

Was Bradman Denied His Prime?



Bachelor-Seminar: Statistik im Sport
Prof. Friedrich Leisch, Manuel Eugster und Sebastian Kaiser

Sebastian Koch

13. Dezember 2009

Inhaltsverzeichnis

1	Motivation	2
2	Die Sportart Cricket	3
2.1	Geschichte	3
2.2	Regeln	3
2.3	Austragungsformen	6
3	Was Bradman Denied His Prime?	7
3.1	Die Legende	7
4	Time Series Clustering	10
4.1	Prinzip	10
4.2	Die Methodik	11
5	Time Series Clustering im Cricket	13
5.1	Die Daten	13
5.2	Die Anwendung im Cricket	13
5.2.1	Beschreiben der Schlageffizienz	13
5.2.2	Modell-Fitting	14
5.2.3	Time Series Clustering	14
5.2.4	Diskussion der Cluster	15
5.2.5	Bradman's Cluster	15
5.2.6	Schätzung über Bradman's Karrierelücke	16
5.3	Ergebnis	17
6	Diskussion des Time Series Clustering	18
7	Ausblick	19
7.1	Cricket	19
7.2	Außerhalb von Cricket	20
8	Fazit	21
Literatur	21
Anhang	23

Kapitel 1

Motivation

„I don't think Don saw it properly. He seemed to have tears in his eyes“. [Bradman Foundation] Donald Bradman ist eine australische Cricket-Legende und wird von den meisten als der größte Cricket-Spieler aller Zeiten bezeichnet. Das Zitat stammt von Hollies, welcher der Bowler bei Bradmans letztem Test war. In seinem letzten Test Match hätte er mit nur vier Runs einen Durchschnitt von 100 in seiner Test-Karriere erreicht. Jedoch schaffte er nicht einen einzigen. Bis heute werden alle herausragenden Cricketspieler mit ihm verglichen. Es gibt aber einen Unterschied zwischen Bradman und den meisten anderen Spielern. Seine internationale Karriere wurde durch den Zweiten Weltkrieg für sechs Jahre unterbrochen. Im folgenden wird auf das Paper „Journal of Quantitive Analysis in Sports, Vol. 5 [2009], Iss. 4, Art. 3: Bracewell et al.: Was Bradman denide His Prime?“ eingegangen, in welchem untersucht wird, wie sich wohl Donald Bradmans Karriere - und damit sein Durchschnitt - entwickelt hätte, wäre seine Laufbahn nicht unterbrochen worden. Hierzu verwendet man Time Series Clustering. Es werden verschiedene Spieler analysiert und anschließend mit Don Bradman verglichen. Hieraus schätzt man anschließend den Verlauf und somit den Durchschnitt, den er erreicht hätte. Es zeigt sich, dass er die 100 Marke geknackt hätte, aber dieses Ergebnis nicht signifikant ist.

Im Folgenden wird sich, soweit nicht anders bemerkt, auf das oben genannte Paper bezogen und alle Daten stammen von der Internetseite www.Cricinfo.com. Nach einem Überblick über Regeln des Cricket und das Leben von Donald Bradman, wird auf die Methodik und Auswertung eingegangen, bevor das Ergebnis betrachtet wird und ein kurzer Ausblick gegeben wird. Ein Fazit beendet schließlich diese Arbeit. Zu Beginn wird auf die Entwicklung dieses faszinierenden Sports eingegangen.

Kapitel 2

Die Sportart Cricket

2.1 Geschichte

Die Geschichte des Cricket begann im Norden Europas. Die Ursprünge liegen vermutlich im Mittelalter, vielleicht auch schon in der Zeit nach dem Römischen Reich. Die ersten schriftlichen Aufzeichnungen liegen aus dem 16. Jahrhundert vor, als bei einem Gerichtsfall von „Kreckett“ berichtet wurde, welches an einer Schule in Guildford ausgeübt wurde. Bauern und Hirten spielten eine Vorform bereits um 1300. Zu dieser Zeit war „creag“ ein beliebter Zeitvertreib von Prinz Edward.

Wie es dazu kam, dass mehr als zwei Spieler spielen, die Punktevergabe und weitere Grundregeln des heutigen Crickets, ist unbekannt. Im Laufe des 17. Jahrhunderts wurde Cricket im Südosten Englands immer populärer und gegen Ende dieses Jahrhundert organisierte man sich in dieser Sportart. Von da an gab es die ersten Profis und im Jahre 1697 wurde ein „great cricket match“ mit 11 Spielern pro Mannschaft in Sussex ausgetragen. Im folgenden Jahrhundert wurden die wesentlichen Grundsätze des Spiels entwickelt und Cricket wurde zum Nationalsport in England.

Das Interesse an dieser Sportart verbreitete sich schnell im British Empire. So kommt es, dass die größten Cricket-Nationen heutzutage unter anderem Süd Afrika, Indien, Australien, Pakistan und England sind. Auch sind regionale Spiele eher uninteressant. Am meisten fiebern die Fans den internationalen Spielen entgegen.

Für die verschiedenen Spielarten gibt es im Detail unterschiedliche Regeln. Doch im Prinzip beruhen alle auf gemeinsame Grundregeln. „Seit seiner Gründung im Jahre 1787 gilt der Marylebone Cricket Club (MCC) als die alleinige Autorität im Aufstellen und Ändern dieser Regeln. Der Club besitzt das Welt-Copyright.“

2.2 Regeln

Am ehesten ist Cricket mit Baseball zu vergleichen, auch wenn die Regeln sehr unterschiedlich sind. Es gibt keine Spielernummern oder Vereinsnamen auf den Trikots. Meist sind es weiße Hemden zu denen man lange weiße Hosen trägt. Ist das Wetter kühler, so trägt man

oft noch einen weißen Pullover mit dem typischen Cricket-Strickmuster. Die weißen Schuhe haben oft Spikes. Die Schutzkleidung besteht unter anderem aus einem Helm, Schienbeinschonern und Handschuhen. Diese schützen den Batsman vor dem harten Ball. Dieser ist für gewöhnlich aus Kork, mit rotem Leder ummantelt und von vier weißen Kordeln umgeben. Der Umfang beträgt in etwa 22,7 cm und das Gewicht liegt um die 160 g.

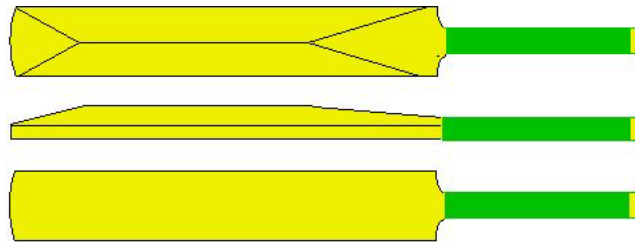


Abbildung 2.1: Der Bat: Hinten, Seite, Vorne

Der Schläger (Bat) besteht aus Weidenholz (Cricketwillow) und darf höchstens 96,5 cm lang sein. Die Schlagfläche (Vorderseite) ist flach und höchstens 10,8 cm breit. Auf der Rückseite ist das Bat durch eine Verdickung zur Mitte hin verstärkt.

Das Spielfeld ist in seinen Ausmaßen nicht festgelegt. Meistens ist das Oval zwischen 100 und 140 Meter lang. Dieses Feld ist mit einem Seil oder einer ähnlichen Absperrung begrenzt. In der Mitte liegt ein genau definierter Bereich namens Pitch.

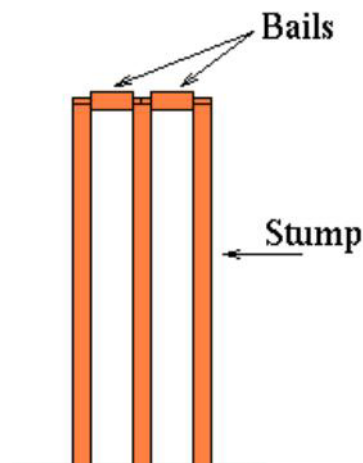


Abbildung 2.2: Das Wicket

Die Pitch ist ein sorgfältig präparierter Bereich. Dieser ist ca. 20 Meter lang. An beiden Enden stehen die sogenannten Wickets. Diese sind der wichtigste Bestandteil des Spiels. Die Wickets bestehen aus drei 71,1 cm hohen Holzstäben (Stumps) mit einem Durchmesser

von 3,6 cm. Diese werden gleichmäßig in den Boden geschlagen und bilden somit das 22,9 cm breite Wicket, durch welches der Ball nicht hindurch kann. In kleinen Vertiefungen auf den drei Stäben liegen als Verbindung zwei Querhölzer, die sogenannten Bails. Diese sind 11 cm lang und stehen im Mittelpunkt des Spiels.

