INSTITUT FÜR STATISTIK
DER LUDWIG–MAXIMILIANS–UNIVERSITÄT MÜNCHEN

BACHELORARBEIT

Analyse des Einflusses von Corona-Einschränkungen auf Fußballspiele der Bundesliga anhand von Spielstatistiken

---

Autor: Dennis Reusch
Betreuer: Prof. Dr. Christian Heumann
Eingereicht am: 17.11.2021

Zusammenfassung

Das Ziel dieser Arbeit ist es, den Einfluss der Einschränkungen, die durch den Coronavirus in Deutschland entstanden sind, auf die Spiele der Fußball-Bundesliga zu untersuchen.
In der explorativen Datenanalyse werden die vorhandenen Daten von vier Saisons der Bundesliga und 2. Bundesliga verglichen. Trotz Corona-Einschränkungen in der Saison 2020/2021, sind die Spieldaten aus dieser Saison, denen vor der Pandemie sehr ähnlich. Das Zeitintervall, welches von besonderem Interesse für diese Arbeit ist, beginnt mit dem ersten Geisterspiel am 11.03.2020 und endet am 28.06.2020. Diese Zeitspanne beinhaltet eine zweimonatige Unterbrechung der Saison, aufgrund des pandemiebedingten Lockdowns in Deutschland. In diese Zeitspanne fällt der wichtigste Unterschied in den Daten auf, dass die Heimsiegquote, die im Mittel bei circa 43 % liegt, in dieser Zeit circa sechs Prozentpunkte niedriger liegt. Um potentielle Gründe hierfür deskriptiv zu finden, werden die Einschränkungen für die Spieler, welche durch die Vorgaben der Bundesregierung und des Hygienekonzepts der Deutschen Fußball Liga entstanden, vorgestellt. Auch eine Übersicht über den zeitlichen Verlauf der Anfangsphase des Coronavirus in Deutschland und Auswirkungen auf den damaligen Zweitligisten Dynamo Dresden als Spezialfall, werden beleuchtet.
Anschließend sollen die Daten statisch modelliert werden, wozu verschiedene Logit Modelle auf ihre Eignung, die jeweilige Fragestellung zu beantworten, verglichen werden. Die erste Frage ist, ob ein Modell anhand von Spieldaten schätzen kann ob dieses Spiel mit oder ohne Zuschauer stattgefunden hat. Die zweite Fragestellung ist, ob ein Modell anhand von Spieldaten und Zeitabschnitt, das Resultat des Spiels korrekt beurteilen kann. Die erarbeiteten Modelle zur Beantwortung der Fragestellungen erreichen eine Vorhersagegenauigkeit von circa 70 %.

Inhaltsverzeichnis

Abkürzungsverzeichnis **iv**

1 Einleitung **1**

2 Daten **2**
2.1 Datenlage 2
2.2 Explorative Datenanalyse 2
2.2.1 Verfügbare Spieldaten 3
2.2.2 Korrelation und Vergleich 6

3 Corona Historie mit Bezug auf die Bundesliga **11**
3.1 DFL Hygienekonzept 11
3.2 Corona Historie in Deutschland mit Bezug auf den Fußball 12

4 Modelltheorie **16**
4.1 Logit Modell 16
4.2 Kumulatives Logit Modell 17
4.3 Generalisierte Additive Modell (GAM) 19
4.4 Gütekriterien 19
4.4.1 Akaike Informationskriterium (AIC) 19
4.4.2 Bayesianisches Informationskriterium (BIC) 20

5 Statistische Modellierung der Daten **21**
5.1 Logit Modell 21
5.2 Generalisiertes Additives Modell (GAM) 23
5.3 Kumulatives Logit Modell 24

6 Fazit und Ausblick **27**

Abbildungsverzeichnis **28**

Tabellenverzeichnis **29**

Literaturverzeichnis **30**

Abkürzungsverzeichnis

Begriffe

AIC	Akaike Informationskriterium
BIC	Bayesianisches Informationskriterium
DFL	Deutsche Fußball Liga
GAM	Generalisiertes Additives Modell
GLM	Generalisiertes Lineares Modell

Variablen

HWR	Heimsiegquote
FTHG	Heimtore bei Spielende
FTAG	Auswärtstore bei Spielende
FTGT	Gesamtzahl Tore bei Spielende
HS	Torschüsse Heimmannschaft
AS	Torschüsse Auswärtsmannschaft
TS	Gesamtzahl Torschüsse
HST	Schüsse auf das Tor Heimmannschaft
AST	Schüsse auf das Tor Auswärtsmannschaft
TST	Gesamtzahl Schüsse auf das Tor
HF	Fouls Heimmannschaft
AF	Fouls Auswärtsmannschaft
HC	Ecken Heimmannschaft
AC	Ecken Auswärtsmannschaft
HY	Gelbe Karten Heimmannschaft
AY	Gelbe Karten Auswärtsmannschaft
HR	Rote Karten Heimmannschaft
AR	Rote Karten Auswärtsmannschaft
Avg_H	Durchschnittliche Wettquote auf Sieg der Heimmannschaft
Avg_A	Durchschnittliche Wettquote auf Sieg der Auswärtsmannschaft

„Für mich steht fest: Ohne den Einsatz unserer Fans hätten wir es niemals geschafft“
Dr. Tim Schumacher [1]

1 Einleitung

In den letzten beiden Spielen der Saison 2016/2017 konnte sich der VfL Wolfsburg mit zwei knappen Siegen gegen Eintracht Braunschweig durchsetzen und die Abstiegsrelegation der Fußball Bundesliga bestehen. Wolfsburg konnte in der Bundesliga verweilen, während Braunschweig im Folgejahr aus der 2. Bundesliga abstieg und den Gang in die 3. Liga antreten musste. Wie an diesem Beispiel zu sehen, können im Fußball einzelne Spiele große Konsequenzen für Spieler, Vereine, Mitarbeiter und Fans haben. Der Sport ist komplex und viele verschiedene Umstände und Einflüsse können den Ausschlag für Erfolg und Misserfolg geben. Dr. Tim Schumacher nannte in dem einleitenden Zitat die Fans als einen entscheidenden Faktor.

Als im Frühjahr 2020 die Covid-19 Pandemie Deutschland erreichte, hatte diese großen Einfluss auf die ganze Gesellschaft und somit auch auf den deutschen Profifußball.
Am 11.03.2020 fand das erste pandemiebedingte Geisterspiel in der Fußball Bundesliga statt. Beim 2:1 Heimsieg von Borussia Mönchengladbach gegen den 1. FC Köln durften keine Zuschauer ins Stadion. Dieses Spiel wurde durch den von der Bundesregierung verhängten Lockdown, das letzte Bundesliga Spiel für circa zwei Monate.
Nach dem Lockdown durften die Mannschaften unter Einhaltung des Hygienekonzepts der Deutschen Fußball Liga (DFL) wieder den Trainings- und Spielbetrieb aufnehmen. Die Fans durften jedoch bis zum Ende der folgenden Spielzeit 2020/2021 nicht ihre Mannschaften im Stadion unterstützen.

In diese Arbeit wird der Einfluss der Corona-Einschränkungen auf die Spiele der Bundesliga und 2. Bundesliga untersucht. Hierfür werden Spieldaten der letzten vier Saisons der beiden Spitzenligen in Deutschland betrachtet. Anhand den vorliegenden Daten sollen statistische Modelle gefunden werden, um verschiedene Fragestellungen zu beantworten. Dafür werden in der explorativen Datenanalyse deskriptiv erste Ergebnisse über den Einfluss der Einschränkungen auf den deutschen Profifußball herausgearbeitet.
Anschließend wird in Kapitel 3 das Hygienekonzept der DFL vorgestellt, um die Einschränkungen der Spieler genauer zu verstehen und ein Blick auf die anfängliche Historie der Covid-19 Pandemie in Deutschland geworfen. Um die statistische Modellierung der Daten in Kapitel 5 durchführen zu können, werden die benutzten Modelle und Gütekriterien in Kapitel 4 definiert und motiviert.
Abschließend folgt in Kapitel 6 das Fazit und ein Ausblick auf fortführende Untersuchungsmöglichkeiten.

[1] Geschäftsführer der VfL Wolfsburg Fußball GmbH, über die Abstiegsrelegation (vgl. Interview, 2017).

2 Daten

Der verwendete Datensatz ist frei zugänglich auf der Internetseite von Football-data[2]. Die Plattform kompiliert Daten und Spielstatistiken für verschieden Ligen. Die Daten die für die Bundesliga und 2. Bundesliga vorliegen stammen, laut Football-data, von Bundesliga.de und wurden stichprobenartig auf Gleichheit und Korrektheit überprüft. Neben den Spieldaten enthält die Datenbank einige Wettquoten von verschiedenen Buchmachern, da die primäre Zielgruppe der Plattform aus dem Sportwettensegment kommt. Für diese Arbeit wurden vier Spielzeiten der Bundesliga und 2. Bundesliga von der Saison 2017/2018 bis einschließlich der Saison 2020/2021 betrachtet.

2.1 Datenlage

Im verwendeten Datensatz befinden sich insgesamt 2448 Spiele. Neben Datum des Spiels, den beteiligten Mannschaften und dem jeweiligen Resultat, stehen charakteristische Spieldaten zur Verfügung. Diese umfassen jeweils einen Wert pro Heim- und Auswärtsmannschaft. Die vorhandenen Spieldaten sind Tore, Torschüsse, Schüsse auf das Tor, Ecken, Fouls, gelbe und rote Karten. Für die weitere Untersuchung der Daten wurden zusätzlich jeweils Summen der einzelnen Mannschaftswerte gebildet.

Die Wettquoten der verschiedenen Sportwettenanbietern sind jeweils Abschlusswettquoten, das heißt die letzte Wettquote bevor das Spiel startete. Da eine Wettquote die vom Buchmacher geschätzte Wahrscheinlichkeit für den Ausgang des jeweiligen Spiels darstellt, wird diese im Verlauf dieser Arbeit ebenfalls genutzt. Um einheitlich und übersichtlich mit den Quoten der verschiedenen Buchmachern umzugehen, wurde über das arithmetische Mittel von fünf Anbietern, welche über alle vier Spielzeiten und über beide Ligen geschlossen vorliegen, eine durchschnittliche Wettquote für jedes Spiel berechnet.

2.2 Explorative Datenanalyse

Bevor die einzelnen Variablen untersucht werden können, muss unterschieden werden, welche Zeitintervalle betrachtet werden sollen. Wie zuvor genannt, fand das erste pandemiebedingte Spiel ohne Zuschauer am 11.03.2020 statt. Das Spiel sollte ursprünglich am 21. Spieltag stattfinden und musste wegen Unwetter verlegt werden (DFL, 2020a). Dieses erste Geisterspiel wird bereits zur Periode unter der Covid-19 Zeit gezählt, auch wenn es noch vor dem Lockdown und der Wiederaufnahme des Spielbetriebs im Mai 2020 stattfand. Der Grund dafür liegt in der Annahme, dass die ungewohnten Umstände für Mannschaften und Spieler nicht zu den Normalbedingungen gezählt werden können.

