Bachelorarbeit

Analyse der Erfolsfaktoren im europäischen Vereinsfußball

Ludwig-Maximillians-Universität München
Institut für Statistik

Nils Wöhl

betreut von
Prof. Dr. Christian Heumann

November 2021

Zusammenfassung

Im Rahmen dieser Bachelorarbeit sollte anhand der fünf größten europäischen Ligen analysiert werden, welche Faktoren für eine erfolgreiche Saison im europäischen Vereinsfußball entscheidend sind. Betrachtet wurden vier aufeinander folgende Saisons von 2017/18 bis 2020/21.
Dafür verwendet wurden Datensätze von FBref mit Variablen, die unterschiedliche Bereiche des Spiels abbilden sollen. Die Umsetzung der Analyse sollte mit Hilfe von Modellen erfolgen. Aufgrund der Vielzahl an Variablen wurde zur Selektion ein Algorithmus entwickelt und angewandt. Dadurch konnten zwei generalisierte additive Modelle geschätzt werden. Beim zweiten Modell wurde im Vergleich zum anderen Modell ein Variablentyp („expected Goals“) weggelassen.
Am einflussreichsten zeigten sich bei beiden Modellen besonders Variablen mit Bezug zum Erzielen oder Verhindern von Toren. So waren beim ersten Modell einige Variablen mit expected Goals Bezug bestimmend. Hingegen beim zweiten Modell zählten die durchschnittliche Zahl an Schüssen pro Spiel, der Ballbesitzanteil und der Anteil gehaltener Bälle zu den bedeutendsten Faktoren. Eine Überschneidung zwischen beiden Modellen gab es bei fünf Variablen, die davon jeweils einflussreichste war der Torschussanteil.
Als finale Erkenntnis offenbarte sich, es gibt keine spezielle Spielweise, durch die sich erfolgreiche Mannschaften auszeichnen. Jedoch ist es immer eine Kombination aus dem Maximieren eigener und dem Minimieren gegnerischen Offensivaktionen.

Inhaltsverzeichnis

1 Einleitung

Daten nehmen in unserem Leben eine immer größere Rolle ein. Davon ist kein Bereich ausgenommen, auch nicht der Fußball. Einige der Statistiken werden schon länger erhoben wie Schüsse, Ballbesitzquote oder Passquote, jedoch kommen auch immer wieder neue Statistiken dazu. Diese Vielzahl an unterschiedlichen Statistiken soll natürlich unterschiedliche Aspekte des Spiels abbilden. Gleichzeitig sorgt diese Vielzahl an Statistiken für Unklarheit, was schlussendlich für den Erfolg ausschlaggebend ist. Kann man das überhaupt an bestimmten Faktoren festmachen? Genau das soll im Rahmen dieser Bachelorarbeit so gut wie möglich beantwortet werden. Da jedoch gerade bei einem einzelnen Fußballspiel der Zufall einen recht großen Einfluss auf den Ausgang haben kann, wird hier der Fokus auf ganze Saisons gelegt, um den Einflussfaktor Zufall so gering wie möglich zu halten. Gleichzeitig sollte auch nicht der Fehler gemacht werden, die erzielten Punkte und die Abschlusstabelle mit der erbrachten Leistung gleichzusetzen. Es gibt zwar den bekannten Spruch „Die Tabelle lügt nicht.“, jedoch gibt es einige Indikatoren, dass dies nicht ganz richtig ist. In seinem Buch „Matchplan - Die neue Fußball-Matrix“ kümmert sich Biermann (2018) in einem ganzen Kapitel, um diese Thematik. Trotzdem bleibt, wer regelmäßig in seinen Spielen eine gute Leistung erbringt, erhöht seine Chance am Ende eine erfolgreiche Saison zu spielen.
Wichtig ist es direkt hier am Anfang zu definieren, was mit Erfolg gemeint ist. Aufgrund der unterschiedlichen finanziellen Möglichkeiten im Fußball definiert jeder Verein für sich selbst unterschiedlich, was eine erfolgreiche Saison ist. An dieser Stelle bedeutet erfolgreicher einfach nur Mannschaft A hat im Schnitt mehr Punkte geholt als Mannschaft B. Es ist also unabhängig von der Erwartung, was eine Mannschaft aufgrund ihres geschätzten Potenzials aus einer Saison herausholen kann. Zur Verdeutlichung ein Beispiel aus der aktuell laufenden Saison 2021/22:
Wenn die SpVgg Greuther Fürth den Klassenerhalt in der ersten Bundesliga schafft, ist das sicherlich höher anzusehen als eine Meisterschaft vom FC Bayern München. Gerade bei einem Blick auf die veröffentlichten Finanzkennzahlen der Deutschen Fußball Liga (DFL) sollte das deutlich werden. Selbst wenn der FC Bayern München sein Ziel verpasst und nur Zweiter wird, aber die SpVgg Greuther Fürth den Klassenerhalt schafft, würde im Rahmen dieser Analyse die Bayern Saison als erfolgreicher angesehen werden.
Man könnte daher die Fragestellung auch umformulieren und fragen: Gibt es Faktoren, die für einen besonders hohen Punkteschnitt sorgen und wenn ja, welche? Um dies zu beantworten, wurden die Saisons von 2017/18 bis 2020/21 von fünf Top-Ligen des europäischen Männer-Vereinsfußballs betrachtet. In Kapitel 2 werden diese Ligen vorgestellt und es folgt eine Begründung, warum genau diese ausgewählt wurden. Anschließend soll es in Kapitel 3 um die Datenquelle und den Umgang mit einigen Variablen gehen. Zudem findet eine deskriptive Analyse der Daten statt. Mit dem Wissen über die Daten kann bei der Theorie angeknüpft werden. Zuerst wird die für die Modelle angewandte Theorie in Kapitel 4 vorgestellt. Da sich auch eine andere Modelltheorie angeboten hätte, findet im Rahmen des Kapitels eine Diskussion zur Modellauswahl statt. Anschließend wird noch ein Algorithmus präsentiert, wie aus der Vielzahl an Variablen die Kovariablen für die Modelle selektiert wurden. Im darauffolgenden Kapitel 5 werden die Ergebnisse der zwei geschätzten Modelle geschildert. Im zweiten Abschnitt 5.2 wird neben der Modellvorstellung ebenfalls erklärt, was die Überlegung war, ein zweites Modell unter Ausschluss einiger Variablen zu schätzen. Das abschließende Fazit wird in Kapitel 6 präsentiert.

2 Die europäischen Ligen

Was ist das Besondere am Jahr 2012? Es ist das letzte Mal, dass eine nicht-europäische Mannschaft die FIFA-Klub-Weltmeisterschaft gewinnen konnte. Bei diesem Wettbewerb treten die Gewinner der kontinentalen Vereinswettbewerbe gegeneinander an, um auszuspielen, welche Mannschaft in diesem Jahr die Beste der Welt sein soll. Europa wird dabei durch den Gewinner der UEFA Champions League vertreten. Auch wenn es in der Einleitung schon thematisiert wurde, dass der Gewinner in einem Wettbewerb nicht automatisch die beste bzw. leistungsstärkste Mannschaft sein muss, kann davon ausgegangen werden, dass bei der FIFA-Klub-Weltmeisterschaft die Teams zu den Besten ihres Kontinents zählen. Die europäische Dominanz dieses Wettbewerbes spricht auf jeden Fall dafür, dass die stärksten Vereine aus dem europäischen Vereinsfußball kommen. Natürlich folgt daraus nicht automatisch die Erkenntnis, die europäischen Ligen sind die besten und leistungsstärksten der Welt, nur weil die Qualität der Spitze des europäischen Vereinsfußballs weltweit nicht übertroffen werden kann. Trotzdem liegt die Vermutung nahe, dass die europäischen Top-Ligen zur Weltspitze gehören. Das heißt, wenn es darum geht, welche Faktoren für den Erfolg auf dem allerhöchsten Niveau des Vereinsfußballs der Männer entscheidend sind, sollte die Wahl, das anhand den europäischen Top-Ligen zu analysieren, nicht vollkommen falsch sein.
Es gilt in Folge dessen noch festzustellen, welche Ligen als Top-Ligen in Europa betrachtet werden können. Eine Möglichkeit stellt hierfür der Verbands-Klubkoeffizient der UEFA dar, der den Erfolg der Vereine aus den jeweiligen Verbänden bzw. Ligen in den kontinentalen Wettbewerben (UEFA Champions League und UEFA Europa League) abbildet. Bei der Betrachtung von diesem Ranking während des Zeitraums von der Saison 2017/18 bis zur Saison 2020/21 fällt auf, dass die selben fünf Verbände mit ihren Ligen auf den ersten fünf Plätzen liegen, nämlich Spanien, England, Deutschland, Italien und Frankreich. Häufig auch als die „großen fünf europäischen Ligen“ bezeichnet, wird der Fokus der Analyse ausschließlich auf ihnen liegen.

2.1 La Liga

La Liga oder auch bekannt als Primera División ist die erste Liga im spanischen Vereinsfußball. Erstmalig wurde dieser Wettbewerb im Jahr 1929 ausgetragen. (Steel, 2021) Seitdem gibt es drei Vereine, die ununterbrochen Teil des Wettbewerbs sind. Dies sind Athletic Bilbao, Real Madrid und der FC Barcelona. Gerade Real Madrid und der FC Barcelona haben diese Liga geprägt, da die beiden Vereine zusammen insgesamt 61-mal in den 90 absolvierten Saisons den spanischen Meister stellten. (Marca, o. D.) Wobei Real Madrid 35-mal die spanische Meisterschaft feiern durfte, während dies beim FC Barcelona nur 26-mal der Fall war. Damit sind die Madrilenen Rekordmeister der La Liga. Neben den beiden konnte auch der bereits erwähnte Athletic Bilbao bereits achtmal Meister werden. Diesem Verein kommt aufgrund seiner Philosophie wahrscheinlich eine einzigartige Rolle im professionellen Vereinsfußball zu. Der baskische Klub hat sich selber auferlegt, dass bei ihnen entweder nur Spieler aus der eigenen Jugendakademie, aus Jugendakademien anderer baskischer Klubs oder im Baskenland Geborene spielen dürfen. (Athletic Bilbao, o. D.) Andere bekannte Vereine aus dieser Liga sind Atletico Madrid, FC Sevilla, FC Valencia, Real Betis Sevilla, FC Villareal und Real Sociedad San Sebastián. Zum Wettbewerb selbst ist noch zu sagen, dass die Liga aus 20 verschiedenen Vereinen besteht und eine einzelne Saison dementsprechend 38 Spieltage lang dauert. Am Ende jeder Saison steigen drei Vereine ab und drei neue Vereine wieder auf.
Etwas, das sicherlich die spanische Liga auszeichnet, ist die Tatsache, dass sie sowohl mit Real Madrid den Rekordsieger der UEFA Champions League als auch mit dem FC Sevilla den Rekordsieger der UEFA Europa League stellt. Das regelmäßige erfolgreiche Abschneiden auf kontinentaler Ebene spiegelt sich auch

im Verbands-Klubkoeffizient der UEFA wieder. Ab der Saison 2012/13 bis zur Saison 2019/20 befand sich der spanische Verband durchgängig auf Platz 1 dieser Rangliste. (UEFA, o. D. b)
In der Liga selbst gibt es aktuell eine klare Struktur an der Spitze. Atletico Madrid, FC Barcelona und Real Madrid machen in der Regel die ersten drei Plätze unter sich aus. Seit der Saison 2012/13 war dies nur einmal nicht der Fall, nämlich 2019/20, da wurde der FC Sevilla dritter, wobei dies nur der Tatsache geschuldet war, dass sie den direkten Vergleich gegen Atletico Madrid gewannen, die punktgleich auf Platz vier lagen. Zudem sind schon 17 Jahre vergangen, seitdem mit dem FC Valencia eine andere Mannschaft als diese drei Meister wurde. Hinter diesen Teams gibt es eine Gruppe, bestehend aus fünf Teams (FC Sevilla, Atletic Bilbao, Real Sociedad, FC Valencia und FC Villareal), die oft untereinander um den Einzug in die kontinentalen Wettbewerbe kämpft. Dahinter gibt es einige Mannschaft, denen es in den letzten zehn Jahren möglich war, sich in La Liga zu etablieren. 13 Vereine schafften es in diesem Zeitraum an mindestens neun von zehn Saisons in La Liga teilnehmen zu können. (transfermarkt, o. D. b)

2.2 Premier League

Seit dem Jahr 1992 ist die Premier League die erste Liga im englischen Fußball. Damals entschieden die 22 Vereine der Football League, die zu der Zeit erste englische Liga, sich aus der Football League zurückzuziehen, um die Premier League zu gründen. (Premier League, o. D.) Es ging ihnen dabei, um Unabhängigkeit vom englischen Fußballverband FA beim Verhandeln von Sponsoren- und TV-Verträgen zu haben. Die Football League selbst wurde 1888 gegründet und fungiert seit der Gründung der Premier League als zweite englische Liga. (Britannica, T. Editors of Encyclopaedia, 2020) Die Liga sagt von sich, sie sei die erste Fußball-Liga der Welt und ein Vorbild für viele Ligen rund um den Globus gewesen. (EFL, o. D.) Zusätzlich dazu bietet der englische Fußball die Besonderheit, dass nicht nur englische Vereine teilnehmen, sondern auch einige walisische Vereine. (Goal, 2019) Es gab eine Zeit, in der es keinen walisischen Fußballverband gab, daher schlossen sich damals einige Vereine dem englischen Fußball an. Selbst mit dem Bestehen eines walisischen Fußballverbandes blieben Vereine im englischen Vereinsfußball und waren auch schon Teil der Premier League. Jedoch gelang keinem die dauerhafte Teilnahme an diesem Wettbewerb, das schafften bisher nur die sechs englischen Vereine Arsenal, Chelsea, Everton, Liverpool, Manchester United und Tottenham Hotspurs. (Premier League, o. D.) Insgesamt treten jede Saison 20 Vereine gegeneinander an, von denen am Ende der Saison drei absteigen müssen. Der englische Rekordmeister in der Premier League, aber auch wenn man die Zeit der Football League als erste englische Liga dazu nimmt, ist Manchester United mit 13 bzw. 20 Meistertiteln.
Aktuell kann sicherlich die Rede davon sein, dass die Premier League zusammen mit La Liga zu den zwei besten Ligen Europas zählt. Dies zeigt auch das UEFA-Ranking, in dem der englische Vereinsfußball von 2017/18 bis zur Saison 2019/20 auf Platz zwei lag, bevor er nach der Saison 2020/21 sogar Platz eins einnehmen konnte. (UEFA, o. D. b) Neben dem Erfolg im kontinentalen Fußball verbindet man mit der Premier League die großen Mengen an Geld, die sowohl für Spielergehälter wie auch für Ablösesummen ausgegeben werden. Begründen lässt sich das durch die vielen finanzkräftigen Investoren und die erzielten Erlöse für die TV-Rechte. Im Zeitraum von 2016 bis 2019 wurden der Liga und den Vereinen für die TV-Rechte 2,3 Milliarden Euro pro Saison bezahlt. (SID, 2018)
In Sachen Struktur der Liga war über längere Zeit von den großen sechs Vereinen (Arsenal, Chelsea, Liverpool, Manchester United, Manchester City und Tottenham Hotspurs) die Rede, die Saison für Saison die ersten sechs Plätze unter sich ausmachten. Bis auf den Überraschungsmeister Leicester City in der Saison 2015/16 konnte keine andere Mannschaft als die ersten fünf genannten seit der Saison 1994/95 englischer

Meister werden. In den letzten Jahren sollte jedoch eher die Rede von einer kleineren Spitzengruppe sein, in der Manchester City tonangebend ist. Ihnen gelang es in den letzten vier Jahren dreimal englischer Meister zu werden. Besonders Arsenal und auch Tottenham Hotspurs taten sich vermehrt schwerer, mit der Ligaspitze mitzuhalten. Als Verfolger dieser sechs Vereine haben sich in der jüngeren Vergangenheit besonders Everton und Leicester City hervorgetan. Dahinter sich zu etablieren scheint auf jeden Fall keine einfache Aufgabe zu sein, da es insgesamt nur zehn Vereine in den letzten zehn abgelaufenen Spielzeiten schafften, an mindestens neun teilzunehmen. (transfermarkt, o. D. d) Das bereits erwähnte Leicester City schaffte das auch nicht, jedoch spielt es seit sieben Jahren ununterbrochen in der Premier League.

2.3 Bundesliga

Die deutsche Bundesliga ist seit ihrem Bestehen 1963 die höchste Liga im deutschen Vereinsfußball. (SPOX, o. D.) Bis auf die Saison 1991/92 bestand sie immer aus 18 Vereinen. In dieser Saison jedoch wurde sie durch die beiden ostdeutschen Vereine Hansa Rostock und Dynamo Dresden aufgrund der Wiedervereinigung erweitert, dafür stiegen am Ende der Saison vier statt den ansonsten üblichen drei Mannschaften in die zweite Bundesliga ab und es kamen auch nur zwei neue Vereine hoch. Eine Änderung bezüglich der Abstiegsregelung gab es ab der Saison 2008/09, so dass mittlerweile nur noch zwei Vereine direkt aus der ersten Bundesliga absteigen, während der drittletzte der ersten Bundesliga am Ende der Saison in der sogenannten Relegation in Hin- und Rückspiel gegen den dritten der zweiten Bundesliga antritt. Der Gewinner der Relegation sichert sich für die neue Saison die Teilnahme an der ersten Bundesliga, die andere Mannschaft muss in der neuen Saison an der zweiten Bundesliga teilnehmen.
Dominiert wird die Liga vom Rekordmeister FC Bayern München, der insgesamt 31-mal deutscher Meister werden konnte, wobei sich die Dominanz gerade in den letzten Jahren zeigt, da es ihm gelang, neunmal in Folge die Meisterschaft für sich zu entscheiden. (kicker, o. D.) In der letzten Zeit waren seine engsten Verfolger in der Regel Borussia Dortmund und RaBa Leipzig, die häufig die Plätze zwei und drei unter sich ausgemacht haben. Die Dortmunder Borussia ist dabei auch der letzte andere Verein, dem es gelang, die Meisterschaft für sich zu entscheiden. Dies geschah in den Saisons 2010/11 und 2011/12. Zusammen mit der anderen Borussia aus Mönchengladbach konnten sie die zweitmeisten Titel in der Bundesliga gewinnen, wobei beide fünfmal den Titel gewannen, was zusätzlich den historischen Abstand von Bayern München zum Rest der Liga unterstreicht. Nicht nur dieser Abstand scheint groß zu sein, sondern auch der zwischen erster und zweiter Bundesliga. Aufzusteigen und längerfristig in der ersten Bundesliga mitzuspielen scheint kein einfaches Unterfangen zu sein, wenn man die Tatsache beachtet, dass es 13 Vereine geschafft haben, an neun der letzten zehn Spielzeiten teilzunehmen. (transfermarkt, o. D. a) Zwar haben dies auch kleinere Vereine wie der 1.FSV Mainz 05 und der SC Freiburg geschafft, jedoch ist eine Durchmischung der Liga durch kleinere Vereine eher unüblich. Es gibt dementsprechend einige gefestigte Hierarchien, die nicht so leicht durchbrochen werden können.
Auf europäischer Ebene hat es sich ergeben, dass die Bundesliga im größten europäischen Wettbewerb, die UEFA Champions League, jährlich mit Bayern München nur einen großen Titelfavoriten hat und nicht mehrere, wie das beispielsweise in La Liga und Premier League der Fall ist. Das offenbart sich auch in dem UEFA-Ranking, indem man zwar seit der Saison 2008/09 durchgängig zu den Top vier gehört, aber größtenteils zwischen den Plätzen drei und vier hin und her pendelt. (UEFA, o. D. b)
Hervorzuheben gegenüber den anderen vier Ligen ist die Tatsache, dass bis auf ein paar wenige Ausnahmen die Vereine von ihren Mitgliedern und nicht von Investoren geführt werden. Aufgrund der 50+1 Regelung dürfen Vereine nicht ihre kompletten Anteile an Investoren verkaufen, sondern die Mitglieder

sollen immer mindestens 50% plus eine zusätzliche Stimme an dem Verein halten. (kicker, 2018) Dadurch stellen die Mitglieder stets die Mehrheit und können den Kurs des Vereins bestimmen. Gerade die großen Traditionsvereine haben durch die ermöglichte Teilhabe hohe fünfstellige oder teilweise sogar sechsstellige Mitgliederzahlen und gehören damit zu den mitgliederstärksten Vereinen weltweit. (Wikipedia, o. D.)

