Ludwig-Maximilians-Universität München

Institut für Statistik

Entwicklung eines Bewertungssystems für Leistungen im Doppelkopf

Bachelorarbeit zur Erlangung des akademischen Grades
BACHELOR OF SCIENCE (B. Sc.)

Melissa Schmoll

Matrikelnummer: 11410234

Betreuer:
Prof. Dr. Thomas Augustin
Dr. Georg Schollmeyer

Abgabedatum: 20. August 2018

Zusammenfassung

Ziel dieser Arbeit ist es, ein Bewertungssystem für die Leistungen von Spielern im Doppelkopf, einem Kartenspiel für vier Personen, zu entwickeln. Der Deutsche Doppelkopf-Verband e.V umfasst ca. 1200 Mitglieder und betreut verschiedene Wettbewerbe. Aus den vorliegenden Turnierergebnissen von 1984 bis 2017 konnte ein Bewertungssystem erstellt werden, welches sowohl die erspielten Punkte, als auch die Stärke des Teilnehmerfeldes berücksichtigt. Anhand des Ratings eines Spielers kann für diesen die Punktzahl, die man aufgrund seines Ratings und des durchschnittlichen Ratings seiner Gegner von ihm erwartet, berechnet werden. Die Differenz von erwarteten und tatsächlich erspielten Punkten wird mittels einer hyperbolischen Tangensfunktion getrimmt. Somit wird der Einfluss von starken Abweichungen, welche beispielsweise durch über ein Turnier anhaltendes Glück oder Pech zustande kommen können, verringert. Das Ausmaß der Trimmung kann über einen Parameter c gewählt werden. Anschließend wird das alte Rating eines Spielers zu einem gewissen Grad λ überschrieben. Zur Bestimmung von λ und c werden zum einen die Methode über die Maximierung der Prognosegüte und zum anderen ein Ansatz über eine Streuungszerlegung genutzt. Da diese Verfahren zu unterschiedlichen Ergebnissen führen, wird die Option einer Expertenbefragung anhand einer beispielhaften Umfrage diskutiert, die aber im Fall von Doppelkopf keine Aussage liefert. Abschließend werden weitere Anpassungsmöglichkeiten des Ratingsystems aufgezeigt, wie zum Beispiel die Möglichkeit zur Einbeziehung von Mannschaftswettbewerben durch zwei getrennte Ratings für jede Person oder ein Bonussystem für besonders aktive Spieler.

Inhaltsverzeichnis

Hinweis: Aus Gründen der leichteren Lesbarkeit wird in der vorliegenden Bachelorarbeit die gewohnte männliche Sprachform bei personenbezogenen Substantiven und Pronomen verwendet. Dies impliziert jedoch keine Benachteiligung des weiblichen Geschlechts, sondern soll im Sinne der sprachlichen Vereinfachung als geschlechtsneutral zu verstehen sein.

1 Einleitung

Diese Arbeit befasst sich mit der Entwicklung eines Systems zur Bewertung der Leistung von Doppelkopfspielern. Besonders im Bereich des Sports ist das Interesse an Bewertungssystemen und statistischen Analysen in den letzten Jahren stark gestiegen. Sportmanager können ihre Entscheidungen über Einkäufe und Einsatz ihrer Spieler auf statistische Modelle bezüglich der Spielstärke stützen. Auch die Möglichkeiten zur Prognostizierung von Spielergebnissen, beispielsweise basierend auf Daten aus vorhergehenden Spielen, sind insbesondere im Bereich der Sportwetten von großer Relevanz. Allerdings beschäftigen sich die meisten Modelle mit Aussagen über den direkten Vergleich zweier Mannschaften. Dadurch fließen oftmals nur Informationen über Sieg, Unentschieden oder Niederlage ein. Tatsächlich erspielte Punktedifferenzen, die in den meisten Kartenspielen von Bedeutung sind, werden somit nicht berücksichtigt. Eins dieser Kartenspiele ist Doppelkopf, dessen Regeln und Wettbewerbe im folgenden Kapitel erläutert werden. In Kapitel drei werden zwei bestehende Ratingsysteme vorgestellt, das Pi-Rating und die Elo-Zahl. Diese bilden die Basis für ein Bewertungssystem, welches sich auf Doppelkopf anwenden lässt. Aus den vom Deutschen Doppelkopfverband e.V. zur Verfügung gestellten Daten der Jahre 1984 bis 2017 wird in Kapitel vier das Ratingsystem erstellt und verschiedene Methoden zur Bestimmung der für das System benötigten Parameter erörtert. Für eine der daraus resultierenden Parameterkonstellationen wird das Bewertungssystem berechnet und einige Eigenschaften dargestellt. Im abschließenden Kapitel werden verschiedene Anpassungsmöglichkeiten erläutert, um das System noch genauer auf die Bedürfnisse des Deutschen Doppelkopfverbandes abzustimmen.

2 Das Spiel

In vielen deutschen Regionen ist Doppelkopf ein beliebtes Kartenspiel für vier Personen. Es ist ein Partnerspiel, wobei die Parteizugehörigkeit mit jedem Spiel wechselt. Diese ist jedoch zu Spielbeginn nicht bekannt und somit besteht eine der großen Herausforderungen darin, seinen Partner zu ermitteln und gemeinsam möglichst viele Punkte zu erzielen. Dafür benötigen die Spieler mathematisches Verständnis, Konzentration und logisches Denkvermögen aber auch psychologische Faktoren sind von großer Bedeutung.
In Abbildung 1 werden Gesellschaftsspiele in ein Glück-Logik-Bluff Dreieck eingeordnet. Auf der Logik - Glück Achse sind Spiele zu finden, bei denen alle Spieler den gleichen Informationsgehalt haben. Während bei Spielen wie zum Beispiel Schach oder Mühle kein Glücksfaktor vorhanden ist, entsteht dieser bei Mensch ärger dich nicht oder Backgammon beispielsweise durch das Würfeln. Die Komponente des Bluffs kommt vor allem durch die Unsicherheit über die Karten des Gegners und dessen Aufstellung zustande. Doppelkopf besteht wie die meisten Kartenspiele sowohl aus dem Faktor des Glücks, als auch aus Logik und Bluff. [1, S. 9]

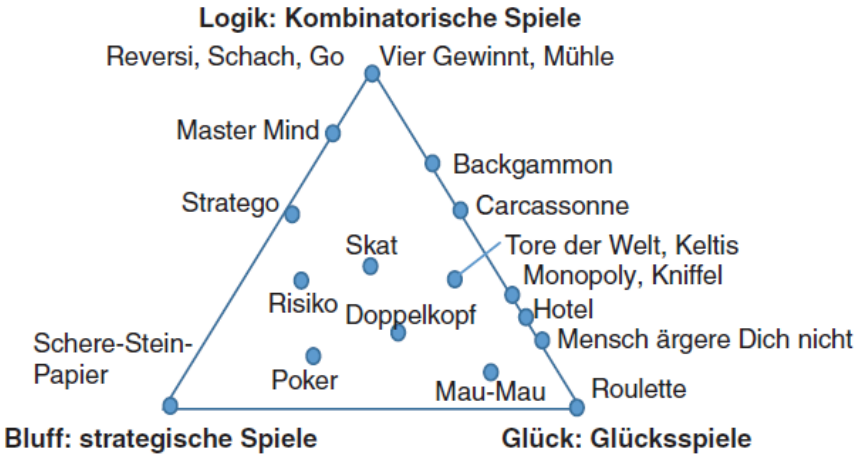

Abbildung 1: [1, S. 10] Doppelkopf beinhaltet wie die meisten Kartenspiele die drei Komponenten Glück, Logik und Bluff.

Es gibt im Doppelkopf zahlreiche Sonderregeln und verschiedenste Spielvarianten. Der am 27. März 1982 in Braunschweig gegründete Deutsche Doppelkopf-Verband e. V. hat ein einheitliches Regelwerk geschaffen und das Spiel somit zu einem gewissen Grad berechenbar gemacht. Zusätzlich haben sich Konventionen entwickelt, die den Spielern weitere Möglichkeiten zur schnelleren Partnerfindung und exakteren Beschreibung der jeweiligen Blätter liefern, um gemeinsam das Spielergebnis zu optimieren. Sowohl das einheitliche Regelwerk, als auch die Konventionen haben dazu geführt, dass sich Doppelkopf gemäß Abbildung 1 aus der Ecke der Glücksspiele entfernt und

in Richtung der Kombinatorischen Spiele bewegt.
Zu Beginn gehörten dem Verband knapp 400 Mitglieder an, bis heute ist die Zahl auf fast 1200 Mitglieder angestiegen. Der Verband ist die Schnittstelle für 71 angeschlossenen Vereine, zudem ist er verantwortlich für die verschiedenen Wettbewerbe und die stetige Weiterentwicklung des Spiels.

2.1 Die Regeln

Im folgenden werden die Regeln des Spiels in einer verkürzten Fassung in Anlehnung an die vom Deutschen Doppelkopfverband e.V. veröffentlichten Kurzregeln erläutert[2].
Das Doppelkopfblatt besteht aus 48 Karten der Farben Kreuz (♣), Pik (♠), Herz (♥) und Karo (♦). In jeder Farbe gibt es je zwei Karten mit den Kartenwerten Neun(0), Zehn(10), Bube(2), Dame(3), König(4), Ass(11). In Klammen angegeben ist der Zählwert, es sind also insgesamt 240 Punkte im Spiel. Es besteht grundsätzlich Bedienpflicht. Nur, wenn eine angespielte Karte nicht bedient werden kann, darf getrumpft oder abgeworfen werden. Eine Runde besteht aus 24 Spielen.
In einem Normalspiel gibt es 26 Trümpfe (jede Karte zweimal vorhanden) in der Reihenfolge

$$\heartsuit 10, \clubsuit D, \spadesuit D, \heartsuit D, \diamondsuit D, \clubsuit B, \spadesuit B, \heartsuit B, \diamondsuit B, \diamondsuit Ass, \diamondsuit 10, \diamondsuit K, \diamondsuit 9$$

Die restlichen 22 Karten sind Fehlkarten in der Reihenfolge Ass, 10, K, 9, mit Ausnahme der ♥10, welche, wie eben beschrieben, die höchste Trumpfkarte ist. Die beiden Spieler, welche die ♣ Damen auf der Hand haben, spielen zusammen als Re-Partei gegen die beiden anderen Spieler (Kontra-Partei).
Eine Ausnahme zum Normalspiel ist das Solospiel. Hier spielt ein Spieler alleine als Re-Partei gegen die drei anderen Mitspieler der Kontra-Partei. Jeder Spieler muss innerhalb einer Runde ein Solo spielen. Diese vier Soli werden Pflichtsoli genannt. In einem Pflichtsolo hat der Solist Aufspielpflicht. Weitere Soli dürfen ohne Aufspielpflicht gespielt werden und nennen sich Lustsolo.
Bei den Soli wird zwischen vier Varianten unterschieden.
Zum einen gibt es das Bubensolo. Hier sind alle Buben Trumpf in der Reihenfolge ♣, ♠, ♥, ♦. Die restlichen Karten sind Fehlkarten in der Reihenfolge Ass, 10, K, D, 9.
Analog dazu gibt es das Damensolo, mit den Damen als Trumpfkarten und Fehlkarten in der Reihenfolge Ass, 10, K, B, 9.
In einem Ass-Solo gibt es keine Trümpfe und die Karten gelten in der Reihenfolge Ass, 10, K, D, B, 9.
Zuletzt gibt es das Farbsolo. Trumpf und Fehl gelten wie im Normalspiel, allerdings können die Trümpfe $\diamondsuit Ass, \diamondsuit 10, \diamondsuit K, \diamondsuit 9$ durch eine beliebige andere Farbe ersetzt werden. In einem Farbsolo

der Farbe Herz bleibt die ♥10 als höchster Trumpf erhalten und es sind somit zwei Trümpfe weniger im Spiel.
Hat ein Spieler beide ♣ Damen auf der Hand nennt man das Spiel Hochzeit. Dieser Spieler bekommt denjenigen zum Partner, der den ersten Stich erspielt. Macht er die ersten drei Stiche selber, so spielt er alleine gegen die anderen drei Spieler, das Spiel wird jedoch nicht als Pflichtsolo gewertet. Zu Beginn des Spiels sagen die Spieler der Reihe nach, beginnend bei dem Spieler links vom Geber, ob sie einen Vorbehalt (Pflichtsolo, Lustsolo oder Hochzeit) haben oder nicht. Bei den Vorbehalten hat das Pflichtsolo höchste Priorität, dann das Lustsolo und niedrigste Priorität die Hochzeit. Haben zwei Spieler einen Vorbehalt gleicher Priorität, so erhält der am weitesten vorne sitzende Spieler das Spielrecht. Meldet ein Spieler eine Hochzeit zu Spielbeginn nicht an, so spielt er alleine und das Spiel wird als Lustsolo gewertet. Nach der Vorbehaltsabfrage beginnt das Spiel. Um ein Spiel zu gewinnen muss die Re-Partei 121 Punkte erreichen, der Kontra-Partei reichen 120 Punkte. Wenn ein Spieler glaubt mit seinem Partner das Spiel zu gewinnen, kann er als zugehöriger der Re-Partei "Re"sagen (entsprechend "Kontra"als zugehöriger der Kontra-Partei).Dies muss geschehen, solange der Spieler noch mindestens 11 Karten auf der Hand hat. Im Fall, dass die Kontra-Partei "Kontra"sagt, braucht sie 121 Punkte um zu gewinnen. Mit jeder weiteren gespielten Karte darf ein Mitglied der Partei die Ansage erhöhen, in den Schritten "Keine 90 Punkte"(mit mindestens 10 Karten), "Keine 60 Punkte"(mit mindestens 9 Karten),"Keine 30 Punkte"(mit mindestens 8 Karten), "keinen Stich"(mit mindestens 7 Karten). Dieses gilt auch im Solospiel. Da bei der Hochzeit die Parteizugehörigkeit erst nach dem ersten fremden Stich geklärt ist, verschiebt sich der Ansagezeitpunkt um je eine Karte, wenn der Klärungsstich der zweite Stich ist und um je zwei Karten, falls der dritte Stich der Klärungsstich ist. Sobald eine Partei "Keine 90"angesagt hat, gewinnt die Gegenpartei sobald sie 90 Punkte erreicht hat, analog dazu bei den anderen Absagen. Auf jede Absage kann die Gegenpartei einen Stich später "Re"oder "Kontra"erwidern. Nach dem Spiel bekommen die Sieger die Spielpunkte positiv und die Verlierer negativ angerechnet. Einen Spielpunkt erhält die Siegerpartei für den Gewinn, zwei für eine Ansage, je einen Spielpunkt für jede weitere Stufe, die die Gegenpartei nicht erreicht hat (90/60/30/Keinen Stich) und je einen Punkt für eine angesagte Stufe. Hat eine Partei gegen eine Absage gewonnen erhält sie für jede weitere Stufe, die sie gegen die Absage erreicht haben, einen weiteren Punkt. Zusätzlich bekommt die Kontra-Partei einen Punkt, wenn sie gegen die Re-Partei gewinnt. Weitere Sonderpunkte bekommt eine Partei, wenn sie einen Stich mit 40 oder mehr Zählpunkten erzielt, sie ein ♦*Ass* des Gegners fängt, oder mit dem ♣ Buben den letzten Stich macht. Diese Punkte werden verrechnet und ergeben die Spielpunkte eines Spiels. Bei einem Solo werden keine Sonderpunkte gewertet. Bei einem Sieg werden dem Solospieler die dreifachen Spielpunkte gutgeschrieben, bei einer Niederlage abgezogen. Die anderen drei Spieler erhalten die einfache Spielpunktzahl mit umgekehrtem Vorzeichen

zum Solisten. Somit ist Doppelkopf ein Nullsummenspiel [3, S. 406]. Die Summe der Punktzahlen aller Spieler ist pro Spiel und auch in der Gesamtwertung immer null.
Das vollständige Regelwerk wird vom Deutschen Doppelkopf-Verband e. V. veröffentlicht[2].

