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

Abschlussbericht zum Projekt
LVS-IR-Taubenstein

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

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

Als beliebtes Ziel für Touristen und Wintersportler besteht im Alpengebiet eine besondere Konfliktsituation zwischen Mensch- und Tierreich. Routen für Spaziergänger, Skifahrer und Skitourengänger grenzen oft direkt an Lebensräume von Wildtieren an und führen so zu Stress für das Wildtierreich. Die vom Deutschen Alpenverein (DAV) in Kooperation mit dem Freistaat Bayern auf den Weg gebrachte Kampagne „Natürlich auf Tour“ soll für eine Sensibilisierung und Informationsgebung rund um das Thema Naturschutz dienen.

Neben der Aufklärung ist ein weiteres Ziel dabei, das Verhalten der menschlichen Besucher zu analysieren um daraus abzuleiten inwiefern man es womöglich steuern kann. In diesem Sinne untersuchte der DAV in Zusammenarbeit mit dem Departement für Geographie der LMU am Berg Taubenstein im Mangfallgebirge rund um den Spitzingsee in der Saison 18/19 und 19/20 den Anteil der Skitourengänger mit sogenannten LVS-Geräten. LVS-Gerät ist die Abkürzung für Lawinenverschüttetensuchgerät, mit Hilfe dieser Geräte können von Lawinen verschüttete Personen schnell gefunden werden. Personen, die ein LVS-Gerät dabeihaben, können mit diesem andere LVS-Geräte suchen und auch selbst gefunden werden.

Anhand der zur Verfügung gestellten Daten zur Saison 18/19 wird durch ein Modell der Anteil der Skitourengänger mit LVS-Gerät in Abhängigkeit von anderen Faktoren (wie z.B. Uhrzeit, Temperatur, Schneehöhe) analysiert . Zudem wird untersucht, von welchen EinflussgröSSen die Messfehler abhängen, welcher Art (Über-/Unterschätzung) sie sind und ob eine Struktur (mögl. Verteilung) vorliegt. Unter Berücksichtigung der Erkenntnisse über die Messfehler werden Hypothesen aufgestellt, in welcher Form die Messfehler die geschätzten Abhängigkeiten beeinflussen.

Noch schreiben, was wir rausgefunden haben.

Quelle: `https://de.wikipedia.org/wiki/Lawinenverschüttetensuchgerät`

2 Datenbasis

2.1 Datengrundlage

Um auf den Anteil der Skitourengängen, die ein LVS-Gerät bei sich hatten, schlieSSen zu können, hat man Checkpoints aufgestellt. Für den Aufstieg am Taubenstein gibt es zwei Routen, eine Nord- und eine Südroute. Die genauen Lagen kann man Abbildung 1 entnehmen. An beiden Routen wurde jeweils ein Checkpoint aufgestellt, an dem vorbeigehende Besucher gemessen werden. Durch Infrarotmessung wird erfasst, ob ein Mensch am Checkpoint vorbeigeht. AuSSerdem werden LVS-Geräte, die auf Sendebetrieb geschaltet sind, erfasst. Für jede einzelne Checkpointmessung liegt das jeweilige Datum mit Uhrzeit vor und an welcher der zwei Routen (N oder S) sie erfasst wurde.