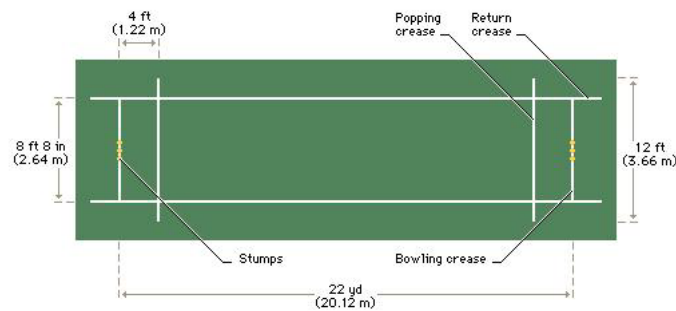


Abbildung 2.3: Die Pitch

Die Enden der Pitch sind in einzelne Bereiche unterteilt. Die Wickets stehen in der Mitte der 2,46 Meter langen bowling crease. 1,22 Meter in Richtung des gegenüberliegenden Wicket ist die Popping Crease. Die sogenannte Return Crease ist 3,66 Meter lang, rechtwinklig zu den anderen beiden Creases und endet frühestens 1,22 Meter hinter dem Wicket. Das Spiel beginnt mit dem Werfen einer Münze, wodurch entschieden wird, welche Mannschaft zuerst schlägt und welche auf dem Feld steht. Eine Mannschaft besteht aus elf Spielern. Neun Spieler der Feldmannschaft verteilen sich gleichmäßig auf dem Oval. Der Bowler steht an einem Ende der Pitch und der Wicket-Keeper steht am anderen Ende. An beiden Enden steht auch ein Batsman der Schlagmannschaft an der Popping Crease (Schlaglinie). Einer von beiden steht vor dem Wicket dem Bowler gegenüber und ist am Schlagen. Dieser wird Striker genannt. Ihm gegenüber steht der Non-Striker neben dem Wicket. Hier steht auch einer der beiden Schiedsrichter (Umpire). Der andere steht auf dem Feld in Höhe des Strikers.

Der Bowler wechselt nach einem Over (6 Bälle). Der Ball wird so geworfen, dass er vor dem Striker einmal aufkommt. Nun versucht der Striker den Ball so zu schlagen, dass er möglichst oft mit seinem Gegenüber (Non-Striker) die Seiten tauschen kann. Damit ein Run zählt, müssen sie hinter die Popping Crease kommen. Hierbei kann auch der Schläger verwendet werden. Ob und wie oft zu einem Run angesetzt wird, entscheiden die Batsmen selber. Es kommt oft vor, dass der Ball zwar geschlagen wird, aber die Batsmen nicht laufen. Sobald der Ball von einem Feldspieler entweder direkt auf das Wicket geworfen wird und somit zerstört oder zum Feldspieler, der am Wicket steht, der dieses dann zerstört, ist der Batsman Out, der näher an einem Wicket ist. Dies ist nur möglich, wenn die Batsmen gerade am laufen sind. Berühren sie den Bereich hinter der Schlaglinie, so wird die Runs dem jeweiligen Batsman und somit seiner Mannschaft gutgeschrieben. Wirft ein Feldspieler am Wicket vorbei, so haben die Batsmen die Möglichkeit weitere Runs zu erzielen. Eine

andere Möglichkeit zu punkten besteht darin, dass der Striker den Ball über die Feldgrenze schlägt. Hierfür bekommt er Sechs Runs, außer der Ball berührt vorher den Boden, dann gibt es vier Runs. Tritt dieser Fall ein, so brauchen die Batsmen auch nicht mehr zu laufen. Der Batsman ist Out, wenn der Bowler bei seinem Wurf das Wicket zerstört. Auch wenn der Striker absichtlich oder unabsichtlich vom Ball getroffen wird ist er Out. Die Schiedsrichter (Umpire) haben das alleinige Sagen auf dem Platz. Die Spieler dürfen keinen Widerspruch geben.

2.3 Austragungsformen

Cricket ist ein Sport, bei dem der Augenmerk hauptsächlich auf internationalen Spielen liegt. Regionale Spiele sind von eher geringem Interesse.

Die traditionell höchste Austragungsform ist das sogenannte Test Cricket. Als Test oder Test Cricket bezeichnet man eine spezielle Form eines internationalen Cricket-Matches. Ein Test geht über fünf Tage. Er wird in jeweils drei rund zweistündigen Blöcken ausgetragen. Tests sind wiederum meist in eine Serie von zwei bis sechs Tests eingebettet, so dass die entsprechenden Duelle sich über mehrere Wochen hinziehen können. Auf nationaler Ebene wird First-Class Cricket gespielt. Diese dauert mindestens drei Tage.

Tests finden nur zwischen wenigen dazu spielberechtigten Nationalmannschaften statt. Aktuell sind dies (in Rangfolge vom 7. Dezember 2009) Indien, Südafrika, Australien, Sri Lanka, England, Pakistan, Neuseeland, Westindische Inseln und Bangladesh.

Anstelle einer Weltmeisterschaft, führt der internationale Cricketverband ICC eine Art Weltrangliste (Test Championship), bei der fortlaufend alle Test-Matches berechnet und die Reihenfolge der Nationen angegeben wird. Der aktuell Führende ist Besitzer der Test Trophy.

Als bedeutenste Test-Match-Serie gelten die Ashes. Hierbei spielen seit 1877 jährlich England gegen Australien um die Ashes (ein Urnenförmiger Pokal). Dieser Pokal enthält die Asche des Wickets, welches die Engländer aus Frust über das erste gegen ein ausländisches Team verlorene Test-Match verbrannten. Diese Team war Australien.

Um den modernen Zeiten und besonders dem Fernsehen gerecht zu werden, wurde ein kürzeres und dramatischeres Format, das One-Day-Cricket, eingeführt. Dieses wird immer populärer, jedoch von Traditionalisten sehr skeptisch begutachtet. Im Gegensatz zum Test Cricket ist das Match hier nicht erst beendet, wenn alle Batsmen out sind, sondern schon nach einer festgelegten Zeit von gewöhnlich 50 Overs. Wie beim Test Cricket wird auch One-Day-Cricket meist in Form einer Serie von drei bis sieben Matches ausgetragen.

Alle vier Jahre wird im One-Day-Modus eine Cricket-Weltmeisterschaft ausgetragen. In den vier Jahren zwischen der WM findet eine Champions Trophy im K.-O.-System statt.

Um den Fernsehgewohnheiten gerechter zu werden, gibt es einen Versuch des englischen Cricket Verbands (ECB). Diese neue Variante heißt Twenty20 Cricket. Hier werden die Innings auf je 20 Overs verkürzt und eine Maximalspieldauer von 75 Minuten je Innings festgelegt. Strafen drohen dem angreifenden Team, welches nicht alle Overs in der festgelegten Zeit schafft. Diese Form findet bislang nur wenig Anklang in der Welt des Cricket.

Kapitel 3

Was Bradman Denied His Prime?

3.1 Die Legende

Donald George Bradman wurde am 27. August 1908, als fünftes Kind von George und Emily Bradman, in Australien geboren. Sein Vater arbeitete von 1911 an in einer Holzfabrik in Bowral. Der Besitzer der Fabrik, Alf Stevens, war Mitglied im örtlichen Cricket Club. Nach kurzer Zeit war auch George dort Mitglied und nahm Don mit zu Spielen. Glebe Oval war nur eine Straße von Bradmans Heim entfernt und wurde 1947 zu Bradman Oval.

Don hatte, abgesehen von den Berichten von den Kriegsschauplätzen des Ersten Weltkrieges, eine behütete Kindheit in dieser ländlichen Stadt. Während der Schulzeit hatte er den ersten Kontakt mit Sport. Er spielte das sehr populäre Tennis und übernahm die Aufgabe, die erzielten Runs zu zählen. Zu Hause entwickelter er ein Einzelspiel, indem er einen Golfball mit einem Stumb gegen das gekrümmte Fundament des Wassertanks am Haus schlug. Er benutzte eine Hausseite als Begrenzung und kreierte in seinem Kopf Test Matches, indem er sich den unberechenbaren Bällen, die vom Tank zurückkamen, als Batsman entgegen stellte. 1924 zog die Familie direkt gegenüber des Cricket Feldes. Zu dieser Zeit war Don Bradman schon in der Region als leistungsfähiger Cricketspieler bekannt, da er erfolgreich an Schulspielen teilnahm. Auch überregional wurde man erstmals auf ihn aufmerksam.

Nebenbei war Donald eine fleißiger Klavierschüler, sang im Chor und half seinem Vater bei diversen Jobs. Er entwickelte eine große Liebe für die Musik. Damals sah er seine Zukunft noch im Malergeschäft.

Im Alter von zwölf Jahren wurde er zu Spielen der Schulmannschaft eingeladen. Bereits bei seinem zweiten Spiel erreichte Don 115 Runs. Das Team erreichte zusammen 150. Im gleichen Jahr lernte er seine zukünftige Frau, Jessie Menzies, kennen und entschied sich, sie zu heiraten. Er wurde immer öfter auch im örtlichen Cricket Club, in welchen sein Vater und zwei seiner Onkel spielten, eingesetzt, wenn Spieler ausfielen.

Im Februar 1921 fuhr Bradman mit seinem Vater zum ersten Test Match. Es war das fünfte Aufeinandertreffen zwischen England und Australien. Von da an war es sein Traum auf

diesem Oval zu spielen.

Mit 14 verließ Donald die Schule und arbeitete für Mr Percy Westbrook in einer Immobilienfirma. Mr Westbrook spielte eine entscheidende Rolle in Bradmans Karriere. Er erlaubte ihm in Sydney zu spielen, als das Angebot kam. In seiner Heimat bekam er immer mehr Aufmerksamkeit. Seine Mutter versprach ihm einen neuen Bat, sollte er bei einem Finale, welches über fünf Sonntage ging, ein Century (100 Runs) schaffen. Er erreichte in einem Innings 300 Runs. Don tat alles für seine Leidenschaft. Um in Sydney spielen zu können stand er um Fünf Uhr auf und kam oft nicht vor Mitternacht wieder zurück. Die Zugfahrt wurde ihm bezahlt. 1932 zog er dann nach Sydney.