2.2 Wettbewerbe

Der Deutsche Doppelkopf-Verband e. V. begleitet verschiedene Wettbewerbe. Im folgenden Kapitel werden zuerst die Einzelwettbewerbe Ranglistenturnier, Regionalmeisterschaft und Deutsche Einzelmeisterschaft vorgestellt und anschließend auf die Mannschaftswettbewerbe Deutsche Mannschaftsmeisterschaft und Bundesliga eingegangen.[4]
Jeder Verein darf einmal im Jahr ein Ranglistenturnier, bestehend aus drei Runden, veranstalten. Das Mindestalter der Teilnehmer beträgt 12 Jahre und es müssen mindestens 40 Spieler teilnehmen. Es ist der einzige Wettbewerb, für den eine Vereinszugehörigkeit nicht notwendig ist. Aus den Ergebnissen der Ranglistenturniere wird zum einen die Rangliste gebildet, zum anderen die Bundesländerwertung. Hat ein Spieler innerhalb des Qualifikationszeitraums von 24 Monaten mindestens 36 Runden auf Ranglistenturnieren gespielt, wird er in die Rangliste aufgenommen. Die Reihenfolge der Rangliste basiert auf dem gespielten Rundenschnitt der Spieler, zuzüglich eines Bonus je nach Anzahl der gespielten Runden. Für die Bundesländerwertung werden nur positive Ergebnisse gewertet und zwar für jeden Spieler das beste Ergebnis, das er in einem Bundesland erzielt hat. Diese Ergebnisse werden aufaddiert und somit die Bundesländerwertung erstellt.
In den drei Regionen Nord, Süd und West wird einmal im Jahr je eine Regionalmeisterschaft veranstaltet. Die Spieler versuchen dort sich innerhalb von acht Runden auf zwei Tage verteilt für die Deutsche Einzelmeisterschaft zu qualifizieren. Es werden 100 Startplätze, anteilsmäßig auf die drei Regionen verteilt, ausgespielt.
Neben der Regionalmeisterschaft gibt es noch drei weitere Möglichkeiten sich für die Deutsche Einzelmeisterschaft zu qualifizieren. Zum einen sind die ersten 32 Spieler der letzten Deutschen Einzelmeisterschaft automatisch im folgenden Jahr zugelassen. Zudem sind die ersten 48 Spieler der Rangliste, welche nicht über die letzte Einzelmeisterschaft qualifiziert sind, spielberechtigt. Ebenfalls dürfen die ersten acht Spieler der Bundesländerwertung, die weder über die Einzelmeisterschaft, noch über die Rangliste qualifiziert sind, teilnehmen. Damit stehen die 188 Teilnehmer für die Deutsche Einzelmeisterschaft fest, welche in acht Runden um den Titel des deutschen Meisters kämpfen. Die erste Einzelmeisterschaft fand 1982, im Gründungsjahr des Deutschen Doppelkopf-Verbandes e. V. statt. Anfangs wurden sowohl bei der Regionalmeisterschaft, als auch bei der deutschen Einzelmeisterschaft nur sechs Runden gespielt, 1996 wurde bei beiden Wettbewerben die Rundenanzahl auf acht erhöht.
Neben all diesen Einzelwettbewerben gibt es auch Mannschaftswettbewerbe. Die Deutsche Mann-

schaftsmeisterschaft wird seit 1984 im K.O-System gespielt. Zudem gibt es einen Bundesliga Wettkampf mit 16 Mannschaften, ausgetragen an fünf Doppelspieltagen. Zusätzlich wird an zwei Doppelspieltagen die Bundesligaqualifikation veranstaltet. Die Anzahl der Auf- und Absteiger am Jahresende richtet sich danach, wie viele Mannschaften an der Qualifikation teilnehmen.

3 Bestehende Ratingsysteme

Ratingsysteme gibt es bereits seit dem 13. Jahrhundert [5, S. 1]. Doch das immer weiter steigende Interesse an Sportergebnissen und deren Prognostizierung, sowie wachsende Kapazität zur Speicherung großer Datenmengen sorgen für eine starke Entwicklung im Bereich der Ratingsysteme.
Jedes Ratingsystem besteht aus drei Phasen. In der ersten Phase, der Evaluierungsphase, werden die Ergebnisse der verschiedenen Teilnehmer über alle Turniere hinweg gesammelt. Darauf folgt die Gewichtungsphase. Hier können den erzielten Ergebnissen der verschiedenen Wettbewerbe unterschiedlich starke Einflüsse zugeordnet werden. In der anschließenden Ratingphase findet die Verknüpfung der Evaluierung und der Gewichtung statt, um das finale Rating zu erhalten. [6]
Im Prozess dieser drei Phasen wird jedem Teilnehmer des Ratings eine Bewertung zugewiesen. Ordnet man diese Bewertungen der Größe nach, so wird aus einem Rating ein Ranking. [5, S. 6]
Unterscheidung, Stabilität und Unabhängigkeit sind drei Faktoren, die nach Franks, D'Amour et al. (2016) besonders wichtig bei der Erstellung und dem Vergleich verschiedener Ratingsysteme sind. Die Unterscheidung sagt aus, wie gut und zuverlässig ein System zwischen verschiedenen Spielern differenzieren kann. Die Stabilität bezieht sich auf das Rating der einzelnen Spieler und deren Konstanz über die Zeit. Unabhängigkeit ist wichtig im Zusammenhang mit anderen Ratingsystemen und soll aussagen, ob das Ratingsystem im Vergleich zu Anderen neue Informationen liefert.[7]
Während Unterscheidung und Stabilität in Kapitel 5.2.2 zur Bestimmung von Parametern als Kriterium verwendet werden können, wird Unabhängigkeit im folgenden nicht weiter betrachtet. Die zahlreichen Ratingsysteme für beispielsweise Basketball, welche sich mit verschiedensten Eigenschaften und Fähigkeiten der Spieler beschäftigen, machen die Untersuchung von Unabhängigkeit in einer solchen Sportart notwendig, was bei Doppelkopf jedoch nicht der Fall ist.
Drayer Barrow (2013) hat verschiedene Ratingsysteme miteinander verglichen, darunter beispielsweise die einfache Verwendung des Anteils an gewonnen Spielen einer Mannschaft oder ein erweitertes Verfahren, welches zuzüglich die Stärke der Gegner, gegeben durch deren Anteil an Siegen berücksichtigt. Das im folgenden Kapitel beschriebene Pi-Rating System basiert auf diesem Prinzip, das Spielergebnis im Zusammenhang mit der Stärke der Gegner zu verwenden.Auch wird die Methode der kleinsten quadratischen Abweichungen untersucht, bei welcher die Ratings so erstellt werden, dass diese möglichst gut mit dem Spielausgang übereinstimmen. Diese wird in Kapitel 5.2.1 zur Paramterbestimmung verwendet. Zudem konnte festgestellt werden, dass im Fußball Ratingsysteme, die Tordifferenzen berücksichtigen, bessere Prognosen für den Ausgang des nächsten Spiels liefern, als die, die lediglich auf Sieg oder Niederlage basieren. Dies soll auch bei Doppelkopf Anwendung finden, indem nicht auf die Platzierung, sondern auf die tatsächlich erspielten Punkte eingegangen wird. [8, S. 200]

3.1 Das Pi-Rating System

Anthony Costa Constantinou und Norman Elliott Fenton (2013) haben ein Bewertungssystem vorgestellt, welches auf den relativen Differenzen in erzielten Toren basiert und Pi-Rating System genannt wird. Es soll auf alle Sportarten angewendet werden können, bei denen die Punktzahl ein Maß für die relative Leistung zwischen den Gegnern ist.[9] Das Pi-Ratingsystem wird im Zusammenhang mit Fußball erläutert. Die Idee des Ratingsystems für Fußballmannschaften soll hier kurz vorgestellt werden. Die genaue Berechnung wird in Kapitel 4 im Zusammenhang mit der Modifizierung für Doppelkopf erläutert. Im Anwendungsfall von Fußball werden drei Anforderungen an das Rating-System gestellt. Erstens soll der sogenannte Heimvorteil berücksichtigt werden. Des weiteren sollen kürzlich erzielten Ergebnissen, im Vergleich zu länger zurückliegenden, eine höhere Bedeutung zu Teil werden, um die aktuelle Spielstärke einer Mannschaft besser darzustellen. Zuletzt soll beachtet werden, dass ein Sieg für eine Mannschaft wichtiger ist, als steigende Tordifferenz. Zur Lösung dieser Anforderungen sollen zunächst getrennte Bewertungen für die Heim- und Auswärtsspielstärke erstellt werden. Jedoch soll jedes Spiel beide Spielstärken beeinflussen, nur unterschiedlich gewichtet. Wie stark sich ein Auswärtssieg auf die Heimspielstärke und umgekehrt auswirkt soll durch eine Gewichtung mit dem Faktor $0 \leq \gamma \leq 1$ geregelt werden. Da Auswärts- und Heimspiele jedoch für die Anwendung an Doppelkopf nicht relevant sind, soll hier nicht weiter darauf eingegangen werden. Auch die zweite Anforderung soll mittels einer Gewichtung erfolgen. Der Parameter λ bestimmt, inwieweit neue Spielergebnisse das alte Rating einer Mannschaft überschreiben. Um zu berücksichtigen, dass ein Sieg für eine Mannschaft wichtiger ist, als steigende Tordifferenz, wird diese Differenz durch eine Funktion modifiziert. Anhand der Ratings der verschiedenen Mannschaften soll nun die erwartete Tordifferenz zwischen den zwei Mannschaften berechnet werden. Nach dem Spiel werden die Ratings der beiden Mannschaften dahingehend aktualisiert, ob sie die von ihnen erwartete Tordifferenz übertroffen oder nicht erfüllt haben.

3.2 Elo-Zahl

Obwohl das Pi-Rating bereits gut auf Doppelkopf anwendbar scheint, ist die Betrachtung eines zweiten Ratingssystems, der Elo-Zahl von Vorteil. Diese ist dem Pi-Rating generell nicht unähnlich, daher kann bei einigen Problemen des Pi-Ratings auf Lösungsvorschläge aus dem System der Elo-Zahl zurückgegriffen werden.
Der Physik Professor Arpad Elo hat ein System zur Bewertung von Schachspielern entwickelt, welches 1970 von der World Chess Federation übernommen wurde und seitdem auch auf viele andere Sportarten übertragen wird [5, S. 54].

Die Berechnung erfolgt, indem zunächst für einen Spieler A die erwarteten Punkte E_A berechnet werden. Ein Sieg bedeutet einen, ein Unentschieden einen halben und eine Niederlage null Punkte.

$$E_A = \frac{1}{1 + 10^{(R_B - R_A)/400}} , \tag{1}$$

wobei R_A und R_B die aktuellen Ratings der Spieler A und B darstellen. Analog wird die erwartete Punktzahl E_B für Spieler B berechnet und die Ratings daraufhin wie folgt aktualisiert:

$$R'_A = R_A + K(S_A - EA)$$

Hier bezeichnet S_A das tatsächlich erspielte Ergebnis von Spieler A. Die Aktualisierung des Ratings von Spieler B erfolgt analog.[10]
Der Wert 400 aus Formel (1) ist wie folgt zu interpretieren. Hat ein Spieler A ein Rating, dass um 400 Punkte höher ist als das eines Spielers B, so ist die Wahrscheinlichkeit, dass Spieler A siegt zehn mal so hoch ist wie die Wahrscheinlichkeit, dass Spieler B gegen Spieler A gewinnt [5, S. 56]. Der Faktor K ist verantwortlich für die angemessene Einbeziehung neuer Abweichungen $(S_A - E_A)$ in das bestehende Rating [5, S. 55]. Da dieser Faktor K die gleiche Funktion hat wie der Parameter λ im Pi-Ratingsystem, werden einige Überlegungen zu dessen Wahl in Kapitel 5.2 im Zusammenhang mit der Bestimmung von λ diskutiert.

4 Ratingsystem für Doppelkopf

Analog zum Fußball kann das Pi-Ratingsystem auf Doppelkopf angewendet werden. Im folgenden Kapitel wird für jeden Schritt zur Erstellung des Ratingssystems zunächst die Vorgehensweise im Zusammenhang mit Fußball erläutert und daraufhin eine mögliche Modifikation für Doppelkopf vorgestellt. Die Notation wird in beiden Systemen gleich gehalten, um die Ähnlichkeiten in der Anwendung besser hervorzuheben.
Genau wie beim Fußball, steigt in einem Pi-Rating System für Doppelkopf jeder Spieler mit einem Rating R von null ein. Sobald sich das Rating eines Spielers um n erhöht, sinkt das Rating von anderen Spielern in Summe um den Wert n und umgekehrt. Somit werden Deflation und Inflation vermieden. Da sich nun die Ratings aller im System aufgenommener Spieler zu null aufsummieren, ist auch der Durchschnitt der Ratings aller Spieler null.
Ausschlaggebend für die Veränderung des Ratings einer Fußballmannschaft ist die Tordifferenz e, die im Spiel zweier Mannschaften erwartet wird, und dem tatsächlichen Ausgang des Spiels. Diese Differenz e wird beim Fußball modifiziert, da ein Sieg für eine Mannschaft wichtiger ist, als steigende Tordifferenz. Constantinou und Fenton verwenden folgende Funktion zur Anpassung dieser Differenz:

$$\psi(e) = c * log_{10}(1 + e) \tag{2}$$

Die Autoren wählen für c ohne genauere Begründung den Wert drei, da es keine Informationen darüber gibt, wie viel wichtiger ein Sieg gegenüber wachsender Tordifferenz ist.
Überträgt man diese Überlegung auf den Anwendungsfall von Doppelkopf, so bedeutet dies, dass es wichtiger ist, ob ein Spieler über oder unter der Erwartung an ihn geblieben ist und weniger wichtig, wie stark er diese Erwartung verfehlt hat. Hier kommt jedoch noch ein anderer Faktor hinzu, der im folgenden immer wieder bedacht werden muss. Auch wenn das Spiel beispielsweise durch Konventionen berechenbarer geworden ist, darf der Faktor des Glücks oder Pechs nicht vernachlässigt werden. Damit ein Spieler seine Geschicklichkeit ausspielen kann, bedarf es einer gewissen Anzahl an Spielen, um Glück und Pech auszugleichen [11, S. 356]. Es ist anzunehmen, dass die Dauer eines Turniers dafür nicht ausreicht und diese Faktoren somit die Ergebnisse beeinflussen. Hat man beispielsweise innerhalb einer Runde das Glück, durch eine gute Kartenverteilung ein hoch gewonnenes Pflichtsolo zu erhalten, kann dies im Vergleich zu einem verlorenen Solo leicht einen Unterschied von 30 Punkten ausmachen. Auch deshalb ist die Verwendung einer Trimmungsfunktion angebracht.
In Abbildung 2 ist die verwendete Funktion $\psi(e) = 3 * log_{10}(1 + e)$ für verschiedene Wertebereiche zu sehen. Links ist der Wertebereich von null bis zehn abgebildet und rechts von null bis 100.

