

Ludwig-Maximilians-Universität München
Institut für Statistik
Wintersemester 2019/2020

**Abschlussbericht zum Projekt
LVS-IR-Taubenstein
im Rahmen des
statistischen Praktikums**

Projektpartner: Sascha Filimon, Roman Ossner
Gruppenbetreuer: Dr. André Klima
Projektgruppe: Alexander Fogus, Lea Vanheyden, Zorana Spasojević

Inhaltsverzeichnis

Inhaltsverzeichnis	2
Abbildungsverzeichnis	3
Tabellenverzeichnis	4
1 Einleitung	6
2 Hintergrund	8
3 Datengrundlage	10
3.1 Erhebungszeitraum 2018/2019	10
3.2 Erhebungszeitraum 2020	21
4 Methodik	22
4.1 Theorie	22
4.1.1 Lineares Regressionsmodell	22
4.1.2 Generalisiertes lineares Regressionsmodell	23
4.1.3 Additives Regressionsmodell	24
4.1.4 Splines	24
4.1.5 Generalisiertes additives Modell	28
4.2 Praktische Herausforderungen	29
4.2.1 Concurvity	29
4.2.2 Überdispersion	30
4.2.3 Autokorrelation	30
5 Angewandte Modelle	32
5.1 Tagesmodell	32
5.2 Zeitmodell	32

5.3	Ergebnisse	33
5.3.1	Ergebnisse des Tagesmodells	33
5.3.2	Ergebnisse des Zeitmodells	35
5.3.3	Signifikanz	36
6	Messfehleranalyse	37
6.1	Messfehler Beschreibung	37
6.2	Deskriptive Analyse der Messfehlerdaten	38
6.3	Messfehlerszenarien im Vergleich	39
6.3.1	Im Tagesmodell	41
6.3.2	Im Zeitmodell	43
6.4	Szenario mit steigender genereller Unterschätzung	44
6.4.1	Im Tagesmodell	44
6.4.2	Im Zeitmodell	46
7	Fazit	48
8	Ausblick	50
9	Umsetzung in R	51

Abbildungsverzeichnis

2.1	Untersuchungsraum: Spitzingsee mit dem Berg Taubenstein (Bildquelle: Bergpixel.de)	9
2.2	LVS-Checkpoints an der Nord- und Südseite des Taubenstein-Parkplatzes (Bildquelle: Roman Ossner DAV)	9
3.1	Gestapeltes Balkendiagramm der absoluten Häufigkeit der Messungen mit und ohne LVS-Gerät nach Datum	12
3.2	Liniendiagramm für den Anteil der Personen mit LVS-Gerät nach Datum	13
3.3	Gestapeltes Balkendiagramm der absoluten Häufigkeit der Personen mit und ohne LVS-Gerät nach Wochentag	14
3.4	Gestapeltes Balkendiagramm der absoluten Häufigkeiten der Personen mit und ohne LVS-Gerät für den Fall, dass es sich um einen Ferientag handelt und für den Fall, dass es sich nicht um einen Ferientag handelt	15
3.5	Liniendiagramm der Schneehöhe in cm nach Datum (ausgegraute Werte an Tagen die aus dem Datensatz entfernt wurden)	16
3.6	Liniendiagramm der Temperatur in °C um 12 Uhr mittags nach Datum (ausgegraute Werte an Tagen die aus dem Datensatz entfernt wurden; horizontale, gestrichelte Linie kennzeichnet 0°C)	17
3.7	Boxplot der täglichen Durchschnittslawinenwarnstufe	18
3.8	Liniendiagramm der täglichen Durchschnittsbewölkung in Prozent nach Datum (ausgegraute Werte an Tagen die aus dem Datensatz entfernt wurden)	19
3.9	Gestapeltes Balkendiagramm der absoluten Häufigkeit der Personen mit und ohne LVS-Gerät nach Uhrzeit der Messung (vertikale Linie kennzeichnet den neuen Tagesbeginn)	20

3.10	Balkendiagramm der absoluten Häufigkeit der Messungen nach Datum, aufgeteilt nach Position des Checkpoints (Nord und Süd) . . .	21
4.1	Schematische Darstellung der Schätzung eines nichtparametrischen Effekts mit B-Splines (? S.307)	26
4.2	Liniendiagramm der Schneedifferenz zum Vortag in cm nach Datum (ausgegraute Werte an Tagen, die aus dem Datensatz entfernt wurden)	29
4.3	Korrelogramm der Autokorrelation nach Verzögerung (Lag)	31
4.4	Korrelogramm der partiellen Autokorrelation nach Verzögerung (Lag)	31
5.1	Zusammenhang der Variablen Wochentag, Lawinenwarnstufe, durchschnittliche Tagesbewölkung, Temperatur, Schneedifferenz und Datum mit dem Anteil der Personen mit LVS-Gerät als glatte Funktionen für das Tagesmodell	33
5.2	Zusammenhang der Variablen Wochentag, Lawinenwarnstufe, stündliche Bewölkung, Temperatur und Schneedifferenz mit dem Anteil der Personen mit LVS-Gerät als glatte Funktionen für das Zeitmodell	35
5.3	Zusammenhang der Variablen Datum und Uhrzeit mit dem Anteil der Personen mit LVS-Gerät als glatte Funktionen für das Zeitmodell (Linien stellen Sonnenauf- und -untergang dar)	35
6.1	Balkendiagramm der absoluten Häufigkeit der gemessenen Kontakte in der manuellen Zählung, aufgeteilt nach Art des Kontakts und Erfassung/Nichterfassung	39
6.2	Gestapeltes Balkendiagramm für das Verhältnis erfasster zu nicht erfasster Kontakte nach Gruppengröße	39
6.3	Zusammenhang der Variablen Wochentag, Lawinenwarnstufe, durchschnittliche Tagesbewölkung, Temperatur, Schneedifferenz und Datum mit dem Anteil der Personen mit LVS-Gerät als glatte Funktionen im Tagesmodell für den originalen Datensatz und jedes Szenario	42
6.4	Zusammenhang der Variablen Wochentag, Lawinenwarnstufe, stündliche Bewölkung, Temperatur und Schneedifferenz mit dem Anteil der Personen mit LVS-Gerät als glatte Funktionen im Zeitmodell für den originalen Datensatz und jedes Szenario	43

6.5	Zusammenhang der Variablen Wochentag, Lawinenwarnstufe, durchschnittliche Tagesbewölkung, Temperatur, Schneedifferenz und Datum mit dem Anteil der Personen mit LVS-Gerät als glatte Funktionen im Tagesmodell für den originalen Datensatz und steigende Werte an allgemeiner Unterschätzung	45
6.6	Zusammenhang der Variablen Wochentag, Lawinenwarnstufe, stündliche Bewölkung, Temperatur und Schneedifferenz mit dem Anteil der Personen mit LVS-Gerät als glatte Funktionen im Zeitmodell für den originalen Datensatz und steigende Werte an allgemeiner Unterschätzung	47

Tabellenverzeichnis

6.1	Tabelle mit Koeffizienten und p-Werten für das Original, Szenario 1 (Generelle Unterschätzung von 22%), Szenario 2 (Unterschätzung nach Gruppengrösse), Szenario 3 (Nächtliche Überschätzung) und Szenario 4 (Unterschätzung nach Temperatur) für das Tagesmodell	41
6.2	Tabelle mit Koeffizienten und p-Werten für das Original, Szenario 1 (Generelle Unterschätzung von 22%), Szenario 2 (Unterschätzung nach Gruppengrösse), Szenario 3 (Nächtliche Überschätzung) und Szenario 4 (Unterschätzung nach Temperatur) für das Zeitmodell	43
6.3	Tabelle mit Koeffizienten und p-Werten der Kovariablen für steigende Werte an allgemeiner Unterschätzung im Tagesmodell	45
6.4	Tabelle mit Koeffizienten und p-Werten der Kovariablen für steigende Werte an allgemeiner Unterschätzung im Zeitmodell	46

Abstract

Im Rahmen dieser Arbeit und in Zusammenarbeit mit der Sektion München und Oberland des Deutschen Alpenvereins und des Departements für Geographie der Ludwig-Maximilians-Universität München wurden die Auswirkungen von Umwelteinflüssen auf die Mitnahme von Lawinenverschüttungssuchgeräten (abgekürzt LVS-Geräte) untersucht.

Ziel der vorliegenden Analyse ist die Untersuchung der Auswirkungen einzelner Kovariablen auf den Anteil von Personen, die ein solches Gerät bei sich tragen. Die Ergebnisse sollen helfen, das Verhalten der Personen besser einschätzen und es in Folge auch besser steuern zu können. Als Datengrundlage dienen die Messungen zweier Checkpoints im Gebiet Spitzingsee und dazugehörige Wetterdaten, welche für die Wintersaison 2018/19 erhoben wurden. Es wurden zwei Modelle, beides Generalisierte Additive Modelle mit der Binomial-Verteilung und der Logit-Link Funktion, geschätzt. Ein Modell lässt die Kovariable Uhrzeit außer Acht, so dass die beiden Modelle die Daten auf unterschiedlichen Ebenen aggregieren. Aus beiden Modellen lassen sich schließlich die Zusammenhänge der Kovariablen mit dem Anteil der Personen mit LVS-Gerät ermitteln. Am Auffälligsten ist dabei eine generelle Abnahme des Anteils über die Saison. Ebenfalls zu beobachten sind bestimmte Uhrzeiten, an denen der Anteil höher ist, und ein steigender zu erwartender Anteil mit der Lawinenwarnstufe. Neben den Daten der Saison 18/19 wurden auch Daten zu zwei Tagen im Februar 2020 analysiert. Das Ziel war es hierbei, mögliche Messfehler der Messgeräte quantifizieren zu können. Es wurde festgestellt, dass die Messgeräte ungefähr 22% der vorbeilaufenden Personen nicht identifizieren. Auf der Grundlage der Erkenntnis über die Messfehler wurden verschiedene Unterschätzungsszenarien simuliert, die Modelle darauf angewandt und dann die Ergebnisse miteinander verglichen. Die Szenarien sind: eine generelle Unterschätzung (dabei wird stufenweise von 10% bis auf 70% vorgegangen), eine Unterschätzung von Gruppen und eine Unterschätzung bei niedrigen Temperaturen. Neben den Unterschätzungen wurde auch ein Überschätzungsszenario durchgeführt und nachts Messungen entfernt, da sie von Wildtieren stammen könnten. Alle Szenarien zeigen kaum Unterschiede in der Art des Zusammenhangs der Zielvariable mit den Kovariablen. Einzig die Höhe des zu erwartenden Anteils insgesamt nimmt ab oder zu.

1 Einleitung

Die Bayrischen Berge sind der Lebensraum zahlreicher Wildtiere. Jedoch nutzen auch Menschen die Alpen. Vor allem im Winter dringen Bergsportler in die Rückzuggebiete der Tiere ein und stören diese.

Der Deutsche Alpenverein versucht, diese Konflikte zu entschärfen. Die Besucher werden vor Ort informiert und sollen durch Lenkungsmaßnahmen gesteuert werden. Dazu ist es nötig, das Verhalten der Besucher zu verstehen. Dafür werden Signale von Lawinenverschüttungssuchgeräten ausgewertet und analysiert. Von Interesse ist deshalb auch, ob und unter welchen Umständen Wintersportler ein solches Gerät bei sich tragen.

Im Rahmen des statistischen Praktikums im Wintersemester 2019/20 und in Zusammenarbeit mit dem Deutschen Alpenverein Sektion München und Oberland und dem Department für Geographie der Ludwig-Maximilians-Universität München wurden in diesem Projekt die Auswirkungen verschiedener Umwelteinflüsse auf den Anteil der Personen, die ein LVS-Gerät bei sich tragen untersucht.

Grundlage der Analyse sind dabei von zwei Checkpoints am Berg Taubenstein erhobene Messungen während der Wintersaison 2018/19. Neben den Checkpointmessungen liegen auch Daten zu Wetterbedingungen vor. Aus diesen Daten werden zwei Modelle geschätzt und analysiert. Dabei handelt es sich um generalisierte additive Modelle mit der Binomialverteilung und einer Logit-Link-Funktion. Die Validität der Modelle ist allerdings nur gegeben, wenn die für sie benutzten Daten keine strukturellen Fehler aufweisen.