Bereits 1928 spielte Bradman sein erstes Test Match für Australien. Da die Wetterbedingungen sehr schlecht waren und ein spezielles Wicket benutzt wurde, was Don nicht kannte, machte er nur sehr wenige Runs. In Folge dessen wurde er zum ersten und einzigen Mal beim nächsten Test nur als Reserver mitgenommen. Schon im nächsten Spiel überzeugte er mit über 100 Runs im zweiten Innings. Im Januar 1930 erzielte er seine höchste Anzahl an Runs. In nur 415 Minuten schafft er 452 ohne out zu sein. Den Rekord bis dahin hielt Bill Ponsford, welcher 437 Runs in 621 Minuten erreichte.

Bei seiner ersten Auslandsreise nach England spielte er weitere Rekorde ein und wurde zu einem Superstar. Es gab Poster und sogar ein Lied über ihn. Er wurde zum australischen Nationalheld und konnte sich vor Presse und Fans kaum noch retten. Er reiste durch das ganze Land.

Auf der zweiten Reise in England hatt sich der Kapitän der englischen Mannschaft eine Art des Werfens überlegt, die später als Bodyline bekannt wurde. Hierbei bowled der Bowler schnell und hart auf den Körper des Strikers. Diese wurden dabei oft durch den harten Ball verletzt. Nach dem Turnier, welches Bradman mit einem Schnitt von nur 56.57 Runs abschloss, schrieben die Australier dem Marylebone Cricket Club, welcher anschließend eine umfassende Regeländerung vornahm. Die Engländer gaben niemals zu, dass diese Art zu bowlen unfair sei.

Alle Test matches wurden für die Zeit des Zweiten Weltkrieges unterbrochen. Als die Spiele wieder aufgenommen wurden, ernannte man Bradman zum Kapitän der australischen Mannschaft. Er beschloss, dass die Tour in England 1948 seine letzten Spiele sein sollen. Er genoss es, seine Karriere als Kapitän einer auf dieser Tour ungeschlagenen Mannschaft zu beenden. Das ganze Team spielte hervorragend.

Das fünfte Aufeinandertreffen von England und Australien sollte das letzte Test Match in Don Bradmans Karriere sein. Als er das Oval betrat bekam er sieben Minuten lang tösenden Applaus. Er brauchte nur 4 Runs um einen Schnitt von 100 Runs in seiner Karriere zu haben. Der zweite Ball vom Batsman Eric Hollies war „googly“ (kullernd) geworfen und Don Bradman schied mit einem „out for a Duck“ (kein Run) aus. Somit hatte er einen durchschnitt von 99.94. Bradman sagte 1950 in einem Interview, dass er sich wohl zu viel Druck gemacht habe.

Don Bradman wurde als einziger australischer Cricketspieler zum Ritter geschlagen, da er sich um den Cricketsport und um die sportliche Zusammenführung des Commonwealth verdient gemacht hat.

Nach seiner sportlichen Karriere arbeitet Bradman erfolgreich als Börsenmakler und Fir-

menmanager. Don Bradman starb am 25. Februar 2001 in Adelaide. Es stellt sich bis heute die Frage, ob Bradman einen Schnitt von 100 oder mehr Runs geschafft hätte, wenn seine Karriere nicht durch den Zweiten Weltkrieg unterbrochen worden wäre. Hierzu versucht man im „Journal of Quantitative Analysis of Sports“ diese Lücke mit Hilfe von Time Series Clustering zu schließen.

Kapitel 4

Time Series Clustering

4.1 Prinzip

Als Zeitreihe (Time Series) bezeichnet man eine Abfolge von diskreten Punkten über die Zeit. Es können allgemeine Trends auftreten, aber auch Saison bedingte Trends. Hierbei sind als Beispiele Börsenkurse oder Wetterdaten genannt.

Beim Clustering geht es darum, die vorhandenen Daten in Cluster (Gruppen) zu unterteilen. Hierbei gilt als allgemeiner Grundsatz: homogen innerhalb der Cluster, heterogen zwischen den Cluster. Es gibt viele verschiedene Arten von Clusteralgorithmen. Die bekanntesten sind Hierarchical Clustering und K Means Clustering.

Bei letzterem handelt es sich um den weitverbreitetsten Algorithmus. Hierbei wird die Anzahl der Cluster vor Beginn festgelegt. Auch muss die Länge der Zeitreihen gleich sein, um euklidische Abstandsberechnungen anzustellen. Hingegen ist man beim Hierarchical Clustering frei vom Festlegen von Parametern. Jedoch ist es auch hier notwendig, dass zum Vergleich die einzelnen Zeitreihen die gleiche Länge vorweisen. Man kann diesen Algorithmus in zwei Richtungen anwenden. Entweder es wird von einer Grundgesamtheit ausgegangen und die nach und nach in kleinere Cluster unterteilt, oder man betrachtet einzelne Daten als Cluster und fügt sie Schritt für Schritt zu größeren Gruppen zusammen, bis man entweder eine genügend kleine Anzahl von Clustern hat oder der Abstand zwischen den Clustern groß genug ist. Hierfür verwendet man üblicherweise die euklidische Distanz:

$$d(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \dots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^n (p_i - q_i)^2} \quad (4.1)$$

An der Formel sieht man schon, dass die Zeitreihen die gleiche Dimension benötigen.

Für die vorliegende Problemstellung wird unter anderem die Methode von Ward dazu verwendet, die untersuchten Spieler in Cluster einzuteilen. Ward benutzt das hierarchische Clustering. Bei dieser Methode werden die Cluster nicht nach den geringsten Abständen zwischen den Clustern eingeteilt, sonder anhand eines Heterogenitätsmaßes. Es wird versucht Cluster, ausgehend von den einzelnen Objekten, also unter Verwendung eines aggro-

merativen Algorithmus, so zu bilden, dass die Varianz innerhalb der Gruppen bei Hinzunahme eines weiteren Objektes am wenigsten erhöht wird. Dies unterstützt die Grundidee von Homogenität innerhalb der Gruppen. Als Maß für die Heterogenität wird das Varianzkriterium (Fehlerquadratsumme) verwendet:

$$V_p = \sum_{i=1}^{n_p} \sum_{j=1}^J (X_{ij} - \bar{X}_{jp})^2 \quad (4.2)$$

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

4.2 Die Methodik

Liao hat die vorigen Methoden in drei große Kategorien unterteilt:

- 1 Annäherung auf Grundlage von Rohdaten: Clustering Methoden, die die Rohdaten umgestalten.
- 2 Modellbasierte Annäherung: Zeitreihen werden von einem Modell beschrieben und Clustering auf Modellparametern ausgeführt
- 3 Auf Eigenschaften basierende Annäherung: Aus den Rohdaten wird eine Reihe von Eigenschaften extrahiert, welche für das Clustering benutzt werden.

Von Wang und anderen wurde ein Vorschlag zur Annäherung gemacht, welcher darauf abzielt eine Methode zum Time Series Clustering bereitzustellen, welche robust gegenüber fehlenden Daten ist. Diese folgt dem Prinzip der Extrahierung von Eigenschaften, welche Gleichheit im Strukturlevel für die Zeitreihendaten abschätzt, beruhend auf globaler Extrahierung von Eigenschaften oder Modellparametern. Es gibt verschiedene Methoden um statistische Eigenschaften herauszufinden, wie zum Beispiel Discret Wavelet Transform (DWT), Discret Fourier Transform (DFT), Auto Regression Moving Average (ARMA) und Compression based Dissimilarity Measures (CDM). Jedoch haben alle Methoden eine hohe Berechnungskomplexität und benötigen bestimmte Bedingungen, welche erfüllt sein müssen, damit diese Methode erfolgreich ist.

Die Extrahierung von Eigenschaften kann ebenso als Mittel betrachtet werden, um eine Reduzierung der Dimensionen in Zeitreihen zu erreichen. Indem man globale Maßzahlen, wie Trend, Saisonabhängigkeit, Periodizität, Reihenkorrelation, Schiefe und Kurtosis zusammen mit erweiterten Maßzahlen, d.h. nicht-lineare Strukturen, Selbstähnlichkeit und Chaos benutzt, kann eine brauchbare Menge an Eigenschaften einer Zeitreihe als Maßzahl gewonnen werden.

Die Gewinnung der zusammengefassten Eigenschaften von Zeitreihen bietet eine sinnvolle Reduzierung der Dimensionen. Damit kann man lange oder verschieden lange Datensätze

auf eine beschränkte Anzahl von Messungen reduzieren, was auch bedeutet, dass die Sensitivität gegenüber dem Rauschen geringer wird. Auf diese Weise bleiben auch die Eigenschaften der ursprünglichen Zeitreihe erhalten und man kann Ähnlichkeiten und Unterschiede zwischen den ursprünglichen Zeitreihen erkennen. Ziel ist es also durch Herausfiltern von Eigenschaften die ursprünglichen Eigenschaften der Zeitreihen nicht zu verlieren, um verschiedene Cluster Techniken anzuwenden. Die typischen globalen Eigenschaften sind im Folgenden beschrieben.

