

## Bachelor-Seminar: Statistik Im Sport

### Was Bradman Denied His Prime?

Sebastian Koch

Institut für Statistik, LMU München  
Prof. Friedrich Leisch, Manuel Eugster und Sebastian Kaiser

WiSe 2009/10

## Gliederung

- 1 Motivation
- 2 Die Sportart Cricket
- 3 Die Legende
- 4 Time Series Clustering
- 5 Time Series Clustering im Cricket
- 6 Ergebnis
- 7 Diskussion
- 8 Ausblick
- 9 Fazit

19.12.2009

Was Bradman Denied His Prime?

1

19.12.2009

Was Bradman Denied His Prime?

2

## Motivation

- Donald Bradman: Eine australische Cricket-Legende (Schnitt von 99.94 Runs pro Innings)
- Zweiter Weltkrieg unterbricht die Karriere
- Wie hätte Bradman in der Zeit gespielt
- Schätzung mit Hilfe von Time Series Clustering

## Geschichte

- Ursprung im Norden Europas im Mittelalter
- Organisation des Sports Ende 17. Jahrhundert
- 18. Jahrhundert: Cricket Nationalsport in England  
1787 Gründung des Marylebone Cricket Club (MCC)
- Cricket Nationen: England, Australien, Indien, Pakistan

19.12.2009

Was Bradman Denied His Prime?

3

19.12.2009

Was Bradman Denied His Prime?

4

# Test Ranking

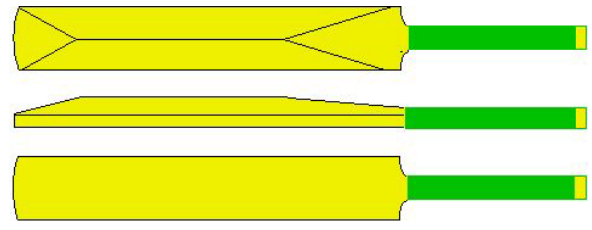
Reliance Mobile Test Championship

Team	Matches	Points	Rating
1 India	32	3957	124
2 South Africa	30	3672	122
3 Australia	31	3600	116
4 Sri Lanka	31	3574	115
5 England	39	4102	105
6 Pakistan	17	1424	83
7 New Zealand	25	2001	81
8 West Indies	25	1910	76
9 Bangladesh	19	255	13

Developed by David Kendix Last Updated: Tue, Dec 15, 2009

# Ausrüstung

- Schutzbekleidung
- Keine Nummern oder Vereinsnamen
- Ball aus Kork mit Leder ummantelt
- Der Schläger (Bat):

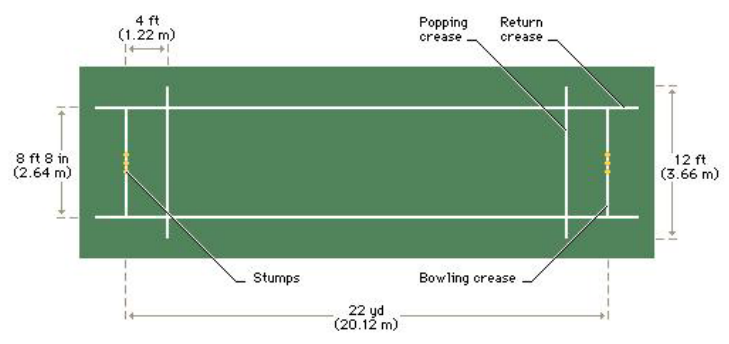


19.12.2009 Was Bradman Denied His Prime? 5

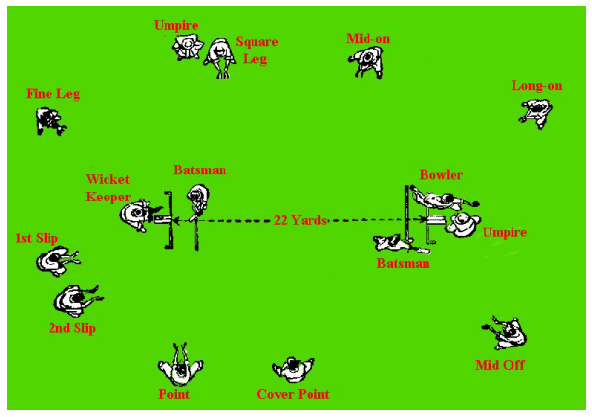
19.12.2009 Was Bradman Denied His Prime? 6