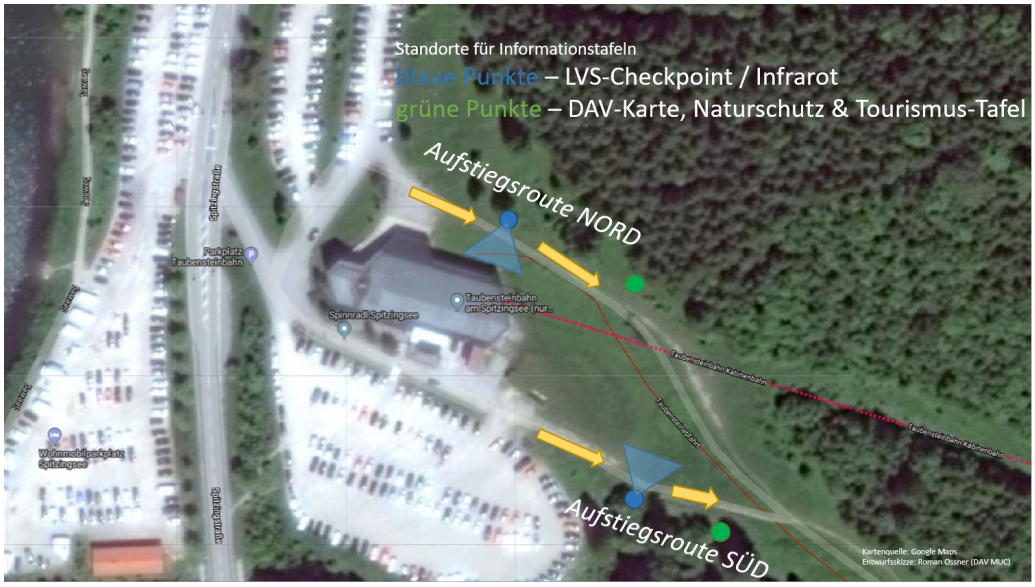

Abbildung 1: Satellitenbild, dass die Lage der zwei Routen und der Checkpoints verdeutlich

Für die erste Untersuchung benutzen wir vorerst nur diese automatisch erfassten Daten zur Saison 18/19. Der umfasste Zeitraum läuft dabei vom 21.12.2018 bis zum 13.04.2019. Anzumerken ist dabei, dass am 23.12. und 24.12. keine Messungen vorliegen, zudem werden Messungen vom 07.01. bis zum 15.01. auSSer Acht gelassen, da

aufgrund von starkem Schneefall die Checkpoints bedeckt waren.

Zusätzlich zu diesen automatischen Messungen wurden manuell Gruppen von (?)Studenten/Mitarbeitern des Departments für Geographie(?) an bestimmten Tagen vor Ort eingesetzt, um durch Befragungen manuelle Daten zu gewinnen. Dabei wurde festgestellt, dass bei den durch die Checkpoints erhobenen Daten Messfehler vorliegen.

Quelle: Folien vom Erstgespräch

Neben der Erfassung von Personen und LVS-Geräten liegen verschiedene weitere Daten vor.Für jeden Tag an dem gemessen wurde gibt es Information zu den Wetterbedingungen bzw. anderen möglichen Einflussvariablen. „snowhight“ bemisst die Schneehöhe in cm. „temperature“ ist die Temperatur des Tages um 12:00 mittags. „solar_radiation“ zeigt die Sonneneinstrahlung in W/cm^2. AuSSerdem sind die Lawinenwarnstufen des jeweiligen Tages angegeben. Es kann vorkommen, dass die Lawinenwarnstufe auf der Spitze des Berges eine andere ist als im Tal, deshalb gibt es zwei Variablen: „avalanche_report_top“ und „avalanche_report_down“. Diese geben die Lawinenwarnstufe an der Spitze und am FuSS des Berges an. Obwohl es Stufen von 1 (niedrig) bis 5 (sehr hoch) gibt, war die in dem beobachteten Zeitraum höchste Stufe eine 4. An Tagen an denen die Stufen unterschiedlich waren wurde auSSerdem in „avalanche_report_border“ der Höhenmeter angegeben, ab dem sich die Lawinengefahr unterscheidet. In „avalanche_report_comment“ ist vermerkt, ob es sich dabei um eine Waldgrenze handelt. „day“ gibt an, um welchen Tag der Woche es sich handelt und „day_weekday“, „day_weekend“ und „holiday“ geben jeweils an, ob der Tag ein Tag unter der Woche oder am Wochenende war und ob er sich innerhalb einer Ferienzeit befunden hat.

2.2 Datenaufbereitung

Der erste Schritt besteht darin, die gegebenen Daten um weitere Bearbeitung der Daten durch uns: (noch mehr schreiben) sunnrise und sunset „count_beacon“ und „count_infrared“ enthalten die Anzahl der gemessenen LVS-Geräte bzw. Infrarotmessungen pro Tag. Umwandlung der Messungen zu Personendaten. Umstellung des Tages von 04:00 bis 04:00. Eine Übersicht über alle Variablen die verwendet wurden ist in Tabelle 1 zu finden.