Trend und Saisonabhängigkeit gehören zu den üblichen Eigenschaften einer Reihe. Man spricht von einem Trend, wenn sich das mittlere Level über eine längere Dauer verändert. Dieser kann durch glatte nicht-parametrische Methoden nachgewiesen werden. Hingegen bedeutet Saisonalität, dass sich der Verlauf periodisch wiederholt. Um diese zu diagnostizieren wird Autokorrelation verwendet.

Periodizität ist eine grundlegende Eigenschaft der Zeitreihendaten, welche uns zeigt, wie oft eine Eigenschaft über die Zeit auftritt.

Reihenkorrelation beschreibt die Beziehung zwischen den aufeinanderfolgenden Werten in der Zeitreihe. Um den Grad der Reihenkorrelation zu ermitteln, bietet sich die Box-Pierce Statistik an.

Schiefe ist ein Hilfsmittel um die Symmetrie zu überprüfen. Sind die Daten normalverteilt, so ist die Schiefe Null. Ist sie negativ, so ist die Verteilung linksschief, ist sie positiv, so ist die Verteilung rechtsschief.

Kurtosis misst wie spitz die Verteilung ist. Ergibt sich ein hoher Wert, so hat die Verteilung einen ausgeprägten Gipfel in der Nähe des Mittelwertes und fällt scharf ab. Ein kleiner Wert weist auf einen flachen Gipfel in der Nähe des Mittelwertes hin.

Selbstähnlichkeit bedeutet, dass bei genauerer Betrachtung bzw. unterschiedlichen Skalen einer Dimension das gleiche Verhalten bzw. Aussehen zu erkennen ist.

Diese Arten von Clustering mit Hilfe von globalen Eigenschaften ist besser als andere Clusterverfahren, da sie nur wenige Bedingungen benötigt, um verwendet zu werden.

In dieser Studie verwendet man polynomiale Funktionen um die Zeitreihen für jedes Individuum zu approximieren, indem man die skalierte mittlere Schlaghäufigkeit pro Kalenderjahr benutzt. Die Benutzung von polynomialen Funktionen erlaubt es, gewichtete Regressionsmethoden zu benutzen, welche eine größere Gewichtung auf die Saison zu legen, in welcher das Individuum mehr Innings spielte. Die geschätzten Parameter dieses Modells, wurden sowohl als Clustering Eigenschaften benutzt, als auch als Reihenkorrelation, Schiefe und Kurtosis. Die Methode, die für diese Analyse gewählt wurde, dient dazu, zunächst die Daten für jeden Cricketspieler zu modellieren und anschließend Cluster Zeitreihen Analyse anzuwenden, welche auf globalen Zeitreiheneigenschaften nach Wang und anderen und den Parametern des gefitteten Modells aufbaut. Auf diesen Eigenschaften aufgebautes Clustering führt zu instinktiveren Ergebnissen.

Kapitel 5

Time Series Clustering im Cricket

5.1 Die Daten

Im Folgenden werden Daten von 20 internationalen Cricketspielern untersucht. Diese Spieler absolvierten mindestens 70 Innings bei einer Karrieredauer von mehr als 17 Jahren und einem Durchschnitt von über 40 Runs zum Stichtag (1. Januar 2009). Aus diesen Daten werden die Höhen und Tiefen der untersuchten Spieler untersucht und anschließend mit den Daten von Donald Bradmans Test-Karriere verglichen. Somit will man einen Schluss auf die Entwicklung dessen Karriere im Zeitraum des Zweiten Weltkrieges ziehen, als internationale Cricket Spiele eingestellt wurden.

5.2 Die Anwendung im Cricket

5.2.1 Beschreiben der Schlageffizienz

Zuerst wird die Schlageffizienz betrachtet. Hierzu berechnet man den Durchschnitt, welcher für den i -ten Player folgendermaßen definiert ist:

$$A_i = \frac{\sum_{j=1}^n R_{i,j}}{n - k} \quad (5.1)$$

$R_{i,j}$ ist die Anzahl von Runs, welche der i -te Spieler im j -ten Innings erzielt hat. n ist die gesammte Anzahl von Innings, in denen der i -te Spieler geschlagen hat und k die Anzahl der Innings in denen der Spieler am Ende nicht out war. Da versucht wird Spieler aus verschiedenen Gegenden zu vergleichen, hängt die Schlaghäufigkeit von den individuellen Einflüssen, wie Beschaffenheit der Pitch oder dem Einfluss unbewachter Wickets, ab. Nichtsdestoweniger wird Zurechnung verwendet um den Schlagdurchschnitt von Bradman, wie oben dargestellt zu berechnen.

Der Beitrag, den ein einzelner Spieler zur Teamleistung beigesteuert hat, wird benutzt, um äußere Einflüsse zu reduzieren. Der Beitrag eines Spielers für ein Innings wird wie folgt

berechnet:

$$C_{i,j} = \frac{S_{i,j}}{S_{T,j}} \quad (i = 1, 2, \dots, 11, j = 1, 2) \quad (5.2)$$

$S_{i,j}$ ist die Gesamtanzahl an Runs, die der i -te Batsman im j -ten Innings erlaufen hat und $S_{T,j}$ die Runs des Teams. Somit ergibt sich als Maß für die Schlageffizienz

$$P_i = \frac{\sum_{j=1}^n C_{i,j}}{n} \quad (5.3)$$

Der Beitrag eines Spielers wird hauptsächlich dazu benutzt, die Funktion zu Glätten. Dies ist ein grundlegender Aspekt von Cricket, auch wenn hier ein Einfluss der Mitspieler vorhanden sein kann. Da es sich um einen Teamsport handelt, müssen die Spieler zusammenarbeiten, wollen sie das Spiel gewinnen. Das Augenmerk liegt hier auf dem Einfluss der verschiedenen Batsman, wofür der Vergleich der Teamkameraden mit Hilfe deren Beitrags angebracht ist.

5.2.2 Modell-Fitting

Um die Höhen und Tiefen der Schlageffizienz in der Karriere eines Cricketspielers zu approximieren, verwendet man gewichtete Kleinste Quadrate Regression. Hiermit modelliert man den skalierten mittleren Beitrag pro Kalenderjahr, P , für die untersuchten Spieler. Der durchschnittliche Beitrag wird so skaliert, dass der Wert eines jeden Individuum zwischen 0 und 1 liegt. Vier polynomiale Terme werden verwendet: T^{-2} , T^{-1} , T und T^2 , wobei T für das Jahr in der Karriere des Batsman steht. Als Gewichtung benützt man die Anzahl der gespielten Innings in jedem Jahr. Ein Intercept ist in dem Modell nicht berücksichtigt. Ausgehend von diesem Modell wird der erwartete mittlere Beitrag benutzt um einen Bereich für andere globale Eigenschaften zu erzeugen: Schiefe, Wölbung und Autokorrelation. Hierzu wird keine zusätzliche Gewichtung verwendet. Abbildung 1 im Anhang zeigt die Ergebnisse.

Beispiele der Ergebnisse sind in Abbildung 2 als Bubbleplot dargestellt, in welche das gefittete Polynom eingezeichnet ist. Die Größe der Bubbles zeigt die relative Anzahl an Innings, welche in diesem Jahr der individuellen Karriere gespielt wurden.

5.2.3 Time Series Clustering

Anhand der sieben Variablen, welche die Eigenschaften der Karrierefortschritte, wie in Tabelle 1 gezeigt, beschreiben, unterteilt man die 20 Spieler in Cluster. Hierzu verwendet man Ward's Minimum Varianz Methode. Die Cluster wurden heuristisch erstellt, indem man die Spieler zusammenfasst, die optisch am ähnlichsten sind. Das Dendogramm in Abbildung 3 zeigt, wie die Cluster eingeteilt wurden.

Man sieht, dass sich daraus sechs Cluster ergeben. Zwei Spieler, Steve Waugh und Len Hutton, verbleiben ohne Clusterzugehörigkeit.

Wichtig anzumerken ist, dass diese Cluster auf der individuellen Leistung im Vergleich zu ihresgleichen, über die ganze Karriere hinweg, basieren. Jeder Spieler hat eine besondere Test-Karriere, was an 17 oder mehr Jahren deutlich wird. Dadurch weisen diese Cluster darauf hin, wann die Batsman wohl ihren größten Einfluss auf das Spiel hatten.

Die Plots in Abbildung 4 zeigen wie gut die einzelnen Cluster passen. Hierbei wurden die einzelnen gefitteten polynomialen Funktionen übereinander gelegt.

5.2.4 Diskussion der Cluster

Geht man nach dem Dendogramm, so sind die Cluster 3a und 3b eng verbunden. Die Plots zeigen jedoch einen deutlichen Unterschied im Verlauf der gefitteten Kurve. Hieran sieht man, dass die Einteilung der Cluster gut gewählt wurde. Für den visuellen Zweck wurde Mitchell's Kurve um vier Jahre nach vorne verschoben.