Es wurde jedoch erkannt, dass die Messgeräte, welche die Daten über die Mitnahme der LVS-Geräte liefern, fehlerhaft sind. Um das Ausmaß des Messfehlers feststellen zu können, wurden im Februar 2020 manuelle Zählungen von Studierenden durchgeführt. Diese werden analysiert und auf Basis der Ergebnisse und Überlegungen der Projektpartner werden verschiedene Szenarien für fehlerhafte Messungen identifiziert. Diese werden an den Daten simuliert und die Modelle werden darauf angewandt. So möchte man untersuchen, welche Auswirkungen

verschiedene Arten von Messfehlern auf die Ergebnisse des Modells haben. Im Folgenden wird kurz auf den Hintergrund des Projekts genauer eingegangen. Danach werden die Daten beschrieben, bevor zur Theorie hinter den Modellen übergegangen wird. Die praktischen Schwierigkeiten werden besprochen und die angewandten Modelle aufgezeigt. Im Anschluss werden die Ergebnisse beider Modelle besprochen und verglichen. Danach folgt die Messfehleranalyse. Zuerst werden die Zählungen der Studierenden deskriptiv betrachtet. Daraus werden dann vier Szenarien abgeleitet und für beide Modelle durchgeführt. Die Ergebnisse werden dann verglichen. Das Szenario der generellen Unterschätzung wird anschließend noch mit steigender Unterschätzung in 10%-Schritten angewandt und verglichen. Abschließend wird ein Fazit gezogen und ein Ausblick gegeben.

2 Hintergrund

Für viele Besucher sind die Bayrischen Alpen ein beliebtes Gebiet, im Sommer sind Wanderer anzutreffen und im Winter vorrangig Skitouren- und Schneeschuhgeher. Jedoch stellt das Alpengebiet auch das Territorium von verschiedenen Tierarten dar. Es ist zu beobachten, dass es immer wieder zu räumlichen Konfliktsituation zwischen Menschen und Tieren kommt. Vor allem im Winter dringen die Sportler tief in die Rückzugsgebiete der Tiere ein und zwingen diese zur Flucht.

Um die räumlichen Konflikte zu entschärfen, werden die Besuchenden vor Ort informiert und durch Lenkungsmaßnahmen gesteuert. Mit dem Ziel, einen nachhaltigen Tourismus gewährleisten zu können, wurde ein Aktionstag seitens des Alpenvereins initiiert. Der Aktionstag ist ein Bestandteil der Kampagne „Natürlich auf Tour“ und dient der Sensibilisierung von Besuchern für einen verantwortungsvollen Umgang mit der Natur. (?)

Um das Besucherverhalten besser zu verstehen und beeinflussen zu können, werden die Signale des LVS-Geräts, das ein Skitourengeher in der Regel dicht am Körper trägt, ausgewertet und analysiert.

Ein Lawinenverschüttungssuchgerät oder auch LVS-Gerät gehört zur Ausrüstung von Sportlern die in Schneegebieten unterwegs sind. Mit Hilfe eines LVS-Geräts können Personen, die von einer Lawine verschüttet worden sind, geortet werden. (?)

Die Untersuchung über die Mitnahme von LVS-Geräten wurde anhand von Checkpoints gemessen. An den LVS-Checkpoints kann der Wintersportler sein Gerät auf dessen Funktionalität testen. Wenn eine Person ein eingeschaltetes LVS-Gerät bei sich trägt und an einem Checkpoint vorbei läuft, wird dies als ein LVS-Signal registriert und das Gerät leuchtet grün auf. Falls eine Person, die kein Gerät dabei hat, bzw. eins mit sich führt, dieses jedoch nicht eingeschaltet ist, am Checkpoint vorbeiläuft, dann wird dies als ein Infrarot-Signal (abgekürzt: IR-Signal) identifiziert und das Messgerät leuchtet rot auf.

Die dabei anonym erhobenen Daten werden für die weitere Untersuchung betrach-

tet.

In diesem Projekt wird nicht das ganze Alpengebiet als Untersuchungsort betrachtet, sondern es wird sich auf den Berg Taubenstein am Spitzingsee beschränkt.

Abbildung 2.1: Untersuchungsraum: Spitzingsee mit dem Berg Taubenstein
(Bildquelle: Bergpixel.de)

Das Gebiet Spitzingsee ist ein beliebter Anlaufpunkt für Skitouren- und Schneeschuhgeher, aber auch das Revier von Wildtieren und daher als Untersuchungsraum für das Projekt geeignet. Am Taubenstein-Parkplatz sind zur Hochsaison zwischen 500 bis 1000 Besucher anzutreffen, weshalb dieser Punkt für das Aufstellen von zwei Checkpoints genutzt wurde. (?)

Abbildung 2.2: LVS-Checkpoints an der Nord- und Südseite des Taubenstein-Parkplatzes (Bildquelle: Roman Ossner DAV)

Es gibt zwei Checkpoints, die auf der Nord- und Südseite des Parkplatzes kurz vor Aufstieg des Bergs aufgestellt sind.

Im Fokus der Analyse steht der Zusammenhang zwischen dem Anteil der Besucher mit LVS-Gerät an allen Besuchern und verschiedenen Umweltfaktoren. Die von den Checkpoints generierten Daten und die einzelnen Kovariablen werden im nächsten Kapitel näher beschrieben.

3 Datengrundlage

In diesem Kapitel wird ein Überblick über die Datengrundlage des Projekts LVS-IR-Taubenstein gegeben.

Zuerst werden die im Zeitraum 2018/19 erhobenen Daten beschrieben. Da bekannt war, dass die Messgeräte fehlerhaft aufzeichnen, wurde im Februar 2020 seitens des Departments für Geographie eine manuelle Erhebung durchgeführt. Die daraus gewonnenen Daten werden anschließend kurz beschrieben, im Kapitel 6 Messfehleranalyse wird genauer darauf eingegangen.

Die Messdaten wurden vom DAV München und Oberland und vom Department für Geographie der LMU zur Verfügung gestellt.

3.1 Erhebungszeitraum 2018/2019

In der Wintersaison 2018/2019 wurden die Daten von zwei Checkpoints an der Nord- und Südseite des Spitzingsee-Parkplatzes erfasst. Die Checkpoints registrieren dabei mithilfe von Infrarotsignalen eine Person ohne LVS-Gerät bzw. mit ausgeschaltetem LVS-Gerät, sowie mittels eines LVS-Signals Personen mit eingeschaltetem LVS-Gerät.

Im Messzeitraum vom 21.12.2018 bis zum 13.04.2019 wurden insgesamt 37593 Messungen erfasst. Da die Messgeräte zeitweise aufgrund von starkem Schneefall bedeckt waren und nicht richtig messen konnten oder anderweitig nicht valide gemessen haben, wurden die Daten für den Zeitraum vom 21.12.2018 bis zum 25.12.2018 und vom 07.01.2019 bis 15.01.2019 entfernt.

Damit bleiben 37216 Messungen übrig. Darunter wurden 28911 Personen ohne LVS-Gerät und 8305 Personen mit LVS-Gerät von den beiden Checkpoints registriert.

Die Messgeräte lösen bei jeder vorbeigehenden Person aus, wodurch auch Mehrfacherfassungen möglich sind. Das führt dazu, dass eine Person zum Beispiel beim Losgehen und Zurückkommen vom Checkpoint doppelt erfasst wird.

Nachfolgend werden die zur Analyse des Modells erforderlichen Variablen vorgestellt.

Zielvariable:

- Anteil der Personen mit LVS-Gerät

Kovariablen:

- Datum
- Wochentag
- Ferientag
- Schneehöhe (in cm)
- Temperatur (in °C)
- Lawinenwarnstufe
- Bewölkung
- Uhrzeit der Messung
- Position

Anteil der Personen mit LVS-Gerät:

Bei der Zielvariable handelt es sich um den Anteil der Personen mit beigeführtem LVS-Gerät an allen Personen die an den Checkpoints vorbeilaufen.

Datum:

Es liegen wie oben erwähnt Daten vom 25.12.2018 bis zum 06.01.2019 und vom 16.01.2019 bis zum 13.04.2019 vor.

Das Balkendiagramm (3.1) zeigt die absolute Häufigkeit der Messungen nach Datum und Art der Messung. Hier zu erkennen sind die fehlenden Messungen im Zeitraum vom 07.01.2019 bis zum 15.01.2019.

Abbildung 3.1: Gestapeltes Balkendiagramm der absoluten Häufigkeit der Messungen mit und ohne LVS-Gerät nach Datum

Da der Anteil der Personen mit LVS-Gerät von Interesse ist, ist dieser in (3.2) über die Saison hinweg zu sehen. Auch hier sind die fehlenden Tage ersichtlich. Zu erkennen ist ein sinkender Anteil der Personen mit LVS-Gerät über den Verlauf der Saison. Jedoch sind starke Schwankungen zu beobachten.

Abbildung 3.2: Liniendiagramm für den Anteil der Personen mit LVS-Gerät nach Datum

Wochentag:

Anhand der Variable Datum konnte der Wochentag mit den Ausprägungen von Montag bis Sonntag festgelegt werden.

Die absoluten Häufigkeiten der Checkpoint-Messungen nach Wochentag sind in Abbildung (3.3) zu sehen. Es ist zu beobachten, dass am Wochenende die Anzahl der Personen insgesamt (mit und ohne LVS-Gerät) im Vergleich zu den restlichen Tagen am höchsten ist.

Abbildung 3.3: Gestapeltes Balkendiagramm der absoluten Häufigkeit der Personen mit und ohne LVS-Gerät nach Wochentag

Ferientag:

Bei jeder Messung ist außerdem bekannt, ob es sich bei dem Datum um einen Ferientag gehandelt hat. Unter den Begriff Ferientag fallen sowohl Feiertage als auch Schulferientage in Bayern. Insgesamt gibt es 22 Ferientage.

Die absoluten Häufigkeiten der Personen mit und ohne LVS-Gerät nach Ferientag sind in Abbildung (3.4) zu sehen.

Abbildung 3.4: Gestapeltes Balkendiagramm der absoluten Häufigkeiten der Personen mit und ohne LVS-Gerät für den Fall, dass es sich um einen Ferientag handelt und für den Fall, dass es sich nicht um einen Ferientag handelt

Schneehöhe:

Zu jeder Messung liegt die tägliche Schneehöhe in cm vor. Der Verlauf der Schneehöhe über die Wintersaison wird in Abbildung (3.5) dargestellt.

Dabei ist zu erkennen, dass die Schneehöhe von Mitte Dezember bis Mitte Januar tendenziell steigt.

Von Mitte Januar bis Mitte Februar ist der höchste Schnee zu beobachten, wobei Schwankungen sichtbar sind. Ab Mitte Februar ist eine absteigende Tendenz der Schneehöhe zu sehen. Der höchste Wert für die Schneehöhe beträgt 212 cm und der niedrigste Wert 16 cm.

Abbildung 3.5: Liniendiagramm der Schneehöhe in cm nach Datum (ausgegraute Werte an Tagen die aus dem Datensatz entfernt wurden)

Die Kovariable Schneehöhe wurde aufgrund des Concurvity-Problems transformiert, worauf in Unterabschnitt 4.2.1 näher eingegangen wird.

Temperatur:

Die Variable Temperatur wurde um 12 Uhr mittags des jeweiligen Tages in Grad Celsius erhoben. Tendenziell ist in Abbildung (3.6) ein steigender Temperaturverlauf mit starken Schwankungen von Ende Dezember bis Mitte April zu beobachten. Die niedrigste gemessene Temperatur beträgt $-7,9\text{ }^{\circ}\text{C}$ und die Höchste $9,7\text{ }^{\circ}\text{C}$.

Abbildung 3.6: Liniendiagramm der Temperatur in $^{\circ}\text{C}$ um 12 Uhr mittags nach Datum (ausgegraute Werte an Tagen die aus dem Datensatz entfernt wurden; horizontale, gestrichelte Linie kennzeichnet 0°C)

Lawinenwarnstufe:

Die Variable Lawinenwarnstufe besteht aus dem Durchschnitt zweier anderer Variablen, nämlich der Lawinenwarnstufe am Fuß und an der Spitze des Berges, welche täglich erhoben wurden. Es gibt Warnstufen von eins bis fünf, wobei fünf die größte Gefahr bedeutet.