Durch die Unterteilung in ein Zeitintervall vor der Pandemie und ein Zeitintervall ab der Pandemie, erhält man von den ursprünglichen 2448 Spielen 1672 Spiele vor Corona und

[2]Quelle: https://www.football-data.co.uk/germanym.php

776 Spiele während. Neben dem Nachholspiel vom 11.03.2020, wurde ein zusätzliches Spiel zwischen Werder Bremen und Eintracht Frankfurt, aufgrund einer Spielverlegung des Europa League Spiels von Frankfurt, vom 24. Spieltag erst am 03.06.2020 nachgeholt (vgl. DFL, 2020c). Die Anzahl von 776 Spielen ergibt sich somit aus den letzten neun Spieltagen mit jeweils acht Spielen in der Bundesliga und der 2. Bundesliga, zwei Nachholspielen und der vollen Saison 2020/2021 mit 612 Spielen.

2.2.1 Verfügbare Spieldaten

Folgende Spieldaten sind im Datensatz enthalten:

Home Win Ratio (HWR): Die Heimsiegquote wurde aus dem Datensatz erhoben. Heimsiege wurden durch die Gesamtanzahl von Spielen geteilt. Der Mittelwert über den gesamten Datensatz beträgt 42,89 %.

Full Time Home Goals (FTHG): Die Anzahl an Toren der Heimmannschaft bei Spielende. Der Mittelwert beträgt 1,634, das Maximum liegt bei acht Toren. Das Spiel 1. FC Köln gegen Dynamo Dresden in der 2. Bundesliga, am 10.11.2018 endete 8:1. Selbiges gelang drei weiteren Heimmannschaften im Beobachtungszeitraum mit dem VFL Wolfsburg, RB Leipzig und dem FC Bayern München.

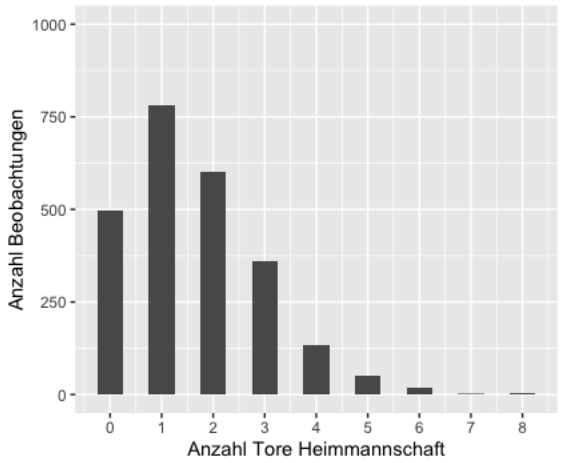

Abbildung 2.1: Histogramm der Toranzahl von Heimmannschaften.

Full Time Away Goals (FTAG): Die Anzahl an Toren der Auswärtsmannschaft bei Spielende. Der Mittelwert beträgt 1,337 und liegt somit etwas niedriger als der Mittelwert der erzielten Tore der Heimmannschaften. Das Maximum hier liegt ebenfalls bei acht Toren, gefallen im Spiel der 2. Bundesliga von Erzgebirge Aue gegen SC Paderborn am 09.05.2021.

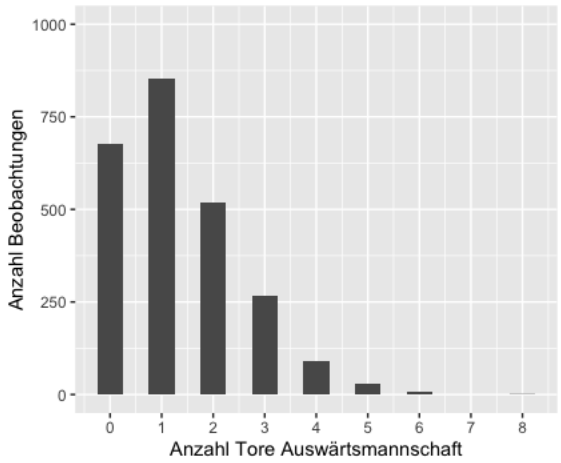

Abbildung 2.2: Histogramm der Toranzahl von Auswärtsmannschaften.

Im Vergleich mit der Anzahl von Toren der Heimmannschaften ist in den Histogrammen zu erkennen, dass die Auswärtsmannschaften mehr Spiele mit null Toren oder einemTor, jedoch weniger Spiele mit mehr als einem Tor haben als die Heimmannschaften.

Full Time Goal Total (FTGT): Die Gesamtanzahl an Toren definiert sich als die Summe der beiden Einzelwerte von Heim- und Auswärtsmannschaften. Der Mittelwert liegt hier bei 2,971 Toren pro Spiel. Das Maximum befindet sich bei elf Toren, das beim bereits genannten Spiel von Erzgebirge Aue gegen SC Paderborn, durch einen Endstand von 3:8 entstand.

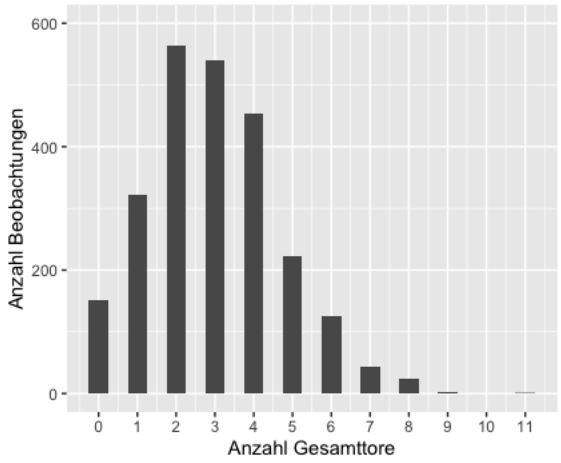

Abbildung 2.3: Histogramm der Gesamttoranzahl .

Home Shots (HS): Die Anzahl der der Torschüsse von Heimmannschaften sind im Mittel 14,005 pro Spiel. Das Maximum in den Daten sind 34 Schüsse von Eintracht Frankfurt im Spiel gegen den SC Freiburg am 26.05.2020.

Away Shots (AS): Für die Torschüsse der Auswärtsmannschaft liegen Mittel und Maximum mit 12,030 und 32 Schüssen unter den Werten der Heimmannschaft. Beim Auswärtsspiel

des FC Bayern München bei Hannover 96, am 15.12.2018, schoss der FC Bayern 32 mal und somit genau so oft wie beim Spiel gegen FC Schalke 04 am 24.01.2021.

Total Shots (TS): Das Total der Schüsse liegt im Mittel bei 26,040. Das Maximum von 46 Schüssen, erfolgte in zwei Begegnungen. Das erste Spiel war das bereits genannte Aufeinandertreffen des FC Bayern München gegen den FC Schalke 04, die zweite Partie fand am 27.04.2018 in der 2. Bundesliga zwischen Arminia Bielefeld und dem 1. FC Kaiserslautern statt.

Home Shots on Target (HST): Der Unterschied zwischen Torschüssen und Schüssen auf das Tor ist, dass ein Torschuss ein beabsichtigter Schuss in Richtung Tor ist. Wird der Schuss geblockt oder geht am Tor vorbei, zählt dieser weiterhin als Torschuss. Ein Schuss auf das Tor hingegen heißt, dass der Ball entweder im Tor landet oder ohne das Eingreifen des Torwarts im Tor gelandet wäre.
Im Durchschnitt brachten Heimmannschaften pro Spiel 4,973 Schüsse aufs Tor. Im bereits genannten Spiel zwischen Eintracht Frankfurt und dem SC Freiburg waren es mit 16 Schüssen der Frankfurter auf das Tor die meisten im Beobachtungszeitraum.

Away Shots on Target (AST): Wie bei den anderen Variablen liegt der Schnitt für Schüsse auf das Tor der Gästemannschaft mit 4,293 ebenfalls niedriger als der Durchschnitt des Wertes der Heimmannschaft. 14 Schüsse von der Auswärtsmannschaft auf das Tor stellt das Maximum dar, welches Holstein Kiel, Bayer 04 Leverkusen und drei mal dem FC Bayern München gelungen ist.

Total Shots on Target (TST): Die Gesamtanzahl an Schüssen auf das Tor in einem Spiel, fielen im Schnitt 9,266. Der Höchstwert von 21 Schüssen auf das Tor wurde zweimal erreicht. Im Spiel von Borussia Dortmund gegen Bayer 04 Leverkusen, am 14.09.2019, und in der mehrfach genannten Begegnung zwischen Eintracht Frankfurt und SC Freiburg.

Durchschnittliche Quote auf Sieg Heimmannschaft (Avg_H): Die durchschnittliche Abschluss Wettquote auf einen Sieg der Heimmannschaft betrag im Mittel 2,615, was einer von den Buchmachern vorhergesagten Siegwahrscheinlichkeit von 38,24 % entspricht. Der Durchschnitt der vorhergesagten Siegwahrscheinlichkeiten liegt deutlich niedriger, als die tatsächliche Heimsiegquote im beobachteten Zeitraum von 42,89 %. Die höchste Heimsiegquote lag beim Spiel 1. FC Nürnberg gegen den FC Bayern München am 28.04.2019 mit einer Quote von 18,070, also einer Gewinnwahrscheinlichkeit von 5,53 % laut Buchmachern.

Durchschnittliche Quote auf Sieg Auswärtsmannschaft (Avg_A): Auf einen Auswärtssieg lag die mittlere Durchschnittswettquote bei 3,876, somit bei einer vorhergesagten Siegwahrscheinlichkeit von 25,80 %. Auch diese liegt deutlich niedriger als die tatsächliche Auswärtssiegquote über die vier betrachteten Saisons von 30,47 %. Hannover 96 hatte beim Auswärtsspiel gegen den FC Bayern München die höchste Wettquote auf einen Sieg

mit 35,020, umgerechnet eine Siegwahrscheinlichkeit von 2,86 %.

Die weiteren Variablen werden in der untenstehenden Tabelle 2.1 mit Mittelwert und Maximum beschrieben.