In Plot 3a wird Cluster 1 dargestellt. Hier fällt der große Einfluss gleich zu Beginn ihrer Karriere, bevor diese langsam nachlässt. Cluster 2 (3b) zeigt einen kurzen Abfall zu Beginn, gefolgt von einem parabelförmigen Verlauf, mit dem Höhepunkt in etwa der Mitte ihrer Karriere. Im nächsten Plot (3c) sieht man einen ähnlichen Start mit einem kleinen Tief, wobei hier der Einfluss von Anfang an viel höher ist als in Cluster 2. Die Batsman in Cluster 3 haben über ungefähr zehn Jahre einen großen Beitrag zur Teamleistung und fallen erst gegen Ende ab. Cluster 4 zeigt einen steilen Anstieg in den ersten Jahren der Karriere und einen recht stabilen, andauernd großen Beitrag zur Gesamtleistung des Teams. Auf diese Cluster wird im nächsten Abschnitt genauer eingegangen. Entgegengesetzt zu den Batsmen in Cluster 4, starten die zwei aus Cluster 5 sehr stark, fallen aber recht schnell ab und bleiben auf einem vergleichsweise niedrigem Niveau. Im letzten Cluster sieht man einen ähnlichen Start wie in Cluster 5. Hier jedoch verweilen die Kurven länger auf einem niedrigen Niveau, bevor sie zum Ende der Karriere wieder leicht ansteigen.

5.2.5 Bradman's Cluster

Da Donald Bradman in Cluster 4 wiederzufinden ist, will man nun auf dieses Cluster etwas genauer eingehen. Neben ihm ist als zweiter Batsman Brian Lara zu finden. Beide sind große Legenden im Cricket. Don Bradman wird in den Meisten Cricketkreisen als der größte Cricketspieler aller Zeiten angesehen. Seit seinem letzten Testmatch 1948 werden viele Spieler mit ihm verglichen, was nur als die Suche nach dem „nächsten Don Bradman“ bezeichnet werden kann. Bis zu seinem Karriereende spielte er mit einer solchen Liebe dieses Spiel und erreichte zahlreiche Rekorde und eine beeindruckende Statistik. Oft wird vergessen, dass seine Karriere durch den Zweiten Weltkrieg um sechs Jahre in der Mitte unterbrochen und somit verkürzt wurde. Dies macht es sehr schwer herauszufinden, wie seine Karriere ohne diese Unterbrechung verlaufen wäre.

Betrachtet man seine Karriere statistisch, so sieht man, dass er seine erste Saison mit einem Durchschnitt von gerade einmal 52 beendet hat. Sein bestes Jahr hatte er 1932 als er einen Durchschnitt von über 400 erreichte. Obwohl Bradman manche schwierige Serie hatte, wie zum Beispiel die berühmte „body line tour“ gegen England, schien es nie so als hätte er

eine längere Krise. Deshalb wurde er in Cluster 4 aufgenommen.

Brian Lara ist mit Tendulkar der größte Spieler seiner Generation. Lara hat sein erstes Test Century gegen Australien erreicht. Er schaffte hier beeindruckende 277 Runs. Tendulkar, welcher in Cluster 2 zu finden ist, erreichte sein erstes Doppel-Century erst nach über zehn Jahren seiner Karriere.

5.2.6 Schätzung über Bradman's Karrierelücke

Da Don Bradman's Karriere am ähnlichsten zu Lara's ist, könnte man annehmen, dass seine Karrierhöhepunkte im zwölften und 14ten Jahr gelegen hätten, also in der Zeit des Zweiten Weltkrieges. In diesem Abschnitt will man herausfinden, wie Bradman in diesem Zeitabschnitt gepunktet hätte und testen, ob dies seinen Schlagdurchschnitt signifikant geändert hätte.

Der erwartete Beitrag aus der polynomialen Funktion, welche Bradman's Karriere fittet, wird benutzt, um seinen Durchschnitt an Runs pro Jahr mit Hilfe von linearer Regression zu schätzen. Bei diesem Modell wurde der Intercept weggelassen. Die mittlere Anzahl an erzielten Runs pro Innings wird verwendet, um einen angemessenen Vergleich zwischen den Jahren, in denen unterschiedlich viele Spiele gespielt wurden, ziehen zu können. Aus diesem Modell benützt man den Standardfehler der geschätzten Steigung, um Konfidenzintervalle zu konstruieren.

Die durchschnittliche Anzahl an Runs muss in den traditionellen Schlagdurchschnitt umgewandelt werden, welcher die Anzahl an Runs pro Dismissal (bis der Batsman out ist) ist. Man erreicht dies, indem man die Anzahl der not outs in einem Kalenderjahr schätzt. Angenommen die Wahrscheinlichkeit, dass man nicht out ist, bleibt während einem Innings konstant. Dann schätzt man mit Hilfe des Anteils an beobachteten not outs eine Ober- und eine Untergrenze der Anzahl an Dismissals pro Jahr.

Um schließlich den wahrscheinlichen Runs in den fehlenden Saisons eine Gewichtung zu geben, wird die beobachtete mittlere Anzahl an Innings pro Jahr auf die nächst kleinere ganze Zahl gerundet, und anschließend dazu verwendet, die erwartete Gesamtanzahl an Runs pro Kalenderjahr zu erhalten. Davon ausgehend ist die untere Grenze für die Zahl an not outs die Zahl der beobachteten not outs, was bedeutet, dass der Batsman in dem berechneten Zeitraum nicht not out gewesen wäre.

Demzufolge ist die Schätzung für den i -ten batsman im s -ten Karrierejahr:

$$\hat{A}_{i,s} = \frac{\hat{T}_{i,s}}{\hat{n}_{i,s} - \hat{k}_{i,s}} \quad (5.4)$$

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

$$\hat{T}_{i,s} = \hat{M}_{i,s} \hat{n}_{i,s} \quad (5.5)$$

Die Schätzung für die Anzahl von gespielten Innings pro Kalenderjahr, $\hat{n}_{i,s}$, ist entweder die Anzahl an gespielten Innings in diesem Jahr oder der Durchschnitt für den i -ten Spieler

pro Kalenderjahr. Der geschätzte Mittelwert für die Runs des i -ten Spielers in der s -ten Saison ist

$$\widehat{M}_{i,s} = k_i \widehat{C}_{i,s} \quad (5.6)$$

k_i steht für den Steigungskoeffizienten und $C_{i,s}$ der geschätzte Beitrag pro Kalenderjahr aus der gefitteten polynomialen Funktion.

Die obere Schätzung für die Gesamtanzahl an Runs benützt die Schätzung für die obere Konfidenzgrenze der erzielten Runs und die obere Konfidenzgrenze für not out (welche die unterste voraussichtliche Anzahl an Dismissals ergibt). Um die unteren Grenzen zu schätzen wird das Gegenteil verwendet.

Dieser Prozess wird für Bradman und Lara durchgeführt. In beiden Fällen war die Regression höchst signifikant ($p < 0.0001$).

Die folgenden Plots zeigen die beobachteten, kumulierten mittleren Runs gegen das Karrierejahr. In diese wurde die obere und untere Grenze des 95%-Konfidenzintervalls, der kumulierten erwarteten Runs, gelegt. In beiden Fällen fallen nach dem ersten Drittel ihrer Karriere die beobachteten mittleren Runs in die Konfidenzintervalle aus dem Schätzungsprozess. Dies zeigt, dass das Modell die Daten angemessen fittet. Somit ergibt sich eine vernünftige Schätzung für Bradman.

Benützt man diese Methode um Bradmans vorraussichtliche Spielweise für 1939-1945 zu bemessen, schätzt man eine Schlagdurchschnitt von 105.41 [95% KI (90.48,123.44)]. Die vergleichbaren Ergebnisse für Lara sind viel geringer. Hier erhält man einen Schlagdurchschnitt von 52.88 mit einem geschätzten 95% KI von (44.81, 67.16). (Siehe auch Abbildung 5 im Anhang)

5.3 Ergebnis

Mit der Schätzung der oberen Grenze des 95% Konfidenzintervalls und dem Mittel um die Standardabweichung der durchschnittlichen erzielten Runs pro Saison zu erhalten, erlaubt es zu testen, ob Donald Bradman einen höheren Schlagdurchschnitt gehabt hätte, wenn der Zweite Weltkrieg seine Karriere nicht unterbrochen hätte.

Vergleicht man den geschätzten Durchschnitt von 105.41 mit dem beobachteten (99.94) so erhält man einen p -Wert von 0.2763. Solange es vermeintlich so scheint, dass der geschätzte Durchschnitt von Bradman höher ist als der beobachtete, gibt es auf dem 5% Signifikanzniveau keine ausreichenden Beweise um darauf zu schließen, dass sein Durchschnitt signifikant höher gewesen wäre, wäre seine Karriere nicht durch den Krieg unterbrochen worden.

Kapitel 6

Diskussion des Time Series Clustering

Diese Methode wurde bewusst verwendet, obwohl bekannt ist, dass es weitere Methoden zur Modellierung der Spielerleistung gibt, mit Rücksicht auf die Entwicklung der Karriere. Bei diesen wird zum Beispiel die Leistung anhand von Alter oder Status der Karriere modelliert. Diese wurden hier aus zwei Gründen nicht berücksichtigt:

Erstens gibt es nur wenige Individuen, die mindestens 17 Jahre auf höchstem Niveau im Cricket absolviert haben. Dies beinhaltet, dass ein Individuum fortlaufend ausreichend gespielt haben muss, um den Einschluss zu gerechtfertigen, um die Auswahl zu stützen. Dies ist an sich wiederum eine interessante Eigenschaft derer Leistung ist. Somit ist es wichtig die Karrieren als Ganzes und nicht als Stückwerk als Funktion ihres Alters oder Status der Karriere zu betrachten.