In Abbildung (3.7) wird anhand eines Boxplots die Verteilung der Variable ersichtlich. Zu beobachten ist, dass der Median bei Warnstufe zwei liegt. Ebenfalls erkennbar ist, dass die Lawinenwarnstufe fünf im vorliegenden Datensatz nicht vorkommt. Die höchste vorkommende Lawinenwarnstufe ist vier und die Niedrigste, die auch das 25%-Quantil darstellt, ist eins.

Abbildung 3.7: Boxplot der täglichen Durchschnittslawinenwarnstufe

Bewölkung:

Die Bewölkung wird als Anteil von 100% angegeben. Eine 100-prozentige Bewölkung bedeutet ein komplett bedeckter Himmel und 0-prozentige Bewölkung ein wolkenfreier Himmel. Im Gegensatz zu den anderen bisher genannten Kovariablen liegt der Anteil der Bewölkung nicht täglich, sondern stündlich vor.

Es wurde neben dem stündlichen Wert auch der tägliche Durchschnittswert errechnet. Dieser ist in Abbildung (3.8) zu sehen, wobei eine starke Schwankung erkennbar ist.

Abbildung 3.8: Liniendiagramm der täglichen Durchschnittsbewölkung in Prozent nach Datum (ausgegraute Werte an Tagen die aus dem Datensatz entfernt wurden)

Uhrzeit der Messung:

Zu jeder Messung registriert das Messgerät neben dem Datum auch die Uhrzeit. Die absolute Häufigkeit der Personen mit und ohne LVS-Gerät nach der Uhrzeit wird in Abbildung (3.9) dargestellt.

Im Zeitraum zwischen 10 Uhr und 20 Uhr ist die Anzahl der Personen insgesamt (mit und ohne LVS-Gerät) am höchsten.

Wie auch schon am Anfang des Kapitels erwähnt, kann dieselbe Person mehrfach registriert werden. So werden Besucher, die erst nach Mitternacht des nächsten Tages von ihrer Tour zurückkehren, erneut registriert. Bei ihrer Entscheidung, ob sie ein LVS-Gerät mitnehmen, haben sie sich jedoch wahrscheinlich an den Umwelt-Bedingungen des Aufbruchtages orientiert.

Deshalb wurde, in Absprache mit den Projektpartnern, beschlossen, den Tag von 4 Uhr morgens bis 4 Uhr morgens des eigentlich nächsten Tages laufen zu lassen. Einer Messung von 0 Uhr bis 4 Uhr werden daher die Kovariablen des Vortags zugeordnet.

Abbildung 3.9: Gestapeltes Balkendiagramm der absoluten Häufigkeit der Personen mit und ohne LVS-Gerät nach Uhrzeit der Messung (vertikale Linie kennzeichnet den neuen Tagesbeginn)

Position:

Neben den bisher erwähnten Variablen ist bei jeder Messung auch der Ort des Messgerätes bekannt, also ob die Person vom Nord- oder Süd-Checkpoint registriert wurde.

Die absoluten Häufigkeiten über die Wintersaison hinweg sind in Abbildung (3.10) zu sehen. Es ist zu erkennen, dass deutlich mehr Personen am Süd-Checkpoint gemessen wurden. Für die folgenden Modelle wurden jedoch die Messungen der beiden Checkpoints zusammengezählt und die Position außer Acht gelassen.

Abbildung 3.10: Balkendiagramm der absoluten Häufigkeit der Messungen nach Datum, aufgeteilt nach Position des Checkpoints (Nord und Süd)

Die Messungen für die Variablen Schneehöhe, Temperatur und Lawinenwarnstufe wurden nicht von den Checkpoints am Spitzingsee gemessen, sondern von dem Gewässerkundlichen Dienst Bayern erhoben. Die Messungen der Variable Bewölkung stammen von Meteoblue.

3.2 Erhebungszeitraum 2020

Da erkannt wurde, dass die Messgeräte teilweise fehlerhaft messen, wurden manuelle Zählungen der Studierenden des Departments für Geografie durchgeführt.

Der Erhebungszeitraum beschränkt sich auf den 27.02.2020 und 28.02.2020. An diesen Tagen wurden nur Zählungen am Checkpoint-Nord durchgeführt (da der Checkpoint-Süd ausgefallen ist).

Diese sollen dabei helfen mögliche Strukturen des Messfehlers zu erkennen. Dabei haben die Studierenden beobachtet, ob eine vorbeilaufende Person vom Checkpoint registriert wird. Jede vorbeigehende Person wurde notiert und ebenso wurde vermerkt, ob diese vom Checkpoint erkannt wurde. Dabei wurde nicht zwischen Personen mit und ohne LVS-Gerät unterschieden.

Die Studierenden haben außerdem angegeben, ob es sich bei der Person um einen Skitourengänger oder einen anderen Kontakt handelt.

Neben den Daten der Studierenden liegen auch die vom Checkpoint erhobenen Daten vor. Insgesamt wurden von den Studierenden 208 Personen gezählt und von den Checkpoints 170 Messungen erfasst. Die aus diesen Daten gewonnen Erkenntnisse und deren Auswirkungen werden in Kapitel 6 beschrieben.

4 Methodik

In diesem Kapitel wird als erstes die Theorie hinter den Modellen erklärt. Danach wird auf die aufgetretenen Schwierigkeiten und die angewandten Lösungen eingegangen.

4.1 Theorie

In folgendem Kapitel soll die Theorie hinter den Modellen erklärt werden. Für das bessere Verständnis wird der Theorie-Teil von einem einfachen Modell bis hin zu dem von uns verwendeten Modell aufgebaut.

Die mathematischen Formeln und Konzepte aus diesem Kapitel basieren, wenn nicht anders vermerkt, aus dem Buch „Regression-Modelle, Methoden und Anwendungen“ von Ludwig Fahrmeir et al..

4.1.1 Lineares Regressionsmodell

Das einfachste Regressionsmodell ist das lineare Regressionsmodell. Hierbei wird der Erwartungswert einer Zielvariable durch die Linearkombination von Einflussgrößen, auch Kovariablen genannt, beschrieben. Das lineare Regressionsmodell nimmt dabei folgende Form an:

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} + \epsilon_i \quad (4.1)$$

Die Zielgröße, y_i , muss stetig und approximativ normalverteilt sein. Die Kovariablen werden durch x_{i1}, \dots, x_{ik} gekennzeichnet und ϵ_i stellt den Störterm dar.

In Anwendungen, bei denen es sich nicht um eine normalverteilte Zielvariable handelt, ist das lineare Regressionsmodell unzureichend. (? S.20 ff.)

Dieses Problem löst das generalisierte lineare Regressionsmodell, auf welches im nächsten Abschnitt eingegangen wird.

4.1.2 Generalisiertes lineares Regressionsmodell

Das generalisierte lineare Regressionsmodell eignet sich für die Modellierung von Zielvariablen, die nicht normalverteilt sind. Die abhängige Variable kann eine Verteilung aus der Exponentialfamilie (z.B. Binomial, Poisson, Multinomial, Normal,...) annehmen. Durch eine Link-Funktion wird sie dann in eine normalverteilte Variable transformiert.

Bei den später verwendeten Modellen lässt sich die Zielvariable wie folgt beschreiben:

$$y_i = \begin{cases} 1 & \text{,wenn Person mit LVS-Gerät gemessen wird} \\ 0 & \text{,wenn Person ohne LVS-Gerät gemessen wird.} \end{cases} \quad (4.2)$$

Die beobachtete Zielvariable ist Binomial-verteilt mit $y_i|x_i \sim B(1, \pi_i)$.

Bei der generalisierten Regression mit binärer Zielvariable, wie in diesem Fall, wird der Effekt der Kovariablen durch die (bedingte) Wahrscheinlichkeit π_i beschrieben.

$$\pi_i = P(y_i = 1|x_{i1}, \dots, x_{ik}) = E(y_i|x_{i1}, \dots, x_{ik}) \quad (4.3)$$

Für den Erwartungswert der Zielvariable gilt:

$$E(y_i) = \pi_i = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)} = h(\eta_i). \quad (4.4)$$

Der Erwartungswert der Zielvariable wird mit dem linearen Prädiktor η_i mittels der Responsefunktion $\pi_i = h(\eta_i)$ verknüpft. Die Logit-Linkfunktion g ist die Umkehrung der Responsefunktion, daher gilt:

$$g(\pi_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right). \quad (4.5)$$

Die hier verwendeten Modelle machen Gebrauch von dem Logit-Link.

Das generalisierte lineare Regressionsmodell beschreibt lineare Einflüsse der Kovariablen auf die Zielvariable. In praktischen Anwendungen ist dies oft unzureichend, da neben den linearen Einflüssen auch nicht-lineare, flexible Einflüsse von Kovariablen auf die abhängige Variable wirken können. (? S.189 ff.) Daher wird im nächsten Kapitel die additive Regression näher erläutert.

4.1.3 Additives Regressionsmodell

Das additive Modell stellt ein semi-parametrisches Regressionsmodell dar. Neben linearen Einflüssen können auch nicht-lineare flexible Einflüsse von Kovariablen auf die abhängige Variable modelliert werden.

Das Standardmodell der additiven Regression hat die Form:

$$y_i = \underbrace{\beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik}}_{\text{parametrische Effekte}} + \underbrace{f_1(z_{i1}) + \dots + f_q(z_{iq})}_{\text{nichtparametrische Effekte}} + \epsilon_i \quad (4.6)$$

Der parametrische Teil des Standardmodells hat die gleiche Form, wie schon im linearen Modell aufgezeigt. Die Funktionen $f_1(z_1), \dots, f_q(z_q)$ sind sogenannte Glättungen und schätzen die nichtparametrischen Effekte. (? S.44 ff.)

Auf die Theorie hinter den glatten Funktionen f wird im nächsten Kapitel eingegangen.

4.1.4 Splines

Die Glättungsfunktion f wird durch sogenannte Splines modelliert. Splines stellen bestimmte Funktionen dar, die stückweise aus Polynomen eines bestimmten Grades zusammengesetzt werden. Es gibt verschiedene Arten von Splines. (? S. 291 ff.)

In den Modellen wurden die Penalisierten Splines, die Zyklische Splines und die Thin-Plate Splines verwendet. Im Folgenden werden diese erläutert, wobei zum einfacheren Verständnis zuvor noch Polynom- und Basis-Splines erklärt werden.

Polynom-Splines

Eine Möglichkeit, die Funktion f abzubilden, sind Splines, die über Polynome modelliert werden und Polynom-Splines, oder auch Regressions-Splines, genannt werden. Das polynomiale Modell hat die Form:

$$f(z) = \gamma_0 + \gamma_1 z + \dots + \gamma_l z^l, \quad (4.7)$$

wobei das Polynom vom Grad l durch den Einfluss von z auf y modelliert wird und γ_j die Regressionskoeffizienten darstellen. Um eine flexible Schätzfunktion gewährleisten zu können, wird der Definitionsbereich von z in Intervalle geteilt und Polynome werden für das jeweilige Intervall geschätzt. Die Zerlegung von z erfolgt stückweise auf Basis von sogenannten Knotenmengen k_1 bis k_m . Ein Problem, welches dabei entstehen könnte, ist, dass an den Intervallgrenzen keine *glatte*

Funktion entstehen. Der Graph einer *glatten* Funktion hat an den Intervallgrenzen keine „Ecken“. Deshalb muss für die „mathematische“ Glattheit die zusätzliche Bedingung, dass die Funktion $f(z)$ an den Intervallgrenzen $(l-1)$ -mal stetig differenzierbar ist, eingeführt werden. Der Grad des Splines und die Anzahl der Knoten können im hohen Maße die Funktion beeinflussen, je höher die Anzahl an Knoten ist, desto *rauer* ist die Funktion $f(z)$. Um Polynom-Splines in der Praxis anwenden zu können, bedarf es einer Darstellung der Menge der Polynom-Splines. Eine mögliche Darstellung eines Polynom-Splines stellt die B-Spline-Basisfunktion dar. (? S.293 ff.)