2.4 Serie A

Der italienische Meister wird seit 1898 in der Serie A bestimmt. Anfangs spielten dabei nur vier Regionalmeister den Titel untereinander aus. (TZ, 2021b) Dies änderte sich erst im Jahr 1929, ab dem Zeitpunkt traten alle teilnehmenden Mannschaften gegeneinander an. Etwas was sich danach noch mehrfach änderte war die Ligakapazität. Seit der Saison 2004/05 sind es 20 Vereine, die pro Saison an der Serie A teilnehmen. Am Ende jeder Saison müssen sich drei Teams Richtung Serie B verabschieden, während der Meister den „Scudetto“ überreicht bekommt. Insgesamt 36-mal konnte der Rekordmeister Juventus Turin den Titel gewinnen, wobei ihm zwei weitere Titel aufgrund eines Wettskandals aberkannt wurde.
In den Saisons 2004/05 und 2005/06 wurden Spiele durch Schiedsrichter-Absprachen manipuliert. Neben mehreren Schiedsrichtern waren auch Funktionäre und neun Vereine beteiligt. Das hatte zur Folge, dass Juventus im Jahr 2006 nachträglich die Meisterschaft aus 2004/05 aberkannt wurde und die Saison offiziell ohne Meister gewertet ist. In der Saison 2005/06 bekam die drittplatzierte Mannschaft Inter Mailand die Meisterschaft zugesprochen, da neben dem Erstplatzierten Juventus Turin auch die zweitplatzierte Mannschaft AC Mailand involviert war. Juventus musste zwangsabsteigen in die Serie B und AC Mailand bekam 30 Punkte abgezogen. Dies war jedoch nicht der erste Wettskandal in der Geschichte der Serie A, sondern Anfang der 80er gab es bereits einen Wettskandal mit den Zwangsabstiegen als Folge für AC Mailand und Lazio Rom. Unabhängig von den Wettskandalen fällt auf, dass neben Juventus Turin einige der Titel an die beiden anderen großen norditalienischen Klubs Inter Mailand (19-facher Meister) und AC Mailand (18-facher Meister) gingen. (Lega Nazionale Professionisti Serie A, o. D.) Seit dem Meistertitel des AS Rom in der Saison 2000/01 ist es keiner anderen Mannschaft als den dreien gelungen italienischer Meister zu werden. Für die letzten zehn Jahre muss jedoch die Rede davon sein, dass Juventus Turin die Liga dominiert hat. Bevor in der letzten abgelaufenen Saison 2020/21 Inter Mailand wieder Meister wurde, gelang es Juventus davor neun Mal in Folge den Scudetto zu gewinnen. Die Mannschaft, die ihnen dabei neben Inter Mailand am dichtesten kam, war der SSC Neapel.
Auf der europäischen Ebene gestaltete sich das dann in letzter Zeit auch häufig ähnlich zur deutschen Situation, indem man mit Juventus fast jährlich einen einzigen Topfavoriten in der UEFA Champions League stellt, jedoch die anderen Teams konkurrenzfähig, aber keine Titelfavoriten waren. Der Unterschied zur Bundesliga hingegen ist, dass der letzte europäische Titel, der an eine italienische Mannschaft ging, deutlich länger her ist, nämlich letztmalig war dies 2009/10 mit dem Gewinn der UEFA Champions League von Inter Mailand der Fall. (UEFA, o. D. a) Das ansonsten ähnliche Auftreten drückt sich im UEFA-Ranking aus, die Serie A ist Dauervertreterin in den Top vier, jedoch lag sie das letzte Mal auf einem der beiden oberen Plätze nach der Saison 2005/06. (UEFA, o. D. b) Eine weitere Ähnlichkeit zur Bundesliga sowie zur La Liga ist, dass es einen gewissen Stamm an Teams zu geben scheint. 14 Vereinen war es möglich an neun von den zehn letzten Saisons teilzunehmen. (transfermarkt, o. D. e)

2.5 Ligue 1

Die Ligue 1 besteht seit dem Jahr 1932 als höchste französische Spielklasse. (TZ, 2021a) Wie auch in La Liga, Premier League und Serie A treten dabei 20 Vereine gegeneinander an. Am Ende der Saison steigen

dann zwei Vereine direkt ab, während der Drittletzte an einer Relegation teilnimmt. Im Vergleich zu den anderen vier Ligen ist sie sicherlich als die kleinste und unbedeutendste Liga anzusehen. Laut TZ (2021a) gehört die Ligue 1 auch erst seit 2002 zu den Top-5-Ligen im europäischen Fußball, was auch damit zu tun habe, dass sich andere Sportarten in Frankreich großer Beliebtheit erfreuen. Ähnliches zeichnet sich im UEFA-Ranking ab. Letztmalig Teil der Top 4 war sie nach der Saison 2007/08. (UEFA, o. D. b) Seit der Saison 2016/17 steht die Ligue 1 ununterbrochen auf Platz 5 in dem Ranking.
Historisch fiel die Ligue 1 besonders durch viele Veränderungen auf, sowohl bei der Ligagröße, Zahl an Absteigern, aber auch durch häufig wechselnde Meister. (TZ, 2021a) Trotz des recht langen Bestehens gibt es keinen Verein, der bisher mehr als zehn Meisterschaften feiern konnte. Wobei es immer wieder Phasen gab, in denen es Mannschaften gelang in mehreren aufeinander folgenden Jahren die Meisterschaft zu gewinnen. Die einzigen Meisterschaften von Olympique Lyon markierten auch die längste Serie einer solchen Phase mit sieben Titeln im Zeitraum von 2002 bis 2008. (transfermarkt, o. D. f) Mittlerweile liegt die Vermutung nahe, dass die Zeit mit vielen wechselnden Meistern vorbei ist. Seit der Übernahme von Paris Saint-Germain (PSG) im Jahr 2011 durch den katarischen Staatsfond "Qatar Sport Investment"dominieren sie die Ligue 1 und gewannen sieben Mal den Titel. (Frerks, 2014 & transfermarkt, o. D. f) Gerade beim Blick auf die Transferausgaben und Transfersalden seit der Übernahme zeigt sich die große Distanz zwischen dem Hauptstadtklub und der restlichen Konkurrenz. Am ehesten kann in dieser Kategorie noch die AS Monaco mithalten, die seitdem auf Platz 13 liegt mit insgesamt 952.11 Millionen Euro an gezahlter Ablösesumme für neue Spieler. (transfermarkt, o. D. g) Anzumerken ist, dass sie in dem Zeitraum sogar mehr eingenommen als ausgegeben haben. Daher kommen sie auf ein Transfersaldo von 25.79 Millionen Euro. Im Vergleich dazu hat PSG am sechstmeisten an Ablösesummen ausgegeben mit 1.40 Milliarden Euro, jedoch mit einem Transfersaldo von -946.35 Millionen Euro. Der restlichen Konkurrenz scheint PSG bei dem Aspekt komplett entwischt zu sein, die beiden nächsten französischen Konkurrenten Olympique Marseille und Olympique Lyon landen in dieser Liste auf Platz 37 bzw. Platz 38 und haben insgesamt jeweils weniger als 400 Millionen Euro für neue Spieler ausgegeben. Aufgrund dessen wirkt gerade die Meisterschaft in der abgelaufenen Saison von LOSC Lille, dem ersten Meister und Gründungsmitglied der Ligue 1, wie eine Überraschung. (TZ, 2021a) Wobei sie neben Olympique Marseille, Olympique Lyon und AS Monaco immer mal wieder zum Verfolgerkreis in den letzten Jahren gehörten. Bis auf die AS Monaco gehören diese Teams alle zu dem Kern der Liga, bestehend aus 10 Teams, die an neun der vergangenen zehn Saisons teilnehmen durften. (transfermarkt, o. D. c) Ein weiteres Team, das dazu gehört, ist der Rekordmeister AS Saint-Étienne mit seinen zehn Meisterschaften. Jedoch ist ihre letzte Meisterschaft 40 Jahre her. (transfermarkt, o. D. f)
Abschließend zur Ligue 1 soll es noch kurz um die Saison 2019/20 gehen, da diese aufgrund der Corona Pandemie nach dem 28. Spieltag abgebrochen wurde. (Redaktion Sportbuzzer, 2020) Dies stellt eine Besonderheit gegenüber den anderen vier Ligen dar, weil sie die einzige mit einem Saisonabbruch ist. Schlussendlich wurde die Tabelle vom 28. Spieltag als Endtablle genommen, anhand derer der Meister und die Absteiger bestimmt wurden.

3 Daten

3.1 Datenherkunft und -aufbereitung

Wie bereits im vorherigen Kapitel bezieht sich die Analyse auf die fünf vorgestellten europäischen Top-Ligen in den vier letzten abgelaufenen Saisons. Für so eine Analyse über mehrere Ligen und Saisons

hinweg braucht es natürlich Daten, die Liga und Saison übergreifend vergleichbar sind. Als einheitliche Datenquelle stand dafür die Internetseite von FBref zur Verfügung, die Daten auf ihrer Seite für unterschiedliche Wettbewerbe im Frauen- und Männerfußball zur Verfügung stellen. (FBref, o. D. a) Hinter FBref selbst steht Sports-Reference, die auch vergleichbare Seiten für andere Sportarten wie z.B. Basketball und Baseball betreiben. Einige ihrer Daten für FBref beziehen sie dabei von StatsBomb, speziell sogenannte „advanced data“, beispielsweise stammen die „expected Goals“ (xG) Werte von StatsBomb. (FBref, o. D. b) Gerade in diesem Bereich kann es zwischen den unterschiedlichen Anbietern zu Unterschieden kommen, da jeder Datenanbieter eigene Modelle hat, mit denen die Werte für „advanced data“ ermittelt werden. FBref (o. D. b) sieht hier ihren Partner StatsBomb für jegliche Analysen auf diesem Gebiet als Anlaufstelle Nummer eins an, weil deren Datensätze nicht mit denen der Konkurrenz vergleichbar sind. Unabhängig davon, ob dies zutreffend ist oder nicht, wird die Analyse um mögliche weitere Ebenen erweitert im Vergleich zu den klassischen Statistiken, die sonst aus den Medien bekannt sind. Wobei mit der Zweikampfquote ein recht populärer Wert bei Fans und in der Sportberichterstattung gar nicht in den Daten vorkommt. Diese Statistik ist jedoch sowieso umstritten, weshalb das kein Problem darstellt. Für alle Interessierten wird in dem Blogbeitrag „Hanno Behrens und das Rätsel der divergierenden Zweikampfquote“ von Zenger (2019) erklärt, warum diese Quote mit Vorsicht betrachtet werden sollte. Davon abgesehen geben die Daten unterschiedliche Aspekte eines Fußballspiels wieder, vom Torwartspiel über defensive Aktionen und Passspiel bis hin zum Torabschluss.

So viel zum Ursprung und Teilaspekten der Daten jedoch können die Daten von der Internetseite nicht direkt für alle Teilaspekte und über alle vier Saisons hinweg als ein Datensatz heruntergeladen werden. Stattdessen kann man für die einzelnen Kategorien in der jeweiligen Saison die Daten in CSV-Dateiform abspeichern. Daher gab es anfangs pro jeweilige Saison viele unterschiedliche Datensätze, jedoch bei jedem entspricht jede Zeile einem unterschiedlichen europäischen Verein. Um damit arbeiten zu können, mussten die Datensätze Saison und Kategorien übergreifend zu einem einzigen Datensatz zusammengefügt werden. Dieser zusammengefügte Datensatz wurde anschließend auch noch um weitere Variablen erweitert, die mit Hilfe der anderen bereits vorhandenen Variablen im Datensatz erstellt wurden. Schlussendlich entstand so ein Datensatz bestehend aus 392 Zeilen und 280 Spalten. Es ist anzumerken, dass nicht jede dieser Spalten eine mögliche Variable darstellt, da einige nur Auskunft über Saison, Wettbewerb oder Mannschaft Auskunft geben. Auch zu den Zeilen soll noch ergänzt werden, wie es zu der Anzahl von 392 kam. Wie bereits im vorherigen Kapitel bei den kurzen Erklärungen zu den unterschiedlichen Ligen erwähnt wurde, treten nicht in allen Ligen die selbe Anzahl an Mannschaften an. Da also vier 20er-Ligen und eine 18er-Liga vertreten sind, treten so pro Saison in diesen fünf Ligen insgesamt 98 unterschiedliche Mannschaften an. Beim Betrachten von vier verschiedenen Saisons kommen 392 Mannschaften und die selbige Zeilenanzahl zusammen.

Ein gerade angeklungener Aspekt, der nicht nur für die Erklärung der Zeilenanzahl entscheidend war, ist die unterschiedlichen Ligengröße. Aufgrund dessen und der abgebrochenen Saison 2019/20 der Ligue 1 war eine Transformation einiger Variablen notwendig, da nicht jede Mannschaft im Datensatz die selbe Anzahl an Spielen absolviert hat. Deshalb galt es für Variablen, bestehend aus sogenannten „total stats“, eine Vergleichbarkeit zwischen den unterschiedlichen Mannschaften und Ligen herzustellen. Zur Erläuterung wird hier das Variablenbeispiel „abgegebene Schüsse einer Mannschaft“ herangezogen. Ursprünglich gab diese im Datensatz an, Mannschaft A hat in der einen Saison eine Gesamtanzahl an b Schüssen abgegeben. In Folge der Transformation wurde diese Variable in die Form gebracht, Mannschaft A hat pro Spiel im Durchschnitt c-mal geschossen. Zusätzlich dazu wurde eine Idee von Knutson (2014) aufgegriffen, bei defensiven Aktionen eine Korrektur basierend auf den Ballbesitzanteilen vorzunehmen.

Die Idee dahinter ist, dass der Eindruck erweckt wird, Mannschaften mit höheren Ballbesitzanteilen bzw. ihre Spieler seien schlechter bei defensiven Aktionen als Mannschaften mit geringeren Ballbesitzanteilen. Dabei haben sie häufig einfach nur seltener die Gelegenheit, defensive Aktionen auszuführen. Daher wurden für alle diese Variablen neue angepasste Variablen erstellt, die in die oben genannten 280 Spalten bereits mit rein zählen. Die Anpassung erfolgte anhand folgender Formel:

$$\frac{\#'Defensive\ \ Aktion'}{Anzahl\ \ der\ \ Spiele} * \frac{50}{100 - Ballbesitzanteil\ \ in\ \ \%} \tag{1}$$

Dadurch soll ein besserer und fairer Vergleich zwischen den Mannschaften ermöglicht werden. In dem Fall wird der Annahme gefolgt, jede Mannschaft stände einem gegnerischen Ballbesitz von 50% gegenüber und dem entsprechend viele bestimmte defensive Aktionen würden sie durchschnittlich pro Spiel ausführen. Ansonsten wurden noch fehlerhafte Werte, die bei der Betrachtung der Variablen bzw. die im Laufe der Modellierungsprozesse aufgefallen sind, korrigiert. Variablen mit fehlerhaften Werten, die nicht korrigiert werden konnten, wurden aussortiert.

3.2 Deskriptive Analyse

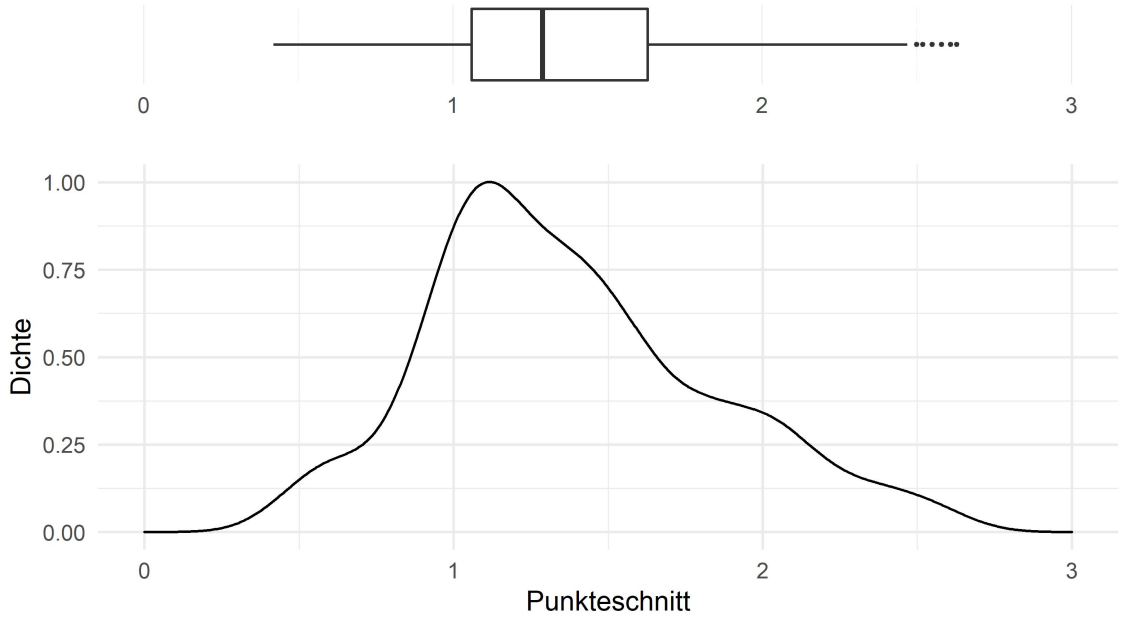


Abbildung 1: Boxplot & Dichteverteilung der Zielvariable (Punkteschnitt)

Nachdem bereits einige Worte über die Herkunft der Daten und den Umgang mit ihnen verloren wurde, soll es nun darum gehen, Teilaspekte der Daten visuell darzustellen, um einen ersten genaueren Überblick zu gewähren. Ein bereits im ersten Abschnitt angesprochener Aspekt ist die unterschiedliche Anzahl an absolvierten Spielen in den jeweiligen Saisons aufgrund der Ligengröße und Corona-Pandemie. Dies ist dann offensichtlich nicht nur für mögliche Kovariablen zu beachten, sondern ist auch ein Thema bei der Zielvariable. Die Wahl dieser war an sich relativ einfach, weil wenn von Erfolg in Ligen die Rede ist, dann ist das natürlich direkt an die Zahl der erspielten Punkte geknüpft. Um nicht allgemein die Bundesliga und die Ligue 1 der Saison 2019/20 gegenüber den anderen Ligen abzuwerten, fiel die Wahl auf den Punkteschnitt pro Spiel. Es handelt sich dabei um eine abzählbare Variable mit einem Wertebereich von

0 bis 3, jedoch ist es für die Analyse sinnvoller, sie als quasi-stetig zu betrachten.
Bei der Betrachtung der Dichte in Abbildung 1 fallen den meisten sicherlich mehrere mögliche Verteilungen ein, mit denen man das darstellen könnte. Um die genaue Verteilung soll es zu einem späteren Zeitpunkt gehen. Allgemein festzustellen ist, dass in den Zeitraum der vier Saisons in den fünf Ligen die Punkteausbeute der Vereine pro Spiel sich fast auf die gesamte Breite des Intervalls verteilt. Nur direkt an den Rändern finden sich keine Beobachtungen, was aber ehrlicherweise gerade für Profiligen wenig überraschend ist, da die Ligen in Sachen Leistungsdichte dann doch in der Regel zu dicht besetzt sind, um Mannschaften zu haben, die entweder alle Spiele einer Saison gewinnen oder verlieren. Interessanterweise illustriert der Boxplot, dass das Mittelfeld der Ligen dichter an den unteren Rängen als an der Spitze dran ist. Für Beobachter des Geschehens sollte das nicht zu überraschend kommen, gerade in Anbetracht einiger Entwicklungen in den letzten Jahren. Noch besser offenbart sich der Abstand der Spitze zum Rest

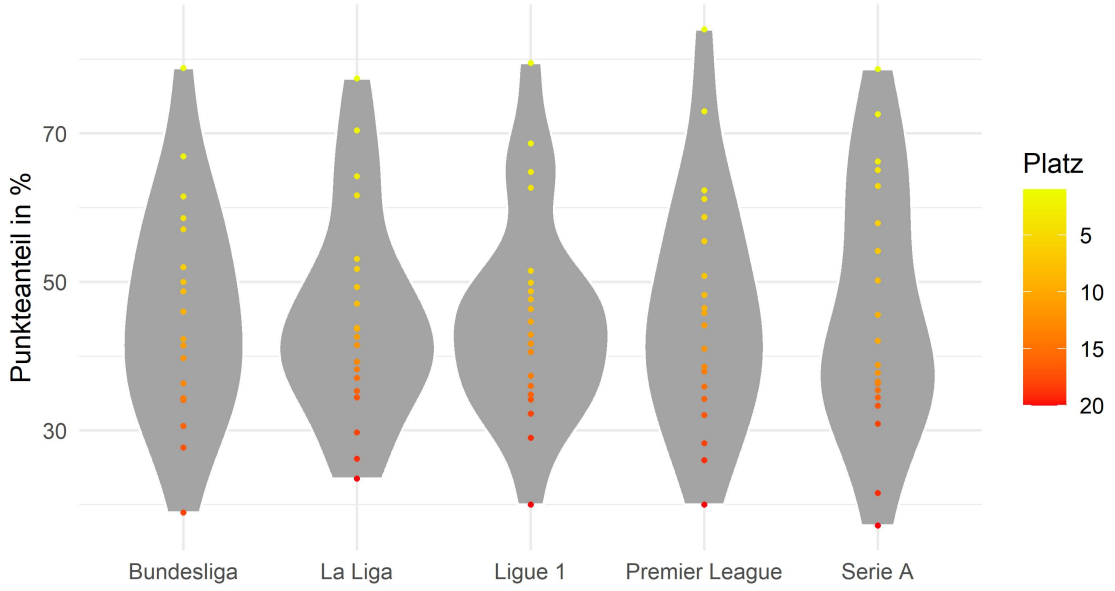


Abbildung 2: durchschnittlich benötigter Anteil an der optimaler Punkteausbeute fürs Erreichen einer bestimmten Platzierung in den jeweiligen Ligen & Darstellung der Punkteverteilung innerhalb der Ligen (Zeitraum: Saison 2017/18 bis 2020/21)

in Abbildung 2. Anknüpfend an Kapitel 2 spiegeln sich hier die dort beschriebenen Ligastrukturen wieder. Auffällig ist La Liga, da im Vergleich zu den anderen Ligen der Tabellenletzte den durchschnittlich deutlich größten Punkteanteil und gleichzeitig der Meister den geringsten Punkteanteil erreicht. Zwar setzen sich die ersten vier Plätze recht eindeutig vom Rest der Liga ab, wie das ansonsten nur in Ligue 1 der Fall ist, aber gerade im Mittelfeld gibt es nur marginale Unterschiede bei der Punkteausbeute. Die Darstellung der Premier League demonstriert auch recht gut die beschriebene Entwicklung, dass der Abstand innerhalb der ersten sechs Plätze mittlerweile ziemlich groß ist. Zudem schafft es ihr Meister durchschnittlich den größten Punkteanteil zu erreichen, verglichen mit den anderen vier Ligen. Die Gemeinsamkeit aller Ligen ist der durchschnittlich verhältnismäßig deutliche Abstand zwischen den Meistern und den Zweitplatzierten. Es gibt zusätzlich noch eine Tatsache, die die Ligen miteinander verbindet, nämlich bei allen ist es recht einfach, die Abstiegsplätze auszumachen. Während auf den Plätzen direkt darüber eigentlich überall ein recht enger Kampf zu beobachten ist, scheinen die Absteiger in der Regel