Der Bereich der linken Grafik ist typisch für Differenzen von erwarteter und tatsächlich erzielter Tordifferenz. Eine Abweichung e von zehn Toren wird durch die Funktion auf ungefähr 3.12 getrimmt. Die rechte Grafik hingegen zeigt einen Wertebereich von null bis 100, da dieses Ausmaß von Abweichungen zwischen erwarteten und tatsächlich erspielten Punkten im Doppelkopf nicht ungewöhnlich ist. Die Trimmung durch die Funktion $\psi(e) = 3 * log_{10}(1 + e)$ ist für diese Werte zu stark, da beispielsweise bei einer Abweichung von 100 Punkten auf 6.01 Punkte getrimmt würde.

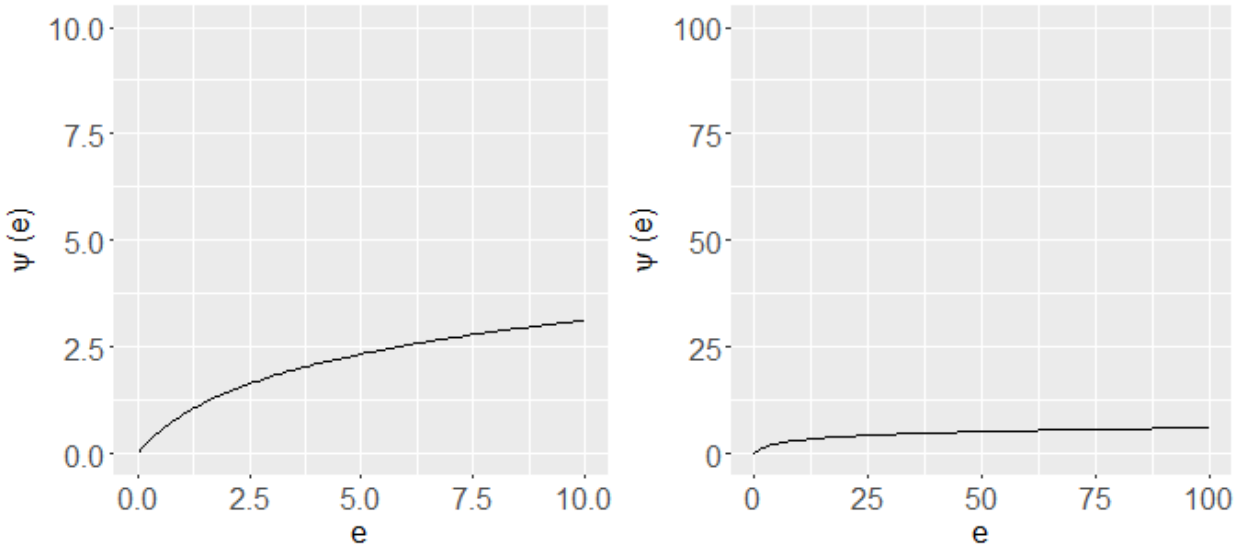

Abbildung 2: In den Grafiken sind für typische Wertebereiche von e (links Tordifferenzen bis zehn Tore, rechts Doppelkopf bis 100 Punkte) die Modifizierung durch die Funktion $\psi(e) = 3 * log_{10}(1 + e)$ zu sehen. Für Doppelkopf ist die Trimmung durch diese Funktion zu stark, beispielsweise wird eine Differenz von erwarteten zu tatsächlich erspielten Punkten von $e = 100$ auf 6.01 getrimmt.

Um eine passende Lösung für Doppelkopf zu finden reicht es nicht, den Wert von c zu erhöhen. Abbildung 3 zeigt in rot die Funktion $\psi(e) = 20 * log_{10}(1 + e)$ und in schwarz die Funktion $\psi(e) = e$. Letztere beschreibt die Werte von $\psi(e)$ ohne Modifizierung. Es lassen sich zwei Probleme bei der Erhöhung von c feststellen. Zum einen liegt selbst für $c = 20$ noch eine starke Trimmung vor. Zum anderen liegen für kleine Werte von e die Funktionswerte von $\psi(e) = 20 * log_{10}(1 + e)$ über der Funktion $\psi(e) = e$, was bedeutet, dass e in diesen Fällen nicht verringert, sondern erhöht wird. Die Autoren gehen auf dieses Problem nicht weiter ein, jedoch sind die Ausmaße dessen in einer Anwendung von Fußball und der Wahl der Funktion $\psi(e) = 3 * log_{10}(1 + e)$ in einem typischen Wertebereich von null bis zehn Tore deutlich geringer.

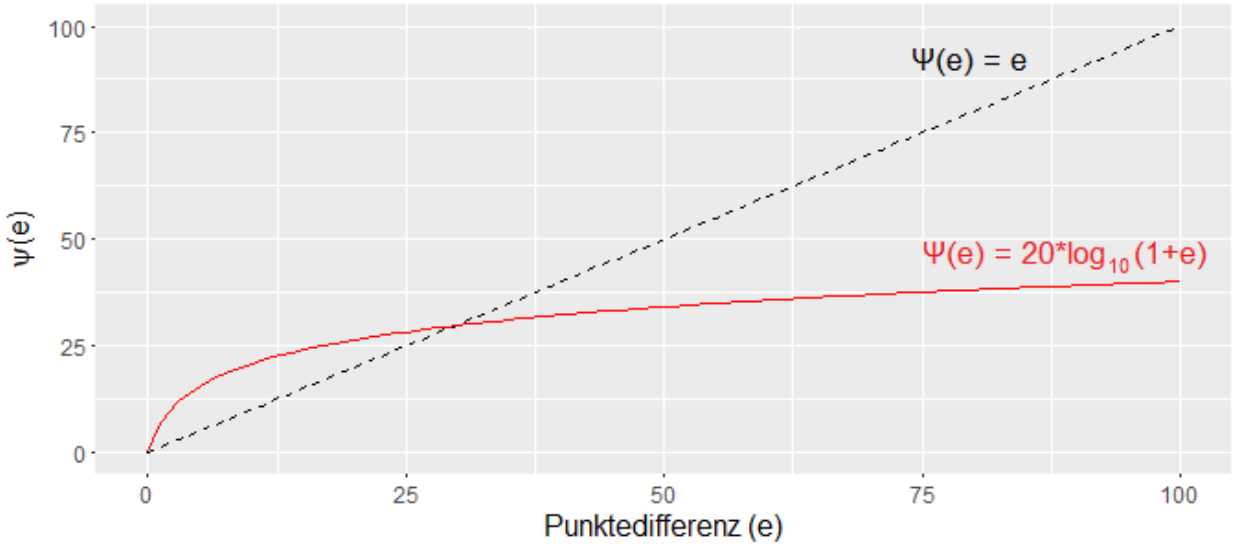

Abbildung 3: Selbst für die Wahl von $\psi(e) = 20 * log_{10}(1 + e)$ ist die Trimmung noch sehr stark. Zudem entsteht für kleine Werte von e durch die Modifizierung mit der Funktion $\psi(e) = 20 * log_{10}(1 + e)$ eine Erhöhung, was daran zu erkennen ist, dass sie in diesem Bereich über der Funktion $\psi(e) = e$ verläuft.

Es empfiehlt sich also eine andere Funktion zu wählen. Diese sollte im Bereich von $e > 0$ nicht über der Winkelhalbierenden $\psi(e) = e$ liegen und den Wert von e angemessen trimmen. Außerdem soll sie sich asymptotisch einem Wert annähern, was bedeutet, dass durch die Funktion $\psi(e)$ eine obere Schranke für die Funktionswerte gegeben ist. Diese Forderung basiert auf den soeben erläuterten Überlegungen zu Glück und Pech. Inhaltlich bedeutet dies, dass egal wie stark ein Spieler die Erwartung verfehlt hat, diese Verfehlung nur bis zu einem maximalen Wert berücksichtigt wird. Zudem muss die Funktion streng monoton steigend sein, sodass das Rating eines Spielers A, der die Erwartungen an ihn stärker übertroffen hat als ein Spieler B, um einen höheren Wert ansteigt, als das des Spielers B. Als Grundlage der Funktion kann der Tangens hyperbolicus gewählt werden. Dieser ist definiert als:

$$tanh\ x = \frac{sinh\ x}{cosh\ x} = \frac{exp(x) - exp(-x)}{exp(x) + exp(-x)} \qquad [12, S.\ 68]$$

Um eine obere Schranke zu gewährleisten kann die Funktion wie folgt angepasst werden:

$$\psi(e) = c * tanh(\frac{e}{c}) \tag{3}$$

Der Faktor c erfüllt den gleichen Zweck wie in der Funktion $\psi(e) = c * log_{10}(1 + e)$. Er bestimmt, wie stark die Differenz e zwischen erwarteten und tatsächlich erspielten Punkten getrimmt wird. Zusätzlich bildet er in der Funktion $\psi(e) = c * tanh(\frac{e}{c})$ die obere Grenze für die Funktionswerte

von $\psi(e)$. Diese Funktion erfüllt alle soeben erwähnten Kriterien und ist in Abbildung 4 für verschiedene Werte von c abgebildet. Sie ist streng monoton steigend und für alle $e > 0$ gilt sowohl $\psi(e) < c$, als auch $\psi(e) < e$. Daher wird sie im folgenden zur Modifizierung von e verwendet.

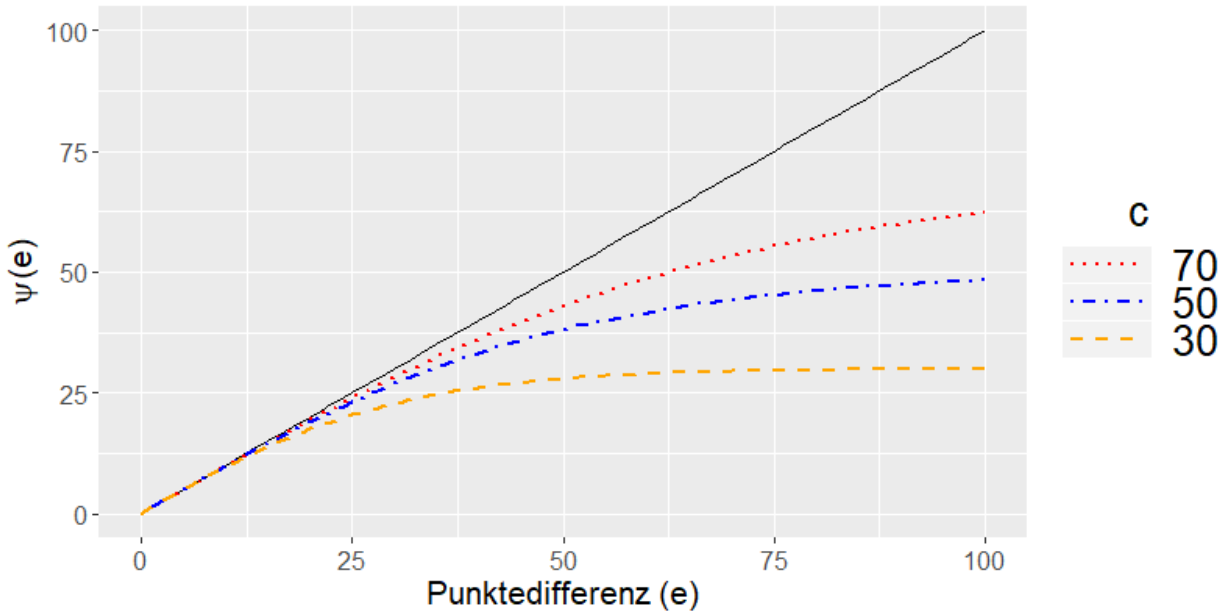

Abbildung 4: Die Funktion $\psi(e) = c * tanh(\frac{e}{c})$ ist streng monoton steigend und liegt für keinen Wert von $e \geq 0$ über der Funktion $\psi(e) = e$ oder übersteigt den Wert c.