Basis Splines

Der B-Spline dient als Basisfunktion für Polynom-Splines. Jedoch wird sich auch der Penalisiertungs-Spline, welcher im folgenden Abschnitt vorgestellt wird, die B-Spline-Basisfunktion zu Nutzen machen. Das Ziel bei der Konstruktion der B-Spline-Basisfunktion ist es, die Polynomstücke des gewünschten Grades an den Knoten ausreichend *glatt* zusammenzusetzen. Die Basisfunktion besteht dabei aus $(l+1)$ Polynomstücken vom Grad l , welche stetig an den Knoten zusammengesetzt werden. Die Funktion $f(z)$ lässt sich durch die Linearkombination der Basisfunktionen:

$$f(z) = \sum_{j=1}^d \gamma_j B_j(z) \quad (4.8)$$

darstellen. Die Koeffizienten γ_j werden mit Hilfe der KQ-Methode geschätzt und dienen zur Skalierung der Basisfunktionen. Die Anzahl der erforderlichen Basisfunktionen $d = m + l - 1$ setzt sich aus der Zahl der Knoten m und den Polynomstücken des gewünschten Grades l , welche $(l-1)$ -mal stetig differenzierbar an den Knoten aneinandergefügt werden, zusammen.

Für B-Splines vom Grad 0 folgt:

$$B_j^0(z) = \mathbb{1}_{[k_j, k_{j+1}]}(z) = \begin{cases} 1 & k_j \leq z < k_{j+1}; \text{ mit } j = 1, \dots, d-1; \\ 0 & \text{sonst.} \end{cases} \quad (4.9)$$

Für B-Splines höheren Grades gilt:

$$B_j^l(z) = \frac{z - k_j}{k_{j+l} - k_j} B_j^{l-1}(z) + \frac{k_{j+l+1} - z}{k_{j+l+1} - k_{j+1}} B_{j+1}^{l-1}(z). \quad (4.10)$$

(? S.303 ff.)

Die Abbildung 4.1 visualisiert die Schätzung eines B-Splines anhand eines fiktiven

Datenbeispiels schematisch. Abbildung (a) demonstriert eine B-Spline Basis vom dritten Grad. In Abbildung (b) wird die Basisfunktion mit dem Kleinst-Quadrate-Schätzer $\hat{\gamma}_j$ skaliert. Die Abbildung (c) zeigt die erhaltene Schätzung, wenn die skalierten Basisfunktionen addiert werden.

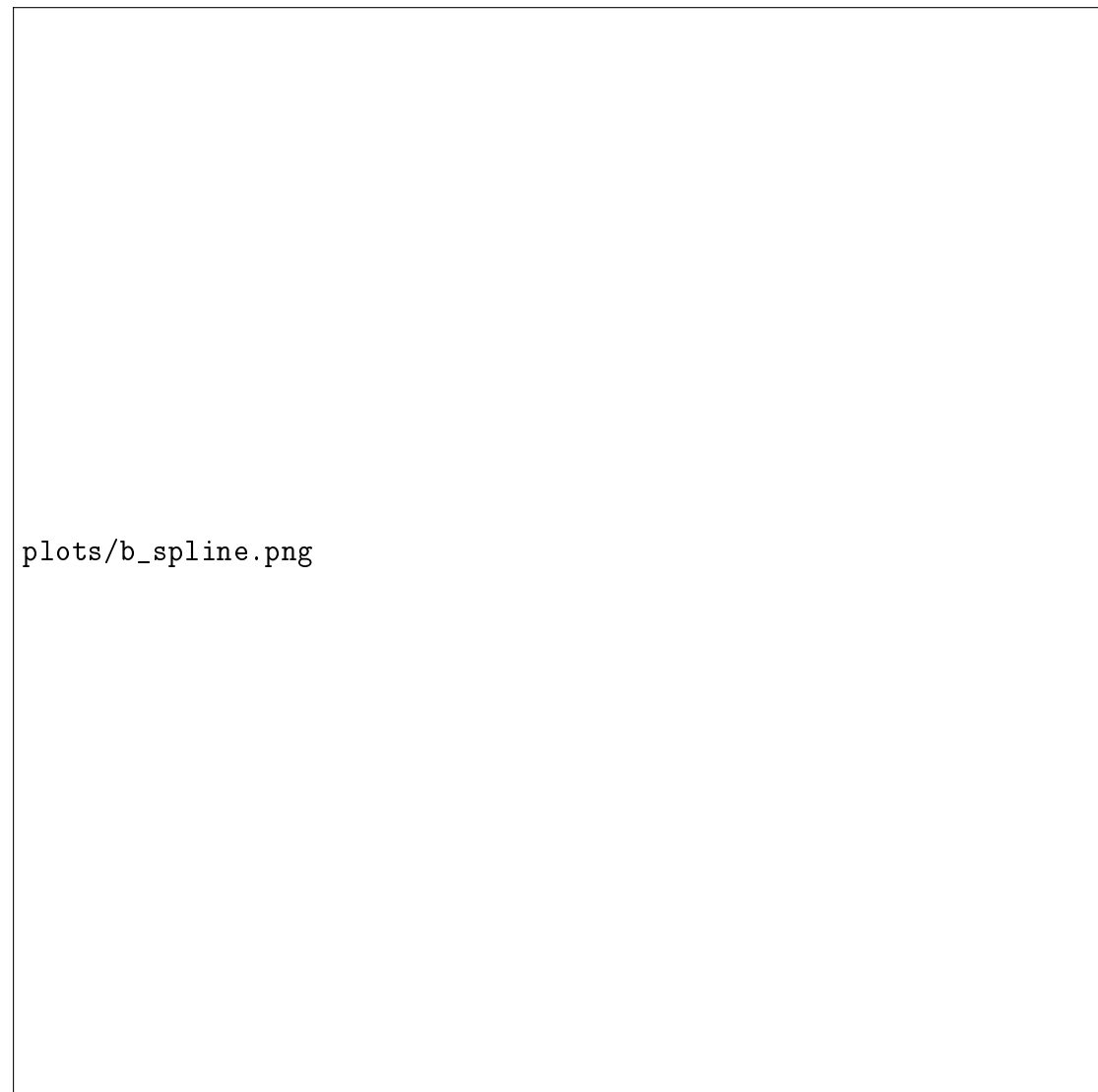


Abbildung 4.1: Schematische Darstellung der Schätzung eines nichtparametrischen Effekts mit B-Splines (? S.307)

Das nächste Kapitel behandelt P-Splines, welche auf B-Splines basieren und durch einen Strafterm erweitert worden sind. Diese werden bei den später erklärten Modellen verwendet.

Penalisierte Splines basierend auf B-Splines

Bereits im vorherigen Abschnitt wurde klar, dass die Knotenmenge einen wichtigen Faktor darstellt. Die Glattheit und Flexibilität der Schätzfunktion $f(z)$ hängt stark von der Anzahl der Knoten ab. Eine zu hohe Anzahl von Knoten führt zu sogenanntem „Overfitting“, das heißt die Funktion ähnelt zu sehr den Daten und ist sehr wacklig. Die Idee der P-Splines ist es, die Funktion $f(z)$ auf Basis einer B-Spline Funktion mit einer großen Zahl an Knoten zu approximieren, um so die Flexibilität der Schätzfunktion sicherzustellen und andererseits zu hohe Variabilität durch einen Strafterm zu sanktionieren. Der Penalisierte-Spline, gekürzt P-Spline, kombiniert somit eine B-Spline-Basis mit einem Strafterm. Der Strafterm für B-Splines wird durch die quadrierte zweite Ableitung der Schätzfunktion $f(z)$ modelliert. Die zweite Ableitung stellt ein Maß für die Krümmung der Funktion dar und ist daher in der Lage die Glattheit einer Funktion einzuschätzen. Der Strafterm:

$$\lambda \int (f''(z))^2 dz. \quad (4.11)$$

In der Anwendung wird für die Approximation der zweiten Ableitungen, die zweite Differenz des Parameters γ_j verwendet. Wenn der Glättungsparameter λ gegen unendlich konvergiert, erhält man eine nahezu lineare Schätzfunktion $f(z)$. Im Fall von $\lambda \rightarrow 0$ konvergiert auch der Strafterm gegen 0 und die P-Splines verhalten sich wie unpenalisierte B-Splines. Der Glättungsparameter λ wird im Modell durch REML (Restriktierte Maximum-Likelihood) geschätzt. (? S.309 f.)

Zyklische P-Splines

Eine Glättungsfunktion heißt zyklisch, wenn die Funktion dieselben Werte und ersten Ableitungen an ihrer oberen und unteren Grenze aufweist. Beispielsweise kann man die Funktion für die Variable Wochentag mit Hilfe des zyklischen P-Splines modellieren. Das Ziel dabei ist es, z.B. den Sonntag mit dem darauffolgenden Montag zu verknüpfen. Dadurch wird sichergestellt, dass die Werte am Ende einer Woche zusammenhängend zu den Werten am Anfang der darauffolgenden Woche sind. Die Schätzfunktion eines zyklischen P-Splines hat die Form:

$$f(z) = \sum_{j=1}^{k-1} B_j(z) \beta_j. \quad (4.12)$$

Der Koeffizient β_j wird durch den transponierten Vektor $\beta^T = (\beta_1, \dots, \beta_{k-1})$ dargestellt, wobei $\beta_j = \beta_k$ gilt. Der Penalisierungsansatz für zyklische P-Splines ist simultan zu der Penalisierung von P-Splines basierend auf B-Splines. (? S.202 ff.)

Thin-Plate Splines

Bisher wurden Funktionen mit nur einer einzigen Kovariable geschätzt. Die Thin-Plate Splines ermöglichen die Schätzung einer Funktion mit mehreren Kovariablen. Bei einem der hier angewandten Modelle handelt es sich um eine Funktion mit zwei Variablen, nämlich dem Datum und der Uhrzeit. Die beiden Variablen wurden in einer Funktion geschätzt, da sie miteinander interagieren. Der Vorteil von Thin-Plate Splines ist, dass nun eine Auswahl von Knotenpositionen oder Basisfunktionen nicht notwendig ist, da sich diese durch die mathematische Darstellung des Glättungsparameters ergibt. Die Glattheit, wenn z zweidimensional ist, wird durch die Minimierung der Funktion von f erzielt, die die folgende Form hat:

$$\|y - f\|^2 + \underbrace{\lambda}_{\text{Glättungsparameter}} \underbrace{J_{md}(f)}_{\text{Strafterm}}. \quad (4.13)$$

Dabei stellt y den Vektor von y_i und $f = [(f(z_1), \dots, f(z_n))]^T$ dar. Die Funktion f wird durch einen Glättungsparameter multipliziert mit einem Strafterm für raue bzw. wackelige Funktionen ergänzt. (? S.215 ff.)

4.1.5 Generalisiertes additives Modell

Das generalisierte additive Modell verbindet das additive Regressionsmodell mit einem generalisierten Modell. Es eignet sich um nicht-lineare Effekte von metrischen Kovariablen auf eine nicht unbedingt normalverteilte Zielvariable zu beschreiben. Wie schon im additiven Modell erläutert, setzt sich der additive Prädiktor (η_i) aus einem parametrischen- und einem nichtparametrischen Teil zusammen.

$$\eta_i = \underbrace{\beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik}}_{\text{parametrische Effekte}} + \underbrace{f_1(z_{i1}) + \dots + f_q(z_{iq})}_{\text{nichtparametrische Effekte}} + \epsilon_i \quad (4.14)$$

(? S.47 f.)

Bei den beiden angewandten Modelle handelt es sich um generalisierte additive Modelle mit einer Logit-Link-Funktion. Die glatten Funktionen werden mithilfe von P-Splines, zyklischen P-Splines und Thin-Plate-Splines geschätzt. Die Knoten

für die Penalisierten Splines wurden in den angewandten Modellen nach Perzentilen gesetzt. Das dient dazu, die Knoten dahin zu setzen wo vielen Beobachtungen sind, um an diesen Stellen mehr Flexibilität zu erreichen. Bevor im Kapitel (...)die konkreten Modelle vorgestellt werden, geht es im nächsten Abschnitt um aufgetretene Schwierigkeiten.

4.2 Praktische Herausforderungen

In diesem Teil der Arbeit wird ein Überblick über die Praktische Herausforderungen im Modell gegeben. Es wird als erstes das Concurvity-Problem erläutert, dann die Überdispersion und im letzten Abschnitt die Autokorrelation.