Name	**Beschreibung**	**Mittelwert**	**Maximum**
Home Fouls (HF)	Fouls Heimmannschaft	12,514	32
Away Fouls (AF)	Fouls Auswärtsmannschaft	13,042	29
Home Corners (HC)	Ecken Heimmannschaft	5,289	19
Away Corners (AC)	Ecken Auswärtsmannschaft	4,578	15
Home Yellow Cards (HY)	Gelbe Karten Heimmannschaft	1,806	8
Away Yellow Cards (AY)	Gelbe Karten Auswärtsmannschaft	2,064	8
Home Red Cards (HR)	Rote Karten Heimmannschaft	0,067	2
Away Red Cards (AR)	Rote Karten Auswärtsmannschaft	0,098	2

Tabelle 2.1: Beschreibung der weiteren Variablen mit Mittelwert und Maximum.

2.2.2 Korrelation und Vergleich

Nun wird die Korrelation zwischen den Variablen betrachtet. Zur Verwendung kommt hier der Korrelationskoeffizient von Spearman, um den monotonen Zusammenhang zwischen den Variablen zu bestimmen. Die folgende Definition stammt aus dem Lehrbuch „Statistik - Der Weg zur Datenanalyse “ (vgl. Fahrmeir et al., 2016, S. 133ff.).
Beim Spearman Korrelationskoeffizient werden die Realisierungen der Variablen x_i und y_i mit ihrem Rang $rg(x_i)$ und $rg(y_i)$ nach Größe versehen. Der Korrelationskoeffizient ergibt sich aus

$$r_{SP} = \frac{\Sigma(rg(x_i) - r\bar{g}_X)(rg(y_i) - r\bar{g}_Y)}{\sqrt{\Sigma(rg(x_i) - r\bar{g}_X)^2 \Sigma(rg(y_i) - r\bar{g}_Y)^2}}.$$

Der Wertebereich des Spearman Korrelationskoeffizent liegt zwischen $-1 \leq r_{SP} \leq 1$.

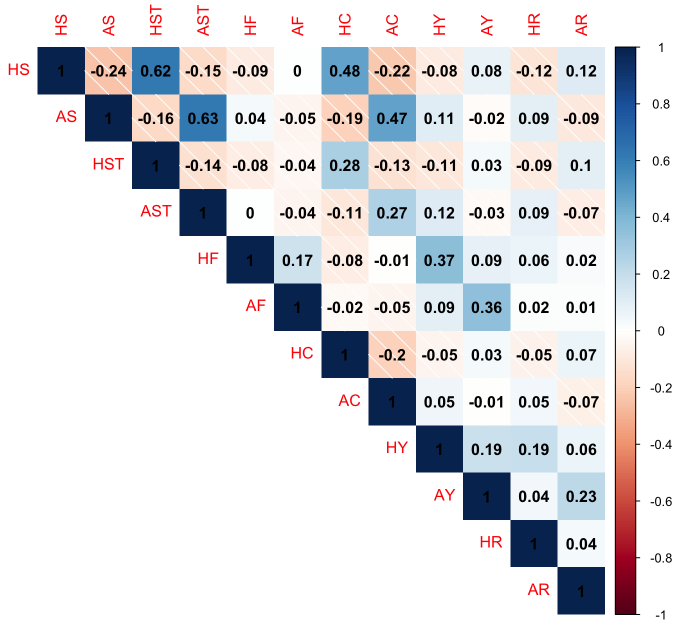

Abbildung 2.4: Spearman Korrelation zwischen den Variablen.

In Abbildung 2.4 ist zu erkennen, dass der stärkste monotone Zusammenhang zwischen AS und AST mit einem Korrelationskoeffizienten von 0,63 besteht. HS und HST besitzen einen minimal kleineren Zusammenhang mit einem Korrelationskoeffizienten von 0,62. Diese Beobachtung überrascht nicht, da HST und AST jeweils abhängig von HS und AS sind, da es ohne einen Torschuss keinen Schuss auf das Tor geben kann.

Einen schwächeren positiven Zusammenhang besitzen die Variablen HS und HC mit einem Wert von 0,48, sowie AS und AC mit einem Korrelationskoeffizienten von 0,47. Auch hier ist ein Zusammenhang nachvollziehbar, da bei mehr Torschüssen mehr Ecken wahrscheinlich sind.

Die letzten zu nennenden Zusammenhänge liegen bei den Variablen HF und HY mit Korrelationskoeffizient 0,37 und AF und AY mit Wert 0,36. Wie bei den bisher genannten Beziehungen mit positivem Zusammenhang, kann auch hier nachvollzogen werden, dass bei einer größeren Anzahl an Fouls, eine größere Anzahl an gelben Karten vorliegt.

Insgesamt ist zu erwähnen, dass die Korrelationen konsistent sind, da die Variablen mit stärkerem Zusammenhang sowohl bei Heim-, als auch bei der Auswärtsmannschaft, vorliegen und der Korrelationskoeffizient jeweils einen ähnlichen Wert annimmt.

Um eventuelle Unterschiede zwischen den beiden Beobachtungsperioden festzustellen wird zunächst ein Blick auf einige Schlüsselwerte aus den Daten geworfen. Hierbei handelt es sich jeweils um das arithmetische Mittel der einzelnen Variablen.

	Beobachtungszeitraum	
Variable	**I**	**II**
HWR	0,431	0,424
FTHG	1,634	1,634
FTAG	1,321	1,372
FTGT	2,955	3,006
HS	14,448	13,049
AS	12,141	11,799
HST	4,991	4,934
AST	4,275	4,334
HF	12,457	12,637
AF	13,167	12,773
HC	5,425	4,996
AC	4,569	4,595
HY	1,747	1,934
AY	2,103	1,979
HR	0,069	0,063
AR	0,099	0,097

Tabelle 2.2: Vergleich der Spieldaten vor Pandemie und während der Pandemie

Beobachtungszeitraum I bezeichnet das Intervall vom ersten Spiel im Datensatz vom 28.07.2017 bis hin zum letzten regulären Spiel mit Zuschauern am 09.03.2020. Beobachtungszeitraum II beschreibt den Start der Geisterspiele am 11.03.2020 bis hin zum letzten Spiel der Saison 2020/2021. In Tabelle 2.2 ist erkennbar, dass fast alle Variablen im Vergleich des Durchschnitts zwischen der Vor-Corona Zeit und der Corona Zeit sehr ähnlich sind. Der größte Unterschied in den Daten ist hier bei den Schüssen der Heimmannschaft, mit einer Differenz von durchschnittlich 1,399 Schüssen weniger pro Spiel. Alle weiteren Spieldaten unterscheiden sich geringfügig.

Um die Unterschiede genauer zu untersuchen, wird nun nicht nur in die Zeit vor der Pandemie und die Zeit während der Pandemie unterteilt, sondern es wird ein gesonderter Blick auf die Zeit nach dem Lockdown geworfen.
Die Saison der Bundesliga und 2. Bundesliga wurde nach dem 25. Spieltag der Saison 2019/2020 unterbrochen. Nach zweimonatiger Spielpause, startete der Spielbetrieb nach strikter Einhaltung des DFL Hygienekonzeptes, auf welches in Kapitel 3 genauer eingegangen wird. Diese sehr ungewohnte Zustand für Verantwortliche und Spieler könnte potentiell stärkeren Einfluss auf den Sport gehabt haben. Um dies zu untersuchen, wird nun in drei Beobachtungszeiträume unterteilt. Zeitraum I bezeichnet wie zuvor die Spanne vor der Pandemie. Zeitraum II.1 beschreibt die neun restlichen Spieltage der Saison 2019/2020 nach dem Lockdown und Zeitraum II.2 beinhaltet die Saison 2020/2021, in welcher Corona-Maßnahmen, wie Geisterspiele, weiterhin griffen.

	Beobachtungszeitraum		
Variable	**I**	**II.1**	**II.2**
HWR	0,431	0,378	0,436
FTHG	1,634	1,549	1,657
FTAG	1,321	1,482	1,343
FTGT	2.955	3,030	3,000
HS	14,448	12,134	13,038
AS	12,141	11,799	11,709
HST	4,991	4,841	4,959
AST	4,275	4,238	4,359
HF	12,457	13,207	12,484
AF	13,167	12,659	12,804
HC	5,425	5,226	4,935
AC	4,569	4,537	4,611
HY	1,747	2,201	1,863
AY	2,103	1,976	1,980
HR	0,069	0,073	0,060
AR	0,099	0,104	0,095

Tabelle 2.3: Vergleich der Spieldaten vor Pandemie, nach dem Lockdown bis zum Saisonende 2019/2020 und der Saison 2020/2021

Nach der Unterteilung in drei Zeitintervalle sind größere Unterschiede festzustellen, als bei der vorherigen Unterscheidung in zwei Abschnitten. Der größte Unterschied liegt in der Heimsiegquote. Während vor der Pandemie und in der Saison 2020/2021 die Heimsiegquote bei circa 43 % liegt, haben Heimmannschaften an den letzten neun Spieltagen der Saison 2019/2020, nach dem Lockdown, nur circa 38 % der Spiele gewonnen. Grafisch kann dies auch in Abbildung 2.5 erkannt werden.

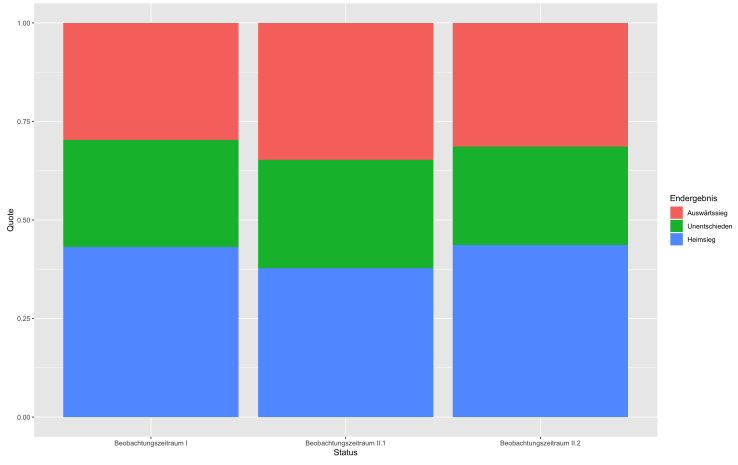

Abbildung 2.5: Gestapeltes Säulendiagramm über die Endergebnisse nach Unterteilung in drei Zeitintervalle.

Zu beachten ist, dass die Verhältnisse der Ausprägungen vor der Pandemie und in der Saison 2020/2021 sehr ähnlich sind. Ebenfalls ist das Auftreten von Spielen, welche Unentschieden enden, bei allen drei Zeitintervallen ähnlich. Der größte Unterschied liegt im Zeitintervall nach dem Lockdown bis zum Saisonende 2019/2020 vor, da es weniger Heimsiege gab, dafür aber mehr Auswärtssiege.