Tabelle 1: Übersicht aller benutzten Variablen mit kurzer Beschreibung

Name	**Beschreibung**	**Werte**
date	Datum	Datum vom 25.12.2018 bis zum 15.01.2019
time	Uhrzeit(minutengenau)	Uhrzeit von 03:59 bis 03:59
position	Route an der gemessen wurde	diskret (2 Ausprägungen): S,N
lvs	Hat die gemessene Person ein LVS-Gerät mitgeführt?	diskret (2 Auspr.): ja, nein
day	Wochentag	diskret (7 Auspr.): Montag, Dienstag,...
snowhight	Schneehöhe in cm (am Tag der Messung)	stetig: 16-212
temperature	Temperatur in řC um 12:00	stetig: (-7.9)-(9.7)
solar_radiation	Sonneneinstrahlung in W/cm^2	stetig: 14-792
avalanche_report	berechnete Lawinenwarnstufe	diskret (7 Auspr.): 1, 1.5, 2,...
holiday	Handelte es sich um einen Ferientag?	diskret (2 Auspr.): ja, nein
sunrise	Uhrzeit des Sonnenaufgangs	Uhrzeit von 05:29 bis 08:02
sunset	Uhrzeit des Sonnenuntergangs	Uhrzeit von 16:24 bis 18:58
day_length	berechnete Länge des Tages (von Sonnenauf- bis Sonnenuntergang)	von 08:24 bis 13:28 h

3 Deskriptive Analyse

Insgesamt 37593 Messungen an 114 Tagen

8468 Beacons, 29125 Infrareds (vor Umkodierung)

nach Umkodierung: 31574 Personen

8468 mit LVS-Gerät, 23106 ohne LVS-Gerät pr Die meisten Leute zwischen 09:00 und 18:00 unterwegs

–

Schneehöhe nimmt bis Mitte Januar stark zu und fällt ab Mitte Februar ab

Temperatur nimmt im Trend bus Mitte Januar ab und steigt danach

Sonnenstrahlung steigt bis März leicht und danach stark

—

Anteil schwankt in den ersten Wochen deutlich

generell viele AusreiSSer, aber kein groSSe Veränderung bei Schneehöhe, Temperatur und Sonneneinstrahlung

Anteile zur Mittagszeit geringer

mit steigender Lawinengefahr steigt die Anzahl

4 Methodik

In folgendem Kapitel soll anhand der zur Verfügung gestellten Daten zur Saison 18/19 in einem Modell der Zusammenhang einer beobachteten abhängigen Variable, in unserem Fall die Skitourengänger mit LVS-Gerät, durch mehrere unabhängige Variablen erklärt werden.
Dabei wird für die Erklärung schrittweise vom binären Regressionsmodell zum generalisierten additiven Modell vorgegangen, welches letztendlich für unsere Regressionsanalyse gewählt wurde.

4.1 Generalisierte lineare Modelle

Das lineare Regressionsmodell eignet sich für stetige Zielvariablen, die sich zumindest approximativ normalverteilt modellieren lassen. Der Erwartungswert einer Zielvariable wird durch die Linearkombination von EinflussgröSSen, auch Kovariablen genannt, beschrieben. In unserem Fall handelt es sich jedoch nicht um eine stetige, sondern eine binäre Zielvariable:
Skitourengänger mit LVS-Gerät: ja/nein.
Daher eignet sich ein generalisiertes lineares Modell, welches Regressionsanalysen für nicht zwangsläufig normalverteilte Zielvariablen ermöglicht.

4.1.1 Binäre Regressionsmodelle

Die binäre Zielvariablen y_i sind 0/1-kodiert und bei gegebenen Kovariablen $x_{i1}, ..., x_{ik}$ (bedingt) unabhängig. In unserem Fall, lässt sich die Zielvariable (Person mit beigeführtem LVS-Gerät) folgendermaSSen darstellen:

$$y_i = \begin{cases} 1 & \text{Person mit LVS-Gerät} \\ 0 & \text{sonst} \end{cases}$$

Des weiteren sind die Wahrscheinlichkeit $\pi_i = P(y_i = 1|x_{i1}, ..., x_{ik})$ und der lineare Prädiktor:

$$\eta_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} = x_i'\beta \tag{1}$$

durch eine Responsefunktion $h(\eta) \in [0, 1]$ miteinander verknüft:

$$\pi_i = h(\eta_i). \tag{2}$$

Im Logit-Modell gilt der Ansatz:

$$\underbrace{\pi = \frac{exp(\eta)}{1 + exp(\eta)}}_{\text{Responsefunktion}} \iff \underbrace{log\frac{\pi}{1-\pi} = \eta}_{\text{Logit-Linkfunktion}} . \tag{3}$$

4.1.2 Gruppierte Daten (hier vllt noch eine Daten-Tabelle)

Bislang sind wir davon ausgegangen, dass die vorliegenden Daten in ungruppierter Form, d.h. als Individualdaten, vorliegen. In unserem Fall bietet es sich an, die Daten nach dem Datum zu gruppieren, da dies zu Rechenzeit- und Speicherplatzersparnis führt. Die transormierte Datenmatrix enthält nun 110 Gruppen (*observations*).