4.2.1 Concurvity

Die Kollinearität beschreibt einen linearen Zusammenhang zwischen den Kovariablen. Die Concurvity kann als Erweiterung der Kollinearität angesehen werden und charakterisiert die nicht-linearen Zusammenhänge von Kovariablen. Ein hohes Maß an Concurvity weist daher auf einen starken Zusammenhang zwischen zwei oder mehreren Kovariablen hin. Durch das Auftreten der Concurvity kommt es zu einer Art Varianz-Inflation der geschätzten Regressionskoeffizienten. Das heißt die Koeffizienten in dem generalisierten additiven Modell werden instabil. (? S.88 f.) Bei der Untersuchung der Kovariablen auf Concurvity zeigen sich einige leicht erhöhte Werte. Deshalb wird die Variable Schneehöhe zu Schneedifferenz transformiert. Die Schneedifferenz des jeweiligen Tages berechnet sich durch Subtraktion der Schneehöhe des vorherigen Tages von der Schneehöhe des jeweiligen Tages. Durch die Transformation der Kovariable Schneehöhe zu Schneedifferenz ist das Maß an Concurvity gesunken. In Abbildung (4.2) ist die neue Variable Schneedifferenz nach Datum zu sehen.

Abbildung 4.2: Liniendiagramm der Schneedifferenz zum Vortag in cm nach Datum (ausgegraute Werte an Tagen, die aus dem Datensatz entfernt wurden)

4.2.2 Überdispersion

Bei den bisher vorgestellten Modellen ist man von Individualdaten ausgegangen. Jeder Beobachtung wird genau ein Individuum aus der Stichprobe zugeordnet. Bei den angewandten Datensätzen im Modell handelt es sich nicht um Individualdaten, sondern um gruppierte Daten. Die Individualdaten werden nach identischen Zeilen in der Datenmatrix gruppiert. (? S.195 f.)

In der Arbeit werden zwei Modelle angewendet, das Tagesmodell und das Zeitmodell.

Im Tagesmodell werden die Daten nach Tag bzw. Datum gruppiert, wodurch die Datenmatrix insgesamt $n = 101$ Zeilen aufweist. Die Kovariable Uhrzeit wird dabei nicht beachtet. Da jedoch auch die Uhrzeit von Interesse für die Datenanalyse ist, wird ein zweites Modell, nämlich das Zeitmodell, aufgestellt. Im Zeitmodell werden die Individualdaten nach der Minute gruppiert und $n = 19334$ Beobachtungseinheiten erzeugt. Das Gruppieren von Individualdaten führt dazu, dass Anteile für Personen mit LVS-Gerät für das jeweilige Datum bzw. das Datum und die Uhrzeit berechnet werden.

Durch das Gruppieren nach Datum oder Minute kann es zur Überdispersion kommen. Das heißt die Daten weisen eine größere Streuung bzw. Varianz auf, als durch das Modell zu erwarten ist. Das Problem der Überdispersion wird durch das Einsetzen eines Dispersionsparameters in die Varianzformel gelöst.

$$Var(y_i|\mathbf{x}_i) = \underbrace{\phi}_{\text{Dispersionsparameter}} \underbrace{\pi_i(1 - \pi_i)}_{\text{Varianz}}. \quad (4.15)$$

Der Dispersionsparameter wird errechnet durch:

$$\phi = \frac{\text{Devianz}}{\text{Freiheitsgrade der Residuen}}. \quad (4.16)$$

((? S.195 f.))

4.2.3 Autokorrelation

Die Autokorrelation beschreibt die Korrelation einer Funktion mit sich selbst zu einem früheren Zeitpunkt. In Zeitreihen kann das Problem der Autokorrelation entstehen, wenn die Annahme der unkorrelierten Störgrößen verletzt ist. Man unterscheidet zwischen der empirischen Autokorrelation (ACF) und der partiellen Autokorrelation (PACF).

Die empirische Autokorrelation wird berechnet durch:

$$\widehat{\text{ACF}}(j) = \frac{\widehat{\text{Cov}}(\epsilon_i, \epsilon_{i-j})}{\widehat{\text{Var}}(\epsilon_i)} \quad \text{mit} \quad \widehat{\text{Cov}}(\epsilon, \epsilon_{i-j}) = \frac{1}{n} \sum_{i=j+1}^n \hat{\epsilon}_i \hat{\epsilon}_{i-j}. \quad (4.17)$$

Dabei sind ϵ_i die Residuen und ϵ_{i-j} die um j verzögerten Residuen. Das ACF gibt die Korrelation zwischen den Störungen ϵ_i und den um j Perioden verzögerten Störtermen an. Das PACF hingegen berechnet die Korrelation zwischen ϵ_i und ϵ_{i-j} , ohne den Einfluss der dazwischen liegenden Störungen zu beachten. Dabei ist:

$$\epsilon_i = \alpha_1 \epsilon_{i-1} + \dots + \alpha_j \epsilon_{i-j} + v_i, \quad (4.18)$$

wobei α_j der Regressionskoeffizient des Modells ist. Die partielle Autokorrelation PACF wird durch:

$$\widehat{\text{PACF}}(j) = \hat{\alpha}_j \quad (4.19)$$

bestimmt. (? S.137 ff.)

Beide Modelle wurden auf Autokorrelation geprüft. Um das Problem der Autokorrelation genauer zu analysieren, können die empirische Autokorrelation und partielle Autokorrelation anhand von sogenannten Korrelogrammen, wie in den Abbildungen (4.3) und (4.4), grafisch dargestellt werden. Die Korrelogramme zeigen niedrige empirische- und partielle Autokorrelation bei beiden Modellen, wodurch keine weiteren Maßnahmen zur Behebung der Autokorrelation erforderlich sind.

Abbildung 4.3: Korrelogramm der Autokorrelation nach Verzögerung (Lag)

Abbildung 4.4: Korrelogramm der partiellen Autokorrelation nach Verzögerung (Lag)

5 Angewandte Modelle

Nach der Erläuterung des theoretischen Hintergrunds folgen nun die beiden bereits angesprochenen Modelle. Zuerst wird der zuvor allgemein erklärte additive Prädiktor für die beiden angewandten Modelle definiert. Danach werden die Ergebnisse der Modelle vorgestellt.

5.1 Tagesmodell

Das Tagesmodell ist ein generalisiertes additives Modell, bei dem die Messungen nach Datum gruppiert sind. Der additive Prädiktor sieht damit wie folgt aus:

$$\begin{aligned}\eta_i = & \beta_0 + \beta_1(\text{Ferientag}_i) + \\ & f_1(\text{Datum}_i) + f_2(\text{Lawinenwarnstufe}_i) + \\ & f_3(\text{Wochentag}_i) + f_4(\text{Temperatur}_i) + \\ & f_5(\text{durchschnittliche Tages-Bewölkung}_i) + \\ & f_6(\text{Schneedifferenz}_i).\end{aligned}\tag{5.1}$$

5.2 Zeitmodell

Das zweite Modell, das Zeitmodell, stellt ebenfalls ein generalisiertes additives Modell dar, welches allerdings um die Variable Uhrzeit erweitert wurde und nach Minute gruppiert ist. Damit ergibt sich folgender additiver Prädiktor:

$$\begin{aligned}\eta_i = & \beta_0 + \beta_1(\text{Ferientag}_i) + \\ & f_{1,2}(\text{Uhrzeit}_i, \text{Datum}_i) + f_3(\text{Lawinenwarnstufe}_i) + \\ & f_4(\text{Wochentag}_i) + f_5(\text{Temperatur}_i) + \\ & f_6(\text{stündliche Bewölkung}_i) + \\ & f_7(\text{Schneedifferenz}_i).\end{aligned}\tag{5.2}$$

Bei der Funktion $f_{1,2}$ handelt es sich um einen sogenannten Interaktionseffekt. Das ist notwendig, wenn eine nicht-lineare Interaktion zwischen zwei oder mehreren Variablen besteht. Hier interagieren die Uhrzeit und das Datum miteinander.

5.3 Ergebnisse

Es werden die Ergebnisse des Tages- und des Zeitmodells vorgestellt und die beiden Modelle miteinander verglichen.

5.3.1 Ergebnisse des Tagesmodells

Das Modell umfasst $n = 101$ Tage, an denen Messungen vorliegen. Der Intercept ist der zu erwartende Anteil von Personen mit LVS-Gerät an einem Tag, der kein Ferientag ist und bei dem alle anderen Kovariablen durchschnittlich sind. Der Intercept des Tagesmodells beträgt 21,67%. An einem Ferientag mit ansonsten durchschnittlichen Variablen wird ein Anteil von Personen mit LVS-Gerät in Höhe von 19,71% erwartet.

Abbildung 5.1: Zusammenhang der Variablen Wochentag, Lawinenwarnstufe, durchschnittliche Tagesbewölkung, Temperatur, Schneedifferenz und Datum mit dem Anteil der Personen mit LVS-Gerät als glatte Funktionen für das Tagesmodell

In Abbildung (5.1) werden die glatten Funktionen der nichtparametrischen Variablen dargestellt. Im Folgenden werden diese einzeln beschrieben.

Wochentag:

An einem Montag, der kein Ferientag ist und an dem alle anderen Kovariablen durchschnittlich sind, liegt der zu erwartende Anteil von Personen mit LVS-Geräten bei etwas über 20%. Dienstags ist der Anteil etwas geringer und mittwochs höher. Am Donnerstag sinkt dieser auf unter 20%, um dann bis Samstag auf knapp 25% anzusteigen. Sonntags ist der Anteil etwas geringer als am Samstag.

Lawinenwarnstufe:

Bei einer Lawinenwarnstufe von 1 ist ein zu erwartender Anteil von Personen mit LVS-Gerät von knapp 20% zu erkennen. Je höher die Warnstufe desto höher ist auch der zu erwartende Anteil. Bei Lawinenwarnstufe 4 ist ein Anteil von ca. 30% zu erwarten. Zu beobachten ist, dass das 95%-Konfidenzintervall ebenfalls

breiter wird. Dies lässt sich darauf zurückführen, dass weniger Tage mit hohen Lawinenwarnstufen vorliegen.

Durchschnittliche Tages-Bewölkung:

Bei der glatten Funktion der Variable Bewölkung ist zu beobachten, dass der zu erwartende Anteil von Personen mit LVS-Geräten stark schwankt. Bei einer Bewölkung von 0% wird ein Anteil von etwas über 20% erwartet. Der Anteil der LVS-Geräte nimmt bis ca. 10% Bewölkung zu und fällt bei etwa 30% Bewölkung auf einen Wert von ungefähr 18%. Auch danach sind Schwankungen mit Hochpunkten bei ca. 45% und 85% Bewölkung ersichtlich. Niedrigere zu erwartende Anteile dagegen erreichen 60% und 100% Bewölkung.

Temperatur:

Der zu erwartende Anteil von Personen mit LVS-Geräten liegt bei negativen Temperaturen bei ungefähr 20%. Nahe dem Nullpunkt (0°C) steigt er und fällt dann wieder ab. Bei ca. 5°C liegt der Anteil erneut bei 20% und steigt danach wieder an.

Schneedifferenz:

Der zu erwartende Anteil steigt zwischen einer Schneedifferenz von 0 cm und 30 cm von etwa 20% auf 30% an. An Tagen mit negativer Schneedifferenz (also weniger Schnee als am Vortag) sinkt der Anteil leicht je näher die Schneedifferenz an 0 cm herankommt. Bei sehr hohen Schneedifferenzen sinkt dann der Anteil wieder. Allerdings zeigt sich dort ein breites Konfidenzintervall. Dies lässt sich durch wenige Tage mit hohen Schneedifferenzen erklären.

Datum:

Der zu erwartende Anteil von Personen mit LVS-Geräten schwankt sehr viel von unter 10% auf über 40%. Trotzdem lässt sich eine generelle Abnahme erkennen. Vorsicht ist bei der Interpretation an den fehlenden Tagen geboten, weshalb die Glättung an diesen Tagen transparenter dargestellt wurde.

5.3.2 Ergebnisse des Zeitmodells

Nachfolgend werden die Ergebnisse des Zeitmodells betrachtet. Hier wurden die Messungen nicht nach dem Tag, sondern nach der Minute gruppiert. Insgesamt gibt es $n = 19334$ Minuten, in denen mindestens eine Messung vorliegt.