Um im nächsten Schritt die erwartete Tordifferenz $\widehat{P_A}$ gegen einen durchschnittlichen Gegner zu bestimmen, verwenden Constantinou und Fenton die Umkehrfunktion von $\psi(e)$. Damit ergibt sich für die von ihnen verwendete Formel $\psi(e) = 3 * log_{10}(1 + e)$ als erwartete Tordifferenz von Mannschaft A mit dem Rating R_A gegen einen durchschnittlichen Gegner:

$$\widehat{P_A} = 10^{\frac{|R_A|}{3}} - 1 \tag{4}$$

Diese muss für das Ratingsystem für Doppelkopf an die gewählte Tangens hyperbolicus Funktion angepasst werden. Für die Umkehrfunktion des Tangens hyperbolicus gilt:

$$-1 < y < 1 \qquad artanh\ y = \frac{1}{2}\ ln\ \frac{1+y}{1-y} \qquad [12, S.\ 74].$$

Durch die Umkehrfunktion der gewählten Trimmungsfunktion $\psi(e) = c * tanh(\frac{e}{c})$ ergibt sich demnach für die zu erwartende Punktzahl eines Spielers A gegen einen durchschnittlichen Gegner :

$$\widehat{P_A} = c * artanh(\frac{|R_A|}{c})$$

Damit können die erwarteten Punkte $\widehat{P_A}$ nur für Spieler mit einem betragsmäßigen Rating kleiner als c berechnet werden damit $\frac{|R_A|}{c} < 1$ gilt.
Für alle Ratings, welche den Wert von c betragsmäßig übersteigen, muss eine andere Lösung gefunden werden. Eine Möglichkeit, die im folgenden verwendet wird, ist es, die erwartete Punktzahl ebenfalls nach oben zu begrenzen. Die erwartete Punktzahl steigt immer weiter an, je näher der Bruch $\frac{|R_A|}{c}$ der eins kommt. Die Grenze wird nun so gewählt, dass der Ausdruck $\frac{|R_A|}{c}$ den Wert 0.99 nicht übersteigt. Der Wert 0.99 liegt nahe an der eins und erlaubt damit bereits einen großen Wert für die erwarteten Punkte. Er kann noch näher an der eins gewählt werden, um die Begrenzung der erwarteten Punkt noch weiter zu erhöhen, oder aber auch gesenkt werden, um die Erwartungen an die Spieler zu verringern. Die erwartete Punktzahl wird damit begrenzt auf $artanh(0.99) * c$.
Für einen Wert von $c = 30$ ergibt dies beispielsweise eine maximal erwartete absolute Punktzahl von 79.4 Punkten gegen einen Durchschnittsgegner. Für $c = 100$ liegt dieser Wert bei 264.7 Punkten. Die Grenze für den Betrag des Ratings R_A, ab dem der Wert von $\widehat{P_A}$ auf den maximal möglichen Wert $artanh(0.99) * c$ gesetzt wird, liegt dementsprechend bei $0.99 * c$.
Somit gilt für die Berechnung von $\widehat{P_A}$:

$$\widehat{P_A} = \begin{cases} c * artanh(\frac{|R_A|}{c}), & |R_A| < 0.99 * c \\ \text{c*artanh(0.99)}, & sonst \end{cases}$$

Da $\widehat{P_A}$ die Erwartungen gegen einen durchschnittlichen Gegner ($R = 0$) angibt, muss das Rating des tatsächlichen Gegners noch berücksichtigt werden. Im Fußball errechnet sich die erwartete Tordifferenz des Gegners gegen eine durchschnittliche Mannschaft analog mit der Formel 4. Da es beim Doppelkopf aber nicht nur einen Gegner gibt, sondern mehrere, muss auch diese Berechnung abgeändert werden. Ein Spieler A spielt auf einem Turnier jede Runde an einem anderen Tisch mit unterschiedlichen Gegnern. Allerdings werden die genauen Paarungen an den Tischen nicht dokumentiert, daher sind die Gegner, gegen die ein Spieler A tatsächlich gespielt hat, nicht zu erkennen und es wird als Gegner das komplette Teilnehmerfeld des Turniers verwendet. Es wird das durchschnittliche Rating R_{GT} der N Turnierteilnehmer, ausgenommen Spieler A, im Turnier T mit dem arithmetischen Mittel bestimmt. Daraufhin wird deren erwartete Punktzahl P_{GT} gegen einen durchschnittlichen Spieler ($R = 0$) ermittelt:

$$\widehat{P_{GT}} = \begin{cases} c * artanh(\frac{|R_{GT}|}{c}), & |R_{GT}| < 0.99 * c \\ \text{c*artanh(0.99)}, & sonst \end{cases}$$

Ist das Rating von Spieler A negativ, so gilt $\widehat{P_A} = -\widehat{P_A}$ und analog bei einem negativen durchschnittlichen Rating der Gegner $\widehat{P_{GT}} = -\widehat{P_{GT}}$.

Die folgenden Schritte sind für Fußball und Doppelkopf gleich und werden hier im Zusammenhang mit Doppelkopf erläutert. Die Punkte , die von Spieler A im Turnier T erwartet werden ergeben sich nun aus:

$$\widehat{P_{AT}} = \widehat{P_A} - \widehat{P_{GT}}$$

Die Einbeziehung der Spielstärke der Gegner ermöglicht es, verschiedene Wettbewerbe miteinander zu vergleichen. Beispielsweise ist eine Deutsche Einzelmeisterschaft oftmals stärker besetzt als eine Regionalmeisterschaft und somit eine hohe Punktzahl in diesem Wettbewerb von größerer Bedeutung. Da von einem Spieler mit hohem Rating bei einem schwächeren Teilnehmerfeld eine höhere Punktzahl $\widehat{P_{AT}}$ erwartet wird, kann so der Schwierigkeitsgrad verschiedener Turniere berücksichtigt werden.
Anschließend kann e berechnet werden durch die Differenz der erwarteten Punkte $\widehat{P_{AT}}$ von Spieler A in Turnier T und den tatsächlich erspielten Punkten P_{AT} von Spieler A in Turnier T

$$e = |P_{AT} - \widehat{P_{AT}}|$$

Gemäß Formel (3) werden nun $\psi_A(e)$ und $\psi_G(e)$ berechnet

$$\psi_A(e) = \begin{cases} \psi(e), \ \widehat{P_{AT}} < P_{AT} \\ -\psi(e), \ sonst \end{cases} \qquad \psi_G(e) = \begin{cases} \psi(e), \ \widehat{P_{AT}} > P_{AT} \\ -\psi(e), \ sonst \end{cases}$$

und anschließend die Pi-Ratings aktualisiert:

$$R'_A = R_A + \psi_A(e) * \lambda \qquad R'_G = R_G + \frac{1}{N} * \psi_G(e) * \lambda$$

Dabei ist zu beachten, dass das Rating von jedem der N Gegner von Spieler A mit Hilfe von $R_G = R_G + \frac{1}{N} * \psi_G(e) * \lambda$ aktualisiert werden muss.

In Abbildung 5 sind die einzelnen Schritte zur Aktualisierung des Pi-Ratings eines Turnierteilnehmers vereinfacht dargestellt. Diese Aktualisierung muss nach Erhalt der Ergebnisse eines Turniers für jeden Turnierteilnehmer vorgenommen werden.

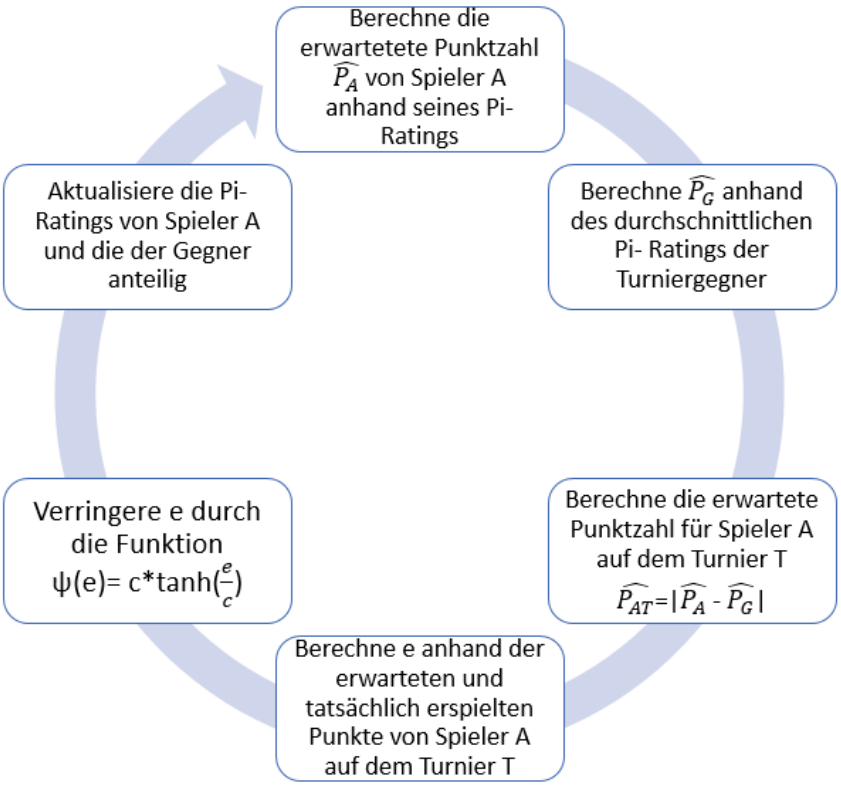

Abbildung 5: In Anlehnung an [9, S. 40]
Die Abbildung zeigt die einzelnen Schritte, die nach einem Turnier T für jeden Turnierteilnehmer durchgeführt werden müssen.

5 Anwendung

5.1 Datengrundlage und Deskription

Um das Ratingsystem zu erstellen stehen Daten der Jahre 1984 bis 2017 zur Verfügung. Dabei werden die Wettbewerbe Deutsche Einzelmeisterschaft, Regionalmeisterschaft und Ranglistenturnier berücksichtigt. Hier liegen Informationen zu den Spielern, welche durch Vor- und Nachnamen eindeutig identifiziert werden können und die jeweiligen im Turnier erspielten Punktzahlen vor. Trotz größter Sorgfalt in der Datenverwaltung entsteht hier eine mögliche Fehlerquelle, da es passieren kann, dass eine Person nach einer Namensänderung, beispielsweise nach einer Heirat oder durch das Erlangen eines akademischen Titels, im Datensatz unter zwei verschiedenen Namen geführt wird. Insgesamt wurden seit 1984 in den drei Wettbewerbsarten 1099 Turniere ausgetragen an denen 4755 verschiedene Spieler teilgenommen haben. Es kann vorkommen, dass ein Turnier durch Ersatzspieler aufgefüllt wird, um eine durch vier teilbare Teilnehmerzahl zu erreichen. Diese spielen bei Regional- und Einzelmeisterschaften außerhalb der Wertung. Daher können sie nicht ins Teilnehmerfeld eingerechnet werden und die erspielten Punkte verfallen. Zwar liegen auch Daten der deutschen Mannschaftsmeisterschaft vor, allerdings eignen sich diese aus verschiedenen Gründen nicht. Zum einen ist durch die Option der Auswechslung nicht bekannt, welcher Spieler wie viele Runden in welchem Stadium der K.O Phase gespielt hat. Damit gibt es keine Informationen darüber, wer wie viele Runden in welchem Teilnehmerfeld und mit welchem Ergebnis gespielt hat. Zum anderen kann sich die Spielweise von Spielern im Mannschaftswettbewerb im Vergleich zum Einzelwettbewerb stark unterscheiden, indem sie beispielsweise weniger Risiko eingehen. Daher scheint ein vermischtes Rating von Mannschafts- und Einzelwettbewerben nicht sinnvoll. Aus dem Bundesliga Wettbewerb liegen keine Daten vor, allerdings sollten diese aus den gleichen Gründen wie bei der Deutschen Mannschaftsmeisterschaft ohnehin nicht verwendet werden.
Die folgenden Boxplots in Abbildung 6 zeigen die von den Spielern erreichten Punktzahlen der Wettbewerbe Deutsche Einzelmeisterschaft, Regionalmeisterschaft und Ranglistenturniere. Die Punkte der Einzelmeisterschaft und Regionalmeisterschaft verhalten sich sehr ähnlich. Das erste Quartil liegt bei -51 beziehungsweise -50 Punkten, das dritte Quartil bei 53 beziehungsweise 52 Punkten. Bei Ranglistenturnieren liegen diese deutlich näher zusammen, das erste Quartil bei -33 und das zweite bei 33 Punkten. Es lässt sich vermuten, dass dies an der höheren Rundenzahl liegt, welche bei Einzel- und Regionalmeisterschaften bei sechs und seit 1996 bei acht Runden liegt, während Ranglistenturniere nur über drei Runden ausgetragen werden. Der Median aller drei Wettbewerbe liegt nahe bei null.

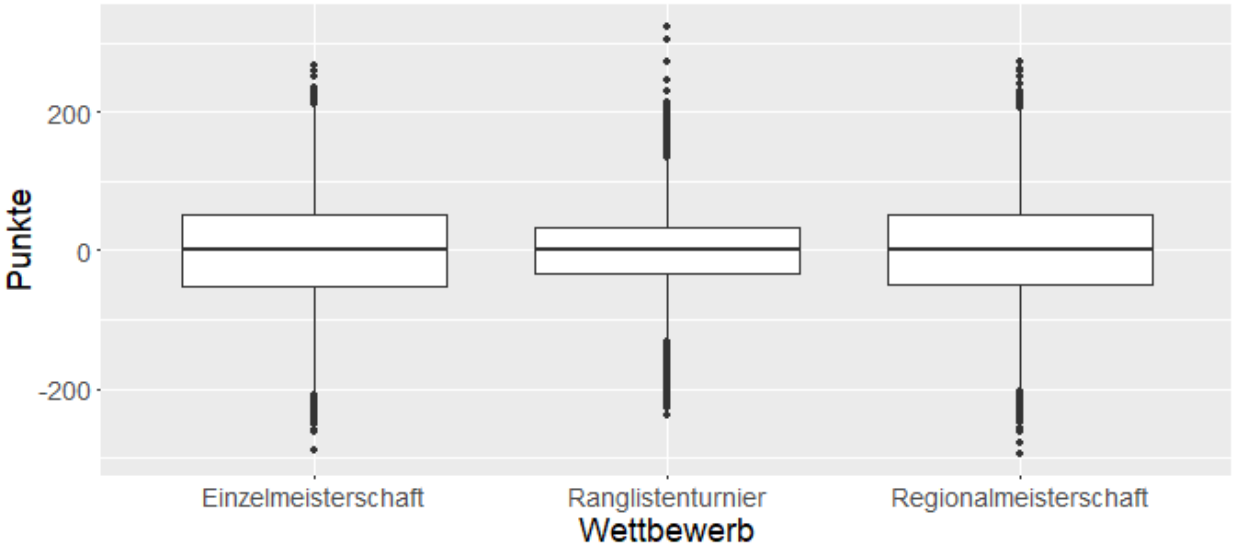

Abbildung 6: Die erspielten Punkte in Einzelmeisterschaft und Regionalmeisterschaft sind ähnlich, das erste Quartil liegt bei ca. -50, das zweite bei ca. 50 Punkten. Die Quartile der Ranglistenturnieren liegen enger zusammen bei -33 und 33 Punkten. In allen drei Wettbewerbsarten liegt der Median ungefähr bei null.

Wie sich die Punkte von einzelnen Spielern verhalten ist in Abbildung 7 zu sehen. Auf der x-Achse sind die Turniernummern dargestellt, welche durch eine chronologische Anordnung und darauffolgende Nummerierung der drei Wettbewerbsarten entstehen. Die y-Achse zeigt die erreichten Punkte in den jeweiligen Turnieren und es sind die vier Spieler mit den meisten gespielten Turnieren abgebildet. Eine sichtbare Struktur lässt sich nicht erkennen. Alle Spieler weisen sowohl positive, als auch negative Spielergebnisse mit ähnlichen Ausmaßen auf. Die Turnierergebnisse liegen in einem Bereich von -200 bis 200 Punkte.