4.1.3 Überdispersion

4.2 Generalisiertes additives Modell (GAM)

Bei den bisher aufgstellten generalisierten linearen Regressionsmodell tritt das Problem auf, dass der Effekt der stetigen Kovariablen auf die Zielvariable möglicherweise nicht-linear ist. Eine Möglichkeit dieses Problem zu lösen ist, ein generalisiertes additives Modell in Betracht zu ziehen. Das generalisierte additive Modell stellt ein nichtparametrisches Regressionsmodell dar.

4.2.1 Standardmodell GAM

Das generalisierte additive Modell stellt ein nichtparametrisches Regressionsmodell dar. Neben den linearen Einüssen können hier also auch nicht-lineare, flexible Einflüsse von stetigen Kovariablen auf die abhängige Variable, hier Personen mit beigeführtem LVS-Gerät, berücksichtigt werden. Dabei wird die Annahme getroffen, dass sich die Effekte der einzelnen Kovariablen additiv auf die Zielvariable auswirken. Des weiteren wird angenommen, dass die Verteilung der auf die Kovariablen bedingte Zielvariable

zur Binomialfamilie gehört. Das Standardmodell für das generalisierte additive Modell hat dabei folgende Struktur:

$$\eta_i = \underbrace{f_1(z_{i1}) + ... + f_q(z_{iq})}_{\text{nicht-parametrische Effekte}} + \underbrace{\beta_0 + \beta_1 x_{i1} + ... + x_{ik}}_{\text{parametrische Effekte}} + \epsilon_i \tag{4}$$

wobei die Funktionen $f_1(z_1), ..., f_q(z_q)$ den nicht-linearen Glättungseffekte der stetigen Kovariablen $z_1, ..., z_q$ beschreiben. Im nächten Schritt werden nicht-parametrische Schätzfunktionen für die Funktion $f(z)$ betrachtet.

4.2.2 Basis Splines (B-Splines)

Eine Möglichkeit, die Schätzfunktionf sinnvoll und flexibel zu modellieren ist die Darstellung von Polynom-Splines mittels der Spline-Basisfunktionen. Der Wertebereich wird in $m - 1$ Intervalle geteilt, zwischen denen m Knoten definiert werden (GLM Vorlesung F.194). Nach der Schätzung von einem Polynom vom Grad l auf jedem Intervall werden die Polynomfunktionen vom Grad l über $l + 2$ Knoten gebildet, um die Polynome der einzelnen Intervalle zusammenzuführen, ohne, dass es an den Knoten-Punkten zu Sprüngen kommt. Anschließend wird $f(z)$ durch $d = m + l - 1$ Basisfunktionen und damit folgende Funktion :

$$f(z) = \sum_{j=1}^{d} \gamma_j B_j z \tag{5}$$

dargestellt.

4.2.3 Penalisierte Splines (P-Splines)

Die Idee von P-Splines ist es die zu schätzdenen Funktionen $f(z)$ durch einen Polynom-Spline zu approximieren und eine hohen Anzahl an Knoten, in der Regel zwischen 20 und 40, festzulegen. Die Anzahl der Knoten bestimmt in hohem Maße die Glattheit der Funktion. Eine große Zahl an Knoten führt zu einer rauen Schätzung, welche für Flexibilität,auch bei stark variierenden Funktionen sorgt.
Eine weitere Idee ist es einen Strafterm für zu hohe Variabiltiät der Schätzung einzuführen.

P-Splines basierend auf B-Splines

P-Splines kombinieren eine B-Spline-Basis verbunden mit einem Strafterm. Um einen Strafterm für B-Splines zu entwickeln bietet es sich an, quadrierte Ableitungen zu verwenden. Diese werden als Maß der Variabiltät angesehen und sind daher in der Lage die Glattheit einer Funktion einzuschätzen. Der Strafterm

$$\lambda \int (f''(z))^2 dz \tag{6}$$

beruht auf der quadrierten zweiten Ableitung der Schätz-Funktion, welche nach z integriert wird und durch den Glättungsparameter λ dargestellt wird. Die Glattheit der Schätzung des Modells im penalisierten Fall, wird primär durch den Glättungsparameter (λ) reguliert, d.h. je größer λ, desto glatter ist die Schätzung.

Zyklische P-Splines

4.2.4 Autokorrelation