# Spielfeld

- Oval ca. 100-140 Meter lang
- Von Seil oder Ähnlichem umrandet
- Feldmannschaft gleichmäßig um Pitch verteilt



# Spielfeld



19.12.2009 Was Bradman Denied His Prime? 7

19.12.2009 Was Bradman Denied His Prime? 8



## Wickets

- Drei Holzstäbe (Stumps) in den Boden geschlagen
- Verbunden durch drei Stäbchen (Bails)
- Bails nur aufgelegt (nicht fest)
- Ball passt nicht zwischen Stumps hindurch

## Ablauf

- Innings besteht aus mehreren Overs (6 Bälle)
- Bowler wechselt nach jedem Over
- Striker wechselt wenn er out ist:
  - Wicket wird zerstört bevor er seinen Run vollendet hat
  - Er berührt den Ball nicht zuerst mit dem Schläger



## Ablauf

- Punkte:
  - Run: Wechseln der Pitch-Seite
  - Ball aus dem Feld schlagen (4/6 Runs)
- Ende eines Innings: Zehn Spieler der Schlagmannschaft sind out → Rollentausch
- Ziel: Mehr Runs als der Gegner erlaufen

## Austragungsformen

- Test: International
- First-Class Cricket: National
- One-Day-Cricket
- Twenty20 Cricket



# Die Legende



Donald Bradman

# Die Legende

- 1908(Cootamundra) - 2001(Adelaide)
- Erster Kontakt während der Schulzeit
- Spielte alleine zu Hause
- Sehr musikalisch
- 115 Runs im zweiten Spiel für Schulmannschaft

# Die Legende

- Ersatz für örtlichen Cricket Club
- 1921 Zuschauer beim Test Match Australien gegen England
- Mit 14 Arbeit → Chef erlaubt Spiele in Sydney
- 1928 erstes Test Match für Australien
- 1932 zog Don nach Sydney

# Die Legende

- Mehrere Rekorde gebrochen
- Superstar: Fans, Poster,...
- 1948 in England letztes Spiel
- Vier Runs benötigt für Durchschnitt von 100 Runs → Out for a Duck
- Wegen seiner Verdienste um Cricket und Australien zum Ritter geschlagen

# Prinzip des Time Series Clustering

- Zeitreihe: Abfolge diskreter Punkte über die Zeit
- Clustering: vorhandene Daten in Gruppen unterteilen homogen innerhalb der Cluster, heterogen zwischen den Cluster
- Bekannteste Clusteralgorithmen:
  - Hierarchical Clustering
  - K Means Clustering

# Prinzip des Time Series Clustering

- K Means Clustering: Vorher festlegen wieviele Cluster
- Hierarchical Clustering: aufteilen/zusammenfügen nach festgelegten Kriterien
- Bei beiden Gleiche Länge der Zeitreihen notwendig

# Hierarchical Clustering

- Entweder Grundgesamtheit in Gruppen aufteilen
- Oder einzelne Beobachtungen zu Gruppen zusammenführen
- Für Abstand der Cluster meistens verwendet: Euklidische Distanz

$$d(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \dots + (p_n - q_n)^2} \quad (1)$$

$$= \sqrt{\sum_{i=1}^n (p_i - q_i)^2} \quad (2)$$

# Methode von Ward

- Einteilung nach Hetrogenitätsmaß (Varianzkriterium)

$$V_p = \sum_{i=1}^{n_p} \sum_{j=1}^J (X_{ij} - \bar{X}_{jp})^2 \quad (3)$$

$X_{ij}$  = Beobachtungswert der Variablen  $j$  ( $j = 1, \dots, J$ ) bei Objekt  $i$  (für alle Objekte  $i = 1, \dots, n_p$  in Gruppe  $p$ )  
 $\bar{X}_{jp}$  = Mittelwert über die Beobachtungswerte der Variablen  $j$  in Gruppe  $p$  ( $= \frac{1}{n_p} \sum_{i=1}^{n_p} X_{ij}$ )