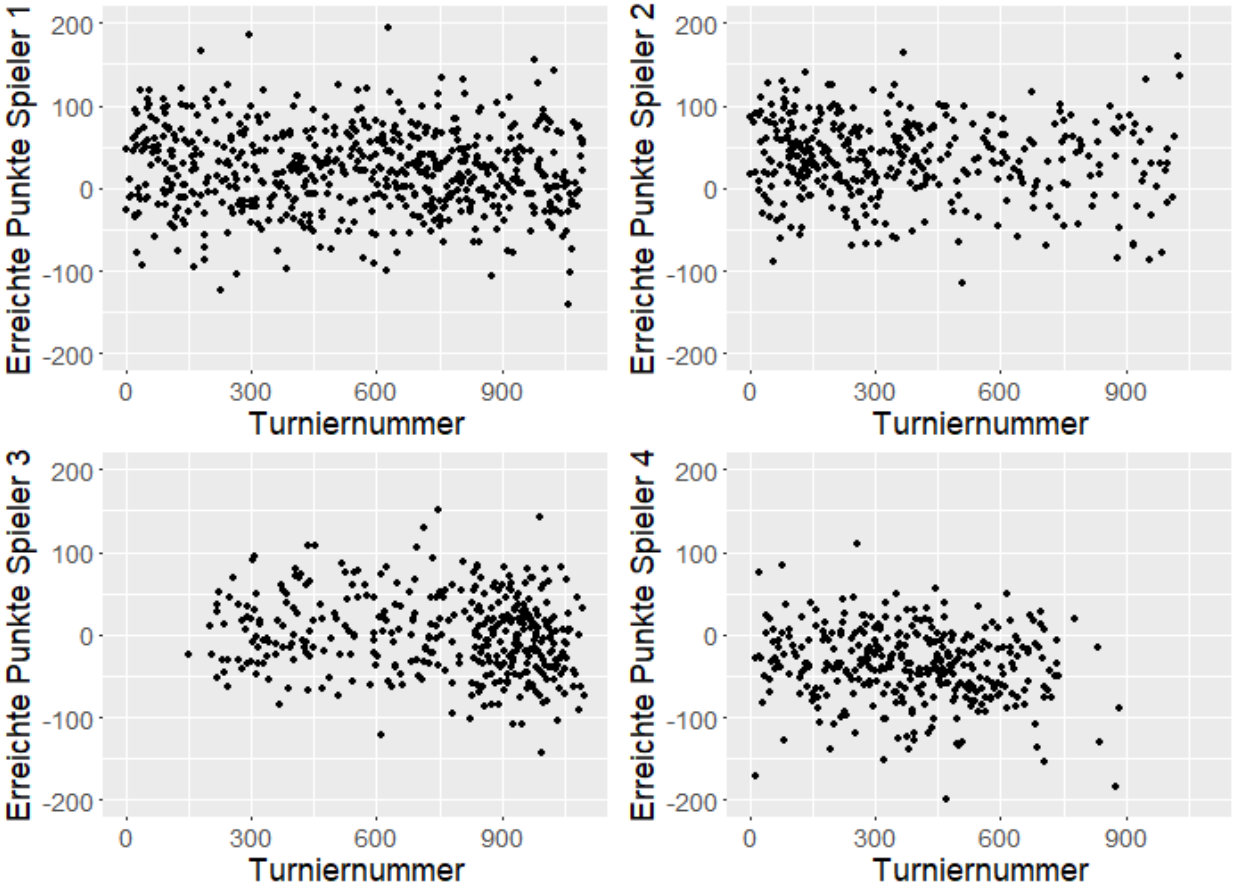

Abbildung 7: Die erspielten Punkte der vier Spieler mit den meisten gespielten Turnieren weisen keine sichtbare Struktur auf.

Ein weiterer Aspekt ist die Anzahl an Turnieren, die ein Spieler absolviert hat. Das Histogramm in Abbildung 8 zeigt in x-Richtung die Anzahl an gespielten Turnieren einer Person, und in y-Richtung die absolute Häufigkeit der Anzahl an gespielten Turnieren. Zusätzlich gibt es 67 Spieler, die mehr als 200 Turniere absolviert haben, welche zur Übersichtlichkeit nicht eingezeichnet werden. Es lässt sich erkennen, dass über 2500 Personen, also mehr als die Hälfte aller im Datensatz vorkommender Personen, fünf oder weniger Turniere gespielt haben. Analog zum Wertungssystem nach Elo sollen Spieler erst aufgenommen werden wenn sie an fünf oder mehr Turnieren teilgenommen haben [13]. In der Anwendung auf den vorhandenen Datensatz, welcher auf historischen Daten basiert, bedeutet dies, dass alle Spieler mit weniger als fünf gespielten Turnieren für das Rating ignoriert werden. Damit verbleiben im Datensatz 2195 Spieler. In einem laufenden System gäbe es verschiedene Möglichkeiten zu verfahren. Zum einen können die Punkte der ersten vier Turniere nachträglich berücksichtigt werden, sobald der Spieler sein fünftes Turnier angetreten hat. Da dies jedoch auch die Ratings aller anderen Teilnehmer der ersten vier Turniere beeinflusst, erscheint diese Methode nicht praktikabel. Die ersten vier Turniere könnten verfallen und somit als

eine Art Eingewöhnung in den Turnierbetrieb angesehen werden und der Spieler würde ab seinem fünften Turnier mit einem Rating von null einsteigen. Eine andere Methode wäre es, analog zu Elo, aus den ersten vier Turnieren eine Spielstärke zu schätzen und diese als Grundlage für weitere Ratings zu verwenden.

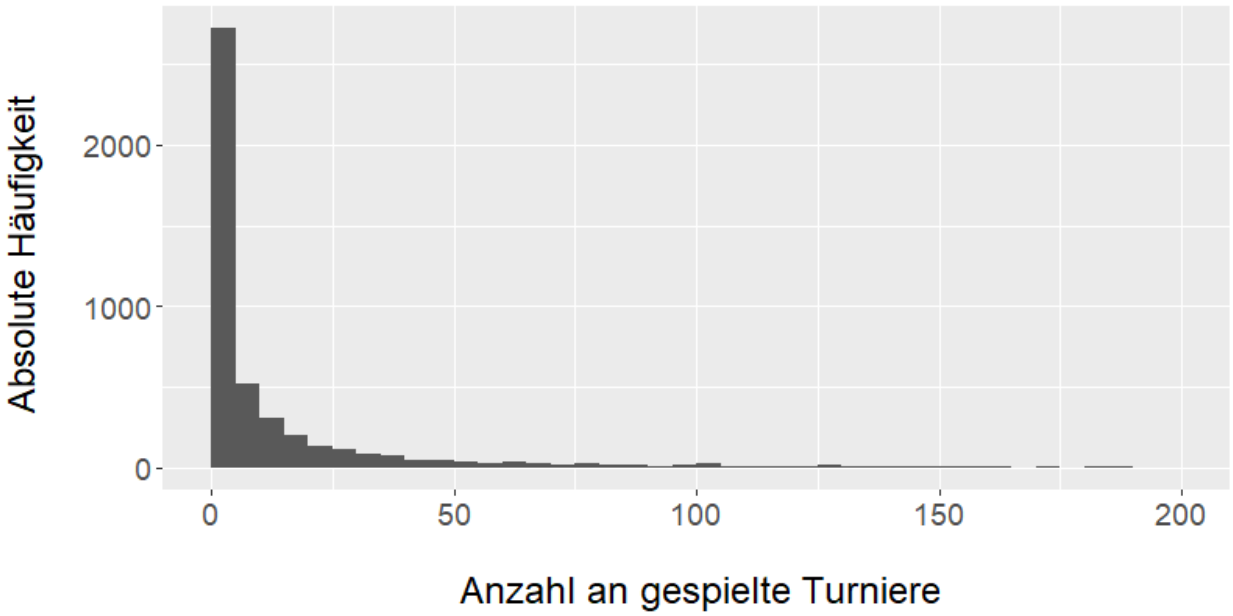

Abbildung 8: Das Histogramm zeigt, dass mehr als 2500 Spieler nur fünf oder weniger Turniere gespielt haben. Das sind mehr als die Hälfte aller aufgenommener Spieler.

5.2 Parameterbestimmung

Um das Rating tatsächlich zu berechnen müssen die Parameter λ und c bestimmt werden.
Letzterer trimmt die entstehende Differenz zwischen der erwarteten und der tatsächlich erspielten Punktzahl eines Spielers A und begrenzt sie auf einen Maximalwert (siehe Abbildung 4). Der Parameter λ regelt, wie stark neue Turnierergebnisse das alte Rating eines Spielers überschreiben. Dies entspricht bei der Berechnung der Elo-Zahl dem K-Faktor. Zu dessen Bestimmungen wurden bereits zahlreiche Überlegungen veröffentlicht, daher sollen die wichtigsten Punkte eines solchen Faktors hier im Zusammenhang mit Schach erläutert werden.
Apard Elo entschied sich bei der Einführung seines Systems dafür, den K-Faktor für Spieler mit einem niedrigen Rating geringer zu wählen, als für Spieler mit hohem Rating. Der Statistiker Jeff Sonas kritisierte daran, dass die Elo-Zahl dann nicht schnell genug auf nachlassende Spielstärke bei Spielern mit hoher Wertung reagieren könne. Mark Glickman hingegen hat ein System vorgeschlagen, bei dem der K-Faktor von der Zuverlässigkeit des Ratings eines Spielers abhängt. Je weniger Spiele bisher in das Rating eines Spielers eingegangen sind und je länger das letzte Spiel

zurück liegt, desto unzuverlässiger ist dessen Elo-Zahl.[10]
In der Fédération Internationale des Échecs(FIDE), dem internationalen Schachverband, wird der K-Faktor momentan wie folgt gewählt:
Für Spieler mit weniger als 30 absolvierten Spielen, oder Spieler unter 18 Jahren und einem Rating unter 2300 wird ein Faktor von 40 verwendet. Spieler mit einem Rating unter 2400 erhalten einen K-Faktor von 20. Hat ein Spieler eine Elo-Zahl von 2400 erreicht und bleibt danach auf diesem Niveau, selbst wenn der Wert unter 2400 fällt, so beträgt der Faktor 10. Übersteigt das Produkt aus der Anzahl der gespielten Spiele eines Spielers und seines Faktors K den Wert 700, so wird K ganzzahlig verringert, bis der Wert des Produkts unter 700 fällt. [13]
Somit werden von der FIDE die Vorschläge von Apard Elo und Mark Glickman kombiniert.

Die soeben dargestellten Überlegungen treffen auch auf den Faktor λ des Pi-Ratings zu. Um den Rahmen der Arbeit jedoch nicht zu überschreiten, soll λ hier für alle Spieler gleich gewählt werden. In Abbildung 9 sieht man die Entwicklung des Ratings eines Spielers (y-Achse) über die ersten 150 Turniere hinweg (x-Achse) für zwei verschiedene Werte von λ bei gleicher Wahl von $c = 70$. Es lässt sich erkennen, dass die Struktur der Verläufe gleich ist. Jedoch hat ein größerer Wert von λ stärkere Veränderungen im Rating einer Person zur Folge.

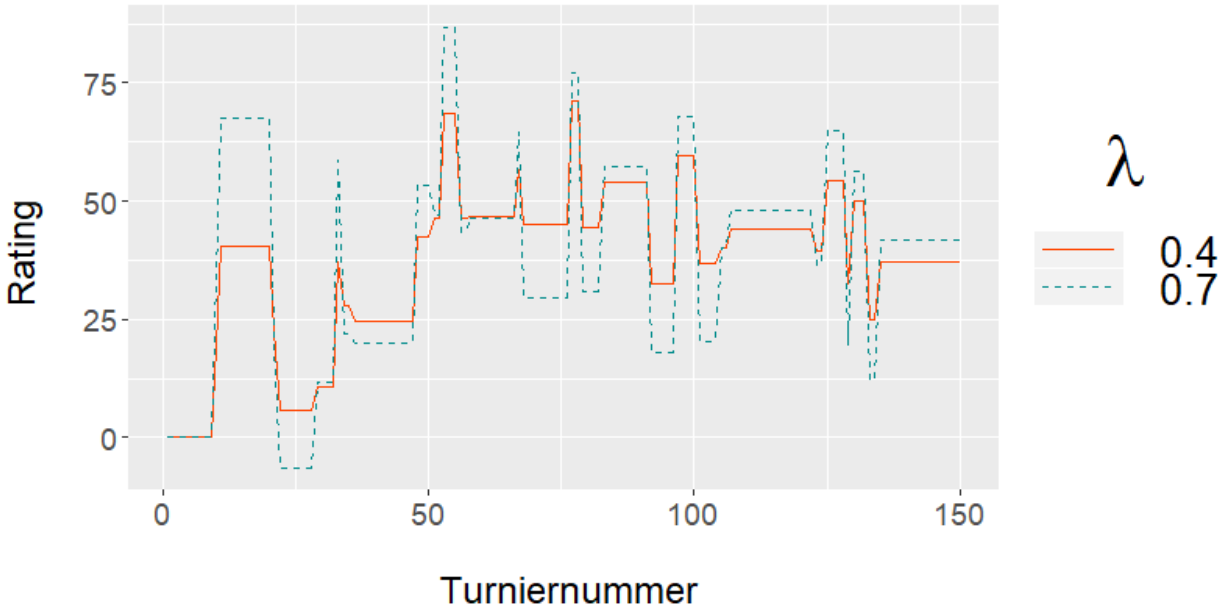

Abbildung 9: Die Struktur der Verläufe ist für die beiden Werte von λ gleich, jedoch sind die Veränderungen im Rating für ein kleineres λ geringer.

In den nächsten zwei Kapiteln werden zwei Verfahren zur Wahl der optimalen Parameter c und λ angewendet. Jedoch sollen zuerst einige Überlegungen zur Wahl der zu testenden Wertebereiche

vorgestellt werden. Wie in Abbildung 7 gezeigt wurde, wiesen die vier Spieler mit den meisten gespielten Turnieren sowohl positive, als auch negative Ergebnisse ohne erkennbare Struktur auf. Somit kommen Wechsel von positiven zu negativen Ergebnissen und umgekehrt vor. Es ist davon auszugehen, dass dieses Phänomen aufgrund des zuvor beschriebenen Einflusses von Glück und Pech nicht nur bei diesen vier Spielern, sondern im kompletten Datensatz zu finden ist. Zusätzlich ist in Abbildung 6 zu erkennen, dass die Interquartilsabstände für Ranglistenturniere bei 66 Punkten, für die Deutsche Einzelmeisterschaft und Regionalmeisterschaften bei knapp über 100 Punkten liegen. Das bedeutet, dass mindestens 50% aller entstehenden Differenzen zwischen zwei Ergebnissen in einem Wertebereich von null bis 110 liegen. Fasst man diese beiden Erkenntnisse zusammen, erscheint ein Wertebereich von null bis 110 für c sinnvoll. Die Differenzen zwischen erwarteten Punkten, welche aus vorhergehenden Leistungen berechnet werden, und tatsächlich erspielten Punkten sollten durch diesen Wertebereich größtenteils abgedeckt sein. Bei Abweichungen von mehr als 110 Punkten ist von einem starken Einfluss von Glück oder Pech auszugehen, weswegen eine Begrenzung angebracht ist. Daher werden für c Werte zwischen null und 110 betrachtet.
In Abbildung 9 wurden die Auswirkungen des Parameters λ dargestellt. Dieser regelt wie bereits beschrieben, wie stark das bestehende Rating von neuen Turnierergebnissen überschrieben wird. Theoretisch liegt λ in einem Bereich von null bis eins. Während ein Wert von null bedeutet, dass neue Turniere gar nicht in das bestehende Rating eingehen, würde bei einem Wert von eins der komplette Wert von $\psi(e)$ in das Rating einfließen. In Anlehnung an Constantinou und Fenton, welche für λ Werte zwischen 0.005 und 0.095 betrachteten [9], wird eine maximalen Einbeziehung neuer Ergebnisse zu 9.5% Prozent gewählt.