Ligen	**Zahl der Vereine, die an ... Saisons teilnahmen**				$\sum$
	4	**3**	**2**	**1**	
La Liga	14	3	4	7	28
Premier League	14	3	5	5	27
Bundesliga	14	2	3	4	23
Serie A	15	3	3	5	26
Ligue 1	14	5	3	3	25
$\sum$	71	16	18	24	129

Tabelle 1: Aufteilung nach den Ligen: Zahl der Vereine mit bestimmter Anzahl an Saisonteilnahmen (Zeitraum: Saison 2017/18 bis 2020/21)

den Anschluss zu verpassen. Ein Thema in dem Zusammenhang könnte die Durchlässigkeit zwischen der jeweiligen ersten und zweiten Liga sein. Noch besser gibt das jedoch wahrscheinlich die Tabelle 1 wieder. In La Liga traten während unseres Beobachtungszeitraums insgesamt 28 unterschiedliche Vereine an, was natürlich für die größte Durchlässigkeit sprechen könnte. Wenn man den Vergleich der Ligen mit einer Größe von 20 Teams bemüht, fällt die Ligue 1 mit der geringsten Anzahl an unterschiedlichen Vereinen in diesem Zeitraum auf. Diese Beobachtungen sind jedoch nicht nur beim Betrachten des sportlichen Wettbewerbs interessant, sondern es spielt auch bei den Überlegungen an die Herangehensweise der Analyse eine Rolle. In den 392 Zeilen des Datensatzes gibt es nur 24 Vereine, die nicht mehrfach auftauchen. Zwar gibt es eigentlich immer innerhalb einer Mannschaft und bei Trainer- und Betreuerstab zwischen zweier Saisons einen gewissen Wechsel, aber selten einen kompletten Austausch. Auch wenn der Wechsel eher geringfügig ausfällt, entwickeln sich Spieler weiter oder machen möglicherweise einen Rückschritt. Trotz allem werden gewisse Gemeinsamkeiten nicht auszuschließen sein. Der Umgang damit wird in dem späteren Kapitelteil 4.2 noch Thema sein. Festzuhalten ist im Rahmen der Analyse, man hat es mit 129 unterschiedlichen Vereinen zu tun.

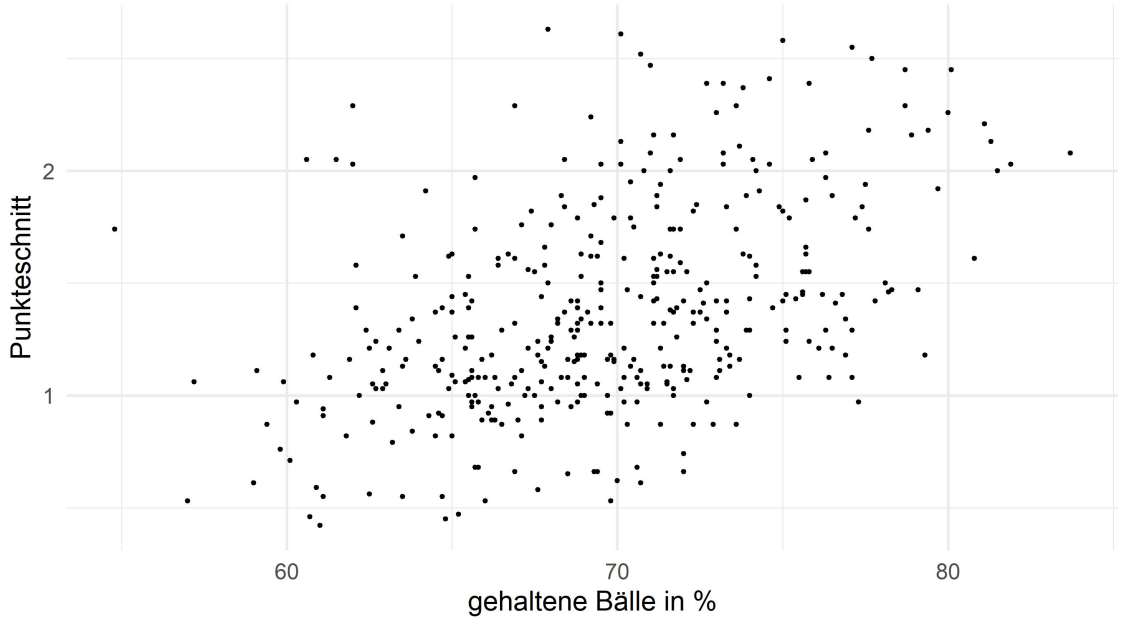


Abbildung 3: Darstellung eines linear anmutenden Zusammenhanges zwischen der Zielvariable & einer potentiellen Kovariable

So viel dazu, nun noch kurz ein Blick auf die möglichen Kovariablen. Im Besonderen soll es jetzt darum gehen, wie sich der mögliche Zusammenhang zwischen diesen und der Zielvariable gestaltet, um für die Wahl des Modelles schon eine grobe Idee zu haben, worauf geachtet werden sollte. Wie die Variable in Abbildung 3 bereits andeutet, im Rahmen der Analyse können möglicherweise lineare Einflüsse eine Rolle spielen. Bei der Betrachtung zu beachten ist, dass aufgrund der Grafik nicht automatisch auf lineare Einflüsse geschlossen werden kann, weil dies natürlich auch von der Verteilungsannahme abhängt. Trotz noch fehlender Verteilungsannahme sollte es für einen ersten schnellen Überblick zu möglichen Zusammenhängen genügen. Es gibt aber auch Variablen, bei denen ein möglicher Zusammenhang auszumachen

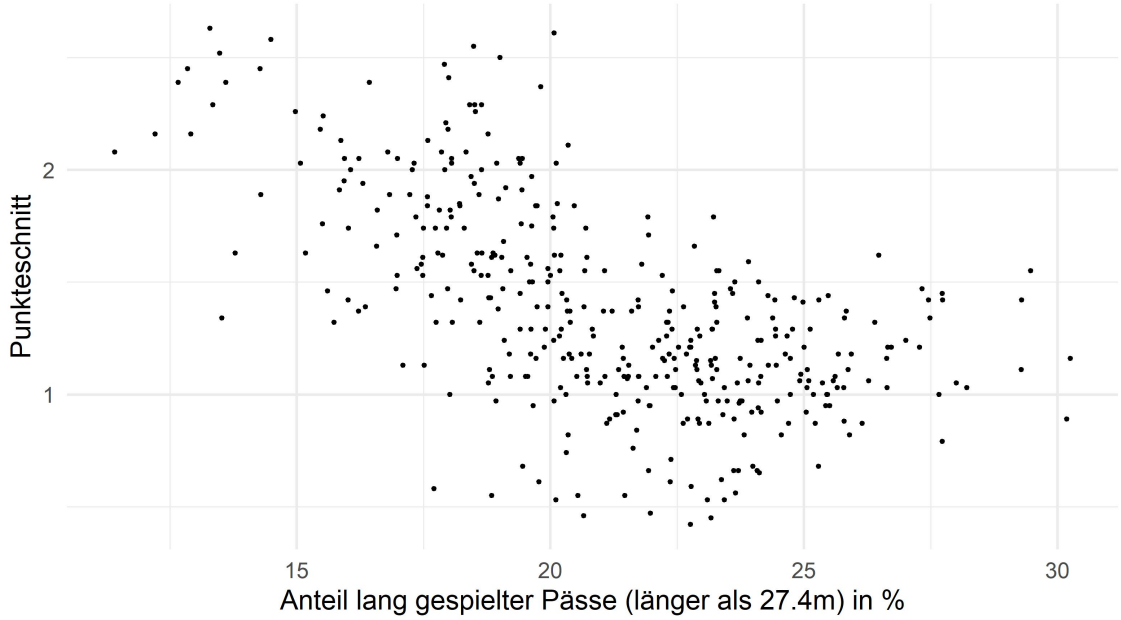


Abbildung 4: Darstellung eines nicht linear anmutenden Zusammenhanges zwischen der Zielvariable & einer potentiellen Kovariable

ist, jedoch scheint dieser nicht linear zu sein. So ein Beispiel ist in Abbildung 4 dargestellt. Beim Betrachten fallen direkt mehrere Wendepunkte auf. Die Frage, wie man damit umgeht, um solche Einflüsse auch in einem Modell wiederzugeben, merken wir uns am besten, damit sie bei den Überlegungen bei der Modellwahl miteinbezogen werden kann.
Im Rahmen dieser deskriptiven Analyse soll noch auf die im ersten Kapitelabschnitt durchgeführte Transformation einiger Variablen mit Bezug zu Defensivaktionen eingegangen werden. Der Hintergrund für die Transformation und die Formel, mit der die Transformation durchgeführt wurde, wurden bereits genaustens vorgestellt, was an dieser Stelle noch gezeigt werden soll, ist die Wirkung der Transformation. Am besten lässt sich das an einem Beispiel demonstrieren, dementsprechend wurde eine dieser Variablen ausgewählt, um zu zeigen, was sich verändert hat. Dafür wurden sowohl die eigentliche und die transformierte Variable zusammen mit der Zielvariable abgebildet (Abbildung 5). Bei dieser direkten Gegenüberstellung zeigt sich sofort, wie sich das Verhältnis zwischen Variable und Zielvariable verändert hat, aufgrund der vorgenommenen Anpassung an den gegnerischen Ballbesitz. Entstand vorher noch der Eindruck, es gäbe einen linear negativen Zusammenhang, hat sich das nach der Transformation komplett geändert. Nun würde die Interpretation wie folgt aussehen, unter der Annahme, jede Mannschaft steht 50% gegnerischem Ballbesitz gegenüber, haben tendenziell die Mannschaften den besseren Punkteschnitt, denen es

häufiger gelingt, den Ball zu blocken. Alleine aus einer rein rationalen Sicht macht der Zusammenhang der transformierten Variable deutlich mehr Sinn, da es aus der defensive Perspektive immer vorteilhaft ist, wenn eine möglicherweise erfolgreiche Aktion des Gegners verhindert werden kann. Da sich bei den transformierten Variablen ein ähnliches Bild ergab, wurden nur diese für Analysen in Betracht gezogen. Auch wenn das hier festgestellt wurde, wird in späteren Kapiteln immer wieder darauf hingewiesen werden, dass es sich bei der jeweiligen Variable um eine Angepasste handelt.

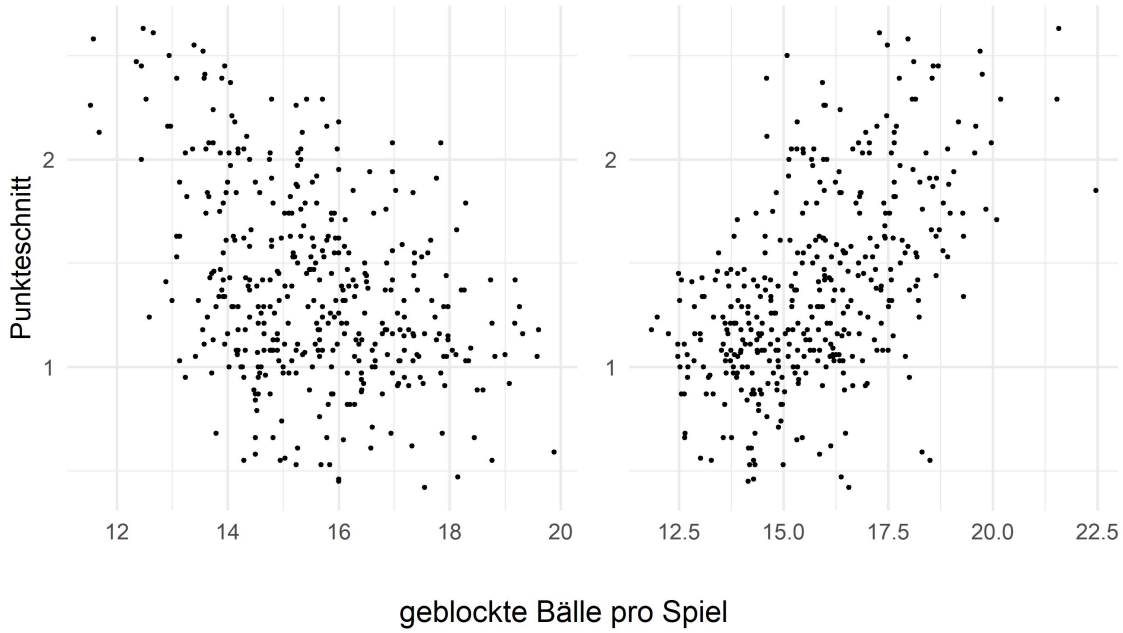


Abbildung 5: Vergleich zwischen der ursprünglichen Variable (links) & der an den gegnerischen Ballbesitz angepassten Version der Variable (rechts)

4 Theorie

4.1 Modellannahmen

Als erster Schritt auf dem Weg zur Beantwortung der Fragestellung kann die Auswahl der jeweiligen Modelltheorie angesehen werden. Die Modelltheorie selbst stellt das Gerüst dar, um das herum die Analyse des gewählten Themas aufgebaut wird. Anhand der Fragestellung ergeben sich dementsprechend die Kriterien, an denen sich orientiert und das Modell ausgewählt werden muss. Im Rahmen dieser Analyse und der gestellten Fragestellung, welche Faktoren haben einen Einfluss auf Erfolg im europäischen Vereinsfußball, liegt der Fokus auf der Interpretation der gewählten Kovariablen. Daher gehörte eine erleichterte Interpretation der Kovariablen zu den zentralen Auswahlkriterien, während eine möglichst genaue Vorhersage von Saisonerfolg eher eine sekundäre Rolle bei der Auswahl spielte, auch wenn natürlich eine sinnvolle Interpretation der Kovariablen von einer akkuraten Schätzung des Saisonerfolges abhängt. Zusätzlich dazu sollte den Kovariablen zugestanden werden, dass ihr Einfluss auch nicht linear sein könne, weshalb schlussendlich die Wahl auf ein generalisiertes additives Modell (GAM) fiel. S. N. Wood (2017, S.161)

beschreibt die allgemeine Struktur eines solchen Modells, wie folgt:

$$\begin{aligned} g(\mu_i) &= A_i\theta + f_1(x_{1i}) + f_2(x_{2i}) + \dots \\ mit\ \ \mu_i &\equiv \mathbf{E}(Y_i)\ \ und\ \ Y_i \sim EF(\mu_i, \phi) \end{aligned} \tag{2}$$

Dabei entspricht Y_i einer univariaten Zielvariable, die einer Verteilung der Exponentialfamilie mit dem arithmetischen Mittel μ_i und dem Skalierungsparameter ϕ folgt. In der Zusammensetzung aus dem Teil der linearen Komponenten, bestehend aus dem A_i stellvertretend für die Reihe i der Modellmatrix für die strikt parametrischen Komponenten und dem dazugehörigen Parameter Vektor θ, sowie dem Teil der glatten Effekte mit den glatten Funktionen f_j der Kovariablen x_k ergibt sich der Wert, den man durch das Einsetzen von μ_i in die Link-Funktion g bekommt. Es soll noch erwähnt werden, ein GAM könnte auch um Interaktionen zwischen zweier Variablen ergänzt werden. Solche Effekte werden jedoch nicht im Rahmen dieser Analyse abgebildet. Daher sind sie auch nicht weiter Thema in diesem und weiteren Kapiteln.
Hier wurde sich darauf festgelegt, dass die Zielvariable, der Punkteschnitt der i-ten Mannschaft, einer Gammaverteilung folgt. Die Gammaverteilung als Teil der Exponentialfamilie bietet den Vorteil, dass Werte kleiner als 0 vollständig und Werte größer als 3 durch entsprechend gewählten Skalierungsparameter fast vollständig ausgeschlossen werden können. Dies ist entscheidend, da der Wertebereich der Zielvariable eigentlich auf $Y_i \in [0, 3]$ beschränkt ist.

$$Y_i \overset{iid}{\sim} Ga(\mu_i, \nu)\ \ mit\ \ i = 1, \dots, n \tag{3}$$

Für Gammaverteilungen stehen laut Fahrmeir et al. (2013, S.300f) unterschiedliche Link-Funktionen zur Auswahl. Bei der Wahl fiel der Blick auf mögliche Bedingungen für die Zusammensetzung aus den linearen Komponenten und den glatten Effekten, die auch durch η dargestellt werden kann. Am Ende wurde die log-Link-Funktion gewählt, weil bei den Alternativen die Bedingung $\eta > 0$ erfüllt werden musste, was zu Restriktionen bei den Parametern führen kann. Die log-Link-Funktion ist an keine solche Bedingungen geknüpft. Sie lässt sich darstellen durch $g(\mu_i) = log(\mu_i)$ bzw. die dazugehörige Response-Funktion durch $h(\eta) = exp(\eta)$.
Ein Aspekt der GAMs, der bisher in diesem Teil ignoriert wurde, ist, wie sie geschätzt werden. Dafür gibt es auch bei den Link-Funktionen unterschiedliche Varianten. Für dieses Modell sollen die nicht-linearen Einflüsse und die damit zusammenhängenden Funktionen mit Hilfe von Penalized Splines (P-Splines) geschätzt werden. Als Hauptideen hinter ihnen beschreibt Fahrmeir et al. (2013, S.431f), dass die Schätzung der Funktion durch polynomische Splines mit einer erhöhten Anzahl an Knoten erfolgt, normalerweise ungefähr zwischen 20 und 40 Knoten. Dadurch soll genügend Flexibilität gewährleistet sein, um hoch komplexe Funktionen darstellen zu können. Des Weiteren wird es ergänzt durch einen zusätzlichen Fehlerterm, dieser wird minimiert und verhindert Overfitting. Bei den P-Splines findet hier das penalized least squares (PLS) Kriterium Anwendung. Der exakte Fehlerterm hängt dabei von dem verwendeten Splines Typen ab. Für dieses Modell basieren die P-Splines auf B-Splines. Für eine einzelne Variable lässt sich die Funktion darstellen durch das Aufsummieren der um den Faktoren γ_j multiplizierten $d = m+l-1$ Basisfunktionen. (Fahrmeir et al., 2013, S.426f)

$$f(z) = \sum_{j=1}^{d} \gamma_j\ \ B_j(z) \tag{4}$$

Gegenüber den Truncated Polynomen tun sie sich besonders dadurch hervor, dass sie lokal definiert sind. Ein weiterer Aspekt ist die Verteilung der Knoten, die für dieses Modell gleichmäßig über ein ganzes Intervall verteilt wurden. So ergibt sich als Definition für Basisfunktionen der ersten Ordnung (Fahrmeir et al., 2013, S.429):

$$B_j^1(z) = \frac{z - \kappa_{j-1}}{\kappa_j - \kappa_{j-1}} \; I(\kappa_{j-1} \leq z \leq \kappa_j) + \frac{\kappa_{j+1} - z}{\kappa_{j+1} - \kappa_j} \; I(\kappa_j \leq z \leq \kappa_{j+1}) \tag{5}$$

Während für Basisfunktionen höherer Ordnung gilt:

$$B_j^l(z) = \frac{z - \kappa_{j-l}}{\kappa_j - \kappa_{j-l}} \; B_{j-1}^{l-1}(z) + \frac{\kappa_{j+1} - z}{\kappa_{j+1} - \kappa_{j+1-l}} \; B_j^{l-1}(z) \tag{6}$$

Es ergibt sich aufgrund dieser Defintion der Basisfunktionen $2l$ äußere Knoten, die außerhalb des Intervalls [a, b] liegen, zusätzlich zu den inneren Knoten $\kappa_1, \ldots, \kappa_m$. In Folge dessen gibt es eine erwartete Knotenfolge bestehend aus $\kappa_{1-l}, \kappa_{1-l+1}, \ldots, \kappa_{m+l-1}, \kappa_{m+l}$.
So weit zu den B-Splines, da jedoch P-Splines geschätzt werden, werden die B-Splines noch um einen Fehlerterm ergänzt. In dem für die Schätzung des Modells verwendete R-Package mgcv von Simon Wood basierte die Penaltisierung auf der zweiten Abbleitung. (Fahrmeir et al., 2013, S.433)

$$\lambda \int (f''(z))^2 \, dz \tag{7}$$

Die Ermittlung des Glättungsparameters λ erfolgt hierfür per generalisierter Kreuzvalidierung (GCV). Aufgrund der Tendenz vom GCV für Overfitting zu sorgen, schlägt S. N. Wood (2017, S.186) vor den verwendeten Standardwert für γ im GCV-Verfahren von 1 auf ungefähr 1.5 zu korrigieren.
Nachdem die Umsetzung der glatten Funktionen einigermaßen klar sein sollte, soll es noch um die Residuen gehen, die häufig auch eine wichtige Rolle beim Beurteilen der Modelle spielen. (S. N. Wood, 2017, S.112) Für GAMs gilt hier dasselbe wie für generalisierte lineare Modelle. Die von linearen Modellen bekannte Annahme, dass die Residuen unabhängig identisch normalverteilt sind, kann nicht so übernommen werden. Deshalb können die Residuen auch nicht einfach mit $\hat{\epsilon}_i = y_i - \hat{\mu}_i$ berechnet werden. S. N. Wood (2017, S.112) verdeutlicht an dem Beispiel eines Poisson Modells die Abweichung von dieser bekannten Bedingung. Im Fall dieses Beispiels steigt die Varianz der Residuen im direkten Verhältnis zu den geschätzten Werten gegenüber den Residuen mit niedrigeren geschätzten Werten. Daher werden in der Regel für generalisierte Modelle die Residuen standardisiert und ähneln in Folge dessen bei der Varianz und dem Verhalten nach den Residuen der linearen Modelle.