Der Intercept des Zeitmodells liegt etwas niedriger als der des Tagesmodells, nämlich bei 20,32%. Der zu erwartende Anteil an einem Ferientag mit ansonsten durchschnittlichen Variablen ist dafür höher und liegt bei 21,40%.

In Abbildung (5.2) zu sehen sind die glatten Funktionen der nichtparametrischen Variablen (bis auf das Datum und die Uhrzeit der Messungen). Diese werden nun mit den jeweiligen glatten Funktionen des Tagesmodells verglichen.

Abbildung 5.2: Zusammenhang der Variablen Wochentag, Lawinenwarnstufe, stündliche Bewölkung, Temperatur und Schneedifferenz mit dem Anteil der Personen mit LVS-Gerät als glatte Funktionen für das Zeitmodell

Bei Wochentag und Lawinenwarnstufe ist ein ähnlicher Trend wie beim Tagesmodell zu erkennen. Die Glättung der Variable Bewölkung schwankt deutlich weniger und es lässt sich ein positiver Zusammenhang erkennen. Je bewölkter es ist, desto höher ist der zu erwartende Anteil der Personen mit LVS-Gerät.

Auch die glatte Funktion der Temperatur verhält sich ähnlich zum Tagesmodell. Bei Minusgraden ist der zu erwartende Anteil allerdings nicht mehr konstant, sondern nimmt mit sinkender Temperatur zu. Bei der glatten Funktion der Variable Schneedifferenz ist ein Trend zu beobachten. Wie auch beim Tagesmodell steigt der Anteil mit steigender Schneedifferenz. Allerdings schwankt der zu erwartende Anteil deutlich und sinkt nach einer Schneezunahme von 10 cm nochmal ab, bevor er wieder steigt.

In Abbildung (5.3) ist die Glättung der Variablen Uhrzeit und Datum zu erkennen. Über die Saison hinweg kann man eine generelle Abnahme des zu erwartenden Anteils von Personen mit LVS-Gerät erkennen.

Abbildung 5.3: Zusammenhang der Variablen Datum und Uhrzeit mit dem Anteil der Personen mit LVS-Gerät als glatte Funktionen für das Zeitmodell (Linien stellen Sonnenauf- und -untergang dar)

Betrachtet man die Uhrzeit, wird ersichtlich, dass morgens von ungefähr 7:00 bis 11:00 Uhr ein besonders hoher zu erwartender Anteil von bis zu 40% zu beobachten ist. Abends von 19:00 bis 22:00 ist ein geringer Anstieg des Anteils erkennbar. Das zu diesen Uhrzeiten ein erhöhter Anteil von Personen mit LVS-Gerät zu erwarten ist ändert sich auch im Verlauf der Saison kaum und scheint nicht mit der Uhrzeit des Sonnenauf- und -untergangs zusammenzuhängen.

5.3.3 Signifikanz

Neben dem Verlauf der glatten Funktionen werden auch die Signifikanzen der Variablen für beide Modellen betrachtet.

Zu beobachten ist, dass sich die Signifikanzen der einzelnen Variablen (auf einem 5%-Signifikanzniveau) zwischen den Modellen stark unterscheiden. Während beim Tagesmodell nur der Intercept, das Datum und die Lawinenwarnstufe signifikant sind, so sind beim Zeitmodell alle Variablen, bis auf Ferientag, und der Intercept signifikant.

Ob eine Variable signifikant ist, also ob der Einfluss der Kovariable auf die Zielvariable mehr als nur zufällig ist, hängt auch von der Anzahl der Beobachtungen (n) ab. Es ist nicht außer Acht zu lassen, dass das Tagesmodell $n = 101$ Beobachtungen aufweist, wohingegen das Zeitmodell $n = 19334$ Beobachtungen hat.

6 Messfehleranalyse

Wie bereits zuvor erwähnt, haben die Projektpartner festgestellt, dass die Checkpoints in ihrer Messung Messfehler haben.

Messfehler sind Abweichungen der erhobenen Daten vom wahren Wert. (?)

Das bedeutet, dass die Checkpointmessungen nicht exakt den vorbeigehenden Personen mit oder ohne LVS-Gerät entsprechen. Somit ist eine Unter-/Überschätzung des Anteils an LVS-Geräten möglich. Dabei können Messfehler verschiedener Art vorkommen, die jeweils eigene Ursachen haben, und so die Stärke und Richtung der Fehlschätzung unterschiedlich beeinflussen.

Im folgenden Teil werden erst von den Projektpartnern gesammelte Daten über die Genauigkeit der Checkpointmessungen an zwei Tagen untersucht und danach verschiedene Szenarien für mögliche Messfehler erstellt. Anhand des jeweiligen Szenarios werden Messungen aus dem originalen Datensatz ergänzt oder entfernt und es wird verglichen, inwiefern sich die Ergebnisse der zwei Modelle für die gewählten Szenarien unterscheiden.

6.1 Messfehler Beschreibung

Um die Schwere und Art der Messfehler zu untersuchen, wurde eine Gruppe von Studierenden mehrmals zum Aufstiegsort geschickt. Unter Anderem wurden dabei verschiedene Tests angestellt, z.B. wurde die Erfassung der Personen abhängig von unterschiedlichen Entfernungen und Winkeln zum Checkpoint untersucht oder man ließ einen Hund mehrmals am Checkpoint vorbeilaufen und markierte, wann dieser vom Checkpoint registriert wurde und wann nicht. Für die folgende Analyse werden Daten benutzt, die von einer manuellen Zählung einer solchen Gruppe von Studierenden am 27.02.- und 28.02.2020 stammen. Genauer gesagt fanden die Zählungen zu drei unterschiedlichen Zeiträumen statt: am 27.02. von 10:32 bis 12:58 Uhr, am 28.02. von 10:38 bis 12:26 Uhr und von 14:40 bis 16:50 Uhr.

Dabei wurde eingetragen, wie viele Kontakte zu jeder Minute am Checkpoint vor-

beigegangen sind. Zudem werden die Kontakte auf zwei verschiedene Arten unterschieden. Einerseits gibt es eine Unterscheidung zwischen „Skitourengeher“, die eigentlich zu interessierende Art von Besuchern mit eventuellen LVS-Geräten, und „andere Kontakte“, also andere Besucher wie Spaziergänger oder z.B. ein Hund. Andererseits wird zusätzlich zwischen „erfassten“ und „nicht erfassten“ Kontakten unterschieden, wobei ein Kontakt als erfasst gilt, wenn er augenscheinlich vom Checkpoint erkannt und registriert wurde.

Wichtig zu erwähnen ist, dass bei der manuellen Zählung keine Unterscheidung zwischen Besuchern mit und ohne LVS-Gerät gemacht wurde. Zusätzlich zu den manuellen Zählungen liegen die tatsächlich von den Checkpoints erhobenen Messungen zur Saison 19/20 vor, wobei der Checkpoint-Süd früh in der Saison ausgefallen ist und zu den zwei Tagen der manuellen Zählungen somit nur ein Checkpoint aktiv war.

6.2 Deskriptive Analyse der Messfehlerdaten

Die Anzahl der in der manuellen Zählung eingetragenen Kontakte beträgt insgesamt 208. Davon sind, siehe Abbildung (6.1), 122 Skitourengeher und 86 andere Kontakte. Als erfasst vom Checkpoint wurden 161 Kontakte eingetragen, als nicht erfasst somit 47.

Da die angegebenen Uhrzeiten in der manuellen Zählung und der tatsächlichen Registrierung im Checkpoint voneinander abweichen, werden beim genaueren Vergleich von Erfassungen und Nichterfassungen nur die Daten der manuellen Zählung hergenommen. Für den gesamten Zeitraum der manuellen Zählung kann man jedoch die tatsächlich vom Checkpoint eingetragene Anzahl an erfassten Kontakten hernehmen, die mit 170 etwas höher ist als die 161 von den Studierenden angegebenen Nichterfassungen. Für die Berechnung der Unterschätzung insgesamt wird deshalb die Anzahl der vom Checkpoint gemessenen Kontakte, also 170, mit den insgesamt durch die manuelle Zählung festgehaltenen Kontakten, also 208, verglichen. Damit kommt man auf eine allgemeine Unterschätzung des Checkpoints von ca. 22%. Hierbei wird „Unterschätzung“ definiert als der Anteil an vorhandenen Messungen, die man dem Checkpoint hinzufügen müsste, um auf den wahren Wert der vorbeigehenden Kontakte zu bekommen. Dem Checkpoint müsste man ca. 22% von 170 erfassten Kontakten hinzufügen, um auf den wahren Wert von 208 Kontakten zu kommen.

Abbildung 6.1: Balkendiagramm der absoluten Häufigkeit der gemessenen Kontakte in der manuellen Zählung, aufgeteilt nach Art des Kontakts und Erfassung/Nichterfassung

Für die Berechnung der Unterschätzung wurden keine Messungen aus den manuellen Zählungen entfernt. Insbesondere auch keine Fälle, zu denen angegeben war, dass das Gerät durch fälschliche Messungen wie Mehrfacherfassungen einer Person den wahren Wert überschätzt. In diesem Sinne ist die berechnete Unterschätzung keine „reine“ Unterschätzung sondern die Unterschätzung für alle gemessenen Beobachtungen. Um zu testen, wie genau die Erfassung des Checkpoints bei größeren Gruppen an vorbeigehenden Kontakten funktioniert, wurde eine Untersuchung des Anteils an erfassten Kontakten in Abhängigkeit von Gruppengröße durchgeführt. Als Gruppe zählen hierbei alle Kontakte, die zur selben Uhrzeit in der manuellen Zählung vermerkt wurden. Sind also beispielsweise um 15:23 vier Kontakte in der manuellen Zählung vermerkt, so gilt dies als Gruppe von vier Personen. In den Daten kommen jeweils mehrere Gruppen in einer Größe von 1 bis 5 und eine Gruppe von 8 Kontakten vor.

Abbildung 6.2: Gestapeltes Balkendiagramm für das Verhältnis erfasster zu nicht erfasster Kontakte nach Gruppengröße

Wie man in der Abbildung (6.2) sieht, schwankt der Anteil von Nichterfassungen an allen Messungen bei den fünf kleineren Gruppengrößen um einen Wert von 18%, der einer Unterschätzung von 22% ($0.18/0.82 = 0.22$) entspricht. In der Achtergruppe ist der Anteil von Nichterfassungen allerdings über 50%, d.h. es wurden in dieser Gruppe mehr Kontakte vom Checkpoint nicht erkannt als erkannt.

6.3 Messfehlerszenarien im Vergleich

Ursprünglich war mit den Projektpartnern vereinbart, die Daten zu den Messfehlern herzunehmen, um das bereits aufgestellte Modell auf Grundlage der Checkpointdaten der Saison 18/19 anzupassen und das neue, verbesserte Modell an den

Daten zur Saison 19/20 zu testen. Da das Frühjahr 2020, zu dem die manuellen Zählungen geplant waren, von den Wetterbedingungen sehr milde war und somit vergleichsweise sehr wenige Besucher am Untersuchungsort unterwegs waren, ist der Umfang der Daten zu den Messfehlern allerdings deutlich geringer als vorgesehen. Da die meisten Kovariablen im Modell täglich erhoben werden und für die Messfehler nur Daten zu zwei Tagen mit nur insgesamt 208 Messungen vorliegen, ist der Informationsgehalt zu gering für eine sinnvolle Anpassung der Modelle.

Stattdessen wurden in Abstimmung mit den Projektpartnern aus den Erkenntnissen über die Messfehler vier verschiedene Szenarien entworfen. Für jedes Szenario wird der originale Datensatz aus der Saison 18/19 angepasst und die bestehenden Modelle auf den jeweils neu erzeugten Datensatz angewandt. Die Ergebnisse für die unterschiedlichen Szenarien werden mit den originalen Ergebnissen verglichen, um Aussagen darüber treffen zu können, inwiefern sich unterschiedliche Messfehler auf die Ergebnisse des Modells auswirken. Da aus den Tests hervorgegangen ist, dass der Checkpoint bei der Messung von LVS-Geräten sehr genau ist und Messfehler eher bei der Infrarotmessung entstehen, werden in jedem Szenario nur Messungen ohne LVS-Gerät hinzugefügt (bzw. entfernt).