Zweitens sind die Höhen und Tiefen einer Karriere bedingt durch frühere Leistungen und als solche ist es notwendig frühere Leistungen in das zugrunde liegende Modell mit einzu beziehen. In dem vorliegenden Paper werden polynomiale Funktionen verwendet, um die auf ein Jahr umgerechnete Leistung pro Karrierejahr zu modellieren, beinhalten die Interaktion zwischen den Jahren und ohne diese könnte etwas von der Anwendbarkeit verloren gehen.

Um die Spielerleistung über die Zeit zu charakterisieren, arbeiten die polynomialen Funktionen angemessen für den vorliegenden Nutzen (Erlauben der Konstruktion von aussagekräftigen Clustern) und die Untersuchung weiterer Methoden wurde als unnötig betrachtet.

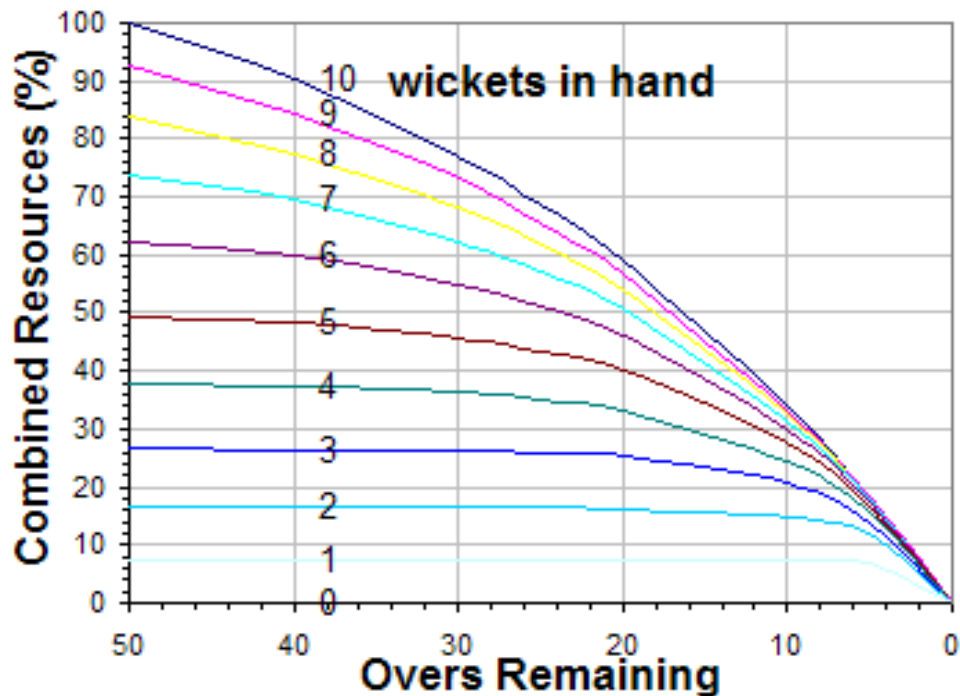
Kapitel 7

Ausblick

7.1 Cricket

In dem gesamten Paper werden nur Spieler aus dem Männer-Cricket betrachtet. Diese Methode lässt sich problemlos auch auf das Frauen-Cricket übertragen. Hier gibt es zwar kein solches Ausnahmetalent wie Don Bradman, aber man kann die einzelnen Karriereläufe, mit den Hochs und Tiefs, sehr gut darstellen und analysieren. Bei den Frauen ist die erfolgreichste Test-Spielerin Janette Ann Brittin, welche einen Durchschnitt von 49.61 hat. Sie spielte zwischen 1979 und 1998 27 Test-Matches und 44 Innings. Davon waren 5 not out. Ebenso wäre es denkbar, erfolgreiche Bowler zu untersuchen. Hierbei würde die Aufmerksamkeit natürlich auf dem niedrigsten Wert liegen. Bei den Männern hält den Rekord George Alfred Lohmann mit einem Durchschnitt von 10.75 in 18 Matches zwischen 1886 und 1896. Bei den Frauen ist es Elizabeth Rebecca Wilson, welche im Zeitraum von 1948 bis 1958 in elf Matches einen Durchschnitt von 11.80 erreichte.

Neben der Betrachtung einzelner Spieler gibt es auch Methoden gesamte Teams oder auch ein Match zu untersuchen. Ein aktueller Versuch ist die Duckworth-Lewis Methode. Hierbei versucht man beim One-Day-Cricket oder dem Twenty20 Cricket den Spielausgang zu schätzen, wenn das Spiel vorzeitig durch Wetter oder andere Umstände beendet wird. Diese Methode wird als sehr gerecht betrachtet, ist aber trotzdem nicht ganz unumstritten. Trotzdem wird die Methode offiziell in den Regeln genannt. Sie versucht anhand von den verbleibenden Overs und Wickets den Endstand zu schätzen, wäre das Spiel nicht unterbrochen worden. Durch statistische Untersuchungen vieler Spiele fand man heraus, dass es eine Korrelation zwischen diesen beiden Größen und dem Anteil an Runs in einem Innings gibt. Die folgende Grafik zeigt die errechneten Kurven:



Diese Grafik basiert auf der sogenannten D/L-Tabelle, welche auf 50 Overs normiert ist (50 Overs entsprechen 100% Ressourcen), und aus den vorher genannten Größen erstellt wurde. Die Abbildung zeigt die „Standard Edition“. In professionellen Spielen wird die „Professional Edition“ verwendet. Diese ist nicht mehr so allgemein, sondern wird für das entsprechende Spiel per Computer berechnet.

7.2 Außerhalb von Cricket

Die statistischen Methoden, die im Cricket angewandt werden, können ebenso auf andere Sportarten übertragen werden. Am nächsten verwandt ist Cricket mit dem Baseball. Auch hier sind Schlageffektivität beim Hitting wichtig. Ebenso wird eine gute Statistik beim Pitching erstrebt.

Auch bei anderen Teamsportarten kann diese Methode angewandt werden, da sie die Leistung des Einzelnen im Bezug auf die Teamleistung als Grundlage hat. Somit sieht man auch, dass Einzelsportarten hiervon ausgeschlossen sind.

Außerhalb von Sport ist die Verwendung in Finanzmärkten oder der Medizin vorhanden. Überall wo Ähnlichkeiten innerhalb von Zeitreihen untersucht werden sollen, findet diese Methode regen Anklang.

Kapitel 8

Fazit

Mit Hilfe des Time Series Clustering wurde gezeigt, dass Donald Bradmans Karriere sehr ähnlich zu Brian Laras ist, doch nicht in allen Punkten übereinstimmt. Es wurden Daten von 20 internationalen Cricketspielern untersucht, welcher in mindestens 17 Jahre über 70 Innings absolviert haben (Stichtag: 1. Januar 2009). Aus diesen wurden zahlreiche globale Eigenschaften bestimmt, um die Höhen und Tiefen einer Karriere zu erkennen.

Die Methode, die für die Anwendung von geclusterten Zeitreihen gewählt wurde, gründet auf einem Modell und dem Herangehen über Eigenschaften. Diese globalen Zeitreiheneigenschaften, wie von Wang et al.(2006) vorgeschlagen, erzeugen instinktive Clusterergebnisse auf der Grundlage von verschiedenen langen Zeitreihen.

Um den relativen Einfluss des einzelnen Batsman auf das Team wiederzuspiegeln, wurden die Daten durch den mittleren Beitrag zur Teamleistung standardisiert. Hierdurch werden, im Gegensatz zur Verwendung der erzielten Runs, zusätzliche Einflussgrößen, wie zum Beispiel die Bedingungen auf dem Pitch, negiert, indem man den die Teamkameraden als Rahmen der Referenzen genommen werden. Der mittlere Beitrag pro Jahr wurde für jeden Batsman durch die Spannweite standardisiert. Aus der geglätteten, standardisierten Datenmenge wurde anschließend eine polynomiale Funktion für jeden Batsman gefittet. Hieraus wurden anhand der Parameter, in Zusammenhang mit Schiefe, Wölbung und Autokorrelation, sechs sinnvolle Cluster gebildet. Da die Cluster durch möglichst ähnliche Karriereverläufe ermittelt wurden, zeigte sich, dass Lara am ähnlichsten zu Bradman ist. Man sah im Vergleich, dass Bradman seine Höhepunkte vermutlich in der Zeit des Zweiten Weltkrieges hatte, als der internationale Sport zum Erliegen kam. Nun wurde versucht auf den mittleren Schlagdurchschnitt von Don Bradman zu schließen, wäre diese Unterbrechung nicht gewesen. Man erhält einen geschätzten Wert von 105.41. Dieser ist jedoch nicht signifikant unterschiedlich zu dem beobachteten Wert von 99.94.

Literaturverzeichnis

Klaus Backhaus, Bernd Erichson, Wulff Plinke, and Rolf Weiber. *Multivariate Analysemethoden. Eine anwendungsorientierte Einführung*. Berlin, 1990.

Deutscher Cricket Bund. Deutscher cricket bund. <http://www.cricket.de>, 2009. [Online; letzter Zugriff: 11. Dezember 2009].

Marylebone Cricket Club. Marylebone cricket club. <http://www.lords.org>, 2009. [Online; letzter Zugriff: 11. Dezember 2009].