Wenn man einen Blick auf Toranzahl pro Spiel in Abbildung 2.6 wirft, ist zu erkennen, dass der Median jeweils bei drei Toren pro Spiel liegt. Ebenfalls sind bei allen drei Beobachtungsintervallen das obere und das untere Quartil identisch mit zwei und vier Toren pro Spiel. Die einzige Unterscheidung, welche aus dem Boxplot festzustellen ist, sind die unterschiedlichen Werte für Ausreisser. Die Maxima für die Zeit vor der Pandemie, nach dem Lockdown und für die Saison 2020/2021 liegen jeweils bei neun, acht und elf Toren in einem Spiel.

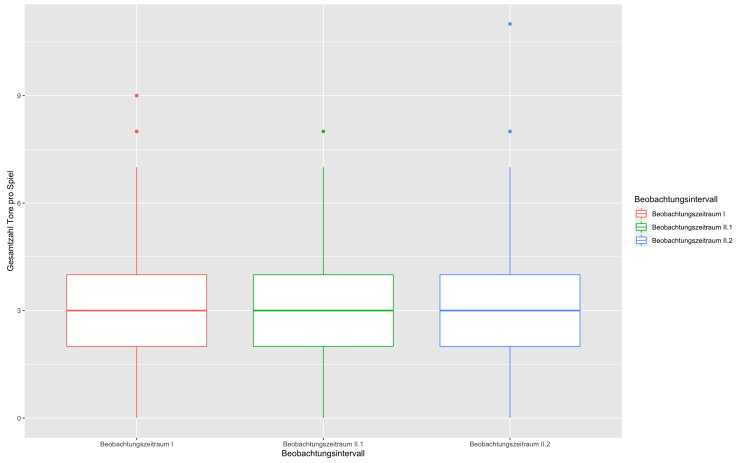

Abbildung 2.6: Boxplot über die Tore pro Spiel nach Unterteilung in drei Zeitintervalle.

Bevor die Grundlagen für die statistische Modellierung der Daten geschaffen wird, ist es wichtig zunächst das Hygienekonzept der DFL vorzustellen und die Historie der Covid-19 Pandemie in Deutschland mit Bezug auf Auswirkungen in Spielen der Bundesliga und 2. Bundesliga zu untersuchen um den deskriptiven Teil der Arbeit abzuschließen. Die explorative Datenanalyse lässt vermuten, dass der Zeitraum vom Spiel am 11.03.2020 bis zum Saisonende 2019/2020 mit besonderem Interesse betrachtet werden sollte. Somit stellt diese Zeitspanne den Hauptaspekt der folgenden Corona-Historie dar.

3 Corona Historie mit Bezug auf die Bundesliga

Bevor auf die Chronik der Corona Pandemie in Deutschland eingegangen wird, soll zunächst das Hygienekonzept der DFL vorgestellt werden. Daraus resultierten die entsprechenden Einschränkungen der Mannschaften und Spieler nach dem Lockdown.

3.1 DFL Hygienekonzept

Das DFL Hygienekonzept entstand durch die so genannte „Task Force Sportmedizin/Sonderspielbetrieb im Profifußball“. Das vorliegende Konzept, welches auf der Internetseite der DFL veröffentlicht wurde, ist die zweite Version des Konzepts vom 01.05.2020.

Die Zielsetzung des Hygienekonzeptes war es, den Spielbetrieb für die ersten drei Ligen der Herren (Bundesliga, 2. Bundesliga und 3. Liga), sowie die Damen Bundesliga und die DFL Pokalwettbewerbe wieder aufzunehmen und in voller Länge zu beenden. Es sollte kein hundertprozentiger Schutz vor der Pandemie gewährleistet werden, da dies nicht möglich sei. Vielmehr sollte aufgrund der wirtschaftlichen und gesellschaftlichen Bedeutung des Fußballs ein verantwortbares Risiko eingegangen werden. Es gab lediglich die Prämisse, dass zur Wiederaufnahme des Spielbetriebs keinen Ressourcenkonflikt zwischen Sport und Gesellschaft geben dürfe. Es wurde in diesem Sinne vorgegeben, dass jeder Spieler mindestens zweimal pro Woche per PCR Test auf eine Infektion mit Covid-19 getestet wird (vgl. DFL, 2020d, S. 3f.).

Für Spieltage wurde eine Zonierung der Stadien in drei Zonen vereinbart. Die erste Zone stellt den Innenraum des Stadions dar, welcher nur von für den Spielbetrieb notwendigen Personen aufgesucht werden durfte. Dazu gehörten unter anderem Spieler, Schiedsrichter, Funktionäre, Sanitäter und Ordner. Die zweite Zone bestand aus den Tribünen des Stadions, auf denen sich beispielsweise Medienvertreter aufhalten durften. Die dritte Zone war das umliegende Stadiongelände bis zur Umfriedung des Geländes. In jeder Zone durften sich maximal 100 Personen aufhalten, somit war es zulässig maximal 300 Personen an einem Spieltag im Stadion unterzubringen. Explizit wurde genannt, dass keine Fans auf dem Stadiongelände zulässig sind (vgl. DFL, 2020d, S. 10ff.).

Einschränkungen für die Spieler am Spieltag waren ebenfalls umfassend. Außerhalb des Spielfelds musste stets ein Mindestabstand von 1,5 Metern eingehalten werden. Dies galt auch für die Anreise zum Stadion. Es wurde empfohlen auf mehrere Busse oder Einzelanreise umzusteigen. Es musste stets ein Mund-Nasenschutz getragen werden, sowie sollten Mannschaften und Spieler so gut wie möglich räumlich und zeitlich voneinander getrennt werden. In der Kabine der Mannschaften, durfte sich jeweils nur die Startelf aufhalten. Ergänzungsspieler und weiteres Personal war in anderen Räumlichkeiten unterzubringen. Jeder Spieler sollte die minimale nötige Zeit in der Kabine verbringen, welche auf 30-40 Minuten geschätzt wurde. Auf den Auswechselbänken mussten Plätze freigelassen werden

um den Mindestabstand einzuhalten. Interviews und Pressekonferenzen sollten lediglich virtuell stattfinden (vgl. DFL, 2020d, S. 25f.).

Vor Wiederaufnahme des Trainingsbetriebs mussten mindestens zwei Testungen aller Beteiligten durchgeführt werden. Alle Räumlichkeiten und Trainingsmaterialien mussten regelmäßig desinfiziert werden. Das Training durfte lediglich unter Ausschluss der Öffentlichkeit stattfinden und es musste abseits des Platzes stets ein Mindestabstand von zwei Metern gewährleistet werden. Mannschaftsräume durften nur wenn zwingend nötig genutzt werden, gemeinsames Speisen der Spieler wurde untersagt. Vor der Wiederaufnahme des Spielbetriebs war ein siebentägiges Trainingslager unter Quarantäne vorgeschrieben (vgl. DFL, 2020d, S. 38f.).

Auch im privatem Umfeld wurden für Spieler Einschränkungen bestimmt. Es durfte kein Kontakt zur Öffentlichkeit oder der Nachbarschaft aufgenommen werden. Die Spieler sollten sich möglichst nur in der eigenen Unterkunft aufhalten. Es durfte kein Besuch empfangen werden und öffentliche Verkehrsmittel wurden den Spielern untersagt. Diese Regelungen umschlossen den gesamten Haushalt des jeweiligen Spielers. Einkäufe durften nur von anderen Personen im Haushalt durchgeführt werden und sollten auf ein Minimum beschränkt werden (vgl. DFL, 2020d, S. 43).

3.2 Corona Historie in Deutschland mit Bezug auf den Fußball

Die folgende Chronik soll den Verlauf der Corona Pandemie in Deutschland mit Querverweisen auf die Fußball Bundesliga und 2. Bundesliga aufführen. Im Speziellen wird der Zeitraum vom ersten Auftreten der Krankheit in Deutschland bis zum Ende der Saison 2019/2020 beleuchtet, da die explorative Datenanalyse nahelegt, dass der sportliche Aspekt die größte Beeinflussung in dieser Phase genommen hat. Die Aufführung der pandemiespezifischen Daten und Ereignisse stammt von der Internetseite des Bundesministerium für Gesundheit (vgl. BMG, 2021).

17. Januar 2020: Die Fußball Bundesliga startet mit dem 18. Spieltag in die Rückrunde der Saison 2019/2020. Vor der ersten Infektion mit Covid-19 in Deutschland und noch mit vollen Stadien.

27. Januar 2020: Der erste Corona-Fall ist in Deutschland aufgetaucht. Ein Mann aus dem Lankreis Starnberg hat sich infiziert und wurde isoliert. Das Risiko für eine Ausbreitung in Deutschland wird zu dieser Zeit als gering eingestuft.

12. Februar 2020: Der Coronavirus hat inzwischen weltweite Auswirkungen. Es gibt zahlreiche Fälle in vielen Ländern der Welt. In Deutschland sind zu diesem Zeitpunkt 16 Fälle bekannt, die alle in Isolation befindlich waren.

26. Februar 2020: Es tauchen die ersten Infektionen mit dem Coronavirus in Baden-Württemberg und Nordrhein-Westfalen auf. Die Bundesliga und 2. Bundesliga haben jeweils den 23. Spieltag beendet. Weiterhin mit voller Stadionkapazität.

10. März 2020: Der gemeinsame Krisenstab des Bundesministerium des Inneren und des Bundesministerium für Gesundheit empfiehlt, alle Großveranstaltungen mit mehr als 1000 Zuschauern abzusagen. Daraufhin entscheiden die Stadt Mönchengladbach, dass das Spiel zwischen Borussia Mönchengladbach und dem 1. FC Köln am Folgetag nur ohne Zuschauer ausgetragen werden darf (vgl. Köln, 2020).

11. März 2020: Das Spiel Borussia Mönchengladbach gegen den 1. FC Köln ist das erste Spiel ohne Zuschauer, bedingt durch den Coronavirus. Unter Betracht der Spieldaten kann nun ein Vergleich angestellt werden.
Mönchengladbach gewinnt das Spiel 2:1 gegen Köln. Drei Tore in dem Spiel liegt genau am Durchschnitt. Zwei Heimtore und ein Auswärtstor liegt am Trend der Daten. Mit 12 Torschüssen und fünf Schüssen auf das Tor liegt Gladbach in diesem Spiel jeweils knapp unter und knapp über dem Durchschnitt aller Spiele. Der FC Köln liegt mit 16 Torschüssen, jedoch nur drei Schüsse auf das Tor überdurchschnittlich und etwas unter dem Mittel. Auch bei den Fouls bewegen sich beide Mannschaften im Bereich des Erwarteten. Im Hinblick auf Ecken mit einer Anzahl von sechs und drei, liegen die Werte erneut knapp über und knapp unter dem Mittelwert der jeweiligen Variable.
Der Tabellenvierte spielte gegen den Tabellenzehnten, die Heimmannschaft gewann und alle Spieldaten befanden sich nahe am jeweiligen Mittelwert. Anhand von den einzelnen Spieldaten kann daher im ersten pandemiebedingtem Geisterspiel kein relevanter Unterschied festgestellt werden.