5.2.1 Maximierung der Prognosegüte

Zur Bestimmung der Parameter γ (zur Gewichtung von Heim- und Auswärtsspielen) und λ im Pi-Ratingsystem, auf dem das System für Doppelkopf beruht, haben Anthony Costa Constantinou und Norman Elliott Fenton für verschiedene Kombinationen dieser Parameter die Summe der quadratischen Abweichungen (e^2) von prognostizierten Tordifferenzen und tatsächlicher Tordifferenz berechnet. Die kleinste quadratische Abweichung ergab sich für $\lambda = 0.035$ und $\gamma = 0.7$. Leonhard Knorr-Held(2000) hat verschiedene Vorgehensweisen zur Bestimmung eines Glättungsparameters in einem Ratingsystem über die Maximierung der Prognosegüte vorgestellt. Neben der oben genannte Verwendung der Summe der quadratischen Abweichungen wird als eine andere Methode die Wahl der Summe der absoluten Abweichungen vorgeschlagen.[14]
Eine andere Methode, die Anzahl der exakten Prognosen zu verwenden, scheint im Fall von Doppelkopf nicht sinnvoll, da diese durch die weit gestreuten Punktzahlen und den beschriebenen

Faktor von Glück und Pech, so gut wie nie erreicht werden kann. In Abbildung 10 wird sowohl der durchschnittliche absolute Fehler e als auch der durchschnittliche quadratische Fehler e^2 für verschiedene Kombinationen von λ und c dargestellt. Für c werden entsprechend den Überlegungen aus Kapitel 5.2 Werte von 10 bis 110 und einer Schrittlänge von 20 betrachtet und für λ Werte von 0 bis 0.095 mit einer Schrittlänge von 0.005. Zwar ist der Wert $\lambda = 0$ inhaltlich nicht sinnvoll, da dadurch neue Turnierergebnisse nicht in das Rating einbezogen werden und sämtliche Ratings der Spieler bei null bleiben, dennoch können durch die Betrachtung dieses Wertes Erkenntnisse gewonnen werden. Läge die maximale Prognosegüte bei einem Wert von $\lambda = 0$ vor, so würde dies bedeuten, dass eine Erhöhung oder Verringerung des Ratings der Spieler durch das Pi-Rating zu einem höheren Prognosefehler führt. Dies könnte zum Beispiel der Fall sein, wenn positive und negative Ergebnisse bei jedem Spieler im Wechsel vorkämen.
Abbildung 10 zeigt, dass dies jedoch nicht der Fall ist. Sowohl der quadratische, als auch der absolute Fehler sind für den Wert $\lambda = 0$ in jeder Kombination mit c maximal. Es ist zu erkennen, dass die Struktur der Verläufe für e und e^2 ähnlich sind. Die Kurven fallen fast alle ca. bis zu einem Wert von $\lambda = 0.05$ und steigen danach wieder an. Lediglich die Kurven mit einem Wert von $c = 10$ fallen kontinuierlich und es ist nicht zu erkennen, ob diese im weiteren Verlauf nochmals ansteigen. Die Minima für λ liegen nah zusammen, für den absoluten Fehler bei $\lambda = 0.045$ und für den quadratischen Fehler bei $\lambda = 0.05$. Sowohl der absolute, als auch der quadratische Fehler erreichen ihr Minimum für einen Wert von $c = 110$.

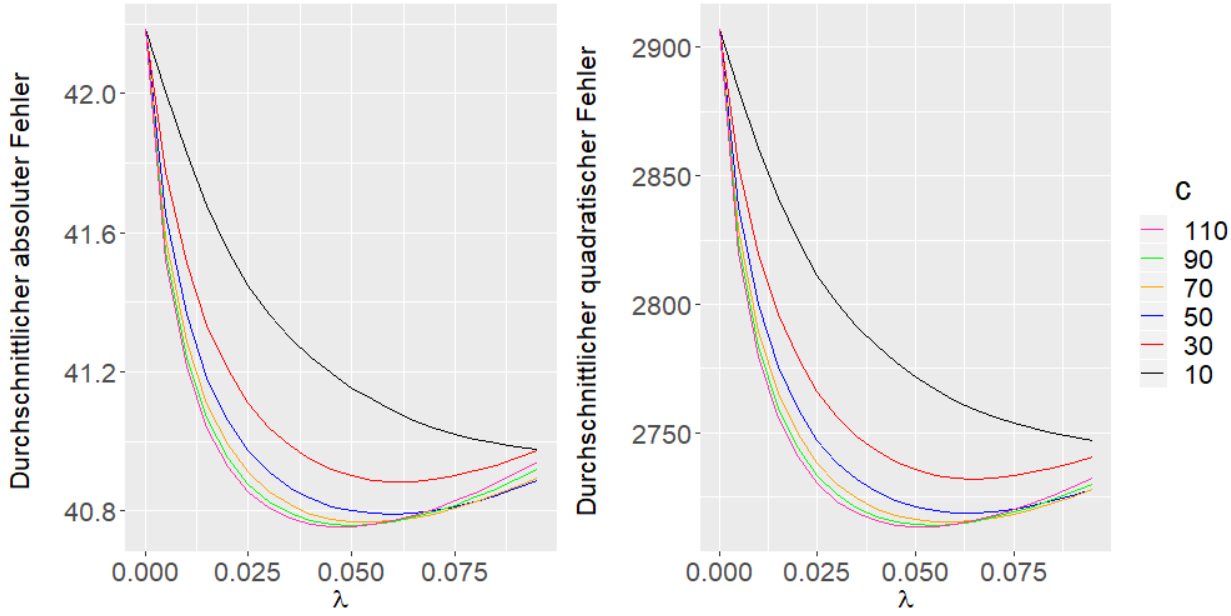

Abbildung 10: Die Grafik zeigt die durchschnittlichen absoluten und quadratischen Abweichungen von erwarteten und tatsächlich erspielten Punktzahlen für verschiedene Parameterkombinationen. Das Minimum liegt in beiden Fällen bei einer Wahl von $c = 110$ und $\lambda = 0.045$ bei absoluten Abweichungen, beziehungsweise $\lambda = 0.05$ bei quadratischen Abweichungen.

Inhaltlich bedeutet dies, dass die Fehler bei einer Wahl von einer hohen Schranke für die hyperbolische Tangensfunktion minimal werden. Die Differenz zwischen erwarteten und tatsächlich erspielten Punkten eines Spielers sollen demnach nur auf Werte unter 110 Punkte begrenzt werden. Der Wert von $\lambda = 0.045$ beziehungsweise $\lambda = 0.05$ bedeutet, dass neue Turnierergebnisse zu 4.5% beziehungsweise zu 5% eingehen sollen.

5.2.2 Unterscheidung und Stabilität

Nach Franks, D'Amour et al. (2016) sind zwei Kriterien für ein Ratingsystem besonders wichtig. Zum einen soll das System zuverlässig zwischen den Spielern unterscheiden können, zum anderen soll es stabil über die Zeit sein [7, S. 152]. Die Grundidee der Autoren, Unterscheidung und Stabilität über eine Streuungszerlegung zu quantifizieren, wird hier in einer abgeänderten Form angewendet, um die Parameter λ und c zu bestimmen.
Die Streuungszerlegung besagt, dass die Gesamtvarianz ausgedrückt werden kann durch die Summe aus der Varianz zwischen den Schichten und der Varianz innerhalb der Schichten [15, S. 73]. Die einzelnen Schichten stellen bei einem Ratingsystem für Doppelkopf die verschiedenen Personen p dar. Mit P wird die Gesamtanzahl der Personen im Ratingsystem bezeichnet. Die Anzahl an vorgenommenen Aktualisierungen von Ratings im Ratingsystem wird mit n bezeichnet und n_p

beschreibt die Anzahl an Aktualisierungen von Ratings einer Person p.
Für die Streuungszerlegung ergibt sich damit:

$$\underbrace{\frac{1}{n}\sum_{p=1}^{P}\sum_{j=1}^{n_p}(x_{pj}-\bar{x})^2}_{Gesamte\ Streuung} = \underbrace{\frac{1}{n}\sum_{p=1}^{P}n_p(\bar{x}_p-\bar{x})^2}_{Streuung\ zwischen\ den\ Personen} + \underbrace{\frac{1}{n}\sum_{p=1}^{P}\sum_{j=1}^{n_p}(x_{pj}-\bar{x}_p)^2}_{Streuung\ innerhalb\ einer\ Person}$$

wobei $\bar{x}_p$ das arithmetische Mittel der verschiedenen Ratings eines Spielers darstellt und $\bar{x} = \frac{1}{n}n_p\bar{x}_p$.
Als Maß für die Unterscheidung wird nun der Anteil gewählt, den die Varianz zwischen den Personen an der gesamten Varianz hat (Z).

$$Z = \frac{\frac{1}{n}\sum_{p=1}^{P}n_p(\bar{x}_p-\bar{x})^2}{\frac{1}{n}\sum_{p=1}^{P}\sum_{j=1}^{n_p}(x_{pj}-\bar{x})^2}$$

Analog dazu wird für das Maß der Stabilität der Anteil der Varianz innerhalb einer Person an der gesamten Varianz gewählt (I).

$$I = \frac{\frac{1}{n}\sum_{p=1}^{P}\sum_{j=1}^{n_p}(x_{pj}-\bar{x}_p)^2}{\frac{1}{n}\sum_{p=1}^{P}\sum_{j=1}^{n_p}(x_{pj}-\bar{x})^2}$$

Damit summieren sich die Maße für Unterscheidung und Stabilität immer zu eins auf. Die Parameter λ und c können nun so gewählt werden, dass der Anteil der Streuung zwischen den Schichten Z möglichst groß ist und der Anteil der Streuung innerhalb einer Person I möglichst klein. Somit sind die Parameter optimal, für die der Quotient $\frac{I}{Z}$ minimal ist.
Das Ratingsystem wurde wiederum für die gleichen Parameterkonstellationen berechnet wie in Kapitel 5.2.1, für λ Werte zwischen 0.005 und 0.095 mit einer Schrittlänge von 0.005 und für c Werte zwischen 30 und 110 mit einer Schrittlänge von 20. Für einen Wert von $\lambda = 0$ ist diese Berechnung nicht sinnvoll, da neue Ergebnisse nicht mit einbezogen werden und somit alle Ratings bei null bleiben und keine Streuung im System entsteht.
In Abbildung 11 sind die Quotienten $\frac{I}{Z}$ für die verschiedenen Werte von c und λ dargestellt. Die Kurven haben alle einen ähnlichen Verlauf. Sie fallen zunächst ab und steigen daraufhin wieder an. Für fast alle Werte von c liegt das Minimum der Kurve für Werte von λ zwischen 0.015 und 0.025 vor. Lediglich für $c = 10$ wird das Minimum erst bei $\lambda = 0.04$ erreicht. Der minimale Wert des Quotienten $\frac{I}{Z}$ liegt für eine Parameterkonstellation von $c = 110$ und $\lambda = 0.015$ vor. Inhaltlich hätte dies zur Folge, dass neue Turnierergebnisse nur zu 1.5% in das bestehende Rating eingehen. Wie schon bei der maximalen Prognosegüte soll für c der Wert 110 gewählt werden und somit eine hohe Schranke für die hyperbolische Tangensfunktion.

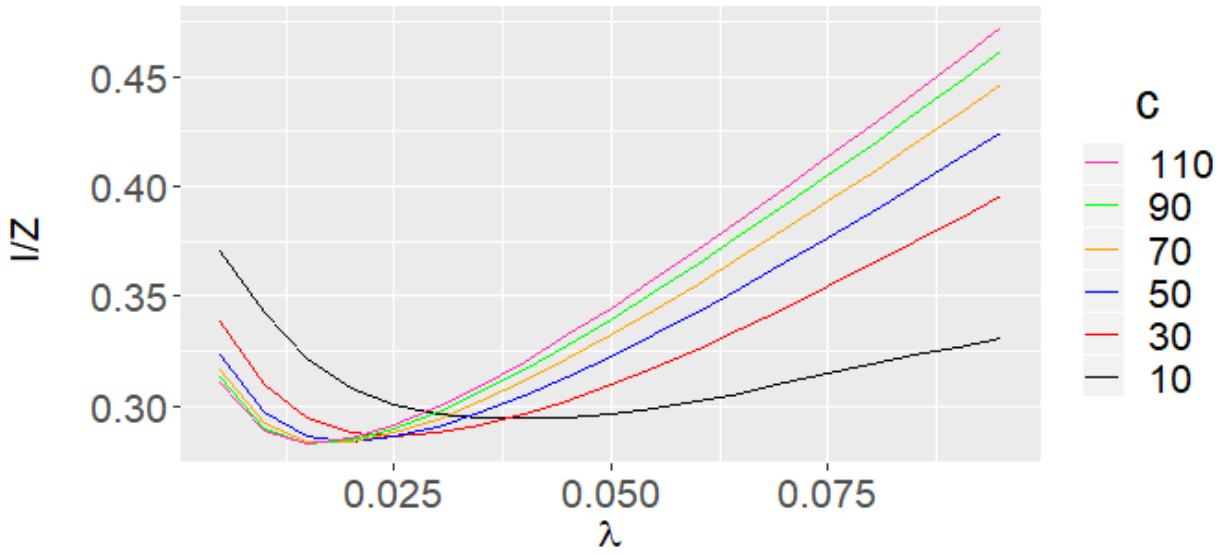

Abbildung 11: Der Quotient aus dem Anteil an Stabilität und dem Anteil der Unterscheidung an der Gesamtvarianz ($\frac{I}{Z}$) ist minimal für die Werte $\lambda = 0.015$ und $c = 110$.

5.2.3 Expertenbefragung

Um zu entscheiden, welcher Wert für λ aus den vorgestellten Verfahren besser geeignet ist, könnte eine Expertenumfrage verwendet werden. Warum diese im Fall von Doppelkopf zu keinem verwertbaren Ergebnis führt, wird im folgenden Kapitel erläutert. In einer beispielhaften Umfrage wurden neun Personen ausgewählt, die in den Ratingsystemen nahe zusammen liegen, sodass verschiedene Parameterkonstellationen zu verschiedenen Rankings dieser neun Personen führen. Es wurden sechs mögliche Anordnungen dieser 9 Personen zur Wahl gestellt. Aus Datenschutzgründen können diese sechs Optionen nicht näher charakterisiert werden. Die Teilnehmer wurden daraufhin gefragt, welche dieser sechs Anordnungen ihrer Meinung nach am ehesten einer Ordnung nach Spielstärke entspricht. Sie wurden gebeten, nur an der Umfrage teilzunehmen, falls sie in letzter Zeit mit all diesen neun Spielern gespielt haben, sodass sie die aktuelle Spielstärke auch beurteilen können. Die Umfrage wurde durch den Deutschen Doppelkopf-Verband e. V. publiziert und es haben 60 Personen daran teilgenommen. Das Ergebnis ist in Abbildung 12 dargestellt. Die Optionen wurden der Größe nach angeordnet und mit eins bis sechs benannt, sodass die Antwortmöglichkeiten für die Teilnehmer nicht mehr nachvollziehbar sind.
Es lässt sich erkennen, dass keine der Optionen eine deutliche Mehrheit erhalten hat. Option 1 wurde 14 mal gewählt (23.3%) und erhielt damit die meisten Stimmen. Für Option 2 stimmte nur eine Person weniger und für Option 3 wiederum eine Person weniger. Die übrigen Optionen erhielten acht, sieben und sechs Stimmen und liegen somit ebenfalls nahe zusammen.

