comune bicicletta viene equipaggiata con un contagiri, che registra accuratamente il numero dei giri della ruota anteriore. La bicicletta compie per prima una distanza standard, almeno un chilo-

metro di strada dritta e piana, per calibrare il contagiri. La distanza standard è misurata da un nastro d'acciaio o dal metro elettronico dell'ispettore ed è quella che viene assunta come esatta a fini statistici. Il contagiri è controllato all'inizio e alla fine della misurazione e la distanza viene calcolata tramite una semplice operazione aritmetica.

Questo metodo può essere usato per tutte le corse su strada. Le principali fonti di errore possono essere le variazioni dovute alle condizioni di calibrazione, per esempio alla variazione di temperatura, e gli errori dei misuratori che non seguono il percorso più breve possibile. Il metodo della bicicletta, se applicato da un esperto misuratore, è un metodo accurato che accetta una tolleranza di 50 m per la maratona.

Il punto di partenza del lavoro degli Autori è la misurazione della maratona delle Olimpiadi del 1984 a Los Angeles. La procedura prevedeva che 13 ciclisti "misuratori" percorressero più volte otto "linee base" (distanze standard), ossia tappe in cui il percorso era stato suddiviso, per calibrare le misurazioni. Oltre alle linee base, erano tenuti sotto osservazione altri intervalli del percorso, per complessivi 25 intervalli.

Il principale scopo dell'analisi degli Autori è di studiare la distribuzione degli errori dei misuratori rispetto agli intervalli di calibrazione esatti, per poi stimare l'errore complessivo medio della misurazione. La stima finale è più accurata grazie all'impiego di numerosi intervalli di calibrazione e al fatto che le misurazioni (i ciclisti) sono indipendenti. Il modello usato per la stima è un modello lineare individuato con il metodo della massima verosimiglianza:

$$y = \alpha + \beta x + \varepsilon$$

dove  $x$  è la quantità che ci interessa misurare e  $y$  una sua misura,  $\alpha$  e  $\beta$  sono costanti ignote e  $\varepsilon$  è un errore casuale. Con questo modello ottengono una stima  $\hat{x} = 30.904,1$  m con un errore standard di 2 m. La stima  $\hat{x}$  rappresenta la lunghezza tota-

le solo delle sezioni in cui si è divisa la corsa per la misurazione, e non dell'intero percorso della maratona; il resto del percorso include gli intervalli di calibrazione, un giro della pista alla fine e poche brevi sezioni non considerate dagli Autori. Analizzando i dati si nota che un elemento di distorsione può essere rappresentato dal fatto che i coefficienti di calibrazione variano nel tempo: le misurazioni iniziali erano, infatti, collegate alle condizioni fredde e umide della mattina, le successive alle condizioni calde e secche del pomeriggio. Questo suggerisce di dividere i dati in due gruppi (un «insieme di mattina» e un «insieme di pomeriggio») e di applicare lo stesso metodo ai due gruppi separatamente. Oppure, in alternativa, di considerare gli effetti meteorologici inserendoli nel modello precedente come variabili casuali e, pertanto, trasformandolo in un modello dinamico. Gli Autori segnalano anche che un così buon risultato (l'errore standard è pari a 2 m) era dovuto alla vasta esperienza dei ciclisti misuratori a Los Angeles. Infatti potrebbe accadere che per un gruppo di misuratori meno esperti si verificano risultati diversi, con un errore più alto.

In sintesi gli Autori giungono con il loro modello ad un errore di 2 m, decisamente piccolo rispetto alla lunghezza della maratona, e ipotizzano che questo potrebbe essere il risultato di tre fattori: a) i ciclisti che misurarono la corsa erano 13, mentre usualmente è uno soltanto; b) i ciclisti erano tutti molto esperti nel metodo di misurazione adottato; c) i numerosi intervalli di calibrazione erano presi nel percorso stesso della maratona.

#### 4. Specialità a prove multiple

In questo contesto si considerano le specialità a prove multiple che rientrano sempre nell'ambito dell'Atletica leggera: l'eptathlon e il decathlon. In un loro lavoro di confronto fra diversi metodi di classificazione, Dawkins B. P., Andreae P. M. e O'Connor P. M. (1994) mettono in evidenza in-

interessanti relazioni multidimensionali fra unità di analisi (atleti) e prove, che mostrano come il risultato finale della gara di eptathlon possa essere previsto sulla base dei risultati ottenuti in alcune specifiche prove. Questa evidenza fa pensare al fatto che tale risultato possa essere legato ad una implicita ponderazione delle prove.

Gli Autori mettono in guardia sulla difficoltà di analizzare insiemi ridotti di dati multidimensionali: soprattutto quando l'insieme dei dati è anche omogeneo, la struttura dei dati è di difficile definizione e spiegazione. Come è ben noto il decathlon consiste in dieci prove e l'eptathlon in sole sette prove atletiche.

Gli Autori descrivono l'applicazione di una tecnica di *clustering* (raggruppamento) *incrementale* ai dati sull'eptathlon delle Olimpiadi del 1992, mettendola a confronto con altre tecniche più tradizionali, dimostrando che la tecnica di clustering incrementale è un metodo efficace per scoprire e studiare le strutture presenti in piccoli insiemi di dati. Il metodo di clustering impiegato dagli Autori è quello chiamato *DySect* che usa metodi di massimizzazione locale non disponibili nei software statistici standard. Tale algoritmo produce, come output, un albero di classificazione multipla, che delucida la struttura dell'insieme dei dati più efficacemente dei metodi classici di clustering e consente una nitida evidenziazione delle diverse tipologie di atleti. L'analisi mostra come alcune prove siano più discriminanti nell'individuare le atleti con i risultati migliori.