4.2 Diskussion - GAM oder GAMM als Modell?

Im ersten Kapitelabschnitt dieses Kapitels wurde bereits die Theorie für die Modelle vorgestellt, die Wahl kann oberflächlich gerade nach der Vorstellung der Daten in Kapitel 3 anfangs eventuell überraschend wirken. Daher soll in diesem Abschnitt diskutiert werden, wieso schlussendlich die Entscheidung für das generalisierte additive Modell (GAM) fiel und was möglicherweise für ein generalisiertes additives gemischtes Modell (GAMM) gesprochen hätte.
Was ist überhaupt der entscheidende Unterschied zu einem GAM? Dies ist die entscheidende Frage, die einen auch direkt zum Kern der Diskussion bringt. Anknüpfen kann man da am Ende des letzten Abschnitts, den Residuen. Wie bereits festgestellt, wird bei einem GAM davon ausgegangen, dass die Residuen bzw. Beobachtungen alle unabhängig voneinander sind. Gegensätzlich dazu gilt bei einem

GAMM diese Annahme eben nicht. Häufig in Zeitreihen- und Clusteranalysen angewendet, erlauben sie die Überlegung, es gäbe Gruppen bzw. Individuen spezifische Effekte, die möglicherweise nicht durch die gemessenen Variablen abgedeckt wurden. (Fahrmeir et al., 2013, S.349f) Dementsprechend gibt es innerhalb der Cluster oder zwischen den Messungen an einem Individuum Korrelationen. Um das im Modell zu berücksichtigen, gibt es unterschiedliche Umsetzungen. Für diese Diskussion soll sich jedoch auf Random Intercept Modelle beschränkt werden. Die Idee hinter diesem Konzept ist, dass es neben dem regulären Intercept für alle Beobachtungen noch einen zweiten gruppenspezifischen Intercept gibt.
Für den noch unkomplizierteren Fall eines linearen gemischten Modells lässt sich $\mu_{it} = \mathbf{E}(y_{it}|\mathbf{b}_i, \mathbf{x}_{it}, \mathbf{z}_{it})$ darstellen mit $\mu_{it} = \mathbf{x}_{it}^T\beta + \mathbf{z}_{it}^T\mathbf{b}_i$, wobei $\mathbf{b}_i$ der zufällige Effekt sein soll. (Groll und Tutz, 2011, S.2f) Außerdem entspricht y_{it} der t-ten Beobachtung im Cluster i, $i = 1, \ldots, n$, $t = 1, \ldots, T_i$ aus dem Vektor $\mathbf{y}_i^T = (y_{i1}, \ldots, y_{iT_i})$. Die Vektoren $\mathbf{x}_{it}^T = (1, x_{it1}, \ldots, x_{itp})$ und $\mathbf{z}_{it}^T = (z_{it1}, \ldots, z_{itq})$ kommen den Kovariablen Vektoren gleich, die mit den festen bzw. zufälligen Effekten in Verbindung stehen. Zusätzlich wird noch angenommen, dass die Beobachtungen y_{it} bedingt unabhängig sind mit der Varianz $var(y_{it}|\mathbf{b}_i) = \phi v(\mu_{it})$. Dabei ist $v(.)$ eine bekannte Varianzfunktion und ϕ ein Skalierungsparameter. Dies kann nun verhältnismäßig einfach zu der GAMM Form erweitert werden.

$$g(\mu_{it}) = \mathbf{x}_{it}^T\beta + \sum_{j=1}^{m} \alpha_{(j)}(u_{itj}) + \mathbf{z}_{it}^T\mathbf{b}_i \tag{8}$$

Wie auch im ersten Kapitelabschnitt steht g für die Link Funktion. Groll und Tutz (2011, S.2) stellen hingegen den linearen Komponententeil als $x_{it}^T\beta$ mit dem Parameter Vektor $\beta^T = (\beta_0, \beta_1, \ldots, \beta_p)$ und die glatten Funktionen als $\alpha_{(1)}, \ldots, \alpha_{(m)}$ mit dem dazugehörigen Kovariablen Vektor $u_{it}^T = (u_{it1}, \ldots, u_{itm})$ dar. Der bereits zuvor erwähnte Cluster abhängige zufällige Effekt $\mathbf{b}_i \sim N(0, \mathbf{Q})$ findet sich natürlich auch in der GAMM Form wieder, wo $\mathbf{Q}$ sowohl eine q x q dimensionale bekannte wie auch unbekannte Kovarianzmatrix sein kann.
Soweit zur Theorie bezüglich eines GAMMs. Wie würde man jedoch ein mögliches Modell passend zur Fragestellung umsetzen? In der Tabelle 1 in Kapitel 3 wurde bereits gezeigt, einige Vereine kommen mehrfach in dem Datensatz vor. Auch wenn gerne die Rede davon ist, dass der Fußball ein schnelllebiges Geschäft ist, wird es nicht möglich sein zu argumentieren, dass es keine Kausalität zwischen den Abschlussergebnissen zweier Saisons in demselben Verein gibt. Selbst wenn der unwahrscheinliche Fall in dem Zeitraum eingetreten ist, dass sich Kader und Betreuerstab komplett geändert haben sollten, gibt es immer noch mögliche Einflussfaktoren wie Infrastruktur und Umfeld, die in der Regel gleichbleibend sind. Genau das war bereits Thema in der Einleitung, unterschiedliche Vereine haben unterschiedliche Voraussetzungen. Ist das Ganze dann nicht eine ziemlich klare Sache? GAMM sollte gegenüber GAM bevorzugt werden. Jedoch können mögliche Einflussfaktoren abseits des Platzes keinen für das Ergebnis spürbaren oder nur geringen Einfluss darstellen oder werden durch das Geschehen auf dem Platz, welches durch die vorhandenen Kovariablen abgebildet werden soll, bereits widergespiegelt. In dem Fall ist ein GAM vollkommen ausreichend. Dies war jetzt erstmal nur eine Überlegung, die natürlich auch nachgewiesen werden müsste. Beim Blick auf die Daten fällt hingegen eine Problematik für das Fitten von einem GAMM auf. Gerade bei den Vereinen, die nur einmal im Datensatz vertreten sind und Cluster der Größe eins entsprechen würden, würde ein Random Intercept einer Korrektur zum echten Wert gleichkommen. Dann hätte man mit hoher Wahrscheinlichkeit ein Modell, das overfittet. Zur Verhinderung davon gäbe es den Ansatz, diese Vereine finden für diese Analyse keine Beachtung. Bei einem eher kleinen Datensatz

wie diesem hätte so eine Reduktion merkliche Auswirkungen auf die Komplexität des Modells und stellt daher eher keine Option dar. Da dies als Argument, sich für das GAM zu entscheiden, nicht ausreichend ist, wurde auch noch ein GAMM zum Vergleich geschätzt. Dafür wurden die Kovariablen von dem finalen Modell mit den expected Goals verwendet und der Datensatz auf die Vereine begrenzt, die mindestens dreimal vorkommen. Es stellte sich heraus, dass die Streuung des Random Intercepts ($\hat{\sigma} = 0.027$) im Vergleich zur sonstigen Streuung ($\hat{\sigma} = 0.103$) zu geringfügig ausfiel, um eine Reduktion des Datensatzes und der Wahl eines GAMMs zu rechtfertigen.

4.3 Variablenselektion

Nachdem in den vorherigen Abschnitten die Frage geklärt wurde, welches Modell Anwendung findet, soll in diesem Abschnitt das Vorgehen erläutert werden, wie die Variablen für die finalen Modelle selektiert wurden. In Kapitel 3 klang es bereits an, dass es eine große Anzahl an möglichen Variablen gibt, auch wenn nicht jede Spalte im Datensatz einer möglichen Variable gleich kommt. So wurden nur numerische Spalten, die keinen direkten Bezug zu Toren bzw. Gegentoren haben, als Kovariablen in Betracht gezogen. Der Ausschluss von Toren und Gegentoren lässt sich wie folgt begründen: Selbst auf eine ganze Saison bezogen ist es offensichtlich, dass die Zahl an geschossenen und kassierten Toren eine wichtige Rolle für den Erfolg spielt. Dafür benötigt es keine tiefgründigere Analyse. Tore können als Teil des Ergebnisses angesehen werden und gehören damit nicht in die Kategorie der Einflussfaktoren, um die es, wie schon in der Einleitung formuliert, hauptsächlich gehen soll. Dazu gehört dann auch so eine Variable, die den prozentualen Anteil an Spielen ohne Gegentor angibt, da diese gleichzeitig mit angibt, in wie viel Prozent der Spiele eine jeweilige Mannschaft mindestens einen Punkt geholt hat. Ansonsten wurden noch numerische Variablen aussortiert, deren Interpretation von vornherein wenig sinnig erschienen, wie z.B. die durchschnittliche Zahl pro Spiel gespielter Pässe.

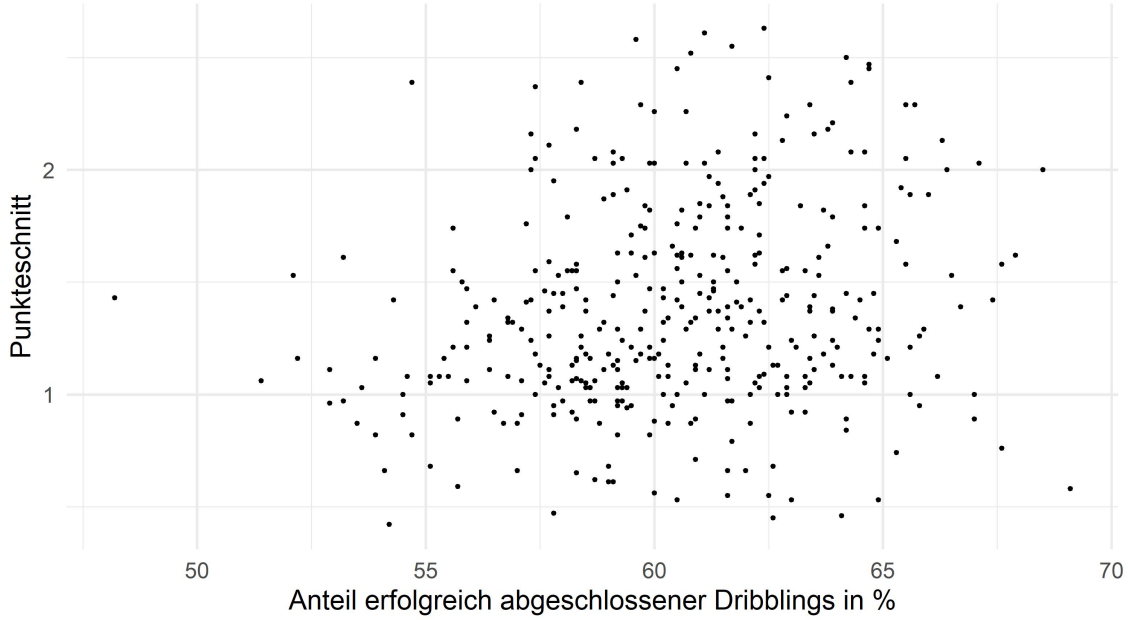


Abbildung 6: deskriptive Analyse des Zusammenhanges zwischen Zielvariable und einer möglichen Kovariable

Nachdem so insgesamt 102 bzw. 94 mögliche Kovariablen für das Modell mit wie auch ohne expected Goals Variablen identifiziert wurden, ging es darum, die Zahl für beide in weiteren Schritten zu reduzieren. Als Erstes wurden hierfür die Variablen einzeln zusammen mit der Zielvariable abgebildet. Alle Variablen, denen kein visueller Zusammenhang mit der Zielvariable nachgewiesen werden konnten (wie in Abbildung 6 dargestellt), wurden aussortiert. Mit diesem Vorgehen konnten recht schnell 37 weitere Variablen für die beiden Modelle ausgeschlossen werden.
Dies war als Selektion nicht ausreichend, daher wurde letztendlich ein Algorithmus zur Variablenselektion entwickelt. Beschreiben lässt sich das Ganze als zweiteiliger Algorithmus, jeweils bestehend aus mehreren sich wiederholenden Phasen. Dabei kommt dem ersten Teil die Rolle zu, für den zweiten Teil Modellformeln zu erstellen, die optimiert werden können. Dementsprechend stellte sich die Frage, wie viele solcher Modellformeln erstellt werden sollten und aus wie vielen Kovariablen sollen sie bestehen. Dies wurde schlussendlich beantwortet durch die zur Verfügung stehende Rechenkapazität. Die Entscheidung fiel auf insgesamt 200 unterschiedliche Modellformeln. Bei der Länge der Modellformeln wurde zwischen den beiden Modellen unterschieden. Für das erste Modell wurde sich auch aufgrund der vorhandenen Kapazitäten auf Modellformeln, bestehend aus 18 verschiedenen Kovariablen, entschieden. Um ein ähnliches Verhältnis zwischen der Gesamtzahl an Kovariablen und der Länge der Modellformeln zu haben, bestanden die Modellformeln für das zweite Modell aus 16 verschiedenen Kovariablen. Erstellt wurden diese Modellformeln nach folgendem Prozess, für jede von ihnen wurde zufällig eine der verfügbaren Kovariablen gezogen, bis die jeweilige für die beiden Modelle festgelegte Anzahl an Kovariablen erreicht wurde. Jedoch gab es bei den zufälligen Ziehungen auch die Bedingung zu beachten, dass zwischen keiner der Kovariablen in einer einzelnen Modellformel der Betrag der Korrelationen den Wert 0.8 überschreiten durfte. Vor jeder Ziehung wurden nach dieser Regel die Kovariablen ausgewählt, die gezogen werden konnten. Zusätzlich gibt es auch noch eine zweite Bedingung, nämlich wurde auch noch eine Liste erstellt, die verhindern soll, dass gewisse Variablen zusammen in ein Modell kommen. Das hat folgenden Hintergrund, es gibt bestimmte Variablen aus denen man zusammen die geschossenen oder kassierten Tore berechnen kann. Wenn dann alle dieser Variablen in ein Modell aufgenommen werden würden, wäre das so, als ob die Tore oder Gegentore direkt mit hineingenommen würden. Zum Beispiel man hat eine Kombination aus drei Variablen, mit denen das möglich ist, dann können zwei der drei Variablen Teil der Modellformel sein, jedoch wird die Aufnahme der dritten Variable unterbunden.
Nach der Erstellung dieser Modellformeln ging es im zweiten Teil, wie bereits erwähnt, um die Optimierung dieser Modellformeln. Inspiriert wurde der Optimierungsprozess von der R-Funktion stepAIC aus dem MASS Paket. (Venables und Ripley, 2002) Eine direkte Umsetzung mit stepAIC war nicht möglich, da diese Funktion keine GAMs optimieren kann. Ähnlich dazu wurden schrittweise die Modellformeln anhand des AICs optimiert, wobei die Optimierung der einzelnen Modellformeln unabhängig voneinander passierte.
Das Akaike Informationskriterium, auch bekannt als AIC, stellt anhand einer Kennzahl dar, wie gut das gewählte Modell zu den Daten passt. (Fahrmeir et al., 2013, S.148) Ein geringerer Wert des AICs steht für eine bessere Anpassung. Definiert wird es folgendermaßen:

$$AIC = -2 * l(\hat{\beta}_M, \hat{\sigma}^2) + 2(|M| + 1) \tag{9}$$

Zum besseren Verständnis eine Erläuterung zu den zwei Hauptkomponenten in der Formel. $l(\hat{\beta}_M, \hat{\sigma}^2)$ steht für den maximalen Wert, den die dazugehörige log-likelihood annimmt, wenn die Maximum Likelihood Schätzer $\hat{\beta}_M$ und $\hat{\sigma}^2$ eingefügt werden. Hingegen $|M| + 1$ steht für die Zahl der Parameter. Hierzu ist

anzumerken, dass die Varianz der Fehler σ^2 als zusätzlicher Parameter zählt.
Im ersten Schritt der Optimierung wurde für jede Modellformel ein dazugehöriges Modell geschätzt, um das jeweilige AIC ermitteln zu können. Jedes Modell folgte natürlich den in Kapitelabschnitt 4.1 vorgestellten Modellannahmen. Schrittweise wurde dann für jede Kovariable getestet, ob sich das AIC verbesserte, wenn diese aus der Modellformel rausgelassen wurden. Sobald dies für eine Kovariable zutraf, wurde die Modellformel dementsprechend angepasst und das gerade beschriebene Prozedere begann von Neuem, nur dass der AIC-Wert der angepassten Modellformel den neuen Referenzwert darstellte. Die Optimierung der Modellformel nahm dann ein Ende, wenn das AIC nicht mehr durch das Entfernen von Kovariablen verbessert werden konnte. Abschließend wurde aus den 200 optimierten Modellformeln als finale Modellformel für das jeweilige Modell, die mit dem geringsten AIC ausgewählt. Zusätzlich musste beachtet werden, dass die effektiven Freiheitsgrade der jeweiligen Modellformel nicht den Wert 39 überschreiten durften, da ansonsten sehr wahrscheinlich eine Überanpassung an die Daten stattgefunden hat.

5 Modelle

In diesem Kapitel geht es nun um die Umsetzung und das Zusammenführen der Theorie mit den Daten. Zur Umsetzung selbst ist noch wichtig zu erwähnen, dass mehrere Funktionen aus dem R Package mgcv von Simon Wood eine entscheidende Rolle bei der Umsetzung hatten. (S. Wood et al., 2016) Die Vorstellung der Ergebnisse beginnt mit einer Einführung in die finalen Modellformeln und die selektierten Variablen. Weitere Aspekte in diesem Kapitel sind die Interpretation der geschätzten Modelle wie auch das Beantworten der Forschungsfrage, welche Faktoren für den Saisonerfolg ausschlaggebend sind. Außerdem wird noch offenbart, wie gut die Anpassung der jeweiligen Modelle an die Daten ist. Im zweiten Abschnitt wird zusätzlich noch erklärt, wieso ein zweites Modell geschätzt wurde, bei dem bewusst Variablen mit Bezug zu expected Goals Modellen rausgelassen wurden.

5.1 Modell mit expected Goals

Wie der Titel bereits sagt, soll es in diesem Abschnitt um das Modell mit den expected Goals Variablen gehen. Das Modell folgt natürlich der in Kapitel 4.1 vorgestellten Theorie. Dementsprechend sollte es auch keine Überraschung mehr sein, dass der Punkteschnitt die Zielvariable ist. Bei der Interpretation soll ein Verständnis dafür erworben werden, wie die europäischen Mannschaften es schaffen, eine möglichst hohe Punktzahl in ihrer Liga zu erreichen und damit zum Erfolg zu kommen. Dafür ist nicht nur wichtig zu wissen, wie groß möglicherweise der Einfluss gewisser Faktoren ist, sondern davor mussten auch erstmal diese Faktoren bzw. Variablen bestimmt werden. Dieser Prozess wurde bereits beschrieben und der dazugehörige Algorithmus vorgestellt. Nach Durchführung vom Algorithmus zur Variablenselektion kam schlussendlich folgende Modellformel dabei raus:

$$\begin{aligned} g(\mu_i) = {} & \beta_{Intercept} + \beta_{SCAPassLive\%} SCAPassLive\%_i + \beta_{PSxG/Torschuss} PSxG/Torschuss_i \\ & + \beta_{\varnothing Schussdistanz} \varnothing Schussdistanz_i + \beta_{PressTouchMittelfeld\%} PressTouchMittelfeld\%_i \\ & + \beta_{OffPSxG/Torschuss} OffPSxG/Torschuss_i + \beta_{Torschuss\%} Torschuss\%_i \\ & + f(Strafraumflanken\%) + f(xG90 + /-) + f(PSxG90 + /-) + f(abgCarries\%) \\ & + f(Klärungsaktionen90) \end{aligned} \tag{10}$$

Diese besteht aus elf Kovariablen, wobei sechs Kovariablen linear aufgenommen wurden. Ursprünglich wurden die linearen Kovariablen auch als glatte Funktionen aufgenommen, jedoch kam beim Schätzen heraus, dass der Einfluss strikt oder nahezu linear ist. Aufgrund der erleichterten Interpretation wurden diese nicht als glatte Funktionen in der Modellformel aufgenommen.