22. März 2020: Der zuvor angekündigte erste Lockdown tritt in Kraft. Das Treffen von Bekannten soll auf ein Minimum beschränkt werden, während Gastronomien und Dienstleistungsbetriebe geschlossen werden. Ursprünglich sind die Maßnahmen für mindestens zwei Wochen geplant (vgl. Bundesregierung, 2020). Der deutsche Profifußball muss seine Saison entsprechend unterbrechen.

09. April 2020: Auch über die Osterfeiertage sollen Kontaktbeschränkungen eingehalten werden. Somit kann auch noch nicht an die Wiederaufnahme des Spielbetriebs in den Bundesligen gedacht werden.

15. April 2020: Die Bundesregierung verlängert den Lockdown bis mindestens 03.05.2020.

01. Mai 2020: Das in Kapitel 3.1 vorgestellte Hygienekonzept der DFL wird vorgestellt. Das Ziel ist den Spielbetrieb zwischen Mai und Juni 2020 wieder aufzunehmen.

06. Mai 2020: Die Bundesregierung ermöglicht es der DFL den Spielbetrieb in der zweiten Maihälfte, unter Einhaltung des Hygienekonzeptes, wieder aufzunehmen (vgl. DFL, 2020b). Auch für die Bevölkerung gibt es schrittweise Erleichterungen des Lockdowns ab Mai.

16. Mai 2020: Die Bundesliga und die 2. Bundesliga setzen nach circa zwei Monaten Pause die Saison fort, um die letzten neun Spieltage auszutragen. Zweitligist Dynamo Dresden kann nicht zum Termin der Saisonfortsetzung antreten, da zur Trainingsrückkehr mehrere Coronainfektionen auftraten und die gesamte Mannschaft und der Trainerstab sich in Quarantäne begeben mussten (vgl. faz.net, 2020b).

Der Spezialfall von Dynamo Dresden

Dynamo Dresden war zur Unterbrechung der Saison 2019/2020 Tabellenletzter der 2. Bundesliga mit vier Punkten Abstand zu einem Nichtabstiegsplatz.
Als es zur Wiederaufnahme des Spielbetriebs kam und Dynamo Dresden mit positiven Coronafällen nicht pünktlich starten konnte, ergab sich hieraus ein Nachteil für die Sachsen. Während die unbeteiligten Mannschaften vom 16.05.2020 bis zum 28.06.2020 die verbleibenden neun Spiele austrugen, konnte Dynamo Dresden erst am 31.05.2020 starten. Daher mussten die Dresdener bis hin zum vorletzten Spiel am 21.06.2020 acht Spiele im dreitages-Rhytmus absolvieren. Sie holten nur acht von 27 möglichen Punkten und stiegen als Tabellenletzter ab.
Die Frage stellt sich, ob ein Zusammenhang mit den Einschränkungen durch die Coronainfektionen und dem sportlichen Misserfolg zu erkennen ist oder ob der Abstieg auch ohne diese wahrscheinlich war. In den letzten neun Spielen holte Dynamo Dresden im Schnitt 0,889 Punkte pro Spiel. Vor der Unterbrechung hatten sie 24 Punkte in 25 Spielen erreicht, was einem Schnitt von 0,960 Punkten pro Spiel entspricht. Der Abstand zu den Nichtabstiegsplätzen betrug vier Punkte. Gegen direkte Konkurrenten spielte Dynamo Dresden an den letzten neun Spieltagen nur ein Mal. Dieses Spiel gegen Wehen-Wiesbaden gewannen die Dresdener. Dresden erhielt an den letzten neun Spielen 17 Gegentreffer, während sie sieben Tore erzielten. Dies entspricht einer Tordifferenz von -10 in neun Spielen und dadurch einer Differenz von -1,111 Toren pro Spiel. Vor dem Lockdown war die Tordifferenz bei -16 nach 25 Spielen. Pro Spiel betrug die Differenz somit -0,64.
Zusammenfassend kann nicht klar festgestellt werden, dass die Benachteiligung durch die Coronainfektionen schuldig waren am sportlichen Misserfolg. Sie haben es Dynamo Dresden mindestens erschwert die Klasse zu halten, anhand der vorherigen Situation war ein Abstieg ohne diese Umstände ebenfalls wahrscheinlich.

Die deskriptive Analyse der Daten zeigt, dass generelle Unterschiede zwischen der Zeit vor der Corona Pandemie und der Zeit während der Pandemie schwer festzustellen sind. Größere Unterschiede in einzelnen Aspekten können zwischen der Zeit vor der Pandemie und der Anfangszeit der Pandemie, bis zum Saisonende 2019/2020 erkannt werden. Für die

Saison 2020/2021 gleichen sich die gemessenen Werte wieder an die Zeit vor der Pandemie an, obwohl noch Einschränkungen vorhanden sind.

Im nächsten Kapitel werden die theoretischen Hintergründe für die statistische Modellierung der Daten vorgestellt. Die Modellierung der Daten in Kapitel 5 soll allgemeine Fragestellungen zum Datensatz beantworten.

4 Modelltheorie

In Kapitel 5 sollen statistische Modelle gefunden werden um die vorliegenden Daten zu beschreiben. Dafür wird zunächst auf die Theorie der verwendeten Modelle eingegangen. Im Speziellen wird ein Blick auf logistische Regressionsmodelle, generalisierte additive Modelle und kumulative Logit Modelle geworfen. Ebenfalls werden Gütekriterien vorgestellt, welche zur qualitativen Überprüfung und zur Auswahl des jeweils besten Modells verwendet werden.

4.1 Logit Modell

Für eine binäre Zielgröße y, bei welcher y_i definiert ist als

$$y_i = \begin{cases} 1 \\ 0 \end{cases} , \forall i \in \{1, ..., n\}$$

kann zur Modellierung der Regression kein lineares Regressionsmodell genutzt werden.
In einem solchen Fall kann auf ein logistische Regressionsmodell zurückgegriffen werden. Die folgende Definition des Logit-Modells stammt aus dem Lehrbuch „Regression“ (vgl. Fahrmeir et al., 2007, S. 34f.).

Der Erwartungswert von y ergibt sich aus

$$E(y) = P(y = 0) \cdot 0 + P(y = 1) \cdot 1 = P(y = 1).$$

Das Ziel der Regressionsanalyse bei binären Zielvariablen ist es, den Erwartungswert $E(y)$, beziehungsweise die Wahrscheinlichkeit

$$P(y = 1) = P(y = 1|x_1, ..., x_k) = \pi$$

in Anwesenheit von Kovariablen zu modellieren. Probleme des linearen Regressionsmodells für binäre Zielvariablen wie, dass das lineare Modell Werte $\pi_i < 0$ und $\pi_i > 1$ akzeptiert, können umgangen werden unter Annahme, dass das Modell

$$\pi_i = P(y_i = 1) = F(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik})$$

eine beschränkte Definitionsmenge auf das Intervall $[0, 1]$ besitzt. Ebenfalls ist es für die Interpretierbarkeit des Modells sinnvoll, dass für F Funktionen gewählt werden, welche monoton steigend sind. Somit sind Verteilungsfunktionen eine natürliche Wahl für F. Mit der logistischen Verteilungsfunktion

$$F(\eta) = \frac{exp(\eta)}{1 + exp(\eta)}$$

ergibt sich das Logit Modell

$$\pi_i = P(y_i = 1) = \frac{exp(\eta_i)}{1 + exp(\eta_i)},$$

mit dem linearen Prädiktor

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}.$$

Wie beim linearen Regressionsmodell, sind die binären Zielvariablen (bedingt) unabhängig, gegeben den Kovariablen $x_i = (x_{i1}, ..., x_{ik})'$. Auch wenn der lineare Prädiktor linear ist, unterscheidet sich die Interpretation von der des linearen Modells: Wenn sich der Wert des linearen Prädiktors von η zu $\eta + 1$ erhöht, steigt die Wahrscheinlichkeit von $y = 1$ nichtlinear von $F(\eta)$ zu $F(\eta + 1)$. Alternativ kann durch auflösen der Gleichung des Logit Modells nach η, mit der inversen Funktion

$$\eta = log(\frac{\pi_i}{1 - \pi_i}),$$

vorgegangen werden. Dadurch ergibt sich

$$log(\frac{\pi_i}{1 - \pi_i}) = log(\frac{P(y_i = 1)}{1 - P(y_i = 1)}) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}$$

oder alternativ, da $exp(a + b) = exp(a) \cdot exp(b)$, die Darstellung

$$\frac{\pi_i}{1 - \pi_i} = \frac{P(y_i = 1)}{P(y_i = 0)} = exp(\beta_0) exp(\beta_1 x_{i1}) \cdot \ldots \cdot exp(\beta_k x_{ik}).$$

Der Quotient aus der Wahrscheinlichkeit für $y = 1$ und der Wahrscheinlichkeit für $y = 0$ bezeichnet die Odds. Entsprechend werden die logarithmierten Odds aus der Lösung der Logit Modellgleichung als log Odds bezeichnet. Wenn sich somit der Wert von der Kovariable x_{i1} um eine Einheit erhöht, führt dies zu einer Veränderung der Odds um den Faktor $exp(\beta_1)$. Ist $\beta_1 > 0$ erhöhen sich die Odds, für $\beta_1 = 0$ gibt es keine Veränderung und bei $\beta_1 < 0$ sinken die Odds.

4.2 Kumulatives Logit Modell

Wie bereits beim Logit Modell, stammt die Definition des kumulativen Modell aus dem Lehrbuch „Regression“ (vgl. Fahrmeir et al., 2007, S. 334ff.). Das kumulative Modell lässt sich herleiten über die Annahme einer latenten (unbeobachteten) Variable, welche die Entscheidungen der beobachteten Alternativen treibt. Das Übertreffen eines gewissen Schwellenwerts auf der latenten Skala resultiert in einer beobachtbaren Kategorie der Zielvariable.
Die unbeobachtete latente Variable u wird entsprechend des Wertes der Kovariable x_i (ausgeschlossen β_0) angenommen als

$$u_i = -x_i'\beta + \epsilon_i,$$

mit β als Parametervektor und ϵ_i als Fehlervariable mit kumulativer Verteilungsfunktion F. Die Verbindung zwischen beobachteter und latenter Variable wird beschrieben durch den Schwellenwert Mechanismus

$$Y_i = r \iff \theta_{r-1} < u_i \leq \theta_r,$$

mit $r = 1, ..., c+1$ und $-\infty = \theta_0 < \theta_1 < \ldots < \theta_{c+1} = \infty$ als latente geordnete Schwellenwerte platziert auf dem latenten Kontinuum. β_0 muss aus dem Prädiktor $x_i\beta$ gehalten werden, um ein identifizierbares Modell zu erhalten, da sonst bei Verschiebung des Achsenabschnitts und eine entsprechende Anpassung der Schwellenwerte ein äquivalentes Modell entstehen könnte.