Szenario 1 geht von einer generellen Unterschätzung von 22% aus, nach dem Wert aus der Analyse der Messfehlerdaten. Hier werden dem Datensatz zu zufälligen Zeitpunkten insgesamt 22% neue Beobachtungen von Kontakten ohne LVS-Gerät hinzugefügt.

Szenario 2 geht von einer ansteigenden Unterschätzung bei zunehmender Gruppengröße aus. Hier wird für jede Beobachtung mit einer bestimmten Wahrscheinlichkeit eine neue Beobachtung hinzugefügt, so dass bei größeren Gruppen im Durchschnitt auch mehr neue Beobachtungen hinzukommen.

Szenario 3 geht von einer nächtlichen Überschätzung aus. Die Tests haben gezeigt, dass teilweise auch ein vorbeilaufender Hund vom Checkpoint erfasst wird. Da nachts wenige Menschen unterwegs und besonders in den Dämmerungszeiten vermehrt Tiere unterwegs sind, wird nachts ein fester Anteil an Beobachtungen ohne LVS-Gerät aus dem Datensatz entfernt, der für irrtümlich erfasste Tiere steht.

Szenario 4 geht von einer wachsenden Unterschätzung bei fallenden Temperaturen aus. Hier wird davon ausgegangen, dass der Checkpoint bei größerer Kälte fehleranfälliger wird, und es wird für sinkende Temperaturen ein bestimmter Anteil an neuen Beobachtungen hinzugefügt.

Die Ergebnisse für die Szenarien im Vergleich zum originalen Datensatz werden

erneut erst für das Tagesmodell, danach für das Zeitmodell verglichen.

6.3.1 Im Tagesmodell

Die Tabelle (6.1) zeigt im Tagesmodell die Werte für die Koeffizienten von Intercept und Ferientag, die p-Werte und Signifikanzen für alle Kovariablen für den originalen Datensatz und jedes Messfehlerszenario im Vergleich.

	Original			Szenario 1			Szenario 2			Szenario 3			Szenario 4		
	Koeffizient	p-Wert		Koeffizient	p-Wert		Koeffizient	p-Wert		Koeffizient	p-Wert		Koeffizient	p-Wert	
(Intercept)	0.217	<2e-16	***	0.158	<0.01	***	0.174	<2e-16	***	0.232	<2e-16	***	0.172	<2e-16	***
Ferientag	0.197	0.748		0.162	0.94		0.156	0.701		0.216	0.81		0.159	0.785	
s(Datum)		0.005	**		<0.01	***		0.005	**		0.02	*		0.007	**
s(Lawinengefahr)		0.030	*		0.03	*		0.035	*		0.06	.		0.030	*
s(Wochentag)		0.198			0.05	.		0.177			0.15			0.207	
s(Temperatur)		0.713			0.70			0.681			0.69			0.639	
s(Bewölkung)		0.371			0.85			0.278			0.24			0.365	
s(Schneedifferenz)		0.635			0.48			0.647			0.72			0.651	

Tabelle 6.1: Tabelle mit Koeffizienten und p-Werten für das Original, Szenario 1 (Generelle Unterschätzung von 22%), Szenario 2 (Unterschätzung nach Gruppengrösse), Szenario 3 (Nächtliche Überschätzung) und Szenario 4 (Unterschätzung nach Temperatur) für das Tagesmodell

Da in den Szenarien nur Beobachtungen ohne LVS-Gerät hinzugefügt bzw. entfernt werden, sinken bzw. steigen die Koeffizienten für die zwei parametrischen Variablen entsprechend. Szenario 1 mit genereller Unterschätzung hat die höchste Menge an hinzugefügten Beobachtungen, deswegen ist der Intercept, also der Anteil an Messungen mit LVS-Geräten bei konstant gehaltenen Kovariablen und keinem Ferientag, mit einem Wert von 15,8% am geringsten. Szenario 3 mit nächtlicher Überschätzung ist das einzige Szenario, bei dem Beobachtungen entfernt werden, deswegen steigt hier der Koeffizient für den Intercept auf einen Wert von 23,2%. Analog gelten die Aussagen für den Anteil an Ferientagen bei konstant gehaltenen Kovariablen. Bei der generellen Unterschätzung ist der Wert für Ferientage mit 16,2% höher als der Wert für den Intercept.

Bei einem Signifikanzniveau von 5% haben nur der Intercept und das Datum bei allen Szenarien wie im Original einen statistisch signifikanten Zusammenhang mit dem Anteil an LVS-Geräten. Die Kovariable für Lawinengefahr hat im Original und bei jedem Szenario außer der nächtlichen Überschätzung einen statistisch signifikanten Zusammenhang. Alle anderen Kovariablen, also Ferientag, Bewölkung, Wochentag, Temperatur, Bewölkung und Schneedifferenz, haben wie im Original

keinen statistischen Zusammenhang mit dem Anteil auf dem gewählten Signifikanzniveau.

Die Abbildung (6.3) zeigt im Tagesmodell den Zusammenhang zwischen Anteil an LVS-Geräten und allen nonparametrischen Kovariablen im Original und den verschiedenen Messfehlerszenarien.

Abbildung 6.3: Zusammenhang der Variablen Wochentag, Lawinenwarnstufe, durchschnittliche Tagesbewölkung, Temperatur, Schneedifferenz und Datum mit dem Anteil der Personen mit LVS-Gerät als glatte Funktionen im Tagesmodell für den originalen Datensatz und jedes Szenario

Daran lässt sich feststellen, dass die Zusammenhänge der einzelnen Kovariablen mit dem Anteil an LVS-Geräten für alle Szenarien zumeist einen ähnlichen Verlauf nehmen wie im Original. So hat die glatte Funktion z.B. für den Wochentag bei jedem Szenario wie im Original höhere Werte für Montag, Mittwoch und Samstag als für Dienstag, Donnerstag oder Freitag. Auch für die anderen Kovariablen unterscheidet sich der Verlauf des Zusammenhangs insbesondere für Szenarien 2 bis 4 nur geringfügig vom Original.

Einzig in Szenario 1, der generellen Unterschätzung, sind an mehreren Stellen größere Unterschiede zu bemerken. So schwankt der Anteil im Zusammenhang mit der Bewölkung weniger stark als im Original und den anderen Szenarien. Für niedrige Temperaturen bis -3 Grad sinkt der Anteil im Gegensatz zu den anderen Szenarien, verläuft danach aber ähnlich. Bei einer Schneedifferenz von 5 bis 20 cm steigt der Anteil geringer als in den anderen Szenarien bis nur ca. 15%, für Werte zwischen 20 und 32 cm dagegen steigt der Anteil in Szenario 1 deutlich stärker und sinkt danach auch deutlich stärker ab als in den anderen Szenarien, mit einem Höchstwert von ca. 28%.

Wie schon erwähnt ist die Unsicherheit in diesem Bereich allerdings erhöht, da nur wenige Tage mit hohen Werten für die Schneedifferenz vorkommen. Beim Datum unterscheidet sich Szenario 1 im Bereich vom 01. März bis zum 15. März, da der Anteil dort fast konstant bei ca. 15% liegt und nicht wie in den anderen Szenarien ansteigt und wieder abfällt.

6.3.2 Im Zeitmodell

Die Tabelle (6.2) zeigt im Zeitmodell die Werte für die Koeffizienten von Intercept und Ferientag, die p-Werte und Signifikanzen für alle Kovariablen für den originalen Datensatz und jedes Messfehlerszenario im Vergleich.

	Original			Szenario 1			Szenario 2			Szenario 3			Szenario 4		
	Koeffizient	p-Wert		Koeffizient	p-Wert		Koeffizient	p-Wert		Koeffizient	p-Wert		Koeffizient	p-Wert	
(Intercept)	0.203	<2e-16 ***		0.127	<2e-16 ***		0.163	<2e-16 ***		0.218	<2e-16 ***		0.119	<2e-16 ***	
Ferientag	0.214	0.2		0.126	0.9		0.170	0.3		0.229	0.2		0.122	0.5	
s(Datum,Uhrzeit)	x	<2e-16 ***		x	<2e-16 ***		x	<2e-16 ***		x	<2e-16 ***		x	<2e-16 ***	
s(Lawinengefahr)	x	6.00E-13 ***		x	<2e-16 ***		x	5.00E-13 ***		x	1.00E-12 ***		x	<2e-16 ***	
s(Wochentag)	x	2.00E-10 ***		x	3.00E-15 ***		x	6.00E-10 ***		x	6.00E-10 ***		x	1.00E-10 ***	
s(Temperatur)	x	5.00E-10 ***		x	<2e-16 ***		x	4.00E-10 ***		x	1.00E-08 ***		x	1.00E-10 ***	
s(Bewölkung)	x	8.00E-16 ***		x	5.00E-13 ***		x	4.00E-16 ***		x	5.00E-16 ***		x	<2e-16 ***	
s(Schneedifferenz)	x	6.00E-08 ***		x	7.00E-09 ***		x	2.00E-08 ***		x	1.00E-07 ***		x	5.00E-08 ***	

Tabelle 6.2: Tabelle mit Koeffizienten und p-Werten für das Original, Szenario 1 (Generelle Unterschätzung von 22%), Szenario 2 (Unterschätzung nach Gruppengrösse), Szenario 3 (Nächtliche Überschätzung) und Szenario 4 (Unterschätzung nach Temperatur) für das Zeitmodell

Ähnlich wie beim Tagesmodell verhalten sich die Koeffizienten für den Intercept und Ferientag. In den Szenarien mit Unterschätzung sinken beide Anteile im Vergleich zum Original bei einem Tiefstwert von 12,7% für den Intercept bei der generellen Unterschätzung und einem erhöhten Wert von 21,8% beim Szenario der nächtlichen Überschätzung, analog für den Anteil an Ferientagen. Bei der generellen Unterschätzung ist der Anteil an Ferientagen mit 12,6% geringer als an Tagen, die keine Ferientage sind.

Wie im Original sind in allen Szenarien alle Variablen bis auf Ferientag statistisch signifikant.

Abbildung 6.4 zeigt im Zeitmodell den Zusammenhang zwischen Anteil an LVS-Geräten und allen nonparametrischen Kovariablen im Original und den verschiedenen Messfehlerszenarien.

Abbildung 6.4: Zusammenhang der Variablen Wochentag, Lawinenwarnstufe, stündliche Bewölkung, Temperatur und Schneedifferenz mit dem Anteil der Personen mit LVS-Gerät als glatte Funktionen im Zeitmodell für den originalen Datensatz und jedes Szenario

Wie im Tagesmodell verlaufen die Zusammenhänge der einzelnen Kovariablen mit dem Anteil an LVS-Geräten bei allen Szenarien zumeist ähnlich zum Original.

Größere Unterschiede bestehen bei der Bewölkung, wo für das Szenario mit genereller Unterschätzung und einer Unterschätzung bei niedrigen Temperaturen ein konstanter Anstieg bei wachsender durchschnittlicher Bewölkung zu beobachten ist, und bei der Schneedifferenz, wo im Bereich von 10 bis 20 cm der Anteil im Szenario mit Unterschätzung bei niedrigen Temperaturen verhältnismäßig weniger Schwankungen unterliegt.

Auffällig ist, dass der Anteil bei Werten für die Schneedifferenz von 20 cm und mehr für das Original und Szenarien 2 bis 4 nun im Vergleich zum Tagesmodell stärker ansteigt und wieder sinkt und somit dem Verlauf von Szenario 1 ähnlicher ist.

6.4 Szenario mit steigender genereller Unterschätzung

Für das Szenario mit genereller Unterschätzung wird im obigen Vergleich eine Unterschätzung von 22% angenommen.

Im Folgenden wird verglichen, wie sich die Ergebnisse der beiden Modelle für das Szenario mit allgemeiner Unterschätzung verhalten, wenn man unterschiedliche Werte für die Unterschätzung annimmt. Dafür wird die Unterschätzung in 10%-Schritten bis zu einem Wert von 70% erhöht.

6.4.1 Im Tagesmodell

Die Tabelle (6.3) zeigt im Tagesmodell die Werte für die Koeffizienten von Intercept und Ferientag, die p-Werte und Signifikanzen für alle Kovariablen für den originalen Datensatz und steigender allgemeiner Unterschätzung.