5.1.1 Variablenvorstellung

So viel dazu, um die Schätzung ordentlich interpretieren zu können, müssen noch die Kovariablen vorgestellt werden. Zwar sollten die Variablennamen für eine grobe Idee aber noch nicht für ein tieferes Verständnis sorgen.
Da wäre die *∅Schussdistanz*, die Distanz, aus der sich die Mannschaften durchschnittlich entscheiden zu schießen. Dies kann ein Indikator dafür sein, ob eine Mannschaft eher schneller den Abschluss sucht oder es bevorzugt Aktionen auszuspielen. Von diesem Wert ausgeschlossen sind Elfmeter. Weitere lineare Kovariable im Zusammenhang mit Schüssen sind sowohl *PSxG/Torschuss* und *OffPSxG/Torschuss*, dabei steht „PSxG“ für „Post-Shot Expected Goals“. Wie der englische Name bereits sagt, geht es um die Torwahrscheinlichkeit nach Abgabe des Schusses, also beispielsweise ein Schuss, der nicht aufs Tor geht, hat einen Wert von 0. Der Wertebereich von dieser Art Variable geht von 0 bis 1. Zur Berechnung dieser Werte fließen unterschiedliche Parameter mit ein, wie die Flugbahn und Geschwindigkeit des Balles aber auch noch weitere. (Goodman, 2018) Das kann sowohl genutzt werden, um die Abschlussqualitäten einer Mannschaft bzw. von einem bestimmten Spieler wiederzugeben wie auch zu zeigen, wie herausfordernden es für einen Torhüter ist die Schüsse abzuwehren. Im Falle unserer beiden Variablen im Modell geht es darum, wie die durchschnittliche Abschlussqualität pro Schuss aufs Tor in der Offensive (*OffPSxG/Torschuss*) ist und auch mit was für Abschlüssen der Torwart in der Defensive (*PSxG/Torschuss*) zu tun hat. Eine weitere Kovariable mit Bezug zum Torabschluss ist *Torschuss%*. Sie gibt an, wie viel Prozent der Schüsse aufs Tor gehen. Da insgesamt in öffentlichen Debatten häufig der Begriff „Torschuss“ austauschbar für sowohl Schüsse wie aber auch nur für Schüsse aufs Tor verwendet wird, soll klargestellt werden, wenn in dem Text von Torschuss die Rede ist, dann sind damit nur die Schüsse gemeint, die aufs Tor gingen. Eine lineare Kovariable, die sich eher mit der Entstehung von Schüssen beschäftigt, ist *SCAPassLive%*. „SCA“ steht für „Shot-Creating Actions“ und damit für Aktionen, die im Vorhinein eines Schusses passiert sind, um diesen vorzubereiten bzw. einzuleiten. Die genaue Definition sagt, dass es um die zwei Aktionen direkt vor dem Abschluss geht. Die kreierenden Aktionen werden auch unterschiedlichen Kategorien zugeordnet. Diese Variable gibt den Anteil wieder, den Pässe aus dem laufenden Spiel an den kreierenden Aktionen für die jeweiligen Mannschaften ausmachen. Dann wäre da noch die lineare Variable *PressTouchMittelfeld%*, die zeigt, wie viel Prozent der Ballberührungen im Mittelfeld unter Druck stattfinden. Zu Ballberührungen ist noch zu ergänzen, dass die Ballannahme, mit dem Ball laufen bis hin zur Ballabgabe alles zusammen nur als eine Ballberührung zählt. Auch das Mittelfeld ist noch genauer zu definieren. Wenn man das Spielfeld in der Vertikalen in Drittel unterteilt, dann steht das mittlere Drittel für das Mittelfeld.
Anknüpfend an die linearen Kovariablen soll es um die glatt geschätzte Kovariable *PSxG90+/-* gehen. Sie unterscheidet sich zu den beiden bereits vorgestellten Kovariablen insofern, dass es darum geht, ob die Torhüter einer Mannschaft im Schnitt in den 90 Minuten mehr oder weniger Tore kassieren, als sie nach PSxG kassieren sollten. Wenn der Wert positiv ist, hat ein Torhüter weniger Gegentore zugelassen, als von ihm erwartet wurde. Dies kann natürlich für eine erhöhte Qualität der Torhüter eines Vereins beim Abwehren von Torschüssen sprechen. Nachdem es nun zum zweiten Mal um PSxG ging, wird nun

„xG“ bzw. die non-lineare Kovariable *xG90+/-* betrachtet. Was ist überhaupt „xG“? „xG“, oder auch „Expected Goals“, soll eine objektive Einordnung zur Qualität der Abschlusssituationen abgeben. Wie auch die PSxG bewegen sich die Werte im Bereich von 0 bis 1. Im Moment des Schusses werden die Distanz, die Positionierung von Gegenspielern und auch noch weitere Parameter in die Berechnung mit einbezogen. Der Abschluss selbst spielt für die Berechnung keine Rolle, daher tendieren Offensivspieler auf Weltklasse-Niveau dazu, im Vergleich zu dem Wert über eine ganze Saison hinweg zu überperformen. Zusätzlich ist noch zu erwähnen, dass Abschlüsse nach Standardsituationen Teil dieses Wertes sind. Es gibt jedoch auch einen Wert, der bewusst Elfmeter außen vor lässt, da das Bekommen von einem Elfmeter ein sehr zufälliges Ereignis ist und nicht unbedingt die Stärke beim Herausspielen von Chancen wiedergibt. Da hier gerade das Thema xG und Elfmeter angeschnitten wurde, ein Elfmeter bekommt in der Regel je nach Modell einen Wert von 0.7 - 0.8 zugeordnet. In dem Falle unseres Datensatzes und dem dort inkludierten xG Modell hat ein Elfmeter den Wert von 0.76. Zurück zu der Variable selbst, *xG90+/-* soll angeben, ob eine Mannschaft sich pro Spiel mehr qualitativ hochwertigere Abschlusssituationen erspielt als sie defensiv zulässt. Genau wie bei der Variable *PSxG90+/-* ist ein positiver Wert besser für eine Mannschaft, weil das bedeutet, es wird erwartet, dass eine Mannschaft auf Grundlage der Abschlusssituationen mehr Tore erzielt als sie kassieren sollte. Ansonsten gibt es keine weitere Kovariable mehr, die so einen direkt Bezug zum Abschluss hat, jedoch gibt es eine Kovariable, die sich damit beschäftigt, wie erfolgreich eine Mannschaft mit einem bestimmten Mittel in den Strafraum kommt. Es handelt sich dabei um *Strafraumflanken%*, wie der Name bereits sagt, geht es hier um Flanken. Explizit davon ausgeschlossen sind Flanken nach Standardsituationen, aber ansonsten geht es darum, wie viel Prozent der Flanken als Ziel den Strafraum hatten. Dann hätten wir noch die glatt geschätzte Kovariable *Klärungsaktionen90*, die zeigt, wie oft eine Mannschaft pro Spiel den Ball einfach nur wegschießt, um eine Aktion zu klären. Abschließend wurde noch die Kovariable *abgCarries%* für das finale Modell ausgewählt. Beginnen wir mit der Erklärung, was bedeuten „Carry“ im Fußball. Ein einzelnes Wort im Deutschen gibt es nicht dafür, jedoch kann man es als führen bzw. treiben des Balles bezeichnen. Die Problematik ist eine Abgrenzung zum Dribbeln zu finden. Hierfür dient das jeweilige Ziel der beiden Aktionen als Parameter zur Verdeutlichung. Beim Dribbeln soll erreicht werden, mit dem Ball an einem gegnerischen Spieler vorbeizukommen, während es bei einem „Carry“ ausschließlich darum geht, einen gewisse Strecke mit dem Ball am Fuß zurückzulegen. Auch wenn nicht direkt ein Duell mit einem gegnerischen Spieler gesucht wird, kann es einer von ihnen schaffen, den ballführenden Spieler vom Ball zu trennen. Wie oft das prozentual einer Mannschaft passiert, stellt die Kovariable *abgCarries%* dar.

5.1.2 Interpretation

So viel zur Vorstellung der Variablen. Wie sehen nun die dazugehörigen Einflüsse aus? Zur Interpretation ist noch anzumerken, dass gegensätzlich zu Formel 10 hier die Response-Funktion $h(\eta) = exp(\eta)$ verwendet wird, damit die Interpretation direkt auf der Ebene der Zielvariable stattfindet. Das muss dann jeweils auf die entsprechenden Koeffizienten bzw. Werte der glatten Funktionen angewendet werden, da die hier angegebenen Werte unverändert sind. Zusammen mit dem Intercept befinden sich die Koeffizienten der linearen Variablen in Tabelle 2. Ein Wert aus dieser Tabelle, der durchaus überrascht, ist der Koeffizient der durchschnittlichen Schussdistanz. Die Interpretation von diesem besagt: Unter Konstanthalten aller anderen Kovariablen erhöht sich durchschnittlich der Punkteschnitt um den Faktor $exp(0.016) \approx 1.016$, wenn sich die durchschnittliche Schussdistanz um einen Meter erhöht. Diese Aussage

Variable	Koeffizient	Standardabweichung
Intercept	0.587	0.270
SCA Pass Live %	-0.004	0.002
PSxG/Torschuss	-1.262	0.229
∅ Schussdistanz	0.016	0.009
Offensiv PSxG/Torschuss	0.827	0.256
Torschuss %	0.020	0.002
Ballberührungen unt. Druck Mittelfeld %	-0.007	0.002

Tabelle 2: Koeffizienten der linearen Kovariablen

ist durchaus überraschend, da unabhängig vom Fußball, es normalerweise einfacher ist ein Ziel zu treffen, das näher dran ist. Einordnend zu erwähnen wäre die Tatsache, dass alle beobachteten Werte zwischen 13.4 Metern und 18.8 Metern liegen. Dementsprechend wäre der Schluss man versucht nur noch von der Mittellinie Tore zu schießen wahrscheinlich unsinnig. Trotzdem ist es schwierig dafür eine Erklärung zu finden. Die beste Erklärung für dieses Phänomen bieten vermutlich Freistöße. Das mag im ersten Moment etwas kurios klingen, aber in Anbetracht des Modells und der anderen Variablen könnte das logisch erscheinen. Neben der ∅ Schussdistanz wurden auch xG90+/- für das Modell selektiert, es ist bekannt, dass die Schussdistanz einer der Einflussfaktoren für die Schätzung dieser Werte ist. Häufig verringert sich der xG-Wert, wenn die Distanz zum Tor größer wird. Jetzt fließen die xG für die Offensive nicht alleine ein, sondern es geht um die Differenz zu den defensiv Zugelassenen. Da für die Interpretation die anderen Werte als konstant angenommen werden, verändert sich die Differenz auch nicht. Gehen wir nun weiter davon aus, die Schussdistanz wird sowohl auf die defensiven wie offensive xG keine Auswirkungen haben und die Werte bleiben auch gleich. Es stellt sich nun die Frage: Wie bleibt das offensive xG gleich, wenn eine erhöhte Schussdistanz häufig für einen schlechten xG Wert sorgt? An der Stelle kommen die direkten Freistöße mit rein. Vergleicht man den xG Wert von einem Schuss aus dem Spiel heraus und einem Freistoß aus derselben Schussdistanz hat normalerweise der Freistoß den höheren Wert, da gegnerische Spieler neun Meter wegstehen müssen und nicht direkt im Weg stehenden den Ball blocken können. So kann es auch sein, dass ein Freistoß den gleichen Wert hat wie ein Schuss, der eigentlich näher am Tor abgegeben wurde. Bleibt noch zu erklären, wieso der Punkteschnitt nach oben geht. Häufig sind es bessere Teams, die mehr Freistöße in Nähe des Tores zugesprochen bekommen. Das treibt gleichzeitig den Durchschnitt der Schussdistanz nach oben, weil diese immer mindestens 16 Meter weg vom Tor sind. Die Variable würde dann repräsentativ dafür stehen, in mehr gefährlichere Räume unabhängig von den Abschlusssituationen zu kommen, gerade unter der Bedingung der Konstanzhaltung der anderen Variablen. Gleichzeitig muss angemerkt werden, der Umgang mit der Variable sollte vorsichtig sein, da eine Chance besteht, dass der Wert nicht signifikant ist. Nach Betrachtung der Standardabweichung wäre die Antwort diesbezüglich eindeutig, jedoch muss das mit einer gewissen Vorsicht betrachtet werden aufgrund der Post-Selektion Inference, die besagt p-Werte und Signifikanz sind nach einer Variablenselektion nicht mehr valide. Zu dem Thema gibt es mittlerweile mehrere Werke, eines stammt von Rügamer (2018).
Nachdem es jetzt schon einige Zeit um diese einzelne Kovariable ging, sollen auch noch die restlichen linearen Variablen betrachtet werden. Vorab kann festgestellt werden, dass bei keiner dieser Variablen Zweifel bezüglich der Signifikanz bestehen. Zu Beachten gilt es den jeweiligen Definitionsbereich, besonders bei den beiden PSxG Variablen. Für diese gilt [0, 1] als Definitionsbereich. Das muss auch bei der Interpretation der Variablen berücksichtigt werden, wie man das am Beispiel PSxG/Torschuss zeigen kann. Unter Konstanthalten aller anderen Parameter verringert sich der Punkteschnitt um den Faktor

$exp(-1.262 * 0.01) \approx 0.987$, wenn sich die PSxG pro Torschuss um den Wert 0.01 erhöhen. Interessant ist der direkte Vergleich zwischen den offensiven und defensiven PSxG/Torschuss. Vorab hätte man vermutlich gemeint, dass wenn beide aufgenommen werden sich die Koeffizienten betragsmäßig ähneln. Dies ist jedoch offensichtlich nicht der Fall. Das heißt, es ist bedeutender, wie qualitativ hochwertig die Abschlüsse sind, mit denen es der eigene Torwart zu tun hat, als die Qualität der Abschlüsse der eigenen Mannschaft. Die Tatsache, dass vermieden werden sollte, bei Ballberührungen im Mittelfeld unter Druck zu stehen, überrascht daher nicht. Es gibt vielleicht Leute, die sich fragen, inwiefern eine Mannschaft beeinflussen kann, ob sie im eigenen Ballbesitz unter Druck kommt oder nicht. Hierzu wäre zu sagen, dass ein gutes Positionsspiel und die Fähigkeit, schnelle Entscheidungen zu treffen, sich hilfreich gegen ein gegnerisches Mittelfeldpressing erweisen. Ebenfalls eher wenig überraschend ist ein höherer Anteil an Torschüssen erhöht den erwarteten Punkteschnitt bei Gleichbleiben der anderen Kovariablen. Zu guter Letzt wäre da noch der Anteil, den Pässe aus dem Spiel an den SCA ausmachen. Auch wenn das von allen linearen Einflüssen der Geringfügigste ist, kann die Aussage getroffen werden, es ist besser, wenn es mehr als ein Mittel gibt, mit dem Schusschancen kreiert werden können. Anknüpfend muss noch ergänzt werden, dass trotzdem die Pässe aus dem Spiel heraus auf jeden Fall den größten Anteil ausmachen. Es gibt nicht eine Mannschaft, die auch nur annäherend unter die 50% fallen würde, der geringste Wert liegt immer noch bei 61.9%.

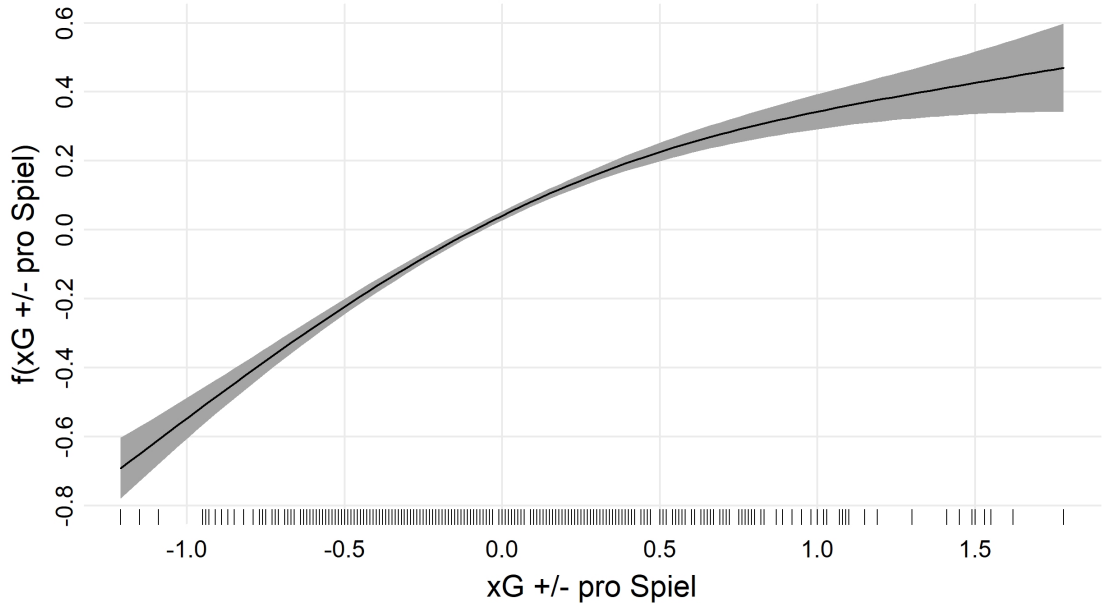
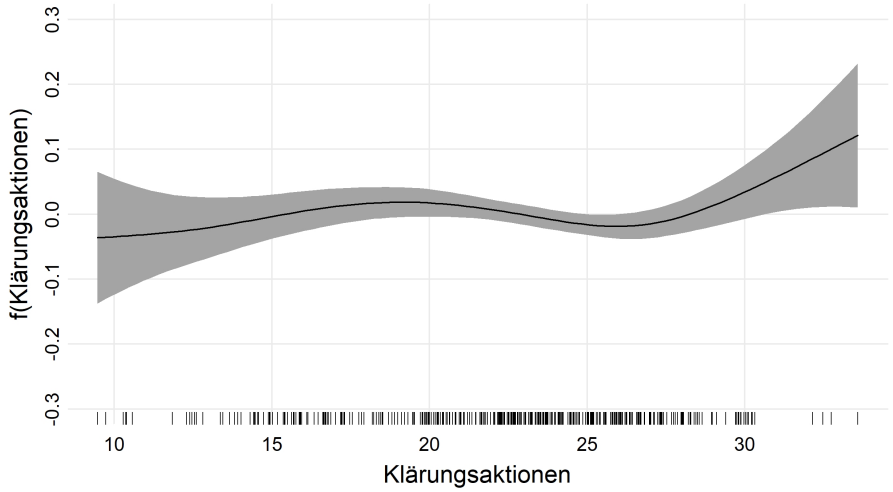


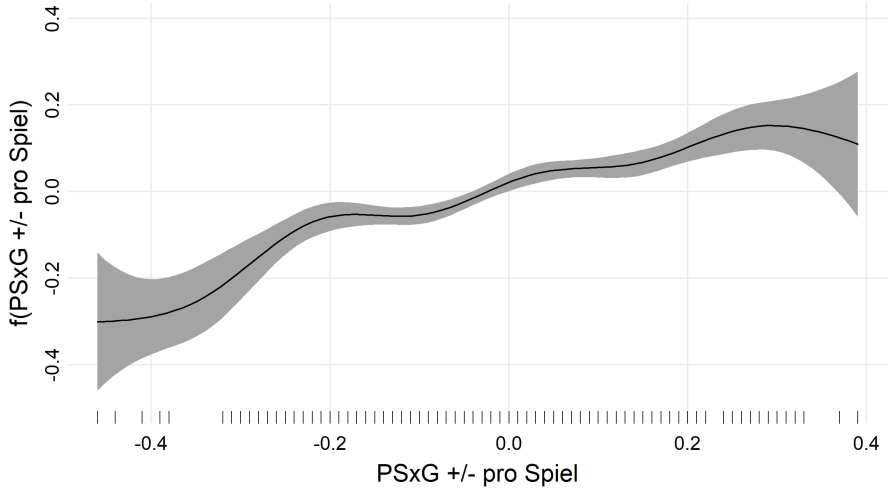
Abbildung 7: Schätzung der glatten Funktion von durchschnittlichen xG +/- pro Spiel