ESPNcricinfo. Cricinfo. <http://www.cricinfo.com>, 2009. [Online; letzter Zugriff: 11. Dezember 2009].

Bracewell et al. *Was Bradman denide His Prime?* Journal of Quantitative Anaylsis in Sports, 2009.

Don Bradman Foundation. Don bradman foundation. <http://www.bradman.com.au>, 2009. [Online; letzter Zugriff: 11. Dezember 2009].

Gerhard Tutz, Ludwig Fahrmeir, and Iris Pigeot Rita Künstler. *Statistik. Der Weg zur Datenanalyse*. München, 2006.

Wikipedia. Duckworth-lewis method. http://en.wikipedia.org/wiki/Duckworth-Lewis_method, 2009. [Online; letzter Zugriff: 12. Dezember 2009].

Anhang

Name	T⁻²	T⁻¹	T	T²	Auto r	Skew	Kurt	P-Value	R-Sq	Clust#
AR Border	-1.71	1.78	0.13	-0.01	0.89	-1.09	0	<.0001	0.94	1
IVA Richards	-1.29	1.42	0.05	0	0.71	-0.66	-0.63	<.0001	0.68	1
RB Kanhai	-1.87	1.85	0.07	0	0.75	-0.77	-0.74	<.0001	0.86	1
G Boycott	-0.52	0.84	0.06	0	0.92	-0.12	-1.6	<.0001	0.85	2
RB Simpson	0.82	-0.68	0.15	-0.01	0.94	-0.76	-0.55	<.0001	0.91	2
SR Tendulkar	-0.07	0.46	0.13	-0.01	0.92	-0.44	-1.09	<.0001	0.93	2
B Mitchell	0.02	-0.18	0.16	-0.01	0.98	-1.4	1.42	<.0001	0.95	3
GS Sobers	-0.96	0.9	0.1	0	0.8	-1.32	1.21	<.0001	0.86	3
JB Hobbs	-0.58	0.66	0.07	0	0.94	-1.21	0.53	<.0001	0.69	3
J Miandad	-0.05	0.9	0.12	-0.01	0.83	-0.72	1.59	<.0001	0.95	3
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	-

Abbildung 1: Zusammenfassung der globalen Eigenschaften von internationalen Batsmen mit langen Karrieren

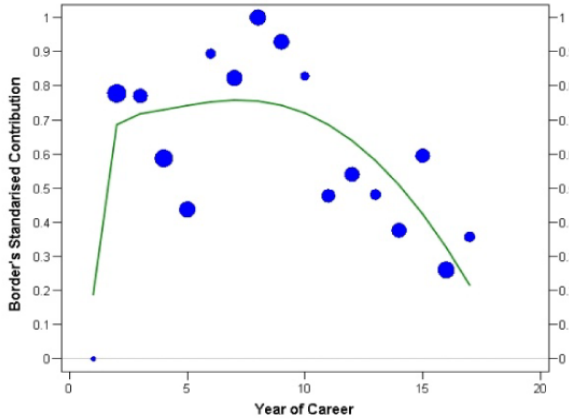


Figure 1a: Bubble Plot overlaid with fitted polynomial function showing Border's relative batting contribution by year.

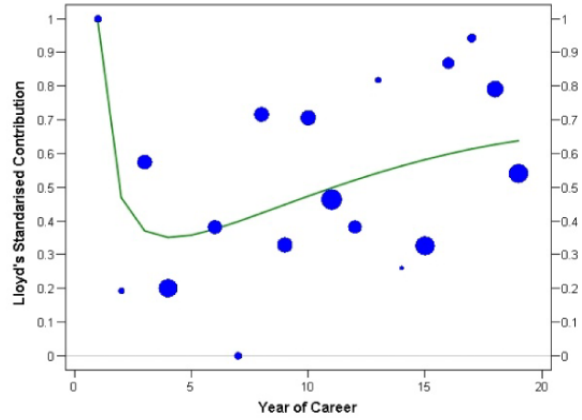


Figure 1d: Bubble Plot overlaid with fitted polynomial function showing Lloyd's relative batting contribution by year.

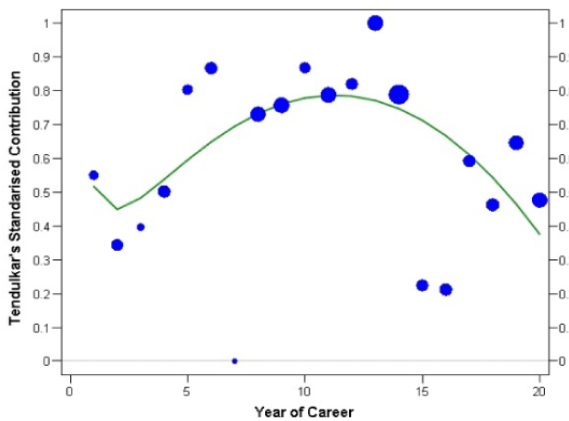


Figure 1b: Bubble Plot overlaid with fitted polynomial function showing Tendulkar's relative batting contribution by year.

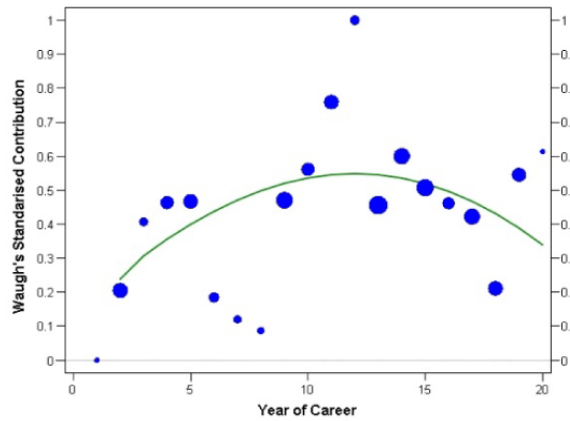


Figure 1e: Bubble Plot overlaid with fitted polynomial function showing Waugh's relative batting contribution by year.

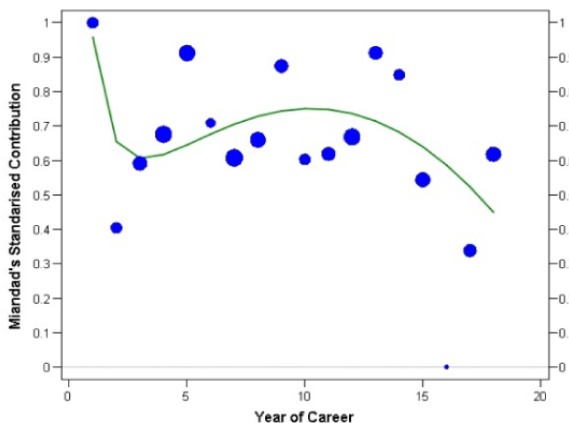


Figure 1c: Bubble Plot overlaid with fitted polynomial function showing Miandad's relative batting contribution by year.

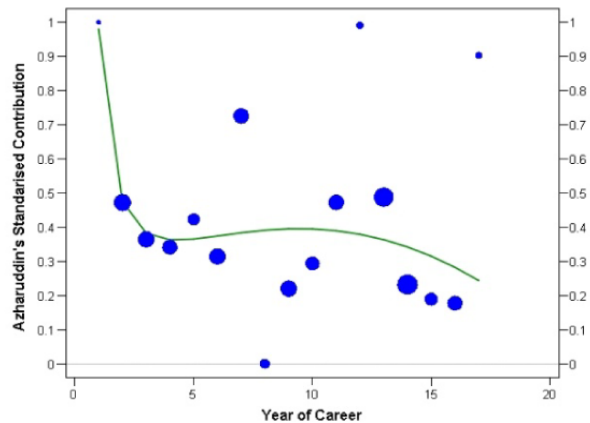


Figure 1f: Bubble Plot overlaid with fitted polynomial function showing Azharuddin's relative batting contribution by year.