### 5. Le matrici dei dati nello studio dell'Atletica leggera

Alla base della metodologia statistica presentata c'è la formalizzazione dei dati di partenza, molto differente a seconda del tipo di specialità e del tipo di situazione che si prende in esame. Le serie storiche delle prestazioni costituiscono il naturale punto di partenza di quasi tutti i contributi; al

di fuori del caso proposto della maratona, è poco frequente trovare applicazioni con riferimento ai fattori (personali, ambientali, ecc.) che possono influenzare la prestazione stessa.

La matrice dei dati è lo strumento fondamentale per la registrazione del dato statistico e sulla quale applicare le metodologie proposte. Una matrice di dati è un insieme di informazioni ottenute dall'intersezione di due insiemi omogenei al loro interno, uno organizzato per righe e l'altro per colonne. Statisticamente parlando, l'insieme di riga è costituito dalle unità di rilevazione e quello di colonna dalle variabili su di esse rilevate. Lo schema della matrice dei dati è anche quello che sottostà alle modalità di impostazione di alcuni *software* per la registrazione dei dati stessi, come ad esempio EXCEL, ACCESS, DBASE, ecc.

Senza pretendere di essere esaustivi e limitandosi ad alcune delle situazioni relative ai lavori precedentemente esaminati, si possono identificare alcune matrici di dati di particolare interesse.

La **previsione dei record** è il filone classico degli studi statistici applicati all'Atletica leggera. In tal caso le variabili da registrare sono i record (o le migliori prestazioni stagionali o annuali), mentre l'unità statistica è il tempo (in genere l'anno); è interessante, anche se complesso, prendere in esame lo sviluppo della componente tecnologica (ad esempio nel salto con l'asta e nei lanci) o della componente ambientale (altitudine, vento e così via). Questi dati sono, in genere, trattati come serie storiche, con la presenza di dati troncati o con la distribuzione del valore estremo come risposta.

Per lo studio delle *prestazioni degli atleti negli sport individuali* si costruisce una matrice che sia in grado di mettere in relazione la prestazione atletica con i fattori individuali, quelli ambientali e le loro interazioni. Unità statistica è l'atleta, mentre le variabili sono le sue *performance* e una serie di covariate necessarie a spiegare il risultato.

L'analisi dei *risultati delle prove multiple*, come

l'eptathlon e il decathlon, introduce interessanti elementi di analisi per gli statistici, in quanto le variabili in gioco aumentano di numero: l'unità statistica è sempre l'atleta; le variabili rilevate sono i risultati nelle varie prove, con la classifica finale e, anche se raramente, le caratteristiche degli atleti che possono svolgere un ruolo esplicativo.

## 6. Alcune considerazioni conclusive

Una così sommaria presentazione di alcuni lavori statistico metodologici che hanno come oggetto di studio problemi legati all'Atletica leggera è principalmente uno stimolo per il dirigente, il tecnico, l'allenatore e l'atleta ad approfondire tali temi: da un lato acquisendo le nozioni statistiche di base per poter interagire con gli statistici e comprendere le analisi effettuate, dall'altro suggerendo agli statistici nuovi ulteriori spunti per l'applicazione dei metodi più avanzati.

Si possono, peraltro, fare alcune considerazioni su quelli che sembrano gli aspetti più interessanti emersi:

- 1) la maggior parte dei lavori analizzati ha affrontato un problema specifico con una metodologia giudicata dall'Autore ovviamente la più adeguata o, almeno, la più coerente. Difficilmente i lavori sono stati preceduti da una discussione sul tipo di dati raccolti e, soprattutto, sulla metodologia con cui sono stati raccolti, in altre parole sulla "qualità" dei dati stessi.
- 2) è importante suggerire l'applicazione di più metodi allo stesso insieme di dati per verificare la validità dei risultati e confrontarne la coerenza.
- 3) il limite di alcuni lavori in questo campo è di non aver ricercato le necessarie competenze tecniche e scientifiche nell'ambito del mondo sportivo: in prospettiva questo tipo di approccio risulterà più fecondo se vi sarà una stretta collaborazione fra statistici da una parte ed esperti e operatori dell'Atletica leggera dall'altra.

## Bibliografia

- Cerioni A., D'Arcangelo E., Sanna F.M. (1997) Lo studio dell'attività sportiva di alta prestazione: i contributi della metodologia statistica nella letteratura internazionale, *Statistica e sport: non solo numeri*, a cura di A. Mussino, Società Stampa Sportiva, Roma.
- Chatterjee S., Chatterjee S., (1982) New Lamps for Old: An Exploratory Analysis of Running Times in Olympic Games, *Applied Statistics*, 31, 1, pp. 14-22.
- Dawkins B.P., Andreae P. M., O'Connor P.M. (1994) Analysis of Olympic Heptathlon Data, *Journal of the American Statistical Association*, 89, 427, pp. 1100-1106.
- Robinson M.E., Tawn J.A. (1995) Statistics for Exceptional Athletics Records, *Applied Statistics*, 44, 4, pp. 499-511.
- Smith R.L. (1988) Forecasting Records by Maximum Likelihood, *Journal of the American Statistical Association*, 83, 402, pp. 331-338.
- Smith R.L., Corbett M., (1987) Measuring Marathon Courses: An Application of Statistical Calibration Theory, *Applied Statistics*, 36, 3, pp. 283-295.
- Smith R.L., Miller J.E. (1986) A Non-Gaussian State Space Model and Application to Prediction of Records, *Journal of the Royal Statistical Society*, B, 48, 1, pp. 79-88.
- Tryfos P., Blackmore R. (1985) Forecasting Records, *Journal of the American Statistical Association*, 80, 389, pp. 46-50.