Nachdem die linearen Variablen besprochen sind, fehlen noch die Funktionen der glatt geschätzten Variablen. Deren Interpretationen sich durchaus unterscheiden. Anhand einer zu den Variablen gehörigen Grafiken sollen deshalb die Interpretation und die Details der Grafik vorgestellt werden. Hierfür wurde die Variable des durchschnittlichen xG +/- pro Spiel ausgewählt. Das Erste, was bei der Betrachtung der Abbildung 7 auffällt, ist der tatsächlich nicht lineare Einfluss. Anfangs scheint der Graph noch linear zu sein bevor die Steigung ab dem Bereich zwischen -0.5 und 0 immer weiter abflacht. Insgesamt ist ein streng monotoner Trend zu beobachten. Zu den Achsen ist anzumerken:
Auf der x-Achse werden die Werte des durchschnittlichen xG +/- pro Spiel aufgetragen. Die Werte der

Koeffizienten finden sich auf der y-Achse wieder. Zudem wären da die Striche an der x-Achse, die sogenannten „rugs“. Anstelle jedes Strichs findet sich mindestens eine Beobachtung der Kovariable wieder. Es gibt jedoch keine erkennbaren Unterschiede, ob ein Strich nur aus einer oder mehreren Beobachtungen besteht.
Schlussendlich wäre da noch das punktweise geschätzte graue 95%-Konfidenzband, das um den Graphen herum liegt. So viel zur Beschreibung solch einer Grafik. Zur Interpretation muss bei dem jeweiligen Wert der Kovariable die y-Koordinate abgelesen werden, um den Koeffizienten zu erhalten. Im Gegensatz zu den linearen Effekten werden diese nicht mit dem dazugehörigen Wert der Kovariable multipliziert. Zusätzlich besteht noch die Möglichkeit, zwei Werte der Kovariable miteinander zu vergleichen. Dafür wählen wir zwei verschiedene Werte auf der x-Achse aus, die leicht ablesbar sind. Hier bieten sich 0.5 und -0.5 an, deren Koeffizienten ungefähr 0.2 bzw. -0.2 sind. Unter der Bedingung, dass sich die anderen Kovariablen nicht unterscheiden, hat dann die Mannschaft mit dem xG90+/- = 0.5 im Vergleich zu der Mannschaft mit dem xG90+/- = -0.5 durchschnittlich den um den Faktor $exp(0.2 - (-0.2)) = exp(0.4) \approx 1.492$ höheren Punkteschnitt. Regulär wird einfach nur der Trend beschrieben. Vorsicht ist bei der Beschreibung an den Rändern angebracht, da dort die Unsicherheit größer wird. Für diese Kovariable stellt das im Vergleich zu anderen ein eher kleineres Problem dar. Hier gilt global: Mit größer werdenden Werten von xG90+/- bei Konstanthalten der anderen Kovariablen erhöht sich durchschnittlich der Punkteschnitt. Mit dieser recht ausführlichen Erläuterung und Interpretation an einem einzelnen Beispiel sollte es unproblematisch sein, den weiteren Interpretationen zu folgen.
Nachdem nun mindestens einmal sowohl die Interpretation der linearen und nicht linearen Einflüsse durchgeführt wurde, kann für das weitere Vorgehen festgehalten werden, dass die Interpretation immer unter der Bedingung erfolgt, dass nicht beobachtete Kovariablen konstant gehalten werden. Daher gilt ab dieser Stelle, diese Bedingung wird nicht mehr jedes Mal explizit erwähnt, sondern ist für alle kommenden Interpretationen vorausgesetzt.
Fahren wir fort mit den anderen glatten Funktionen. Beginnend mit der Abbildung 8a, die sich einer ähnlichen Problematik gegenüber sieht wie die lineare Kovariable ∅Schussdistanz. In fast jedem Abschnitt der x-Achse liegt die 0 innerhalb des Konfidenzintervalls, daher sollte die Interpretation der Kovariable mit Vorsicht erfolgen, da die Möglichkeit besteht, dass sie keinen Effekt auf die Zielvariable hat. Nur im Abschnitt 30 oder mehr Klärungsaktionen durchschnittlich pro Spiel liegt die 0 nicht innerhalb des Konfidenzbandes. Ähnlich wie auch bei der linearen Variable gilt an dieser Stelle Vorsicht mit der Behauptung, die Kovariable könne daher rausgelassen werden, aufgrund der Post-Selektion Inferenz. Unabhängig davon zeigt sich ein global zunehmender Trend. Wenn man einen der beiden Abschnitte 10 bis 18 bzw. 27 bis 35 betrachtet, dann erhöht sich der erwartete Punkteschnitt mit steigender Zahl an durchschnittlichen Klärungsaktionen pro Spiel. Im Intervall dazwischen dreht sich dieser Effekt um und mit steigender Zahl an Klärungsaktionen verringert sich der erwartete Punkteschnitt. Man könnte sagen, die Funktion gibt den Zwiespalt dieser Aktion gut wieder. Es ist zwar besser, dem Gegner den Ball in einem ungefährlicheren Raum zu geben, trotzdem ist natürlich ein Ballverlust nie etwas Positives. Eine Kovariable, bei der die glatte Funktion keine Zweifel wegen der Aufnahme ins Modell zeigt, ist abgebildet in Abbildung 8b. Abgesehen von dem rechten Rand kann von einem zunehmenden Trend gesprochen werden, heißt mit Zunahme des durchschnittlichen PSxG +/- pro Spiels erhöht sich der erwartete Punkteschnitt. Ausnahmen bilden nur die beiden Plateaus im Bereich von -0.2 bis -0.1 und 0.05 bis 0.15 bei denen sich der Koeffizient nicht verändert. Für den Randbereich von 0.3 bis 0.4 gilt die beschriebene Auswirkung natürlich nicht. Eine Beschreibung davon erweist sich jedoch wegen der verhältnismäßig großen Unsicherheit als wenig sinnvoll. Schlussendlich wären da noch die beiden Kovariablen *abgCarries%* und



(a)



(b)

Abbildung 8: Schätzung der glatten Funktion zweier Kovariablen

Strafraumflanken%, die beide auch mit derselben Problematik wie die Klärungsaktionen zu tun haben, nämlich ob sie wirklich einen Einfluss haben. Daher soll es nur kurz um die beiden gehen. Bei dem Anteil der abgenommen Carries liegt eine konvexe Funktion vor. Im Intervall [1.5, 3.2] verringert sich der erwartete Punkteschnitt, wenn der Anteil an abgenommen Carries steigt. Danach steigt der Graph wieder an, was bedeuten würde für diesen Abschnitt mit einem Anstieg des Anteils der abgenommenen Carries steigt der erwartete Punkteschnitt. Diese Aussage sollte auf jeden Fall hinterfragt werden. Gegensätzlich zu der Funktion gerade sieht man bei dem Anteil der Strafraumflanken eine konkave Funktion mit ihrem Höhepunkt bei 70%. Dementsprechend kann dies dafür sprechen, dass nicht jede Flanke als Ziel den Strafraum haben soll, sondern auch Flanken vor den Strafraum eine Option sind. Aufgrund der Unsicherheit bezüglich des vorhandenen Einflusses sollte dieser Schluss mit Vorsicht betrachtet werden. Die zu den beiden Variablen gehörigen Abbildungen 14 und 15 befinden sich im Anhang A.

Nach der Vorstellung aller Kovariablen und der dazugehörigen Koeffizienten im Einzelnen soll nochmal ein Blick auf das Gesamtbild geworfen werden, das von dem Modell gezeichnet wird. Wie würde nach dem

Modell und den selektierten Variablen eine erfolgreiche Mannschaft aussehen? Zur Beantwortung sollten die meisten der Kovariablen als Repräsentant für gewisse Aspekte des Fußballspiels angesehen werden. Das bestimmende Thema bei Betrachtung der Wahl der Kovariablen ist das Schießen und Verhindern von Toren. Das mag wenig überraschend wirken, jedoch sind es sieben von elf Kovariablen, die direkt damit zu tun haben. Die Haupterkenntnisse sind daher:
In der Offensive ist ein qualitativ hochwertiger Abschluss wichtig. Dies beginnt damit, den Ball häufig aufs Tor zu bringen. Das alleine ist jedoch nicht ausreichend, wie die Variable *OffPSxG/Torschuss* zeigt, sondern die Abschlüsse aufs Tor sollten unter anderem auch gut platziert sein. Zusätzlich sollte eine Mannschaft in der Art und Weise flexibel sein, wie sie diese Situationen vorbereitet. Abgesehen von Pässen aus dem Spiel heraus braucht es auch andere Aktionen, die schlussendlich für Abschlüsse sorgen. Außerdem reicht es für den Erfolg nicht nur das Offensiv- oder Defensivspiel zu beherrschen, sondern es muss eine gewisse Balance zwischen den beiden herrschen. Die Kovariable *xG90+/-* zeigt, eine Mannschaft muss in der Lage sein, selber in gute Schusssituationen zu kommen wie auch den Gegner an Selbigen zu hindern. Zudem ist zur Defensive anzumerken, Gegner sollen zu unplatzierten Abschlüssen gezwungen werden. Ebenfalls braucht es einen Torwart, der überdurchschnittlich gut im Abwehren von Torschüssen ist, wie *PSxG90+/-* zeigt. Abgesehen davon ist die dritte Komponente die Fehlervermeidung. Es gibt gleich mehrere Kovariablen, die dieses Ziel wiedergeben. Dazu gehört auch der Anteil der Ballberührungen unter Druck im Mittelfeld, weil präventiv verhindert werden soll, Spieler in Situationen zu bringen, in denen die Fehlerrate höher ist.

5.1.3 Modellgüte

Es bleibt die Frage der Modellgüte zu beantworten. Wie gut passen also Modell und Daten zusammen? Zur Beantwortung davon können, wie in Kapitel 4.1 erwähnt, die Residuen betrachtet werden. Laut S. N. Wood (2017, S.113) sind hier die Devianz Residuen zu bevorzugen, weil die Verteilung der Pearson Residuen um 0 herum asymmetrisch sein können. Damit sei die Annäherungen an die Residuen eines linearen Modells im Vergleich zu den Devianz Residuen geringer. Dementsprechend fiel die Wahl auch auf die Devianz Residuen, die wie folgt berechnet werden können:

$$\hat{\epsilon}_i^d = sign(y_i - \hat{\mu}_i)\sqrt{d_i} \tag{11}$$

Dabei sei d_i eine Komponente der Devianz, eingebracht von der i-ten gegebenen Größe.

$$\begin{aligned} D &= 2\{l(\hat{\beta}_{max}) - l(\hat{\beta})\}\phi \\ &= \sum_{i=1}^{n} 2\omega_i\{y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)\} = \sum_{i=1}^{n} d_i \end{aligned} \tag{12}$$

Ein paar Anmerkungen zur Notation der allgemeinen Form der Devianz:
$l(\hat{\beta}_{max})$ steht stellvertretend für die maximierte Log-Likelihood eines gesättigten Modells, ein Modell mit einem Parameter pro Datenpunkt. (S. N. Wood, 2017, S.108f) In Folge dessen ist dies der höchste Wert den die Log-Likelihood erreichen kann, gegeben der Daten. Für deren Evaluation reicht es bei einem Modell mit Exponentialverteilung aus $\hat{\boldsymbol{\mu}} = \mathbf{y}$ zu setzen. Die kanonischen Parameter des gesättigten und des beobachteten Modells werden dargestellt durch $\tilde{\boldsymbol{\theta}}$ bzw. $\hat{\boldsymbol{\theta}}$. Für unseren Fall vernachlässigbar ist ω_i, weil keine gewichtete Gamma-Verteilung verwendet wurde, also gilt $\forall_{i=1,\dots,n}\ \omega_i = 1$. Ansonsten lässt sich

die Devianz der Gamma-Verteilung, wie laut S. N. Wood (2017, S.104) folgt darstellen:

$$D = 2 * \{\frac{y - \hat{\mu}}{\hat{\mu}} - log(\frac{y}{\hat{\mu}})\} = \sum_{i=1}^{n} 2 * \{\frac{y_i - \hat{\mu}_i}{\hat{\mu}_i} - log(\frac{y_i}{\hat{\mu}_i})\} = \sum_{i=1}^{n} d_i \tag{13}$$

Mit den Devianz Residuen ist es nun möglich eine Evaluation vorzunehmen wie gut die Anpassung zwischen den Daten und dem Modell sind. Diese Beurteilung soll anhand eines Quantil-Quantil-Diagramms erfolgen, auch bekannt als Q-Q Plot. Eine perfekte Anpassung würde vorliegen, wenn alle Beobachtungen genau auf der eingezeichneten Diagonale liegen würden. Wie die Abbildung 9 zeigt ist dies hier nicht der Fall. Es gibt ein paar Auffälligkeiten. So liegen die Devianz Residuen in der Mitte fast konsequent oberhalb der Diagonalen, während die Werte zu den Rändern hin großteils unterhalb der Diagonale zu finden sind. Besonders ein paar wenige Devianz Residuen direkt am linken Rand fallen auf durch einen größeren Abstand zu der Diagonale als der Rest. Insgesamt kann trotz kleinerer Auffälligkeiten festgestellt werden, die Anpassung des Modells an die Daten ist zufriedenstellend.

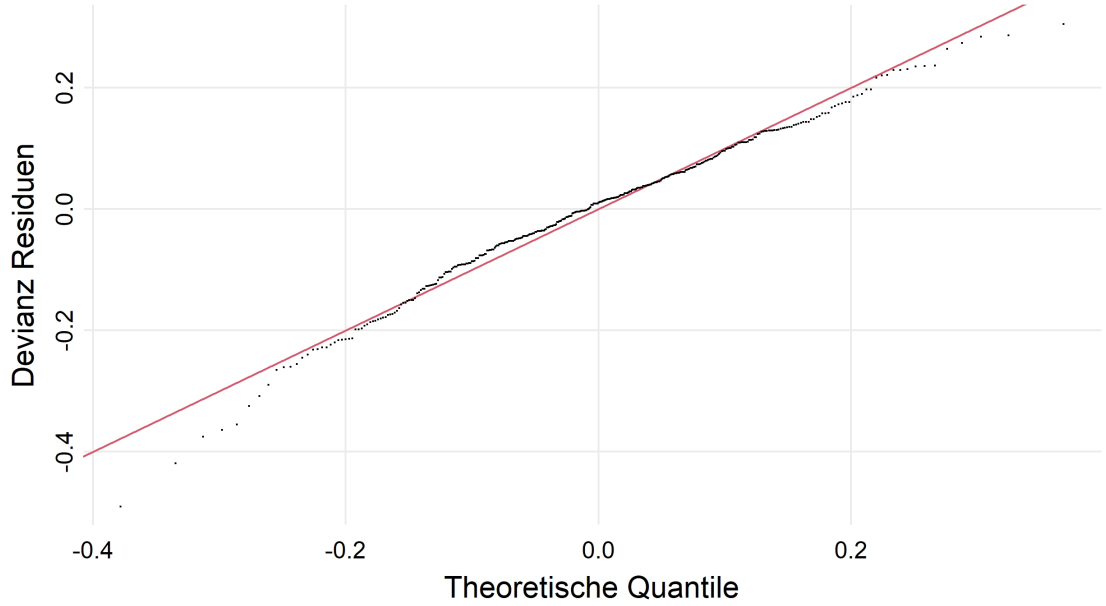


Abbildung 9: Normal Q-Q Plot zu den Devianz Residuen des Modells mit expected Goals

5.2 Modell ohne expected Goals

Nachdem es im Abschnitt gerade um das Modell mit expected Goals Variablen ging, soll nun ein Modell ohne solche Variablen vorgestellt werden. Als erstes stellt sich da direkt die Frage: Was ist die Intention dahinter zwei unterschiedliche Modelle zu schätzen, besonders wenn der größte Unterschied das Weglassen eines Variablentyps ist? Tatsächlich verändert das Weglassen von den einigen wenigen Variablen die Aufgabenstellung an das Modell. Von den Variablen mit dem xG Hintergrund ist bekannt, dass es ihre Stärke ist, die Qualität im Offensiv- wie auch Defensivspiel zu messen. Jedoch sind es keine besonders transparenten Variablen, weil die Anbieter zwar die aufgenommenen Kriterien bekannt geben, aber die genaue Gewichtung ist normalerweise Betriebsgeheimnis. Für tiefgründigere Analysen eignen sie sich daher nur bedingt bis wenig. Daher war die Aufgabenstellung für das erste Modell, Variablen miteinander zu kombinieren, die zum einen für eine gute Anpassung und zum anderen für eine gute Er-

klärbarkeit sorgen. Wie in Kapitelabschnitt 5.1 dargestellt, ist dies nicht so richtig gelungen, da gerade der Erkenntnisgewinn gering ausfiel. Für dieses Modell wurde von Anfang an gesagt, gewisse Abstriche bei der Anpassung werden in Kauf genommen, dafür sollen im besten Fall Variablen ausgewählt werden, die auch noch andere Phasen und Aktionen im Spiel abbilden als das Schießen und Verhindern von Toren. Alles natürlich mit dem Hintergrund ein besseres Verständnis für das Spiel in den europäischen Topligen zu bekommen. Zumindest lässt die Modellformel vermuten, dass es gelungen ist, verschiedene Aspekte mit einzubringen.

$$\begin{aligned} g(\mu_i) = {} & \beta_{Intercept} + \beta_{GegSteilpässe} GegSteilpässe_i + \beta_{SCAPassLive\%} SCAPassLive\%_i \\ & + \beta_{adjBlocks90} adjBlocks90_i + \beta_{PressTouchMittelfeld\%} PressTouchMittelfeld\%_i \\ & + \beta_{gehaltBälle\%} gehaltBälle\%_i + \beta_{\varnothing ProgDist/ProgCarry} \varnothing ProgDist/ProgCarry_i \\ & + \beta_{TW\varnothing Passlänge} TW\varnothing Passlänge_i + f(Schüsse90) + f(Klärungsaktionen90) \\ & + f(Torschuss\%) + f(Ballbesitz\%) + f(Strafraumflanken\%) \end{aligned} \tag{14}$$

Wie auch schon beim Modell im ersten Abschnitt wurden wieder Kovariablen linear aufgenommen, nur dieses Mal sind es sieben statt sechs. Insgesamt besteht dieses Modell aus zwölf Kovariablen, von denen einige nicht mehr vorgestellt werden müssen. *PressTouchMittelfeld%*, *Strafraumflanken%*, *SCAPassLive%*, *Torschuss%* und *Klärungsaktionen90* sind alles Kovariablen, die ebenfalls im Modell mit den xG-Variablen sind und daher als bekannt vorausgesetzt werden. Die jeweiligen Erklärungen kann man sich in Kapitel 5.1 ansehen. Auffällig ist, dass die Variable *Torschuss%* vorher als lineare und nun hier als glatt geschätzte Kovariable aufgenommen wurde.

5.2.1 Variablenvorstellung

Die erste neue Variable, die vorgestellt werden soll, ist die lineare Kovariable *TW∅Passlänge*. Eine längere Erklärung wird hierfür nicht benötigt, es handelt sich um die durchschnittliche Länge der Pässe, die der Torwart spielt, und kann möglicherweise auch dafür stehen, wie von hinten das Spiel aufgebaut wird. Beim Thema Pässe angekommen gäbe es noch eine zweite lineare Kovariable zu dem Thema, nämlich *GegSteilpässe*. Das sind Pässe, die durch die Schnittstelle der Abwehrkette in den offen Raum zwischen Verteidigern und Torwart gespielt werden. Unsere Variable hier beschäftigt sich damit, wie oft es durchschnitttlich der gegnerischen Mannschaft pro Spiel gelingt, so einen Ball anzubringen. Von einer linearen, defensiv geprägten Kovariable zur nächsten bei *adjBlocks90* geht es um die Zahl der im Durchschnitt geblockten Bälle pro Spiel. Zu geblockten Bällen gehören sowohl geblockte Schüsse wie auch geblockte Pässe. Es benötigt keine ausführliche Erläuterung für die Erkenntnis, dass das Verhindern einer möglicherweise erfolgreichen Ausführung einer gegnerischen Aktion positiv ist. Noch zu der Variable zu erwähnen ist, dass sie zu den angepassten Variablen, die schon in Kapitel 3 thematisiert wurden, gehört. Bleiben wir weiterhin im Bereich der Defensive und gehen vielleicht sogar noch ein bisschen weiter nach hinten, wenn das Verteidigen der Feldspieler nutzlos war und es zum gegnerischen Torschuss kommt. In dem Moment geht es nur noch darum, kann der Torwart das Gegentor verhindern oder nicht. Um zu messen, wie gut die Torwarte beim Verhindern sind, gibt es die Statistik der prozentualen abgewehrten Torschüsse. Genau das gibt auch die lineare Kovariable *gehaltBälle%* wieder nur nicht auf einen einzelnen Torwart beschränkt, sondern auf alle Torwarte, die für den jeweiligen Verein in der einen Saison im Einsatz waren. Abschließend zu den linearen Kovariablen wäre da noch *∅ProgDist/ProgCarry*. Im ersten Abschnitt dieses Kapitels wurde bereits definiert, was man unter einem „Carry“ versteht. Die Definition von „Pro-

gressiv Carry“ hat nochmal weitere Einschränkungen. Es beginnt mit der Tatsache, dass alle Carries in den defensiven 40% des Spielfeldes nicht als progressiv gelten. Ansonsten muss noch die Bedingung erfüllt werden, dass der ballführende Spieler um mindestens 4.6 Meter im Vergleich zu Beginn seiner Aktion näher am gegnerischen Tor dran ist oder er den Ball in den Strafraum führen konnte. Bei der progressiven Distanz wird gemessen, wie viel näher ist der Spieler während des Führens des Balles dem gegnerischen Tor gekommen. Dementsprechend soll die Variable abbilden, was für eine progressive Distanz in Metern durchschnittlich pro progressivem Carry zurückgelegt wurde.
Da vier der glatt geschätzten Kovariablen bereits bekannt sind, bleiben nur noch zwei weitere zum Vorstellen, die auch beide recht bekannt und verbreitet sind. Zum einen gäbe es die Kovariable *Ballbesitz%*, die angibt, wie viel Prozent des Ballbesitzes eine einzelne Mannschaft über die ganze Saison hinweg hatte. Die zweite wäre *Schüsse90*. Sie gibt die Zahl der im Durchschnitt pro Spiel abgegebenen Schüsse in der Offensive wieder.