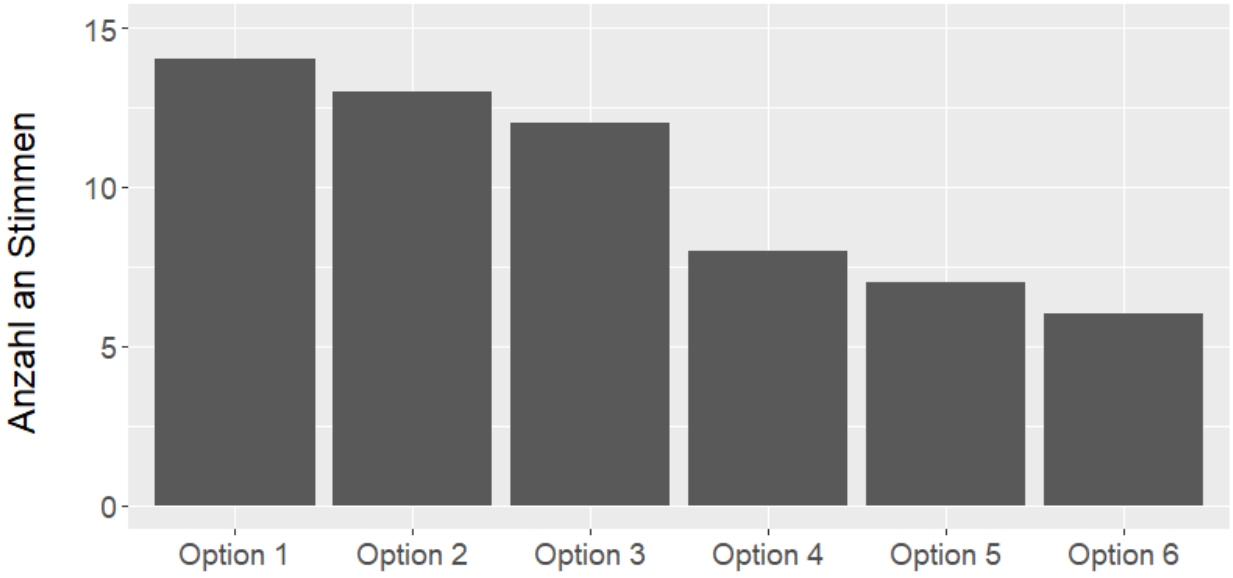

Abbildung 12: Eine Anordnung nach Spielstärke führt in einer Expertenbefragung zu keinem eindeutigen Ergebnis.

Ein Grund für die Unterschiedlichen Wahrnehmungen von Spielstärke könnte sein, dass ein Großteil der Spieler seit vielen Jahren zusammen Doppelkopf spielt. Daher könnten die Einschätzungen der aktuellen Spielstärke durch länger zurückliegende Erfahrungen mit den zur Wahl stehenden Spielern beeinflusst werden. Hat ein Spieler in den letzten Jahren an Spielstärke verloren oder sich im Gegenzug deutlich verbessert, so könnte das Bild, dass ein Teilnehmer von diesem Spieler hat, veraltet sein.
Die eigene Art zu spielen, also der Stil eines jeden Teilnehmers kann sich ebenfalls auf die Beurteilung anderer Spieler auswirken. Diese kommen eventuell besser mit dem Stil eines der zur Wahl stehenden Spielers zurecht als mit anderen und schätzen diesen folglich stärker ein. Auch kann die eigene Leistung während eines Aufeinandertreffens die Bewertung beeinflussen. Hat man mit einem der neun Spieler am Tisch ein besonders gutes Ergebnis erzielt, behält man ihn möglicherweise positiver in Erinnerung, unabhängig von der tatsächlich erbrachten Leistung dieses Spielers. Faktoren wie Sympathie oder Vereinszugehörigkeit können ebenfalls Auswirkungen auf die Einschätzung haben.

5.3 Darstellung des Ratingsystems

Im folgenden Kapitel wird das Ratingsystem für eine bestimmte Parameterkonstellation dargestellt. Es wird die Konstellation aus der Maximierung der Prognosegüte gewählt, da dies ein etabliertes statistisches Verfahren ist und die Methode, die Constantinou und Elliott bei der ur-

sprünglichen Vorstellung des Pi-Ratingsystems verwendet haben. Es wird das Ergebnis des durchschnittlichen absoluten Fehlers verwendet, da hier große Abweichungen, welche beispielsweise durch über ein Turnier hinweg anhaltendes Glück oder Pech zu Stande kommen können, weniger stark ins Gewicht fallen. Im folgenden wird demnach das Ratingsystem für die Parameterkonstellation von $\lambda = 0.045$ und $c = 110$ vorgestellt.
In Abbildung 13 ist der Verlauf der Ratings für zwei Spieler dargestellt. Zum einen für den Spieler mit dem am Ende höchsten Rating und zum anderen für den Spieler mit dem am Ende niedrigsten Rating. Es ist zu sehen, dass beide Spieler Auf- und Abwärtsbewegungen in den Verläufen ihrer Ratings aufweisen. Eine Abwärtsbewegung bei dem Spieler mit hohem Rating bedeutet nicht zwingend, dass er negative Ergebnisse erspielt hat, sondern lediglich, dass er weniger Punkte erspielt hat, als man von ihm erwartet hat.

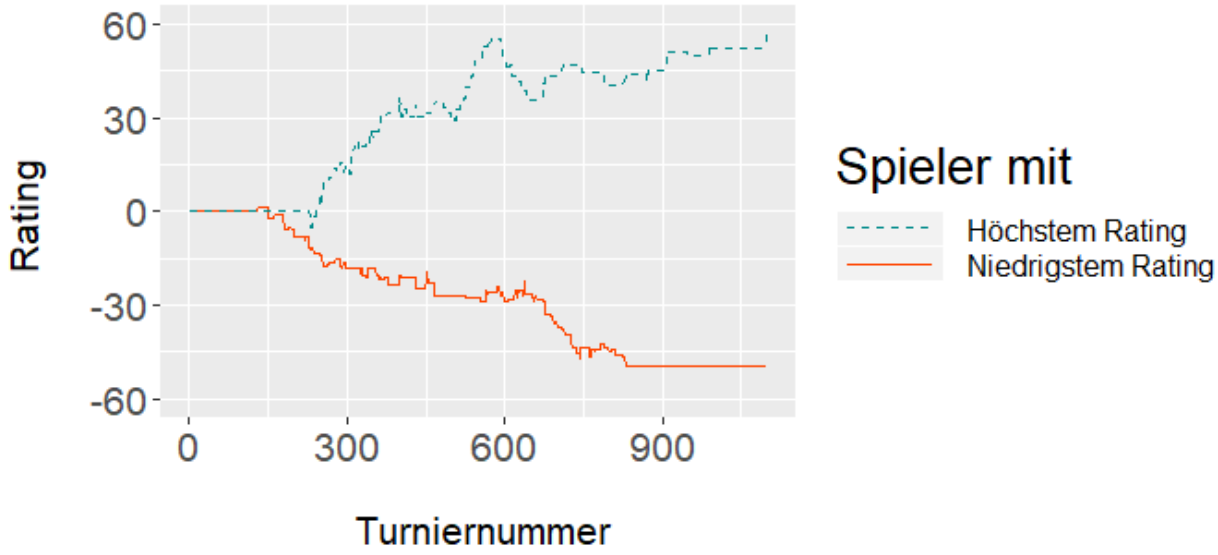

Abbildung 13: Die Grafik zeigt die Verläufe der Ratings für den Spieler mit dem am Ende höchsten und niedrigstem Rating.

Abbildung 14 zeigt die Verteilung der Ratings über die Jahre. Es wurde in jedem Jahr der Zeitpunkt der Deutschen Einzelmeisterschaft gewählt und ein Boxplot von allen sich im Bewertungssystem befindenden Ratings erstellt. Es lässt sich erkennen, dass sich das Ratingsystem bis ca. 1997 aufbaut. Da sich zu Beginn der Daten, also zum ersten Turnier 1984, alle Ratings bei null befinden, dauert es etwas, bis die Spieler das Rating ihrer Spielstärke erreicht haben, da neue Ergebnisse das alte Rating nur zu 5% überschreiben. Auch werden über die Jahre immer mehr Spieler in das Ratingsystem aufgenommen. Ab 1997 bleibt die Box (die mittleren 50% der Daten) in einem ähnlichen Wertebereich und auch die Ausreißer nehmen ähnliche Ausmaße an. Somit sollte ab diesem Zeitpunkt eine Umverteilung der Punkte stattfinden.

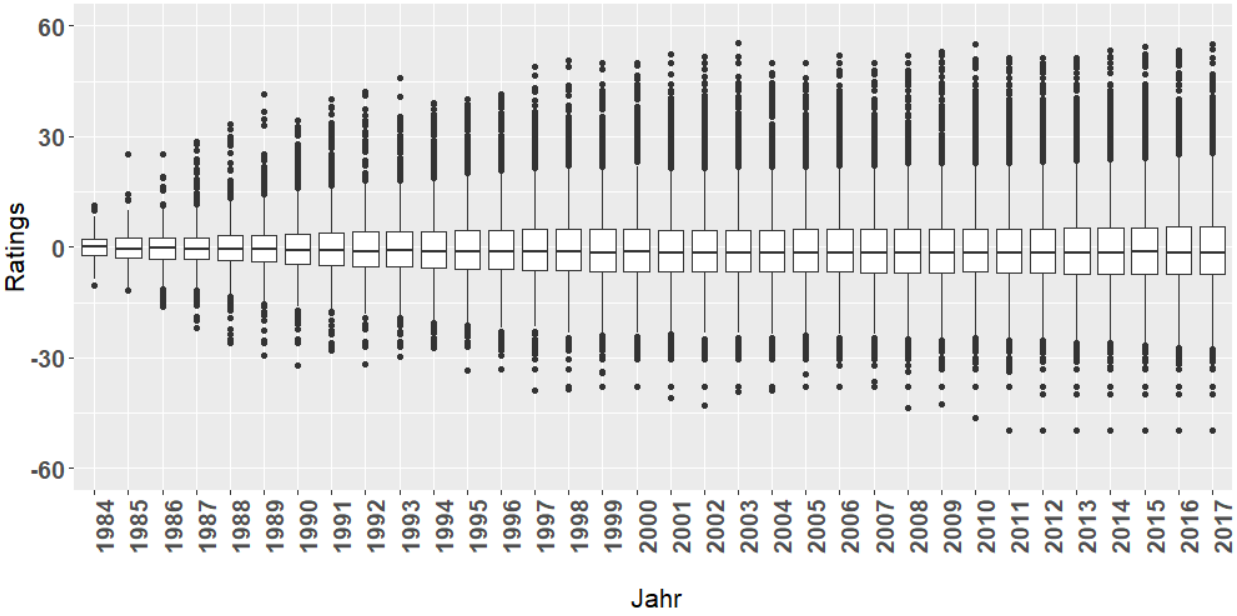

Abbildung 14: Die Ratings bauen sich bis ca. 1997 auf, danach befinden sich die mittleren 50 % der Daten und die Ausreißer in ähnlichen Wertebereichen.

6 Kritik und Optimierungsvorschläge

Das in Kapitel 4 vorgestellte System bietet eine Möglichkeit, die Leistung von Doppelkopfspielern anhand von erspielten Punkten auf verschiedenen Turnieren zu bewerten. Dennoch soll im folgenden Kapitel auf einige Probleme hingewiesen und weitere Optimierungsmöglichkeiten angesprochen werden.
Ein Vorteil des entwickelten Systems ist es, dass sich die Ratings aller Spieler zu jedem Zeitpunkt zu null aufsummieren. Dies verhindert zum einen Inflation und Deflation, zum anderen stellt es eine leichte Kontrollmöglichkeit für mögliche Fehler in der Berechnung dar. Eine Problematik besteht darin, zu entscheiden, wie mit Personen verfahren werden soll, welche den Verband verlassen. Eine mögliche Variante besteht darin, diesen Spieler und sein Rating aus dem System herauszunehmen, wodurch jedoch die Eigenschaft der Aufsummierung aller Ratings zu null verloren ginge. Dieses Problem könnte gelöst werden, indem das Rating des Spielers auf alle anderen Spieler verteilt wird. Dabei bleibt zu bedenken, dass falls ein Spieler mit hohem Rating das System verlässt, die Ratings aller Spieler ansteigen, ohne dass diese eine Leistung dafür erbringen müssen. Dies beeinflusst das aktuelle Ranking nicht (lediglich das Rating), die Bewertungen eines aktiven Spielers zu zwei unterschiedlichen Zeitpunkten sind aber eventuell nicht mehr vergleichbar.
Auch kann überlegt werden, nach einer gewissen Zeit der Inaktivität eines Spielers, dessen Rating erneut auf null zurück zu setzen, da das bestehende Rating den aktuellen Leistungsstand möglicherweise nicht mehr ausreichend zuverlässig repräsentiert. Dabei muss analog zum Verlassen des Verbandes eine Lösung gefunden werden, wie mit dem alten Rating dieses Spielers verfahren wird.
In dem bestehenden System der Rangliste ist ein Bonus für aktive Spieler integriert. Je mehr Runden ein Spieler auf Ranglistenturnieren spielt, umso höher ist sein Bonus auf den Rundenschnitt. Auch die Bundesländerwertung belohnt aktive Spieler, da nur positive Ergebnisse eingehen und man mehr Möglichkeiten hat Punkte zu erzielen, wenn man in möglichst vielen verschiedenen Bundesländern Turniere besucht. Ein ähnliches Bonussystem könnte auch in dem Pi-Rating für Doppelkopf berücksichtigt werden. Eine Möglichkeit zur Umsetzung wäre es, ausschließlich bei Ranglistenturnieren das Rating von teilnehmenden Spielern leicht zu Erhöhen und das der nicht teilnehmenden Spielern so zu senken, dass die Aufsummierung der Ratings zu null bestehen bleibt.
Für die Wettbewerbe Regionalmeisterschaft und Einzelmeisterschaft ist ein Bonussystem nicht sinnvoll. Die Teilnahme eines Spielers an der Regionalmeisterschaft ist dadurch beeinflusst, ob er bereits für die Einzelmeisterschaft qualifiziert ist. Nimmt er trotz einer bestehenden Qualifikation an dem Wettbewerb teil, verfällt diese Qualifikation und er muss sie sich durch das Turnier neu erspielen. Bei der Einzelmeisterschaft ist es einem Spieler nicht freigestellt, ob er teilnehmen möchte oder nicht, da eine Qualifikation nötig ist. Da somit für diese beiden Wettbewerbe die Entschei-