	Original			10%			20%			30%		
	Koeffizient	p-Wert		Koeffizient	p-Wert		Koeffizient	p-Wert		Koeffizient	p-Wert	
(Intercept)	0.217	<2e-16	***	0.182	<2e-16	***	0.162	<2e-16	***	0.145	<2e-16	***
Ferientag	0.197	0.748		0.181	0.984		0.164	0.97		0.149	0.94	
s(Datum)		0.005	**		0.003	**		<1e-03	***		<1e-03	***
s(Lawinengefahr)		0.030	*		0.011	*		0.02	*		0.02	*
s(Wochentag)		0.198			0.117			0.08	.		0.03	*
s(Temperatur)		0.713			0.720			0.69			0.75	
s(Bewölkung)		0.371			0.814			0.89			0.82	
s(Schneedifferenz)		0.635			0.527			0.42			0.40	

	40%			50%			60%			70%		
	Koeffizient	p-Wert		Koeffizient	p-Wert		Koeffizient	p-Wert		Koeffizient	p-Wert	
(Intercept)	0.134	<2e-16	***	0.122	<2e-16	***	0.115	<2e-16	***	0.107	<2e-16	***
Ferientag	0.131	0.95		0.126	0.93		0.113	0.966		0.103	0.91	
s(Datum)		<1e-03	***		<1e-03	***		<1e-03	***		<1e-03	***
s(Lawinengefahr)		0.02	*		0.03	*		0.012	*		0.01	*
s(Wochentag)		0.02	*		0.02	*		0.009	**		0.01	*
s(Temperatur)		0.78			0.62			0.724			0.65	
s(Bewölkung)		0.81			0.78			0.732			0.63	
s(Schneedifferenz)		0.43			0.25			0.214			0.24	

Tabelle 6.3: Tabelle mit Koeffizienten und p-Werten der Kovariablen für steigende Werte an allgemeiner Unterschätzung im Tagesmodell

Da für steigende Unterschätzung ein größerer Anteil an Beobachtungen ohne LVS-Geräte hinzugefügt wird, sinken die Werte für Intercept und Ferientag. Einen höheren Anteil an Ferientagen gibt es bei einer Unterschätzung von 20%, 30% und 50%.

Dies wird allerdings hauptsächlich davon beeinflusst, an welchen Tagen neue Beobachtungen zufällig generiert werden. Wie im Original sind in jedem Fall der Intercept, Datum und Lawinengefahr statistisch signifikant, ab einer Unterschätzung von 30% ist auch Wochentag signifikant.

Abbildung (6.5) zeigt im Tagesmodell den Zusammenhang zwischen Anteil an LVS-Geräten und nonparametrischer Kovariablen im Original und steigenden Werten für den Anteil an Unterschätzung.

Abbildung 6.5: Zusammenhang der Variablen Wochentag, Lawinenwarnstufe, durchschnittliche Tagesbewölkung, Temperatur, Schneedifferenz und Datum mit dem Anteil der Personen mit LVS-Gerät als glatte Funktionen im Tagesmodell für den originalen Datensatz und steigende Werte an allgemeiner Unterschätzung

Die Verläufe des Zusammenhangs zwischen Anteil an LVS-Geräten und den ein-

zelenen Kovariablen sind für alle Anteile an Unterschätzung sehr ähnlich. Eine kleine Auffälligkeit ist ein größeres Abflachen des Anteils an LVS-Geräten bei einer Lawinenwarnstufe ab 3 bei Anteilen an Unterschätzung von 40% und höher. Unterschiede zum Original verhalten sich bei allen Werten für Unterschätzung wie vorher für eine Unterschätzung von 22% beschrieben.

6.4.2 Im Zeitmodell

Die Tabelle (6.4) zeigt im Zeitmodell die Werte für die Koeffizienten von Intercept und Ferientag, die p-Werte und Signifikanzen für alle Kovariablen für den originalen Datensatz und steigender allgemeiner Unterschätzung.

	Original			10%			20%			30%		
	Koeffizient	p-Wert		Koeffizient	p-Wert		Koeffizient	p-Wert		Koeffizient	p-Wert	
(Intercept)	0.203	<2e-16	***	0.165	<2e-16	***	0.133	<2e-16	***	0.107	<2e-16	***
Ferientag	0.214	0.2		0.167	0.7		0.133	0.9		0.105	0.6	
s(Datum,Uhrzeit)		<2e-16	***		<2e-16	***		<2e-16	***		<2e-16	***
s(Lawinengefahr)		<2e-16	***		<2e-16	***		<2e-16	***		<2e-16	***
s(Wochentag)		<2e-16	***		<2e-16	***		<2e-16	***		<2e-16	***
s(Temperatur)		<2e-16	***		<2e-16	***		<2e-16	***		<2e-16	***
s(Bewölkung)		<2e-16	***		<2e-16	***		<2e-16	***		<2e-16	***
s(Schneedifferenz)		<2e-16	***		<2e-16	***		<2e-16	***		<2e-16	***

	40%			50%			60%			70%		
	Koeffizient	p-Wert		Koeffizient	p-Wert		Koeffizient	p-Wert		Koeffizient	p-Wert	
(Intercept)	0.0899	<2e-16	***	0.0771	<2e-16	***	0.0664	<2e-16	***	0.0582	<2e-16	***
Ferientag	0.0851	0.1		0.0735	0.2		0.0620	0.05	*	0.0537	0.02	*
s(Datum,Uhrzeit)		<2e-16	***		<2e-16	***		<2e-16	***		<2e-16	***
s(Lawinengefahr)		<2e-16	***		<2e-16	***		<2e-16	***		<2e-16	***
s(Wochentag)		<2e-16	***		<2e-16	***		<2e-16	***		<2e-16	***
s(Temperatur)		<2e-16	***		<2e-16	***		<2e-16	***		<2e-16	***
s(Bewölkung)		<2e-16	***		<6e-14	***		<6e-14	***		<6e-14	***
s(Schneedifferenz)		<2e-16	***		<2e-16	***		<2e-16	***		<2e-16	***

Tabelle 6.4: Tabelle mit Koeffizienten und p-Werten der Kovariablen für steigende Werte an allgemeiner Unterschätzung im Zeitmodell

Analog wie im Tagesmodell sinken die Koeffizienten für Intercept und Ferientag mit wachsender Unterschätzung. Ab einer Unterschätzung von 30% ist der Anteil an LVS-Geräten an Ferientagen geringer.

Wie im Original sind für alle Werte für die Unterschätzung alle Kovariablen bis auf Ferientag signifikant, ab einer Unterschätzung von 60% ist auch Ferientag signifikant.

Abbildung (6.6) zeigt im Zeitmodell den Zusammenhang zwischen Anteil an LVS-

Geräten und nonparametrischen Kovariablen im Original und steigenden Werten für den Anteil an Unterschätzung.

Abbildung 6.6: Zusammenhang der Variablen Wochentag, Lawinenwarnstufe, stündliche Bewölkung, Temperatur und Schneedifferenz mit dem Anteil der Personen mit LVS-Gerät als glatte Funktionen im Zeitmodell für den originalen Datensatz und steigende Werte an allgemeiner Unterschätzung

Wie im Tagesmodell verlaufen die Zusammenhänge zwischen Anteil an LVS-Geräten und den einzelnen Kovariablen für alle Werte an Unterschätzung sehr ähnlich. Größere Unterschiede gibt es weder untereinander, noch zum Original.

7 Fazit

Um die Auswirkungen von Umwelteinflüssen auf die Mitnahme von LVS-Geräten zu untersuchen, wurden zwei Modelle gebildet: einmal wurde dabei nach Datum gruppiert und einmal nach der Uhrzeit.

Beim nach Datum gruppierten Modell, dem Tagesmodell, wurden dabei Wochentag, Ferientag, Lawinenwarnstufe, tägliche Durchschnittsbewölkung, Temperatur, Schneedifferenz und Datum als Kovariablen berücksichtigt. Als signifikant haben sich der Intercept, die Lawinenwarnstufe und das Datum herausgestellt. Bei der Betrachtung des Zeitmodells wurde statt der täglichen Durchschnittsbewölkung die stündliche Bewölkung verwendet und außerdem Uhrzeit als weitere Variable mit aufgenommen. Hier zeigen sich alle Variablen bis auf Ferientag signifikant. Beide Modelle zeigen eine Abnahme des Anteils von Personen mit LVS-Gerät über die Saison hinweg. Außerdem ist zu erkennen, dass bei höheren Lawinenwarnstufen auch höhere Anteile zu erwarten sind. Das Zeitmodell zeigt, dass Uhrzeit ebenfalls eine Rolle spielt.

Eine Analyse der Daten zu den Messfehlern hat ergeben, dass der wahre Wert an vorbeigehenden Kontakten durch den Checkpoint um ca. 22% unterschätzt wird. Eine besonders hohe Unterschätzung liegt bei einer großen Gruppe an zur selben Minute vorbeigehender Kontakte vor. Ein Vergleich der Modellergebnisse des originalen Datensatzes und verschiedenen Messfehlerszenarien hat ergeben, dass der Verlauf der Zusammenhänge zwischen Anteil an LVS-Geräten und Kovariablen in beiden Modellen für alle Szenarien ziemlich ähnlich ist. Einzig das Szenario mit genereller Unterschätzung zeigt im Tagesmodell mehrere größere Unterschiede, ansonsten kommen diese nur an vereinzelten Stellen vor.

Jedoch ändert sich die Höhe des zu erwartenden Anteils je nach Szenario. Auch ein näherer Blick auf das Szenario mit genereller Unterschätzung zeigt bei steigenden Werten für die Unterschätzung nur kleinere Unterschiede zwischen den Szenarien. Auch hier verlaufen die Zusammenhänge recht ähnlich und unterscheiden sich nur in der zu erwartenden Höhe des Anteils von Personen mit LVS-Gerät. Messfeh-

ler, welche den vorgestellten Szenarien entsprechen, ändern somit zwar den Anteil an Personen mit LVS-Gerät den man zu erwarten hat, jedoch nicht die Richtung des Zusammenhanges.

Dies führt zu dem Ergebnis, dass, auch bei fehlerhaften Messgeräten, über die Saison hinweg weniger Besucher LVS-Geräte bei sich tragen. Bei höheren Lawinenwarnstufen tragen mehr Besucher ein LVS-Gerät bei sich und über die Saison hinweg ändern sich die Zeiten, zu denen mehr Besucher LVS-Geräte bei sich tragen, nicht.

8 Ausblick

Für eine mögliche verbesserte Genauigkeit des Modells und eine eventuelle Reduzierung der Concurvity zwischen Kovariablen könnte es sinnvoll sein, weitere Kovariablen hinzuzufügen oder mit bestehenden Kovariablen auszutauschen. Dazu würden sich z.B. Wetterdaten über Windstärke oder Niederschlag eignen. Außerdem könnten stündliche statt tägliche Werte für die Kovariablen die Modellgüte verbessern.

Für eine weitere Betrachtung der Auswirkung der Messfehler könnten zusätzliche Szenarien erstellt werden. Außerdem könnte man vorhandene Szenarien miteinander verbinden, z.B. eine Unterschätzung nach Gruppengröße und Temperatur zusammen, um einen höheren Grad an Komplexität in den Szenarien und damit eine mögliche Näherung an die tatsächliche Zusammensetzung der Messfehler zu testen.

Weitere Feldstudien zu den Messfehlern und die damit erhöhte Menge an Daten könnte man nutzen, um genauere Informationen über die Richtungen der Messfehler zu sammeln und damit die Modelle selbst anzupassen.

9 Umsetzung in R

Für die technische Umsetzung der beschriebenen Methoden, Berechnungen und die Erstellung der Grafiken wurde das Softwareprogramm R (Version 4.0.0) und die Entwicklungsumgebung RStudio benutzt.

Verwendete Packages:

tidyverse (für die Datenbearbeitung und Grafikerstellung)

patchwork (für das Anzeigen von mehreren Grafiken in Rastern)

GGally (für die Darstellung mehrerer Grafiken)

mgcv (für die Umsetzung der Modelle)

lubridate (für die Datenbearbeitung mit Datum- und Zeitvariablen)

gridExtra (für die Darstellung der glatten Funktionen in einem Raster)

gtools (für die Formattierung von p-Werten)

cowplot (für das Abspeichern einer Legende als eigenes Objekt)

readxl (für das Einlesen von Exceldateien)

Der verwendete R-Code liegt der Abgabe bei.