5.2.2 Interpretation

Nach der Vorstellung der Variablen kann direkt in die Interpretation eingestiegen werden, besonders nach Kapitelabschnitt 5.1 sollte der Umgang vertraut sein. Es wird auch übernommen, bei der Interpretation nicht mehr explizit zu erwähnen, dass alle nicht beobachteten Kovariablen unverändert bleiben. Den An-

Variable	**Koeffizient**	**Standardabweichung**
Intercept	-0.462	0.275
SCA Pass Live %	-0.009	0.003
gehaltene Bälle %	0.020	0.002
Ballberührungen unt. Druck Mittelfeld %	-0.016	0.003
Torwart ∅ Passlänge	0.010	0.002
Progressive Distanz/Progressiven Carry	-0.013	0.007
angepasste Blocks pro Spiel	-0.012	0.007
gegnerische Steilpässe	-0.140	0.027

Tabelle 3: Koeffizienten der linearen Kovariablen

fang machen sollen die Kovariablen mit dem betragsmäßig größten Koeffizienten in der Tabelle 3. Wenn die gegnerische Mannschaft es durchschnittlich schafft pro Spiel einen Steilpass mehr anzubringen, dann verringert sich der erwartete Punkteschnitt um den Faktor $exp(-0.140) \approx 0.869$. Zur Einordnung ist zu erwähnen, dass durchschnittlich ein angekommener Steilpass pro Spiel mehr eine verhältnismäßig große Veränderung ist, weil die Mannschaften im Datensatz nur durchschnittlich zwischen 0.16 und 1.87 pro Spiel zulassen. Unabhängig davon zeigt diese Variable, dass Teams sehr bemüht darum sein sollten, den Gegner an einer erfolgreichen Umsetzung zu hindern. Häufig bleibt bei einem erfolgreichen angebrachten Steilpass nur noch der Torwart, um einen Einschlag des Balles im Tor zu verhindern. Das bringt einen direkt zur nächsten Kovariablen, den Anteil gehaltener Bälle. Wenig überraschend konnte das Modell bestätigen, dass es für den Erfolg hilfreich ist, einen Torhüter zu haben, der einen höheren Anteil an Bällen, die auf sein Tor kommen, halten kann. Daher wollen wir uns gar nicht groß bei dieser Variable aufhalten, sondern gehen zur nächsten Variable über, die sich mit der Defensive einer Mannschaft beschäftigt. Es handelt sich dabei um die angepasste Variable durchschnittliche Blocks pro Spiel, die auch schon in Abbildung 5 im Kapitel 3 gezeigt wurde. Wenn man sich zurückerinnert ist es überraschend, dass sich der erwartete Punkteschnitt um den Faktor $exp(-0.012) \approx 0.988$ verringert, wenn eine Mannschaft

unter der Annahme, sie stehe 50% gegnerischen Ballbesitz gegenüber, durchschnittlich zu einem Block pro Spiel mehr kommt. Dementsprechend liegt die Erklärung nahe, dass der in der Abbildung dargestellte Zusammenhang sich durch andere Kovariablen erklären lässt. Die offensichtliche Variante für dieses Modell ist natürlich der Ballbesitzanteil, der auch mit aufgenommen wurde und um den die Variable korrigiert wurde. Unabhängig davon sollte sowieso eine gewisse Vorsicht bei der Interpretation dieser Kovariable herrschen, da nicht vollends ausgeschlossen werden kann, dass kein Effekt vorhanden ist. Damit sind die linearen Kovariablen bezüglich Spiel gegen den Ball abgehandelt und nun soll es um diese im Spiel mit dem Ball gehen. Beginnen soll das hinten beim Torwart und dessen durchschnittliche Passlänge. Es zeigt sich mit Zunahme der durchschnittlichen Passlänge um einen Meter erhöht sich der erwartete Punkteschnitt um den Faktor $exp(0.010) \approx 1.010$. Laut Modell sollten Torwarte bevorzugen, den Ball hinten lang rauszuspielen. Weiter soll es gehen mit zwei linearen Kovariablen, die auch bereits Teil des anderen Modells waren. Interessanterweise haben sich bei beiden Kovariablen im direkten Vergleich die Tendenzen in diesem Modell verstärkt. Das heißt immer noch je öfter anteilsmäßig die Ballberührungen im Mittelfeld unter Druck geschehen, desto niedrig ist der erwartete Punkteschnitt. Konkret heißt das, wenn der Anteil der Ballberührungen unter Druck um 1% zunimmt, verringert sich der erwartete Punkteschnitt um den Faktor $exp(-0.016) \approx 0.984$. Im Falle der Schusschancen, die durch Pässe aus dem laufenden Spiel vorbereitet werden, nimmt der erwartete Punkteschnitt um den Faktor $exp(-0.009) \approx 0.991$ ab, wenn der Anteil der Pässe aus dem Spiel heraus an den kreierenden Aktionen um 1 % zunimmt. Abschließend zu den linearen Kovariablen gäbe es noch die durchschnittliche progressive Distanz, die pro progressivem Carry, zurückgelegt wird. Die Aussage, die im Zusammenhang mit dem Koeffizienten getroffen werden kann, ist, bei progressiven Carries sollte eine eher kürzere progressive Distanz zurückgelegt werden. Wie auch bei den Blocks sollte, diese Erkenntnis mit gewisser Vorsicht beachtet werden, da es nicht ausgeschlossen werden kann, dass kein solcher Effekt vorliegt.

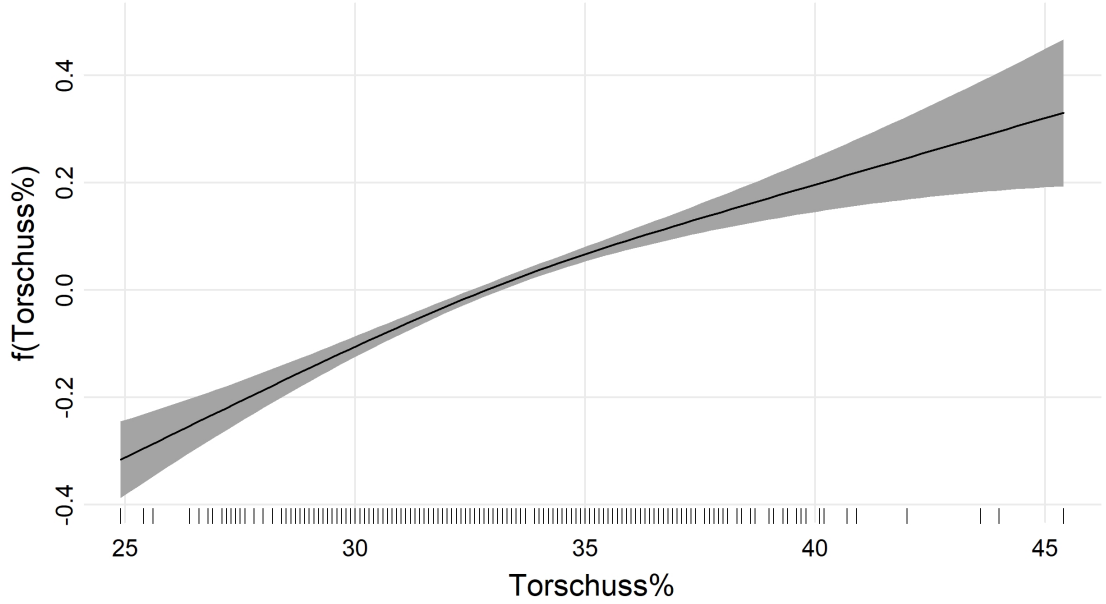


Abbildung 10: Schätzung der glatten Funktion vom Torschussanteil

So viel zu den linearen Effekten nun kann mit den glatten Effekten fortgefahren werden. Dazu ist vorab

zu sagen, ohne zu sehr vorzugreifen, dass es bei allen Kovariablen erfreulicherweise keine Unsicherheit gibt, ob ein Einfluss vorhanden ist. Den Anfang machen soll der Torschussanteil, der zwar auch schon eine Kovariable im anderen Modell war, aber wie vorher beschrieben nur linear aufgenommen wurde. In Abbildung 10 ist ein global monoton steigender Trend sichtbar, der sich in seiner anfänglichen Steigung kaum von dem linearen Koeffizienten aus dem anderen Modell unterscheidet. Ungefähr ab 35% Torschussanteil flacht die Steigung etwas ab. Insgesamt sagen lässt sich mit zunehmendem Torschussanteil erhöht sich der erwartete Punkteschnitt. Weiter geht es mit der nächsten Kovariable, die bereits aus dem anderen Modell bekannt ist, nämlich die durchschnittlichen Klärungsaktionen pro Spiel. Beim Vergleich der beiden Abbildungen(8a & 11) fällt auf, es gibt große Ähnlichkeiten bei der Form jedoch haben sich die Tendenzen an den Rändern verstärkt. Insgesamt bleibt es dabei, dass es sich um einen global steigenden Trend handelt, der sich in drei Intervalle unterteilen lässt. Im Bereich von [10, 19] und [27, 34] lässt sich ein steigender Trend beobachten, während im Intervall dazwischen ein leicht fallender Trend auszumachen ist. Eine glatt geschätzte Kovariable, die im Vergleich zu den beiden anderen Vorgestellten neu ist, wird in Abbildung 12a gezeigt. Die Ränder ausgeklammert aufgrund der geringen Datenlage und großen Unsicherheit sieht man einen global monoton steigenden Trend, der besagt, mit höherem Ballbesitzanteil erhöht sich der erwartete Punkteschnitt. Interessanterweise ist die stärkste Steigung um die 50% herum. Ein Ballbesitzanteil von 51% statt 49% hat damit einen größeren positiven Einfluss auf den erwarteten Punkteschnitt, als wenn sich dieser von 60% auf 62% erhöht.

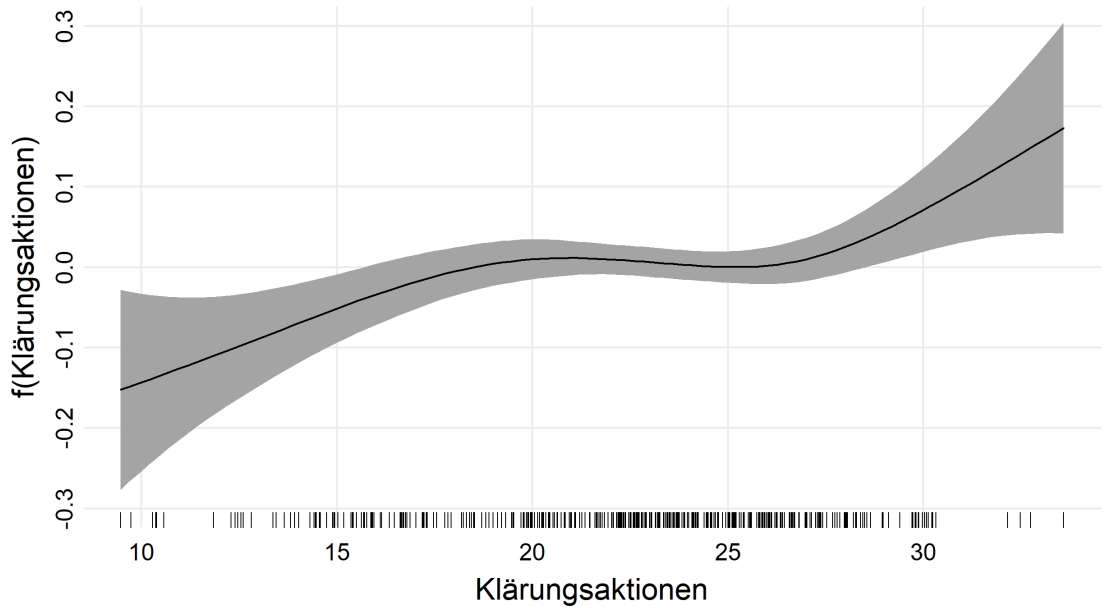
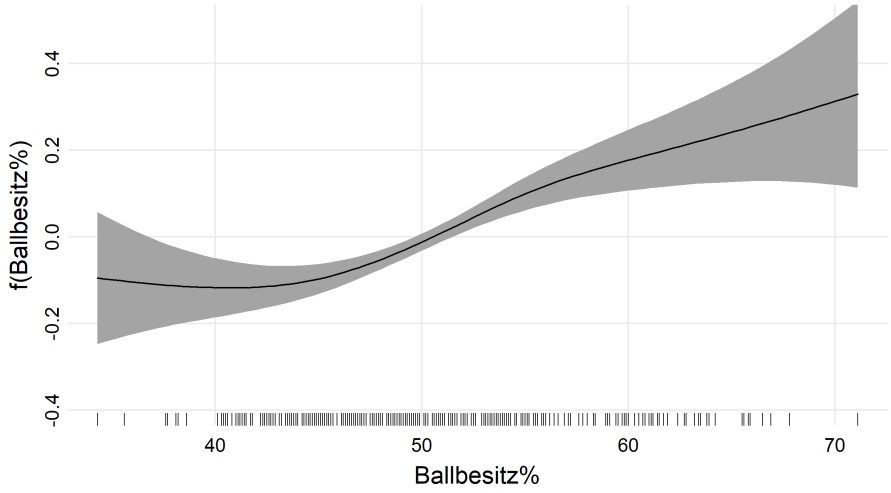


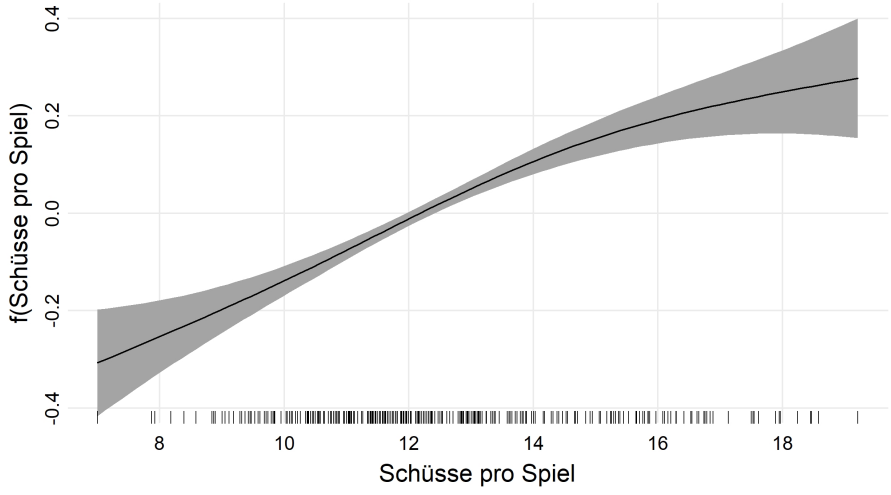
Abbildung 11: Schätzung der glatten Funktion von durchschnittlicher Anzahl an Klärungsaktionen pro Spiel

Dann gäbe es noch die Abbildung 12b zu der durchschnittlichen Anzahl der Schüsse pro Spiel. Auch hier ist, wie bereits mehrfach zuvor, ein global monoton steigender Trend zu sehen. Das bedeutet, wenn die durchschnittliche Anzahl an Schüssen pro Spiel steigt, erhöht sich der erwartete Punkteschnitt. Jedoch sollte es kein Selbstzweck sein mehr zu schießen. Wie bei den anderen Interpretationen auch geschieht diese erwartete Verbesserung des Punkteschnitts nur unter der Bedingung gleichbleibender Kovariablen. Hervorzuheben ist in dem Fall der Torschussanteil, der sich nicht ändert, weil dementsprechend nicht

aus Positionen geschossen werden soll, bei denen es unmöglich scheint den Ball überhaupt aufs Tor zu bringen. Sondern eigentlich geht es um mehr Abschlüsse, die eine ähnlich gute Chance haben wie zuvor, dass der Schuss zu einem Torschuss wird. Schlussendlich geht es nochmal um eine Kovariable, die sich in beiden Modellen findet. Wie bei den jeweiligen anderen zeigt sich beim Vergleich der beiden glatten Funktionen die Form hat sich nicht verändert. Die Koeffizienten des Strafraumflankenanteils bilden immer noch einen konkaven Graphen. Gerade der Bereich von 50% bis 60% lässt sich fast nicht interpretieren, unproblematischer wird es danach. Der Höhepunkt scheint bei ungefähr 69% Strafraumflanken zu liegen und in Folge dessen ist an der Stelle der höchste Punkteschnitt zu erwarten. Danach verringert sich stetig der erwartete Punkteschnitt, wenn sich der Anteil an Strafraumflanken erhöht. Dargestellt ist dies in Abbildung 16 im Anhang A.



(a)



(b)

Abbildung 12: Schätzung der glatten Funktion zweier Kovariablen

Wenn man nun nicht mehr nur einzelne Kovariablen betrachtet, sondern das Große und Ganze des Modells, welche Erkenntnisse kann man dabei gewinnen? Festzustellen ist, dass unterschiedliche Phasen des

Spiels ihren Weg ins Modell gefunden haben. Natürlich finden sich das Schießen und Verhindern von Toren wieder jedoch auch das Vorbereiten der Abschlusssituationen und Aspekte des Ballbesitzspiels. Da sich die Spiele selten nur direkt vor den Toren abspielen, ist das positiv zu bewerten, dass es im Modell auch nicht nur darum geht. Für die Offensive wird folgendes Bild gezeichnet:
Am ehesten stellt sich dort in den fünf Ligen Erfolg ein, wo Mannschaften auch in der Lage sind, mit anderen Alternativen als dem Passspiel zu Schüssen zu kommen. Selbiges bei den Flanken nicht jede soll im Strafraum landen, gleichzeitig sollte der Anteil der Strafraumflanken auch nicht unter ein gewisses Niveau fallen. Erfolgreiche Mannschaften beherrschen also nicht nur eine Variante, vielmehr sind sie in der Lage, auch anders zu ihren Abschlüssen zu kommen. Insgesamt schießen sie auch durchschnittlich mehr pro Spiel als Mannschaften mit weniger Erfolg. Trotzdem sollte das, wie bereits im Kapitel erwähnt, keinem Selbstzweck dienen. Es spielt nämlich auch wenig überraschend eine Rolle, wie hoch der Anteil der Schüsse ist, die überhaupt aufs Tor finden. Natürlich kann man auf die eigenen Abschlussqualitäten vertrauen, den Ball aus 30 Metern aufs Tor zu bekommen, es wird sich wahrscheinlich zeigen, dass dies dann doch nicht so oft gelingt. Daher sollte gerade die Kombination aus Schüssen pro Spiel und Torschussanteil demonstrieren, eine vermehrte Anzahl an Schüssen sollte aus Positionen und Situationen geschehen, in denen die Chance realistisch ist, den Ball auch aufs Tor zu bringen.
Hingegen ist es in der Defensive schwierig, in die ausgewählten Kovariablen mehr rein zu interpretieren. Beim Verhindern von erfolgreichen gegnerischen Steilpässen gibt es unterschiedliche Ansätze, wie das gelingen kann. Ob erfolgreiche Mannschaften wissen, wann sie sich im richtigen Moment zurückzuziehen oder sie den Passgeber stärker unter Druck setzen, um damit die erfolgreiche Umsetzung zu stören, kann nicht beantwortet werden. Dementsprechend können keine tiefer gehenden Erkenntnisse gewonnen werden, Selbiges bei den gehaltenen Bällen. Bei der Kovariable zeigt sich, es ist hilfreich, einen Torwart zu haben, der einen hohen Anteil der Torschüsse hält.
Erfolgreiche Mannschaften zeichnen sich laut Modell noch in einem anderen Bereich aus, nämlich beim Ballbesitz. Nicht nur haben sie mehr von ihm, sondern sie wissen es auch zu vermeiden, ihre Spieler im Mittelfeld in Situationen mit höherer Fehlerwahrscheinlichkeit zu bringen. Das hilft dann natürlich sowohl ihren Defensiven wie auch ihren Offensiven.