## Die Methodik

- Methode zum Time Series Clustering nach Wang et al. (robust gegenüber fehlenden Daten)
- Prinzip der Extrahierung von Eigenschaften, welche Gleichheit im Strukturlevel für die Zeitreihendaten abschätzt (beruhend auf globaler Extrahierung von Eigenschaften oder Modellparametern)
- Verschiedene Methoden zur Gewinnung der Eigenschaften → sehr Rechenaufwendig

## Die Methodik

- Eigenschaften: Trend, Saisonabhängigkeit, Periodizität, Reihenkorrelation, Schiefe und Kurtosis
- Durch Gewinnung sinnvolle Reduzierung der Dimensionen:
  - lange oder verschieden lange Datensätze auf eine beschränkte Anzahl von Messungen reduzieren
  - die Sensitivität gegenüber dem Rauschen wird geringer
  - Eigenschaften der ursprünglichen Zeitreihe erhalten
- Nur wenige Bedingungen nötig

## Time Series Clustering im Cricket

Die Daten:

- 20 internationalen Cricketspieler
- mindestens 70 Innings bei einer Karrieredauer von mehr als 17 Jahren
- Durchschnitt von über 40 Runs zum Stichtag (1. Januar 2009)

## Durchschnitt für $i$ -ten Spieler

$$A_i = \frac{\sum_{j=1}^n R_{i,j}}{n - k} \quad (4)$$

$R_{i,j}$  Anzahl Runs,  $i$ -ter Spieler,  $j$ -tes Innings  
 $n$  Gesamtanzahl an Innings, davon  $k$  not out

# Beitrag eines einzelner Spielers zur Teamleistung für ein Innings Schlageffizienz

$$C_{i,j} = \frac{S_{i,j}}{S_{T,j}} \quad (i = 1, 2, \dots, 11, j = 1, 2) \quad (5)$$

$S_{i,j}$  Gesamtanzahl an Runs, vom  $i$ -ten Batsman im  $j$ -ten Innings  
 $S_{T,j}$  Runs des Teams

$$P_i = \frac{\sum_{j=1}^n C_{i,j}}{n} \quad (6)$$

Beitrag eines Spielers hauptsächlich zur Glättung

## Modellfitting

- Schätzung von Höhen und Tiefen durch gewichtete Kleinste Quadrate Regression
- skaliertes mittlere Beitrag pro Kalenderjahr  $P$
- so skaliert, dass der Wert eines jeden Individuum zwischen 0 und 1

## Modellfitting

- Vier polynomiale Terme werden verwendet:  $T^{-2}, T^{-1}, T, T^2$   
 $T$  entspricht Jahr in der Karriere des Batsman
- Gewichtung: Anzahl der gespielten Innings in jedem Jahr
- Ein Intercept ist in dem Modell nicht berücksichtigt