dung der Teilnahme nicht wie bei einem Ranglistenturnier nur von der Reisebereitschaft und der für Doppelkopf verfügbaren Zeit abhängt, sollte hier kein Bonussystem verwendet werden.
Zwei der in Kapitel 2.2 vorgestellten Wettbewerbe wurden bislang nicht mit eingearbeitet, die Deutsche Mannschaftsmeisterschaft und die Bundesliga. Es wurde in Kapitel 5.1 erläutert, warum ein vermischtes Rating für Einzel- und Mannschaftswettbewerbe nicht sinnvoll ist.
Um die verbleibenden Wettbewerbe der Mannschaftsmeisterschaft und der Bundesliga dennoch mit aufzunehmen, könnten zwei getrennte Ratings für jeden Spieler erstellt werden. Analog zur Heim- und Auswärtsspielstärke im Pi-Ratingsystem könnten die Ergebnisse von jedem Turnier unterschiedlich stark gewichtet in die beiden Ratings einer Person eingehen.
Bisher wurde die Anzahl der gespielten Runden in einem Turnier noch nicht berücksichtigt. Dies sollte bei einer tatsächlichen Anwendung jedoch noch bedacht werden. Die Punkte, die man von einem Spieler erwartet, werden in Kapitel 4 durch das Rating eines Spielers, dem durchschnittlichen Rating seiner Gegner, und der verwendeten Trimmungsfunktion bestimmt. Da beispielsweise eine Einzelmeisterschaft über acht Runden und ein Ranglistenturnier über drei Runden veranstaltet wird, sollten die erwarteten Punkte dahingehend angepasst werden. Da der Anteil der Ranglistenturniere jedoch knapp 90% der vorhandenen Turniere darstellt, lässt sich vermuten, dass dies nur geringe Auswirkung auf die vorgenommenen Analysen hat.
Neben den in Kapitel 5.2 diskutierten Überlegungen zur Wahl des Parameters λ können zusätzlich inhaltliche Aspekte für den Anwendungsfall Doppelkopf einbezogen werden. Es kann durch die Wahl verschiedener Werte für λ in unterschiedlichen Wettbewerben neben der Einbeziehung der Spielstärke der Gegner beispielsweise einer Deutschen Einzelmeisterschaft nochmals mehr Gewicht eingeräumt werden.
Ein Schwachpunkt des Systems bleibt die Verwendung der durchschnittlichen Spielstärke der Gegner zur Berechnung der erwarteten Punkte. Bei einem Ranglistenturnier mit 60 Teilnehmern, spielt ein Spieler A beispielsweise nur gegen neun davon, dennoch werden alle Teilnehmer (ausgenommen Spieler A) zur Berechnung der Erwartungen an Spieler A berücksichtigt. Somit bekommt die Auslosung eine starke Bedeutung. Die Problematik, ob ein Spieler in einem durchschnittlich starken Teilnehmerfeld genau mit neun schwächeren Spielern gespielt hat, ist nur durch Aufzeichnung der genauen Zusammensetzungen an den Tischen möglich. Dies hätte jedoch größeren Aufwand in der Datenverwaltung und auch komplexere Berechnungen des Ratings zur Folge.

7 Zusammenfassung

Die verschiedenen Einzelwettbewerbe im Doppelkopf können in einem Bewertungssystem zusammengefasst werden, um die erbrachten Leistungen darzustellen. Hierfür wurde das bestehende Pi-Rating modifiziert und an die Eigenschaften des Kartenspiels angepasst. Es basiert auf der Idee, die Punktzahl, die man von einem Spieler aufgrund seines Ratings und dem Rating seiner Gegner erwartet, zu bestimmen. Die Differenz der erwarteten und tatsächlich erspielten Punkte wird durch eine hyperbolische Tangensfunktion getrimmt und fließt anschließend in das bestehende Rating ein. Der Parameter c bestimmt in der Trimmungsfunktion zum einen die Stärke der Trimmung, zum anderen bildet er die obere Schranke für die maximal einzubeziehende Abweichung. Wie stark neue Turnierergebnisse in das bestehende Rating einfließen sollen, wird durch den Parameter λ ausgedrückt. Die Parameter c und λ konnten sowohl durch die Maximierung der Prognosegüte, als auch anhand einer Streuungszerlegung bestimmt werden.
Die Prognosegüte wird anhand der entstehenden Abweichungen zwischen erwarteten und tatsächlich erspielten Punkten ermittelt. Diese Abweichungen werden sowohl absolut, als auch quadratisch betrachtet und deren Minima liefern zwei Möglichkeiten zur optimalen Wahl von λ und c. Es ergab sich sowohl für quadratische, als auch für absolute Fehler eine maximale Prognosegüte für den Wert von $c = 110$. Bei der Betrachtung des absoluten Fehlers sollen neue Ergebnisse zu 4.5% einbezogen werden, für den quadratischen Fehler liegt die maximale Prognosegüte für eine Einbeziehung neuer Ergebnisse zu 5% vor.
Für die Streuungszerlegung wurde der Quotient aus der Streuung der Ratings innerhalb der Spieler und der Streuung der Ratings zwischen den Spielern minimiert. Die Parameter wurden demnach so gewählt, dass die Ratings einer Person über die Zeit möglichst konstant bleiben, die Ratings zwischen den Personen aber vorzugsweise unterschiedlich sind. Hierfür ergab sich eine optimale Parameterkonstellation von $c = 110$ und $\lambda = 0.15$.
Der Versuch, die Parameter mittels einer Expertenbefragung zu bestimmen, sodass das Ratingsystem die Spielstärke der Spieler reflektiert, kam zu keinem eindeutigen Ergebnis, da die Spielstärke einzelner Spieler sehr unterschiedlich wahrgenommen wird .
Abschließend wurden weitere Anpassungsmöglichkeiten des Systems diskutiert, wie zum Beispiel die Einbeziehung von Mannschaftswettbewerben durch eine zweite Wertung oder ein Bonussystem für aktive Spieler. Auch konnten einige Schwachstellen des Ratingsystems aufgezeigt werden, die beispielsweise durch unzureichende Informationen über die Zusammensetzungen an den Tischen zustande kommen.

8 Abbildungsverzeichnis

9 Literarturverzeichnis und Methoden

[1] Nils Hesse. *Spielend gewinnen*. Springer Fachmedien Wiesbaden, Wiesbaden, 2015.

[2] DDV-Regeln und Ordnungen. `http://www.doko-verband.de/Regeln__Ordnungen.html`. Abrufdatum: 11.08.2018.

[3] Siegfried K. Berninghaus, Karl-Martin Ehrhart, and Werner Güth. *Strategische Spiele*. Springer, Berlin, Heidelberg, 2010.

[4] DDV-Wettbewerbe. `http://www.doko-verband.de/wettbewerb.html`. Abrufdatum: 11.08.2018.

[5] Amy N. Langville and Carl Dean Meyer. *Who's #1? The science of rating and ranking*. Princeton University Press, Princeton N.J., 2012.

[6] R. T. Stefani. A taxonomy of sports rating systems. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 29(1):116–120, 1999.

[7] Alexander M. Franks, Alexander D'Amour, Daniel Cervone, and Luke Bornn. Meta-analytics: tools for understanding the statistical properties of sports metrics. *Journal of Quantitative Analysis in Sports*, 12(4):151–165, 2016.

[8] Daniel Barrow, Ian Drayer, Peter Elliott, Garren Gaut, and Braxton Osting. Ranking rankings: an empirical comparison of the predictive power of sports ranking methods. *Journal of Quantitative Analysis in Sports*, 9(2), 2013.

[9] Anthony Costa Constantinou and Norman Elliott Fenton. Determining the level of ability of football teams by dynamic ratings based on the relative discrepancies in scores between adversaries. *Journal of Quantitative Analysis in Sports*, 9(1):37–50, 2013.

[10] Paul Lodder. The use of the k-factor in estimating individual ability: Advanced study in individual differences. *University of Amsterdam*, 2012.

[11] Jörg Bewersdorff. *Glück, Logik und Bluff: Mathematik im Spiel - Methoden Ergebnisse und Grenzen*. Springer Fachmedien Wiesbaden, Wiesbaden, 2018.

[12] Eberhard Zeidler. *Springer-Handbuch der Mathematik I*. Springer Fachmedien Wiesbaden, Wiesbaden, 2013.

[13] FIDE - World Chess Federation. http://fide.com/fide/handbook.html?id=197&view=article. Abrufdatum: 11.08.2018.

[14] Leonhard Knorr-Held. Dynamic rating of sports teams. *Journal of the Royal Statistical Society*, 49(2), 2000.

[15] Ludwig Fahrmeir, Rita Künstler, Iris Pigeot, and Gerhard Tutz. *Statistik: Der Weg zur Datenanalyse*. Springer, Berlin [u.a.], 6 edition, 2007.

Methoden

Verwendete R Pakete:

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

Baptiste Auguie (2017). gridExtra: Miscellaneous Functions for ”Grid”Graphics. R package version 2.3. https://CRAN.R-project.org/package=gridExtra

Hadley Wickham (2007). Reshaping Data with the reshape Package. Journal of Statistical Software, 21(12), 1-20. URL http://www.jstatsoft.org/v21/i12/.

David B. Dahl (2016). xtable: Export Tables to LaTeX or HTML. R package version 1.8-2. https://CRAN.R-project.org/package=xtable

10 Eigenständigkeitserklärung

Hiermit versichere ich, dass ich die vorgelegte Bachelorarbeit eigenständig und ohne fremde Hilfe verfasst und die den benutzten Quellen entnommenen Passagen als solche kenntlich gemacht habe. Diese Bachelorarbeit ist in dieser oder einer ähnlichen Form in keinem anderen Kurs und / oder Studiengang als Studien- oder Prüfungsleistung vorgelegt worden.

Ort, Datum

Unterschrift

11 Anhang

Parameterbestimmung

Durchschnittlicher absoluter Fehler

	c					
λ	10	30	50	70	90	110
0	42.19	42.19	42.19	42.19	42.19	42.19
0.005	42.00	41.78	41.65	41.59	41.55	41.52
0.01	41.83	41.51	41.37	41.29	41.24	41.21
0.015	41.68	41.33	41.18	41.11	41.07	41.04
0.02	41.55	41.21	41.06	40.99	40.95	40.93
0.025	41.45	41.11	40.98	40.91	40.88	40.86
0.03	41.36	41.04	40.91	40.86	40.83	40.81
0.035	41.30	40.99	40.87	40.82	40.79	40.78
0.04	41.24	40.95	40.84	40.79	40.77	40.76
0.045	41.20	40.92	40.81	40.78	40.76	40.75
0.05	41.16	40.90	40.80	40.77	40.76	40.75
0.055	41.12	40.89	40.79	40.77	40.76	40.76
0.06	41.09	40.88	40.79	40.77	40.77	40.77
0.065	41.06	40.89	40.79	40.78	40.78	40.79
0.07	41.04	40.89	40.80	40.79	40.80	40.81
0.075	41.02	40.90	40.81	40.81	40.82	40.83
0.08	41.01	40.92	40.83	40.83	40.84	40.85
0.085	41.00	40.93	40.85	40.85	40.86	40.88
0.09	40.99	40.95	40.86	40.87	40.89	40.91
0.095	40.98	40.97	40.89	40.90	40.92	40.94

Durchschnittlicher quadratischer Fehler

	c					
λ	10	30	50	70	90	110
0	2907.12	2907.12	2907.12	2907.12	2907.12	2907.12
0.005	2883.60	2853.94	2838.05	2829.06	2823.63	2820.15
0.01	2860.94	2819.66	2800.00	2789.51	2783.38	2779.55
0.015	2841.44	2796.07	2775.80	2765.40	2759.48	2755.85
0.02	2824.69	2778.87	2759.10	2749.30	2743.87	2740.59
0.025	2811.24	2765.90	2747.06	2738.07	2733.20	2730.33
0.03	2800.24	2756.03	2738.22	2730.07	2725.81	2723.35
0.035	2791.05	2748.63	2731.68	2724.39	2720.71	2718.68
0.04	2783.72	2743.03	2726.88	2720.42	2717.33	2715.70
0.045	2777.37	2738.74	2723.44	2717.77	2715.24	2714.02
0.05	2771.72	2735.71	2721.08	2716.17	2714.19	2713.36
0.055	2766.85	2733.45	2719.61	2715.41	2713.97	2713.52
0.06	2762.51	2732.26	2718.89	2715.37	2714.44	2714.36
0.065	2758.96	2731.95	2718.80	2715.91	2715.49	2715.77
0.07	2756.38	2732.36	2719.28	2716.96	2717.03	2717.67
0.075	2753.95	2733.39	2720.26	2718.45	2718.99	2719.98
0.08	2751.73	2734.59	2721.69	2720.32	2721.32	2722.65
0.085	2749.99	2736.22	2723.44	2722.54	2723.98	2725.64
0.09	2748.43	2738.23	2725.55	2725.06	2726.92	2728.92
0.095	2747.15	2740.66	2728.04	2727.85	2730.13	2732.44

Unterscheidung und Stabilität

$$\frac{I}{Z}$$

	c					
λ	10	30	50	70	90	110
0.005	0.37	0.34	0.32	0.32	0.31	0.31
0.01	0.34	0.31	0.30	0.29	0.29	0.29
0.015	0.32	0.29	0.29	0.28	0.28	0.28
0.02	0.31	0.29	0.28	0.28	0.28	0.29
0.025	0.30	0.29	0.29	0.29	0.29	0.29
0.03	0.30	0.29	0.29	0.29	0.30	0.30
0.035	0.29	0.29	0.30	0.30	0.31	0.31
0.04	0.29	0.30	0.30	0.31	0.32	0.32
0.045	0.30	0.30	0.31	0.32	0.33	0.33
0.05	0.30	0.31	0.32	0.33	0.34	0.34
0.055	0.30	0.32	0.33	0.34	0.35	0.36
0.06	0.30	0.33	0.34	0.36	0.36	0.37
0.065	0.31	0.34	0.35	0.37	0.38	0.38
0.07	0.31	0.34	0.37	0.38	0.39	0.40
0.075	0.31	0.35	0.38	0.39	0.40	0.41
0.08	0.32	0.36	0.39	0.41	0.42	0.43
0.085	0.32	0.37	0.40	0.42	0.43	0.44
0.09	0.33	0.39	0.41	0.43	0.45	0.46
0.095	0.33	0.40	0.42	0.45	0.46	0.47

Inhalt der CD

Auf der beigelegten CD befinden sich folgende Dateien:

- Vorgelegte Bachelorarbeit als PDF
- Erläuterung der allgemeinen Vorgehensweise
- R Code und Workspaces
- Darstellung der Expertenumfrage