Abbildung 2: Bubbleplots mit eingezeichneter gefitteter Polynome

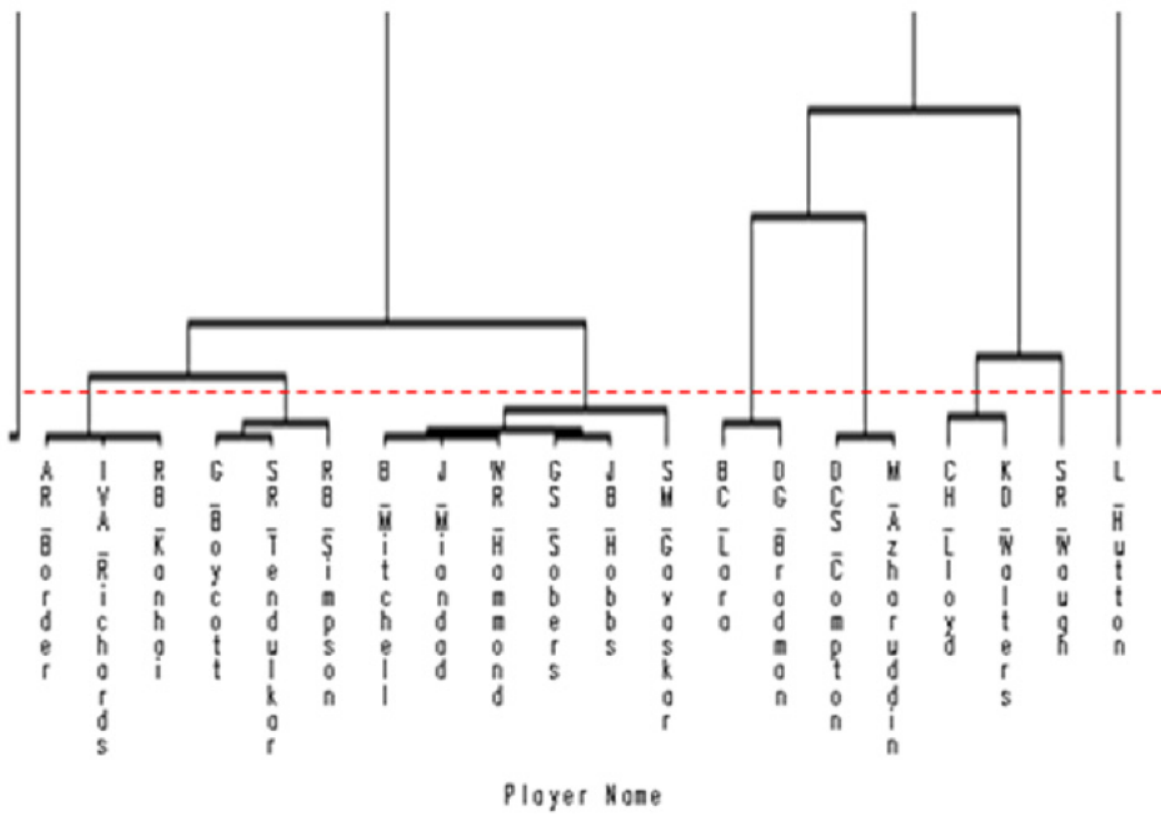


Abbildung 3: Ausschnitt aus dem Dendrogramm mit eingezeichneter Aufteilung der Cluster

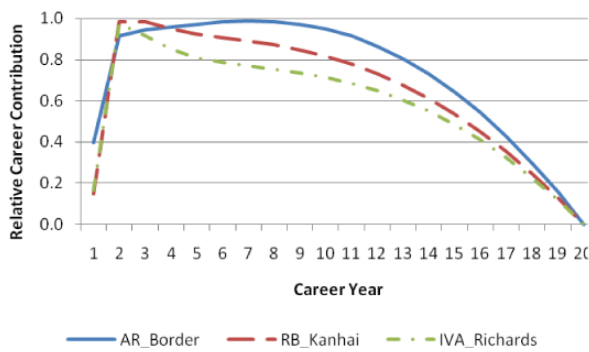


Figure 3a: Line plot displaying the fitted polynomial function for relative batting contribution by year for members of cluster 1.

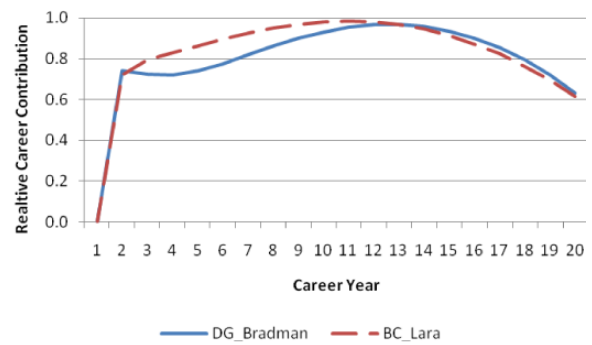


Figure 3d: Line plot displaying the fitted polynomial function for relative batting contribution by year for members of cluster 4.

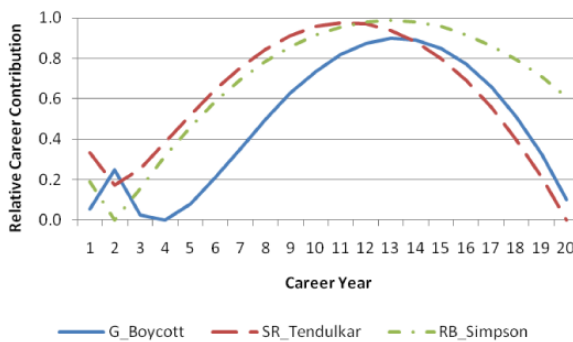


Figure 3b: Line plot displaying the fitted polynomial function for relative batting contribution by year for members of cluster 2.

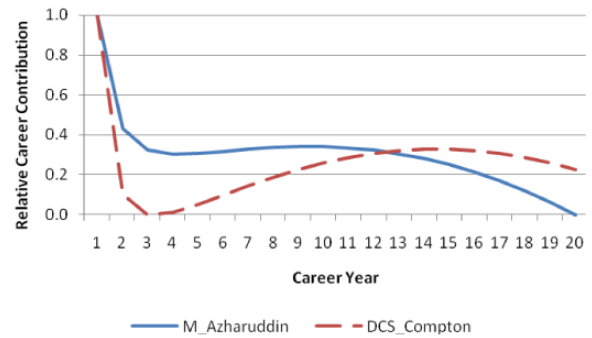


Figure 3e: Line plot displaying the fitted polynomial function for relative batting contribution by year for members of cluster 5.

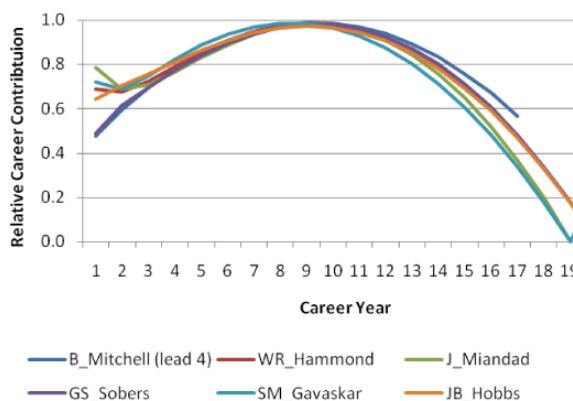


Figure 3c: Line plot displaying the fitted polynomial function for relative batting contribution by year for members of cluster 3.

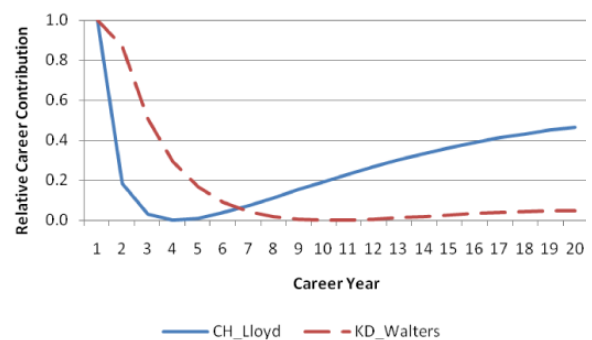


Figure 3f: Line plot displaying the fitted polynomial function for relative batting contribution by year for members of cluster 6.

Abbildung 4: Die sechs Cluster mit den gefitteten Polynomen

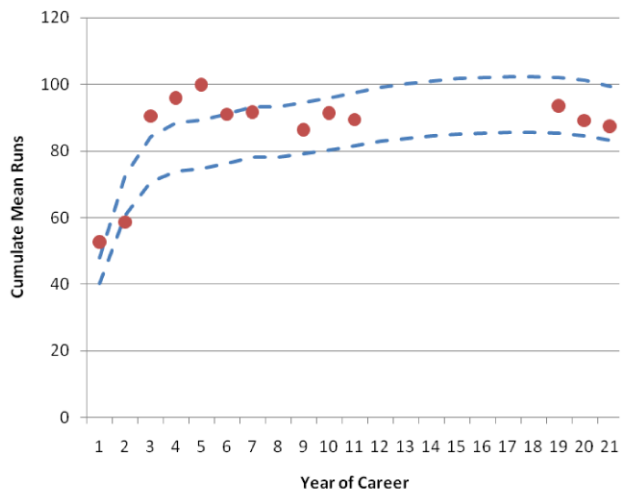


Figure 4: Plot of the observed cumulative mean runs against the career year of Bradman with 95% confidence interval limits from the estimation model overlaid.

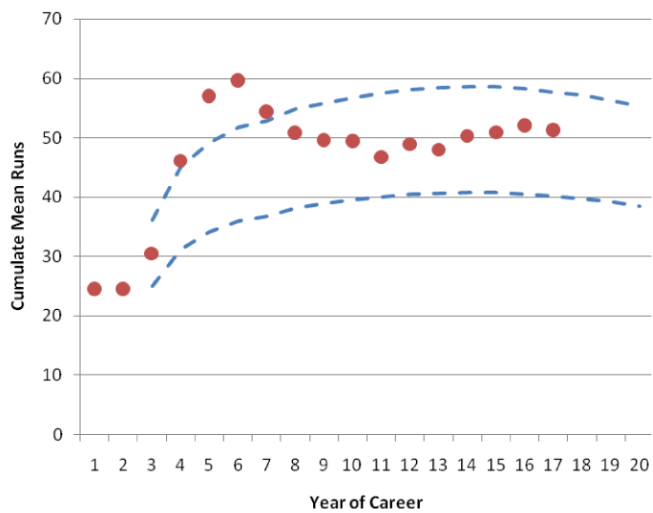


Figure 5: Plot of the observed cumulative mean runs against the career year of Lara with 95% confidence interval limits from the estimation model overlaid.

Abbildung 5: Beobachtete kumulierte mittlere Runs mit Konfidenzintervallgrenzen