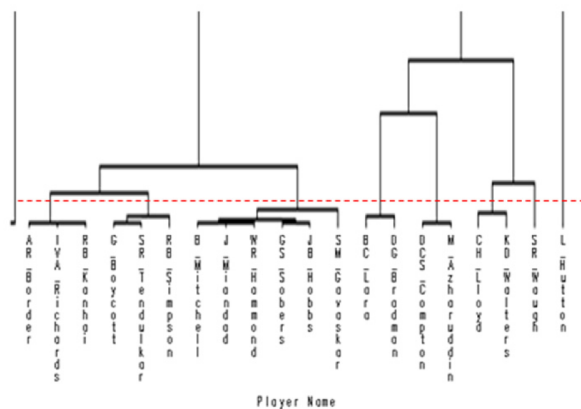


# Time Series Clustering

Name	T <sup>2</sup>	T <sup>1</sup>	T	T <sup>2</sup>	Auto r	Skew	Kurt	P-Value	R-Sq	Clust#
SM Gavaskar	1.19	-0.33	0.12	-0.01	0.75	-0.17	1.85	<.0001	0.91	3
WR Hammond	0.03	0.66	0.14	-0.01	0.94	-1.38	1.36	<.0001	0.89	3
BC Lara	-1.31	1.18	0.08	0	0.65	-2.96	10.54	<.0001	0.8	4
DG Bradman	-0.97	1.13	0.06	0	0.58	-2.58	8.61	<.0001	0.84	4
DCS Compton	0.77	0.05	0.06	0	0.28	2.53	9.67	<.0001	0.79	5
M Azharuddin	0.4	0.51	0.06	0	0.74	2.27	8.9	<.0001	0.82	5
CH Lloyd	0.4	0.54	0.05	0	0.56	1.56	4.11	<.0001	0.88	6
KD Walters	-1.37	2.24	0.03	0	0.97	2.22	4.06	<.0001	0.86	6
L Hutton	-2.32	2.22	0.07	0	0.12	-4.31	18.97	<.0001	0.86	-
SR Waugh	-0.54	0.44	0.08	0	0.92	-1.99	4.97	<.0001	0.91	-

- Sieben Variablen
- $p$ -Wert immer  $<0.0001$
- Einteilung der Cluster mit Ward's Minimum Varianz Methode

# Time Series Clustering



Einteilung der Cluster



# Time Series Clustering

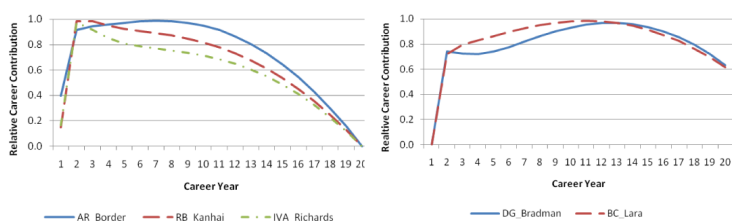


Figure 3a: Line plot displaying the fitted polynomial function for relative batting contribution by year for members of cluster 1.

Figure 3d: Line plot displaying the fitted polynomial function for relative batting contribution by year for members of cluster 4.

Zwei der sechs Cluster

- Sechs Cluster
- Gefittete polynomiale Funktionen übereinander gelegt

# Bradman's Cluster

- Bradman in Cluster vier mit Lara
- Bradman und Lara große Legenden im Cricket
- Steiler Start und nur schwaches Abfallen



### Schätzen über Bradman's Karrierelücke

- Da ähnlich zu Lara, vermutlich Karrierhöhepunkte währen des Zweiten Weltkrieges
- Durchschnitt an Runs pro Jahr mit Hilfe von linearer Regression schätzen
- Mittlere Anzahl an erzielten Runs pro Innings verwendet, um angemessenen Vergleich zwischen den Jahren mit unterschiedlich vielen Spielen ziehen zu können
- Standardfehler der geschätzten Steigung benutzt, um Konfidenzintervalle zu konstruieren
- Umwandeln in traditionellen Schlagdurchschnitt

### Schätzen über Bradman's Karrierelücke

Schätzung für den  $i$ -ten Batsman im  $s$ -ten Karrierejahr:

$$\hat{A}_{i,s} = \frac{\hat{T}_{i,s}}{\hat{n}_{i,s} - \hat{k}_{i,s}} \tag{7}$$

$\hat{T}_{i,s}$  ist entweder die beobachtete Anzahl an erzielten Runs,  $T_{i,s}$ , oder

$$\hat{T}_{i,s} = \hat{M}_{i,s} \hat{n}_{i,s} \tag{8}$$



### Konfidenzintervalle

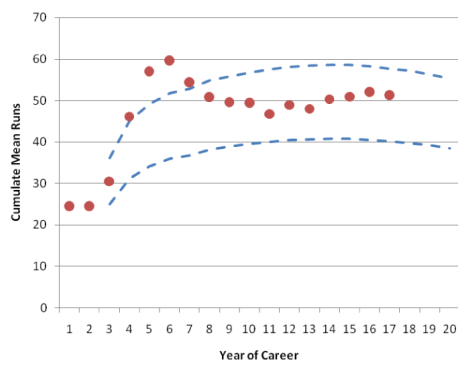


Figure 5: Plot of the observed cumulative mean runs against the career year of Lara with 95% confidence interval limits from the estimation model overlaid.