5.2.3 Modellgüte

Nach der Einordnung der Interpretationen im Gesamtkontext des Modells soll noch die Modellanpassung an die Daten gezeigt werden. Wie bereits beim ersten Modell werden auch hierfür wieder die Devianz Residuen und der Q-Q Plot verwendet. Da es keine Unterschiede bei der Theorie gibt, ist diese nachlesbar in Kapitel 5.1.3.
Die Abbildung 13 demonstriert, die Anpassung des Modells an die Daten ist in Teilen akzeptabel. Es gibt ein paar Auffälligkeiten, die angesprochen werden sollten. Die Werte in der Mitte liegen alle sehr nah an der Achse dran, aber fast konsequent oberhalb. Das ist sicher nicht ideal, jedoch insgesamt wenig problematisch besonders im Vergleich zu den Rändern. Am unteren wie auch oberen Rand finden sich einige Residuen, die deutlich von der gezogenen Achse abweichen. Die Tendenzen gab es auch bereits beim anderen Modell, aber keineswegs in dem Ausmaße. Aufgrund der vorhandenen systematischen Abweichung der Achse sollten zukünftig mögliche andere Verteilungen für das Modell ohne die xG-Variablen nicht ausgeschlossen werden.

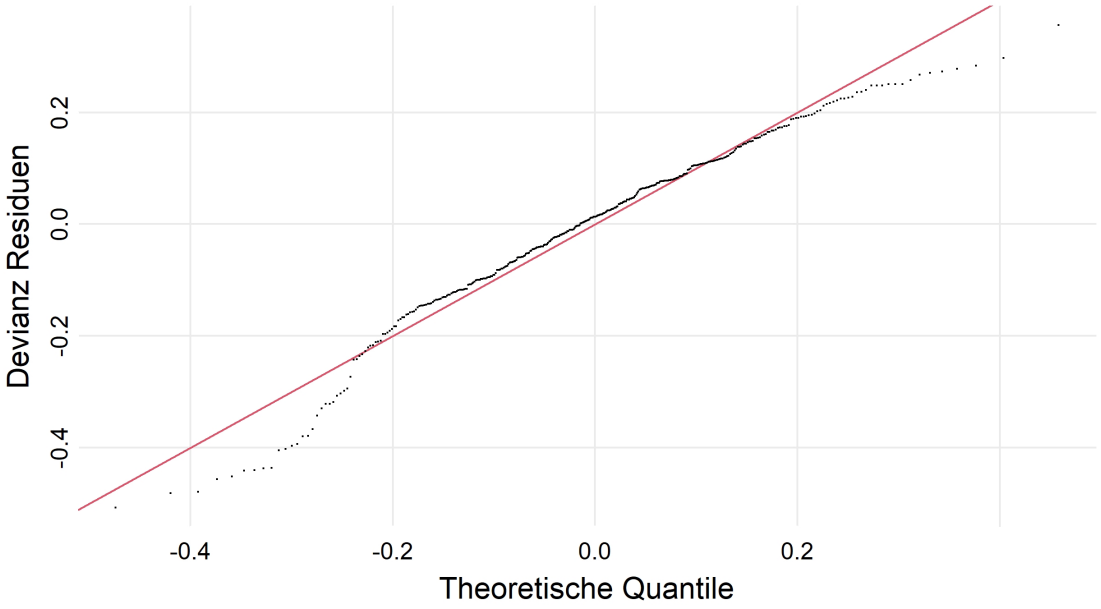


Abbildung 13: Normal Q-Q Plot zu den Devianz Residuen des Modells mit expected Goals

6 Fazit

Was lässt sich nach Abschluss der Analysen über die Erfolgsfaktoren im europäischen Vereinsfußball sagen? Es ist anzumerken, neue Erkenntnisse zu gewinnen stellt sich als schwierig heraus. Wenn man sich diese erhofft hat, ist das Ergebnis vermutlich enttäuschend.
So wurden bewusst Tore und Gegentore als Variablen der Modelle nicht ausgewählt, trotzdem hatten die Modelle einen gewissen Fokus darauf, wie die Tore erzielt und Gegentore verhindert werden können. Besonders ausgeprägt zeigte sich das im Modell mit den xG-Variablen. Daher gab es bei dem Modell auch weniger Überlegungen anzustellen, ob die Kovariablen stellvertretend für nicht genannte Aspekte im Modell stehen könnten. Demzufolge ist die Haupterkenntnis, erfolgreiche Teams schießen mit einer höheren Abschlussqualität und aus besseren Abschlusssituationen, gleichzeitig wissen sie dies beim Gegner zu verhindern. Dazu gehört auch, die Schüsse so oft wie möglich aufs Tor zu bringen. Fernab davon tat sich das Modell mit den xG-Variablen schwer, andere Erkenntnisse zu gewinnen. Das offenbart etwas über die Aussagequalität der xG-Variablen, erfüllt aber nicht wirklich die Hoffnung, Erklärungen zu bringen.
Das zweite geschätzte Modell ist in diesem Aspekt vielversprechender. Interessanterweise gibt es eine gewisse Überschneidung, so tauchen fünf Variablen in beiden Modellen auf. Dementsprechend kommt ihnen wahrscheinlich eine gewisse Rolle beim Erklären von Erfolg zu, egal ob es darum geht, wie häufig Bälle geklärt, Schüsse anteilsmäßig durch andere Aktionen als Pässe aus dem Spiel heraus vorbereitet oder anteilsmäßig Ballkontakte unter Druck im Mittelfeld vermieden werden. Ergänzend dazu kam der Ballbesitzanteil, dem zumindest laut Modell auch eine entscheidende Rolle beim Erreichen von Erfolg zukommt. Auch dieses Modell kommt nicht ohne Variablen aus, die sich mit Erzielen und Verhindern von Toren beschäftigen. So ist es ähnlich wichtig, dass eine Mannschaft viele Schüsse hat wie auch das diese häufig aufs Tor gehen. Während auf der anderen Seite des Balles es eine entscheidende Rolle spielt, einen Torwart zu haben, der viele Torschüsse abwehrt.
Allgemein zu den Modellen ist anzumerken, dass die Anpassung der gewählten Modelle in Ordnung war, aber trotzdem zumindest Überlegungen zu anderen Verteilungen angestellt werden könnten. Als richtig erwiesen hat sich die Herangehensweise, Kovariablen nicht nur linear aufzunehmen, dies sollte bei weiteren Analysen so übernommen werden.
Abschließend kann festgestellt werden, dass es nicht eine bestimmte Spielweise gibt, durch die sich erfolgreiche Mannschaft auszeichnen. Es muss jedoch immer eine gewisse Balance zwischen Offensive und Defensive herrschen. Wenn nicht die Umsetzung von beidem auf hohem Niveau passiert, ist es schwierig, erfolgreich zu sein. Eine Anregung für künftige Analysen wäre zu prüfen, ob es bezüglich des Erfolges wirklich keinen Unterschied zwischen den Spielweisen gibt.

Literatur

Athletic Bilbao. (o. D.). *Club's Philosophy*. https://www.athletic-club.eus/en/philosophy/what-is-it

Biermann, C. (2018). *Matchplan - Die neue Fußball-Matrix*. Kiepenheuer & Witsch.

Britannica, T. Editors of Encyclopaedia. (2020). English Football League. *Encyclopedia Britannica*. https://www.britannica.com/topic/Football-League

EFL. (o. D.). About the EFL. https://www.efl.com/-more/all-about-the-efl/

Fahrmeir, L., Kneib, T., Lang, S. & Marx, B. (2013). *Regression - Modells, Methods and Applications*. Springer-Verlag.

FBref. (o. D. a). *All About FBref.com*. https://fbref.com/en/about/#site_menu_link

FBref. (o. D. b). *StatsBomb*. https://fbref.com/en/statsbomb/

Frerks, O. (2014). Und morgen die Weltherrschaft. *SPOX*. https://www.spox.com/de/sport/fussball/championsleague/1402/Artikel/paris-saint-germain-nach-der-qsi-uebernahme-ligue-1-champions-league-bayer-leverkusen-ibrahimovic.html

Goal. (2019). Explained: Why Welsh teams play in the 'English' Premier League. https://www.goal.com/en/news/explained-why-welsh-teams-play-in-the-english-premier-league/5ch7uidrrtr41a9rcl7ov7ten

Goodman, M. (2018). A New Way to Measure Keepers' Shot Stopping: Post-Shot Expected Goals. *StatsBomb*. https://statsbomb.com/2018/11/a-new-way-to-measure-keepers-shot-stopping-post-shot-expected-goals/

Groll, A. & Tutz, G. (2011). *Variable Selection for Generalized Additive Mixed Models by Likelihood-based Boosting*. Ludwig-Maximillians-Universität München. https://epub.ub.uni-muenchen.de/12286/1/TR110.pdf

kicker. (2018). 50+1: Die exakte Regelung. https://www.kicker.de/502b1_die-exakte-regelung-679442/artikel

kicker. (o. D.). Bundesliga Titelträger. https://www.kicker.de/bundesliga/titeltraeger

Knutson, T. (2014). Introducing Possession-Adjusted Player Stats. *StatsBomb*. https://statsbomb.com/2014/06/introducing-possession-adjusted-player-stats/

Lega Nazionale Professionisti Serie A. (o. D.). Seria A TIM - Honours List. https://www.legaseriea.it/en/serie-a/roll-of-honour

Marca. (o. D.). Clasificación histórica Liga Santander - Primera División. https://www.marca.com/futbol/primera-division/clasificacion-historica.html

Premier League. (o. D.). *Premier League Origin*. https://www.premierleague.com/history/origins

Redaktion Sportbuzzer. (2020). Offiziell: Saison in der Ligue 1 abgebrochen - PSG ist Meister, zwei Absteiger. *Sportbuzzer*. https://www.sportbuzzer.de/artikel/ligue-1-saison-abbruch-psg-paris-meister-absteiger-aufsteiger-reaktionen/

Rügamer, D. (2018). *Estimation, model choice and subsequent inference: methods for additive and functional regression models*. Ludwig-Maximilians-Universität München. http://nbn-resolving.de/urn:nbn:de:bvb:19-223947

SID. (2018). Neuer TV-Vertrag: Premier League steuert auf neuen Rekord zu. *Fokus Online*. https://www.focus.de/sport/fussball/premierleague/england-neuer-tv-vertrag-premier-league-steuert-auf-neuen-rekord-zu_id_8463481.html

SPOX. (o. D.). Bundesliga: Geschichte, Regeln, Rekorde. https://www.spox.com/de/sport/fussball/bundesliga/bundesliga-geschichte-regeln-modus-rekorde.html

Steel, A. (2021). Have Barcelona and Real Madrid ever been relegated from La Liga? *Goal*. https://www.goal.com/en/news/have-barcelona-real-madrid-been-relegated-from-la-liga/1aha3d8yt25b31al5alcffj76l

transfermarkt. (o. D. a). Ewige Tabelle: Bundesliga von 2011/12 bis 2020/21. https://www.transfermarkt.de/bundesliga/ewigeTabelle/wettbewerb/L1/plus/?saison_id_von=2011&saison_id_bis=2020&tabellenart=alle

transfermarkt. (o. D. b). Ewige Tabelle: La Liga von 2011/12 bis 2020/21. https://www.transfermarkt.de/laliga/ewigeTabelle/wettbewerb/ES1/plus/?saison_id_von=2011&saison_id_bis=2020&tabellenart=alle

transfermarkt. (o. D. c). Ewige Tabelle: Ligue 1 von 2011/12 bis 2020/21. https://www.transfermarkt.de/ligue-1/ewigeTabelle/wettbewerb/FR1/plus/1?saison_id_von=2011&saison_id_bis=2020&tabellenart=alle

transfermarkt. (o. D. d). Ewige Tabelle: Premier League von 2011/12 bis 2020/21. https://www.transfermarkt.de/premier-league/ewigeTabelle/wettbewerb/GB1/plus/?saison_id_von=2011&saison_id_bis=2020&tabellenart=alle

transfermarkt. (o. D. e). Ewige Tabelle: Serie A von 2011/12 bis 2020/21. https://www.transfermarkt.de/serie-a/ewigeTabelle/wettbewerb/IT1/plus/?saison_id_von=2011&saison_id_bis=2020&tabellenart=alle

transfermarkt. (o. D. f). Französischer Meister. https://www.transfermarkt.de/ligue-1/erfolge/wettbewerb/FR1

transfermarkt. (o. D. g). Transfereinnahmen und -ausgaben von 2011/12 bis 2021/22. https://www.transfermarkt.de/transfers/einnahmenausgaben/statistik/plus/0?ids=a&sa=&saison_id=2011&saison_id_bis=2021&land_id=&nat=&kontinent_id=&pos=&altersklasse=&w_s=&leihe=&intern=0&plus=0

TZ. (2021a). Ligue 1: Geschichte, Vereine, Meister – alle Infos zur Liga in Frankreich. https://www.tz.de/sport/fussball/ligue-1-fussball-tabelle-spielplan-teams-psg-ligue-2-franzoesische-liga-90218234.html

TZ. (2021b). Serie A: Geschichte, Vereine, Meister – alle Infos zur Fußball-Liga in Italien. https://www.tz.de/sport/fussball/serie-a-italien-alle-infos-geschichte-vereine-meister-torjaeger-ikonen-90211412.html

UEFA. (o. D. a). Champions League - Geschichte. https://de.uefa.com/uefachampionsleague/history/

UEFA. (o. D. b). *Länder-Koeffizienten*. https://de.uefa.com/memberassociations/uefarankings/country/#/yr/2021

Venables, W. N. & Ripley, B. D. (2002). *Modern Applied Statistics with S* (Fourth) [ISBN 0-387-95457-0]. Springer. https://www.stats.ox.ac.uk/pub/MASS4/

Wikipedia. (o. D.). Liste der mitgliederstärksten Sportvereine. https://de.wikipedia.org/wiki/Liste_der_mitgliederst%C3%A4rksten_Sportvereine

Wood, S. N. (2017). *Generalized Additive Models - An Introduction with R*. Taylor & Francis Group.

Wood, S., N., Pya & Säfken, B. (2016). Smoothing parameter and model selection for general smooth models (with discussion). *Journal of the American Statistical Association*, *111*, 1548–1575.

Zenger, F. (2019). Hanno Behrens und das Rätsel der divergierenden Zweikampfquote. *Clubfans United*. https://www.clubfans-united.de/2019/08/20/hanno-behrens-und-das-raetsel-der-divergierenden-zweikampfquote-fcn/

A Grafiken

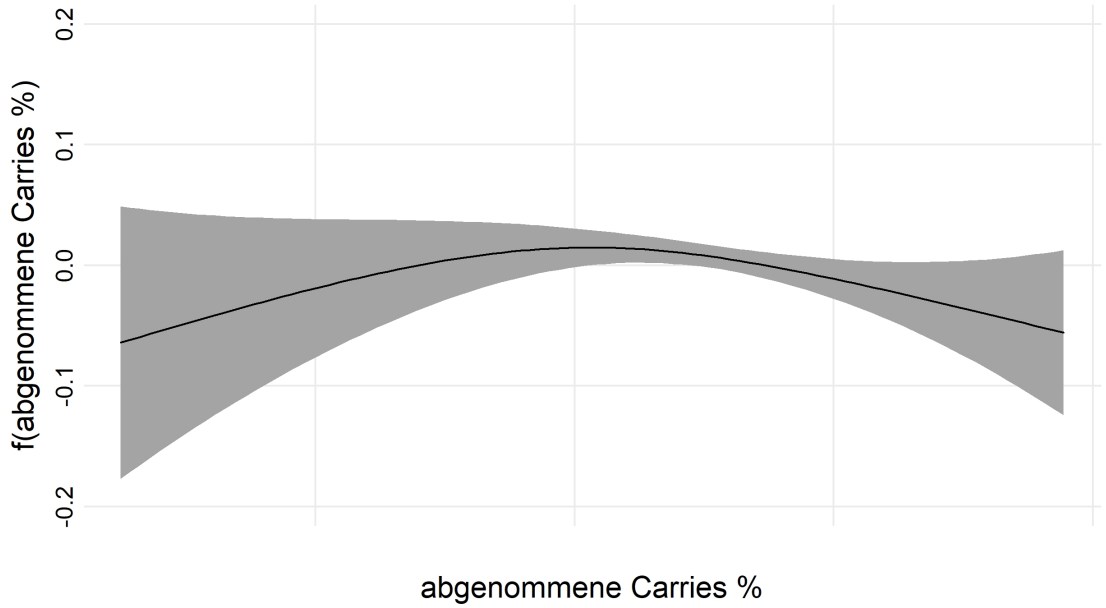


Abbildung 14: Schätzung der glatten Funktion vom Anteil abgenommener Carries (Modell mit expected Goals)

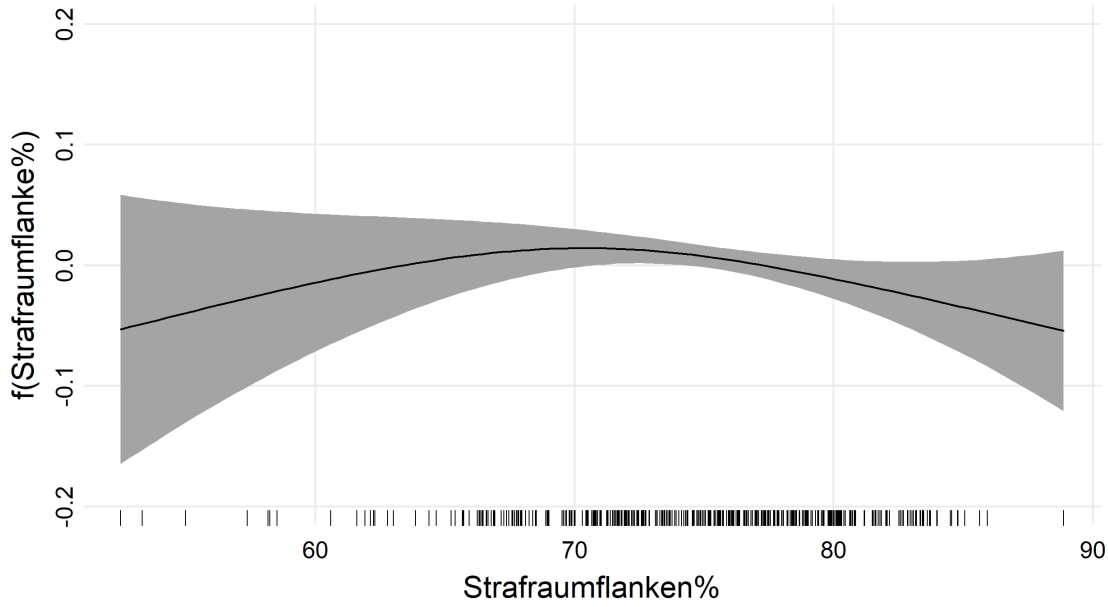


Abbildung 15: Schätzung der glatten Funktion vom Strafraumflankenanteil (Modell mit expected Goals)

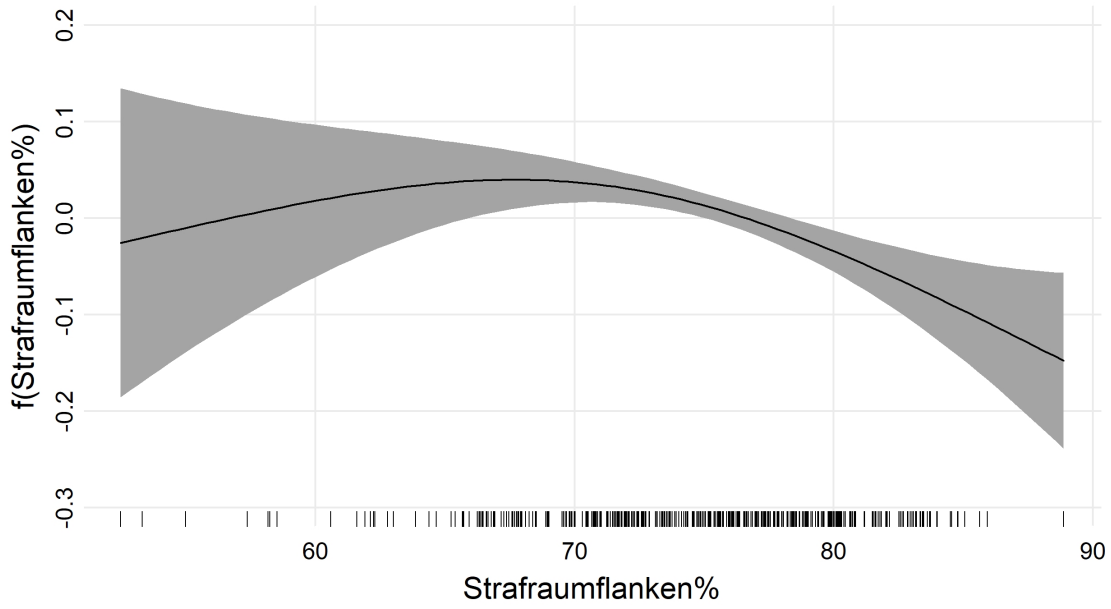


Abbildung 16: Schätzung der glatten Funktion vom Strafraumflankenanteil (Modell ohne expected Goals)

Eigenständigkeitserklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Alle sinngemäß und wörtlich übernommenen Textstellen aus fremden Quellen wurden kenntlich gemacht.

Ort, Datum	Nils Wöhl