Kategorie r ist beobachtet, falls die latente Variable zwischen den Schwellenwerten θ_{r-1} und θ_r liegt. Als binäre Entscheidung interpretiert, erhält man eine Zielvariable $Y_i \leq r$ falls $u_i \leq \theta_r$, das heißt falls die latente Variable unter dem Schwellenwert θ_r liegt, und $Y_i > r$ falls $u_i > \theta_r$.

Aus den bisherigen Annahmen von u_i und Y_i erfolgt das kumulative Modell mit Verteilungsfunktion F durch

$$\begin{aligned} P(Y_i \leq r) &= P(u_i \leq \theta_r) \\ &= P(-x_i'\beta + \epsilon_i \leq \theta_r) \\ &= P(\epsilon_i \leq \theta_r + x_i'\beta) \\ &= F(\theta_r + x_i'\beta) \qquad \text{mit } r = 1, ..., c+1. \end{aligned}$$

Der Name des kumulativen Modells bezieht sich auf die Spezifikation von kumulativen Wahrscheinlichkeiten $P(Y_i \leq r) = P(Y_i = 1) + \ldots + P(Y_i = r)$ auf der linken Seite der Modellgleichung. Das Modell selbst beinhaltet die latente Variable nicht mehr und kann als Regressionsmodell mit Regressoren x_i und Parametern $\theta_1, ..., \theta_c$ und β betrachtet werden. Die Auftretenswahrscheinlichkeiten sind gegeben durch

$$P(Y_i = r) = F(\theta_r + x_i'\beta) - F(\theta_{r-1} + x_i'\beta) \quad \text{mit } r = 1, ..., c+1.$$

Durch die Wahl der logistischen Verteilungsfunktion F erhält man das kumulative Logit Modell mit

$$P(Y_i \leq r) = \frac{exp(\theta_r + x_i'\beta)}{1 + exp(\theta_r + x_i'\beta}$$

oder äquivalent

$$log\frac{P(Y_i \leq r)}{P(Y_i > r)} = \theta_r + x_i'\beta.$$

Das kumulative Logit Modell wird auch Proportional Odds Modell genannt. Proportional bezieht sich darauf, dass das Verhältnis der kumulativen Odds für Teilpopulationen

charakterisiert durch x_i und $\tilde{x_i}$ gegeben ist durch

$$\frac{P(Y_i \leq r|x_i)/P(Y_i > r|x_i)}{P(Y_i \leq r|\tilde{x_i})/P(Y_i > r|\tilde{x_i})} = exp((x_i - \tilde{x_i})'\beta).$$

Das Verhältnis ist unabhängig von Kategorie r und somit sind die kumulativen Odds proportional über alle Kategorien.

4.3 Generalisierte Additive Modell (GAM)

Das generalisierte additive Modell stellt die nichtparametrische Erweiterung des generalisierten linearen Modells dar. Es erweitert

$$g(\mu(x_i)) = \eta_i = \beta^\top x_i = \beta_1 x_{i1} + \ldots + \beta_k x_{ik}$$

zu

$$g(\mu(x_i)) = \eta_i = \beta_1 + f_2(x_{i2}) + \ldots + f_k(x_{ik})$$

als Summe von geglätteten Funktionen der Kovariablen (vgl. Yee, 2015, S. 81).

Zur Glättung können verschiedene Methoden verwendet werden. Ein Beispiel hierfür, welches in Kapitel 5 eingesetzt wird, sind penalisierte Splines (P-Splines). Diese sollen die Funktion f_z per Polynomial Splines mit einer großzügigen Anzahl an Knoten annähern. Somit sollen auch komplexe Funktionen f_z dargestellt werden können. Zusätzlich wird ein Strafterm eingeführt, welcher overfitting verhindert und das penalisierte kleinste Quadrate Kriterium (PLS) minimiert. (vgl. Fahrmeir et al., 2007, S. 431f.). Das PLS ist definiert als

$$PLS(\beta) = (y - X\beta)'(y - X\beta) + \lambda \cdot pen(\beta) \rightarrow \min_{\beta}$$

mit $pen(\beta)$ als Strafterm und λ als Glättungsparameter (vgl. Fahrmeir et al., 2007, S. 202).

4.4 Gütekriterien

Zur Modellauswahl können verschiedene Gütekriterien verwendet werden, um das bestgeeignetste Modell zu identifizieren. Im Folgenden werden das akaiken Informationskriterium (AIC) und das bayesianische Informationskriterium (BIC) näher erläutert.

4.4.1 Akaike Informationskriterium (AIC)

Das AIC stellt eine Maßzahl dar, in welcher die Log-Likehood ($log(L_M)$) des Modells negativ und die Anzahl der Parameter positiv einfließt. Es wird dargestellt als

$$AIC(M) = -2 \cdot log(L_M) + 2 \cdot dim(M).$$

Die Log-Likelihood eines Modells ist umso größer, je besser das Modell die abhängige Variable anhand der verwendeten Kovariablen erklärt. Der zweite Summand in der Formel

stellt die verdoppelte Anzahl der verwendeten Kovariablen dar. Wenn mehrere Modellkandidaten mit dem AIC verglichen werden, so entscheidet man sich für das Modell, welches das AIC minimiert.
Die Entscheidung anhand des AIC stellt somit einen Trade-off zwischen einem möglichst hohen Erklärungswert der Zielvariable und der Komplexität des Modells dar (vgl. Emmert-Streib and Dehmer, 2019, S. 16f.).

Für Modellselektion anhand des AIC eignet sich die `R`-Funktion `stepAIC`, welche Modelle anhand deren AIC vergleicht und die Modellgleichung schrittweise von einem Nullmodell aus um die Kovariable erweitert, welche den AIC minimiert. Wenn man in `stepAIC` die Richtung der Schritte auf `'both'` setzt so wird in jedem Schritt die Kovariable zur Modellgleichung hinzugefügt oder entfernt, je nachdem welche Änderung der Modellgleichung den kleinsten Wert des AIC zur Folge hat. Falls weder durch Hinzufügen, noch durch Entfernen einer Kovariable eine Verringerung des AIC erreicht werden kann, so stellt das Modell mit der Modellgleichung zu diesem Zeitpunkt das laut dieser `R`-Funktion beste Modell dar.

4.4.2 Bayesianisches Informationskriterium (BIC)

Das BIC ähnelt in der Form dem AIC. Auch beim BIC wird das Modell mit dem geringeren BIC Wert als besser angesehen. Das BIC kann allgemein dargestellt werden als

$$BIC(M) = -2 \cdot log(L_M) + p \cdot log(n).$$

Der hauptsächliche Unterschied zwischen AIC und BIC ist, dass komplexere Modelle beim BIC über den Penalisierungsterm stärker bestraft werden, als beim AIC. Somit bevorzugt das BIC einfachere Modelle (vgl. Fahrmeir et al., 2007, S. 149f.). Ebenfalls geht die Stichprobengröße n nicht in den AIC ein, während der BIC diesen logarithmiert und mit dem Penalisierungsterm multipliziert.
Die Grundphilosophien des AIC und BIC unterscheiden sich laut der Publikation von Kuha. Während das BIC das Modell, welches die höchste Wahrscheinlichkeit hat, das wahre Modell zu sein, finden möchte, gehe das AIC davon aus, dass kein wahres Modell gefunden werden kann (vgl. Kuha, 2004, S. 216f.).

5 Statistische Modellierung der Daten

Zur statistische Modellierung der Daten werden die, in Kapitel 4 vorgestellten Modelle verwendet. Es sollen zwei Fragestellungen mit zwei Zielvariablen untersucht werden:

- Kann ein Modell anhand von Spieldaten korrekt beurteilen, ob das Spiel mit oder ohne Zuschauern stattgefunden hat?
- Kann ein Modell anhand von Spieldaten den Ausgang des Spiels (Heimsieg, Unentschieden, Auswärtssieg) korrekt bestimmen?

Für die erste Fragestellung werden jeweils zwei konventionelle Logit Modelle und generalisierte additive Modelle mit geglätteten Effekten vorgestellt und auf Eignung getestet. Da die Zielvariable bei der zweiten Fragestellung ordinal mit drei Ausprägungen ist, wird hier das kumulative Logit Modell verwendet.

5.1 Logit Modell

Die dichotome Zielgröße y ist in diesem Modell definiert als

$$y_i = \begin{cases} 1, & \text{falls Spiel mit Zuschauern} \\ 0, & \text{falls Spiel Geisterspiel} \end{cases}, \quad \forall i \in \{1, ..., n\}.$$

Die folgende Tabelle zeigt die Modellformel und die Werte der Gütekriterien des AIC und BIC.

Modell	Modelltyp	Modellformel	AIC	BIC
I	GLM binomial	$Zuschauer \sim Avg_H + Avg_A + FTHG + FTAG + HS + AS + HST + AST + HC + AC + HF + AF + HY + AY + HR + AR + (HS * HST) + (AS * AST)$	2977	3087
II	GLM binomial	$Zuschauer \sim Avg_H + HS + AS + HST + AST + AF + HY + AY + HR$	2963	3021

Tabelle 5.1: Vergleich der GLM Logit Modelle nach AIC und BIC

Das erste Modell beinhaltet alle gegebenen Einflussvariablen, welche linear in das Modell eingehen. Ebenfalls enthalten sind Interaktionsterme der untereinander korrelierten Variablen HS und HST, sowie AS und AST. Über die `R`-Funktion `stepAIC` wurden die für den AIC wichtigsten Variablen in Modell II behalten. Das zweite Modell beinhaltet für die Einflussvariablen lediglich die durchschnittliche Quote auf Heimsieg, Torschüsse und Schüsse auf das Tor für beide Mannschaften, Fouls der Auswärtsmannschaft, gelbe Karten beider Mannschaften und rote Karten der Heimmannschaft. Beide Interaktionsterme scheinen für den AIC nicht aussagekräftig zu sein.

Nachdem die `stepAIC` Funktion, das Modell nach dem AIC optimiert, ist es klar, dass das zweite Modell einen geringeren AIC Wert besitzt und somit zu bevorzugen ist. Der BIC

Wert ist ebenfalls geringer, was durch die Verringerung der Komplexität nachzuvollziehen ist.