### Konfidenzintervalle

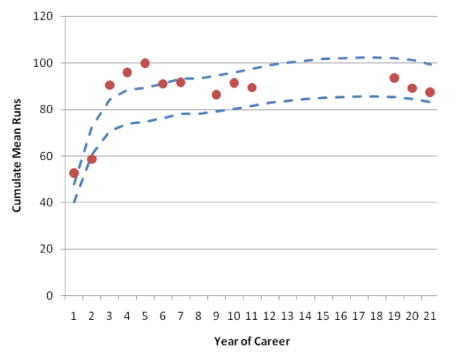


Figure 4: Plot of the observed cumulative mean runs against the career year of Bradman with 95% confidence interval limits from the estimation model overlaid.

## Ergebnis

Mit den Schätzungen ist es möglich zu testen, ob Bradman einen höheren Schlagdurchschnitt gehabt hätte, wäre seine Karriere nicht durch den Zweiten Weltkrieg unterbrochen worden

- Geschätzter Schlagdurchschnitt: 105.41
- $p$ -Wert von 0.2763 → nicht signifikant

## Diskussion des Time Series Clustering

- Diese Methode bewusst anderen bevorzugt
- Nur wenige Spieler mit über 17 Jahren Test Erfahrung
- Karriere als Ganzes und nicht als Stückwerk (Alter, Status der Karriere)
- Höhen und Tiefen bedingt durch frühere Leistungen (Interaktion zwischen den Jahren)
- Polynomiale Funktionen hier nützlich (andere Methoden nicht betrachtet)

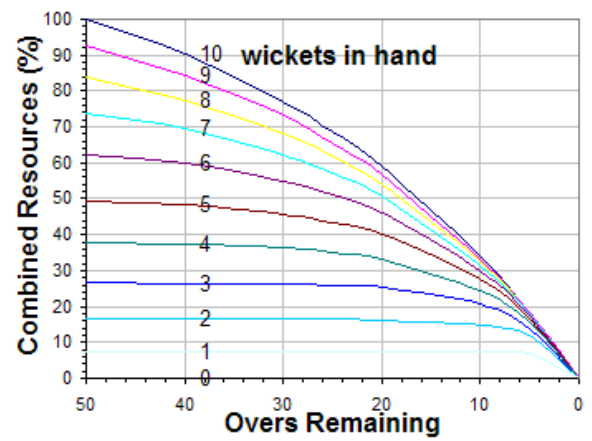
## Ausblick: Cricket

- Übertragbar für Frauen im Cricket
- Ebenso für Bowler
- Methoden zur Untersuchung ganzer Teams oder eines Matches → Duckworth-Lewis Methode

## Duckworth-Lewis Methode

- Anwendung im Tenty20 Cricket und One-Day-Cricket
- Schätzen des Spielausgangs, wenn das Spiel abgebrochen wird
- Variablen: verbleibenden Overs und Wickets  $\sim$  Runs in einem Innings
- D/L-Tabelle, welche auf 50 Overs normiert ist (50 Overs entsprechen 100% Ressourcen)
- In professionellen Spielen wird die „Professional Edition“ verwendet (Nicht so allgemein, wird für entsprechendes Spiel per Computer berechnet)

## Duckworth-Lewis Methode



## Ausblick: Außerhalb von Cricket

- Anwendung im Baseball (Schlageffektivität beim Hitting, gute Statistik beim Pitching)
- Andere Teamsportarten, da Leistung des Einzelnen im Bezug auf die Teamleistung → Keine Einzelsportarten
- Finanzmärkte und Medizin
- Allgemein Untersuchung von Ähnlichkeit von Zeitreihen

19.12.2009 Was Bradman Denied His Prime? 41 19.12.2009 Was Bradman Denied His Prime? 42

## Fazit

- Daten von 20 internationalen Spielern untersucht (mindestens 17 Jahre, über 70 Innings) (Stichtag: 1. Januar 2009)
- Time Series Clustering zeigt: Bradman's Karriere ähnlich zu Lara's
- Methode erzeugt instinktive Clusterergebnisse
- Standardisierung durch mittleren Beitrag zur Teamleistung → Negierung von Einflüssen wie Bedingung der Pitch

## Fazit

- Mittlerer Beitrag pro Jahr durch Spannweite standardisiert
- Polynomiale Funktion → Cluster anhand von Eigenschaften
- Bradman:
  - Höhepunkte vermutlich während Zweitem Weltkrieg
  - Geschätzter Wert 105.41 (nicht signifikant unterschiedlich zu 99.94)