GLM Logit Modell			
	Koeffizenten	Standardabweichung	P-Wert
Intercept	-0,534	0,279	$0,056$
Avg_H	-0,064	0,030	$0,033$
HS	0,106	0,013	$5,5 \cdot 10^{-16}$
AS	0,062	0,013	$2,76 \cdot 10^{-6}$
HST	-0,125	0,024	$1,49 \cdot 10^{-7}$
AST	-0,051	0,025	$0,039$
AF	0,018	0,012	$0,118$
HY	-0,141	0,035	$5,66 \cdot 10^{-5}$
AY	0,054	0,037	$0,145$
HR	0,321	0,182	$0,078$

Tabelle 5.2: Koeffizentenschätzungen, Standardabweichungen und P-Werte des besten GLM Logit Modells

Die Koeffizienten β_i gehen linear in die $log(Odds)$ ein, beziehungsweise als $exp(\beta_i)$ in die Odds. Im Modell sind vor allem die Torschüsse und Schüsse auf das Tor, sowie die Anzahl der gelben Karten, jeweils für die Heimmannschaft, und die Torschüsse der Auswärtsmannschaft, statistisch signifikant. Auch die durchschnittliche Wettquote auf Sieg der Heimmannschaft besitzt einen P-Wert $< 0,05$, ebenso wie die Schüsse auf das Tor von der Gastmannschaft. Den größten positiven Einfluss auf die $log(Odds)$ hat die Anzahl der Torschüsse der Heimmannschaft, da sowohl Standardabweichung als auch P-Wert der rote Karten Anzahl für die Heimmannschaft vergleichsweise hoch ist. Dies lässt vermuten, dass je öfter die Heimmannschaft auf das Tor schiesst, desto wahrscheinlicher schätzt das Modell auf ein Spiel mit Zuschauern, bei festhalten der anderen Einflüsse. Mit Blick auf Tabelle 2.2 aus Kapitel 2.2, kann erkannt werden, dass die Vorzeichen der Koeffizienten plausibel sind.

Um die Genauigkeit der Modellvorhersagen zu überprüfen wurde mit der `R`-Funktion `predict` Werte für das Modell über den kompletten Datensatz von 2448 Spielen geschätzt und anschließend mit den tatsächlichen Ausprägungen der Variable Zuschauer verglichen. Der Schwellenwert für die Wahrscheinlichkeit des Modells wurde bei 0,5 gesetzt.

$Zuschauer_i = 1$	Geschätzt	
Tatsächlich	Falsch	Wahr
Falsch	61	715
Wahr	44	1628

Tabelle 5.3: Konfusionstabelle des ausgewählten GLM Logit Modells

In Tabelle 5.3 ist zu erkennen, dass das Modell weniger als 10 % der Spiele, die ohne

Zuschauer stattfanden, korrekt vorhersagen konnte. Insgesamt beträgt die prozentuale Genauigkeit der richtigen Vorhersagen circa 69 %, dies allerdings nur, da der Großteil der Spiele mit Zuschauern korrekt vorhergesagt wurde und dies auch die meisten Spiele darstellt.

5.2 Generalisiertes Additives Modell (GAM)

Mit dem generalisierten additiven Modell sollen nun die Wettquoten auf Heim- und Auswärtsmannschaft als penalisiertem Spline eingehen.

Modell	Modelltyp	Modellformel	AIC	BIC
I	GAM binomial	$Zuschauer \sim s(Avg_H) + s(Avg_A) + FTHG + FTAG +$ $HS + AS + HST + AST + HC + AC + HF + AF +$ $HY + AY + HR + AR + (HS * HST) + (AS * AST)$	2836	3021
II	GAM binomial	$Zuschauer \sim s(Avg_H) + s(Avg_A) + HS + AS +$ $HST + HY + AY + HR$	2824	2950

Tabelle 5.4: Vergleich der GAM nach AIC und BIC

Das erste GAM in Tabelle 5.4 beinhaltet, wie Modell I der GLM Logit Modelle, alle Einflussvariablen. Der Unterschied zwischen diesen beiden Modellen ist allerdings, dass im GAM die durchschnittlichen Wettquoten auf beide Mannschaften per penalisierter Splines einfließen. Dies verbessert das Modell in Hinblick auf AIC und BIC. Modell II der GAMs wurde manuell im Hinblick auf den AIC verbessert. Zum zweiten GLM Logit Modell unterscheidet sich, neben den geglätteten Termen, dass hier beide Durchschnittsquoten enthalten sind, wohingegen AST und AF nicht mehr im finalen Modell enthalten sind. Modell II des GAM stellt das Modell mit den besten AIC und BIC Werten und somit das Modell das unter den vier gezeigten Modellen gewählt werden sollte.

GAM Logit Modell			
	Koeffizenten	Standardabweichung	P-Wert
Intercept	-0,722	0,232	$0,002$
HS	0,113	0,014	$< 2^{-16}$
AS	0,060	0,011	$5,28 \cdot 10^{-8}$
HST	-0,123	0,025	$5,55 \cdot 10^{-7}$
HY	-0,146	0,036	$6,08 \cdot 10^{-5}$
AY	0,068	0,036	$0,059$
HR	0,377	0,188	$0,045$

Geglättete Terme			
	EDF		P-Wert
$s(Avg_H)$	7,359	-	$< 2^{-16}$
$s(Avg_A)$	7,453	-	$< 2^{-16}$

Tabelle 5.5: Koeffizentenschätzungen, Standardabweichungen und P-Werte des besten GAM Logit Modells

Die Interpretation der Koeffizienten in Tabelle 5.5 funktioniert beim GAM Logit Modell analog zum vorherigen GLM Logit Modell. Die Koeffizienten der Einflussvariablen, die in beiden Modellen enthalten sind, sind sehr ähnlich. Auch die P-Werte befinden sich auf einem ähnlichen Niveau, sind beim GAM tendenziell jedoch etwas kleiner.
An den geschätzten Freiheitsgraden $EDF > 7$ der beiden geglätteten Variablen, ist zu erkennen, dass die beiden Terme nicht linear, sondern deutlich komplexer sind.

$Zuschauer_i = 1$	Geschätzt	
Tatsächlich	Falsch	Wahr
Falsch	173	603
Wahr	137	1535

Tabelle 5.6: Konfusionstabelle des ausgewählten GAM Logit Modells

Die Konfusionstabelle 5.6 zeigt auf, dass das GAM weniger Spiele mit Zuschauern richtig schätzt, als das GLM. Allerdings verbessert dieses Modell Vorhersagen für Spiele ohne Zuschauer. Insgesamt beläuft sich die Genauigkeit des Modells auf circa 70 %. Die Genauigkeit ist ähnlich gut, wie die des GLM, jedoch ist das GAM hier, aufgrund der besseren Werte von AIC und BIC, sowie der verbesserten Vorhersage für Spiele ohne Zuschauer, zu bevorzugen.

5.3 Kumulatives Logit Modell

Für die zweite Fragestellung, ob über die Spieldaten das korrekte Resultat des Spiels ermittelt werden kann, muss ein kumulatives Logit Modell verwendet werden, da die Zielvariable y kategorial ist mit Ausprägungen

$$y_i = \begin{cases} H, & \text{falls Resultat Heimsieg} \\ U, & \text{falls Resultat Unentschieden} \\ A, & \text{falls Resultat Auswärtssieg} \end{cases}, \qquad \forall i \in \{1, ..., n\},$$

mit Ordnung $A < U < H$.

Modell	Modelltyp	Modellformel	AIC	BIC
I	Logit Ordinal	$Resultat \sim Status + Avg_H + Avg_A + FTHG + HS + AS + HST + AST + HC + AC + HF + AF + HY + AY + HR + AR$	3134	3245
II	Logit Ordinal	$Resultat \sim Avg_H + Avg_A + FTHG + HS + AS + HST + AST + HC + AC + HF + HR + AR$	3131	3212

Tabelle 5.7: Vergleich der kumulativen Logit Modelle nach AIC und BIC

In Model I der Tabelle 5.7 befinden sich alle Einflussvariablen, ausser FTAG, welches nicht

in das Modell einfließen konnte, aufgrund von Konvergenzproblemen im Algorithmus der `polr` Funktion in `R`. Status ist eine kategorielle Variable, die die drei Beobachtungszeiträume aus Kapitel 2.2, mit Beobachtungszeitraum II.1 als Referenzkategorie, beschreibt. Diese fällt nach Anwendung der `stepAIC` Funktion jedoch aus dem Modell.

Kumulatives Logit Modell			
	Koeffizenten	Standardabweichung	P-Wert
Intercept A|D	-1,745	0,361	$1,34 \cdot 10^{-6}$
Intercept D|H	0,614	0,360	$0,087$
Avg_H	-0,256	0,047	$6,18 \cdot 10^{-8}$
Avg_A	0,049	0,026	$0,059$
$FTHG$	1,880	0,071	$< 2^{-16}$
HS	-0,065	0,014	$5,40 \cdot 10^{-6}$
AS	0,101	0,016	$9,95 \cdot 10^{-11}$
HST	0,069	0,030	$0,020$
AST	-0,677	0,033	$< 2^{-16}$
HC	-0,056	0,020	$0,005$
AC	0,114	0,022	$3,05 \cdot 10^{-7}$
HF	-0,028	0,013	$0,031$
HR	-1,281	0,194	$4,22 \cdot 10^{-11}$
AR	0,736	0,360	$0,088$

Tabelle 5.8: Koeffizentenschätzungen, Standardabweichungen und P-Werte des besten Kumulativen Logit Modells

Die Koeffizienten des kumulativen Logit Modells in Tabelle 5.8 sind, außer den Intercepts, ähnlich zu interpretieren wie die aus dem regulären Logit Modell. Die einzelnen Koeffizienten gehen linear in die $log(Odds)$ ein. Auch hier bedeuten positive Koeffizienten ein steigen der Wahrscheinlichkeit in eine höhere Kategorie zu gelangen, ein negativer Koeffizient bedeutet ein sinken der Wahrscheinlichkeit in einer höheren Kategorie zu landen, jeweils bei festhalten der anderen Einflussvariablen. Die zwei Intercepts des Modells stellen Schwellenwerte, dass y_i mindestens in Kategorie r liegt. Liegt der Ausgabewert über 0,614, beurteilt das Modell, dass y_i entweder in Kategorie Unentschieden oder Kategorie Heimsieg liegt.

$Zuschauer_i = 1$	Geschätzt		
Tatsächlich	Auswärtssieg	Unentschieden	Heimsieg
Auswärtssieg	539	182	25
Unentschieden	166	322	164
Heimsieg	21	157	872

Tabelle 5.9: Konfusionstabelle des ausgewählten kumulativen Logit Modells

Tabelle 5.9 zeigt die Modellschätzungen für den Datensatz in Hinblick auf das Resultat des Spiels. Von 2448 Spielen schätzte das Modell 1733 Resultate korrekt ein. Dies entspricht einer Genauigkeit von circa 71 %. Durch die Eigenschaft des kumulativen Logit Modells

mit r Kategorien, dass es $r-1$ Intercepts besitzt und dem entsprechenden Schwellenwerte um mindestens in $Kategorie_i$ zu liegen, schätzt das Modell kaum Spiele in die entfernteren Kategorien ein, also tatsächliche Heimsiege zu Auswärtssiegen oder anders herum. Die meisten Fehler macht das Modell bei Unentschieden wo es circa die Hälfte aller Unentschieden richtig schätzt und jeweils ungefähr ein Viertel in die beiden anderen Kategorien.

Insgesamt schätzen die Modelle die Daten nicht optimal. Jedes der drei vorgestellten Modelltypen hat seine Schwächen. Die Genauigkeiten liegen bei allen drei Modellen innerhalb von zwei Prozentpunkten. Beide Fragestellungen können nicht vollständig zufriedenstellend beantwortet werden. Für die Zielvariable Resultat ist zu beachten, dass der Status, also in welchem Zeitabschnitt die Beobachtung lag, für den AIC nicht aussagekräftig zu sein scheint, um in einem Modell das Resultat eines Spiels vorherzusagen. Ebenfalls konnten bei der ersten Fragestellung nur wenige Spiele ohne Zuschauer korrekt vorhergesagt werden.

6 Fazit und Ausblick

Insgesamt kann festgestellt werden, dass der Einfluss der Corona-Einschränkungen auf die Spiele der Fußball Bundesliga nicht stark ist. Bereits in der explorativen Datenanalyse konnte erkannt werden, wie ähnlich sich die Spieldaten der vor der Corona Zeit und die der Saison 2020/2021 waren. Lediglich in der Anfangsphase der Corona-Pandemie in Deutschland waren markantere Veränderungen, so wie die niedrigere Heimsiegquote an den letzten neun Spieltagen der Saison 2019/2020 erkennbar. Das Fehlen der eigenen Fans im Stadion als Grund für diese Veränderung scheint unwahrscheinlich, da die Heimsiegquote der Folgesaison wieder auf ähnlichem Niveau zum Durchschnittswert lag. Ein wahrscheinlicherer Grund, der jedoch schwer messbar ist, sind die (inneren) Umstände für die Spieler. Auch professionelle Sportler sind Menschen mit Gefühlen und Sorgen. Mit Hinblick auf Kapitel 3 und den Einschränkungen in den Alltag der Spieler, auch zwischenmenschlicher Natur, können es schwierig gemacht haben, die gewohnte Leistung zu zeigen. Eine Erklärung für die Auswärtsstärke in dieser Zeit könnte sein, dass Spieler auf Auswärtsreisen weniger ungewohnte Einschränkungen hatten, als Spieler für Heimspiele. Diese durften weder ihre Nachbarn empfangen, noch zum Einkaufen gehen. Auch das Zusammensein mit ihren Mannschaftskollegen war in dieser Zeit schwieriger durch Abstandsregeln, Kontaktverboten und Überwachung. Die Fans im Stadion scheinen für die Spieler weniger nötig für die Leistung zu sein, als gut für das Gefühl. So drückte es auch der deutsche Nationalspieler Leon Goretzka nach einem Weltmeisterschafts-Qualifikationsspiel 2021 gegen Armenien aus: „Die Begeisterung der Fans ist für uns Balsam auf die Seele.“[3]

In weiteren Schritten könnte man die Ergebnisse aus Deutschland mit den Werten aus anderen Ländern vergleichen, um zu untersuchen, ob sich die Trends anhand der jeweiligen Einschränkungen von Verbänden und Regierungen unterscheiden. Ebenfalls waren die Herangehensweisen an die Wiederaufnahme in den Ländern unterschiedlich. Beispielsweise wurde die Saison der Ligue 1, der obersten französischen Fußballliga, abgebrochen und nicht fortgesetzt (vgl. faz.net, 2020a).
Interessant könnte hier beispielsweise sein, ob zur neuen Saison nach dem Abbruch in Frankreich ein ähnlicher Trend zu erkennen ist, wie in Deutschland zur Fortsetzung der Saison nach dem Lockdown, oder ob die Werte denen der Vorsaison ähneln.
Auch ein Blick auf neuere Spielmetriken wie erwartete Tore (xG), die jeder Torchance eine Torwahrscheinlichkeit zwischen 0 und 1 zuordnen und diese über das Spiel hinweg summieren (vgl. Schwerdtfege, 2021), könnte interessant für Analysen und Vergleiche sein.

Insgesamt ist es nicht einfach Sport in Metriken zu quantifizieren, aber über neue Methodiken werden immer mehr Daten erhoben um dies umzusetzen. Somit können in Zukunft hoffentlich viele Fragestellungen der Spieler, Verantwortliche und Fans bestmöglich beantwortet werden.

[3] Quelle: https://www.dfb.de/news/detail/goretzka-begeisterung-der-fans-ist-fuer-uns-balsam-auf-die-seele-231675/

Abbildungsverzeichnis

2.1 Histogramm Anzahl Tore Heimmannschaft 3
2.2 Histogramm Anzahl Tore Auswärtsmannschaft 4
2.3 Histogramm Anzahl Gesamttore 4
2.4 Korrelationsplot 7
2.5 Gestapeltes Säulendiagramm Endergebnis 9
2.6 Boxplot Tore pro Spiel 10

Tabellenverzeichnis

2.1 Weitere Variablenbeschreibung 6
2.2 Vergleich der Spieldaten vor Pandemie und während der Pandemie 8
2.3 Vergleich der Spieldaten vor Pandemie, nach dem Lockdown bis zum Saisonende 2019/2020 und der Saison 2020/2021 9
5.1 Vergleich der GLM Logit Modelle nach AIC und BIC 21
5.2 Koeffizientenschätzung, Standardabweichung und P-Wert GLM Logit Modell 22
5.3 Konfusionstabelle GLM Logit Modell 22
5.4 Vergleich der GAM nach AIC und BIC 23
5.5 Koeffizientenschätzung, Standardabweichung und P-Wert GAM Logit Modell 23
5.6 Konfusionstabelle GAM Logit Modell 24
5.7 Vergleich der kumulativen Logit Modelle nach AIC und BIC 24
5.8 Koeffizientenschätzung, Standardabweichung und P-Wert kumulatives Logit Modell 25
5.9 Konfusionstabelle kumulatives Logit Modell 25

Literatur

BMG (2021). Coronavirus-Pandemie (SARS-CoV-2): Chronik bisheriger Maßnahmen und Ereignisse. https://www.bundesgesundheitsministerium.de/coronavirus/chronik-coronavirus.html. [Online; zuletzt besucht 13.11.2021].

Bundesregierung (2020). Besprechung der Bundeskanzlerin mit den Regierungschefinnen und Regierungschefs der Länder vom 22.03.2020. https://www.bundesregierung.de/breg-de/themen/coronavirus/besprechung-der-bundeskanzlerin-mit-den-regierungschefinnen-und-regierungschefs-der-laender-vom-22-03-2020-1733248. [Online; zuletzt besucht 08.11.2021].

DFL (2020a). Neuansetzung: Borussia Mönchengladbach gegen 1. FC Köln auf 11. März terminiert. https://www.bundesliga.com/de/bundesliga/news/neuansetzung-borussia-monchengladbach-1-fc-koln-21-spieltag-nachholspiel-9924. [Online; zuletzt besucht 05.11.2021].

DFL (2020b). Politik ermöglicht Saison-Fortsetzung ab der zweiten Mai-Hälfte – Statement von Christian Seifert. https://www.dfl.de/de/aktuelles/politik-ermoeglicht-saison-fortsetzung-ab-der-zweiten-mai-haelfte-statement-von-christian-seifert/. [Online; zuletzt besucht 13.11.2021].

DFL (2020c). SV Werder Bremen gegen Eintracht Frankfurt wird neu angesetzt. https://www.bundesliga.com/de/bundesliga/news/dfl-sv-werder-bremen-eintracht-frankfurt-24-spieltag-ansetzung-neuer-termin-10209. [Online; zuletzt besucht 05.11.2021].

DFL (2020d). TASK FORCE SPORTMEDIZIN/SONDERSPIELBETRIEB IM PROFIFUSSBALL | VERSION 2. https://media.dfl.de/sites/2/2020/05/2020-05-01-Task-Force-Sportmedizin_Sonderspielbetrieb.pdf. [Online; zuletzt besucht 12.11.2021].

Emmert-Streib, F. and Dehmer, M. (2019). Evaluation of regression models: Model assessment, model selection and generalization error. *Machine learning and knowledge extraction*, 1(1):521–551.

Fahrmeir, L., Heumann, C., Künstler, R., Pigeot, I., and Tutz, G. (2016). *Statistik: Der Weg zur Datenanalyse*. Springer-Verlag.

Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. (2007). *Regression*. Springer.

faz.net (2020a). Corona-Abbruch in Frankreich. https://www.faz.net/aktuell/sport/fussball/frankreich-beendet-fussball-profiligen-wegen-corona-krise-16745935.html. [Online; zuletzt besucht 14.11.2021].

faz.net (2020b). Dynamo Dresden schickt alle Profis in Quarantäne. https://www.faz.net/aktuell/sport/wegen-positiver-corona-tests-dynamo-

dresden-schickt-profis-in-quarantaene-16762505.html. [Online; zuletzt besucht 13.11.2021].

Interview, V. W. H. (2017). Abschlussinterview mit Dr. Tim Schumacher – Teil 1. https://www.vfl-wolfsburg.de/newsdetails/news-detail/detail/news/ohne-die-fans-haetten-wir-es-nicht-geschafft/. [Online; zuletzt besucht 04.11.2021].

Kuha, J. (2004). Aic and bic: Comparisons of assumptions and performance. *Sociological methods & research*, 33(2):188–229.

Köln, . F. (2020). Corona-Virus - KEINE ZUSCHAUER IN GLADBACH. https://fc.de/de/fc-info/news/detailseite/details/keine-zuschauer-in-gladbach/. [Online; zuletzt besucht 08.11.2021].

Schwerdtfege, V. (2021). xGoals - was ist das überhaupt? https://www.kicker.de/xgoals-was-ist-das-ueberhaupt-798633/artikel. [Online; zuletzt besucht 14.11.2021].

Yee, T. W. (2015). *Vector generalized linear and additive models: with an implementation in R*. springer.

Eigenständigkeitserklärung

Ich versichere, dass ich die vorgelegte Bachelorarbeit eigenständig und ohne fremde Hilfe verfasst, keine anderen als die angegebenen Quellen verwendet und die den benutzten Quellen entnommenen Passagen als solche kenntlich gemacht habe.
Diese Bachelorarbeit ist in keinem anderen Kurs in dieser oder einer ähnlichen Form vorgelegt worden.

München, den 17.11.2021

.....

Dennis Reusch