technische universität
dortmund

Modellwahl bei einem Brückenmonitoring

Bachlorarbeit von Erik Adam
Betreuerin: Prof. Dr. Christine Müller

Abgabetermin: 27. April 2020

Inhaltsverzeichnis

1 Einleitung **3**

2 Datenmaterial und Brückenmonitoring **4**
2.1 Beschreibung der Brücke und des Monitorings 4
2.2 Bedingungen des Monitorings 4
2.3 Datenmaterial 5
2.4 Problemstellung 6

3 Statistische Methoden **7**
3.1 Statistische Modellierung 7
3.1.1 Notation und Voraussetzungen 7
3.1.2 Lineares Modell 8
3.1.3 Elastisches Netz 12
3.1.4 k-fache Kreuzvalidierung 17
3.1.5 Varianz des Varianzschätzers bei Normalverteilung 17
3.2 Statistische Tests 18
3.2.1 Lilliefors-Test 18
3.2.2 Runs-Test 20
3.3 Praktische Anwendung der Methoden 21

4 Erstellung, Bewertung und Auswahl von Modellen **22**
4.1 Vorplanung und erster Modellierungsversuch 22
4.1.1 Auswahl der Variablen 22
4.1.2 Modellierungsansatz 23
4.1.3 Anpassung der ersten Modelle und Diskussion 25
4.2 Wiederholung des Vorgehens mit reduzierten Daten 28
4.3 Modellwahl zu WON2 30
4.3.1 Angehen des Problems der Heteroskedaszität 31
4.3.2 Reduzierung der Variablen durch Variation der Parameter 37
4.4 Abschließende Diskussion und Ausblick 41

Anhang **43**
A Weitere Grafiken 43
B Wichtige Ausschnitte des Programmcodes 43

Literaturverzeichnis **47**

1 Einleitung

Bei einer Brücke bei Bochum sind im Rahmen einer Routineuntersuchung Risse an der Oberfläche entdeckt worden. Daraufhin wurde ein Brückenmonitoring installiert. Hierbei wurden die Breite der Risse alle 2 Sekunden gemessen, ebenso wie die Temperaturen unterhalb und oberhalb der Brücke. Die Erhebung der Daten dauerte etwa 2 Jahre.

Um Brückeneinstürze und ähnliche Unglücke zu vermeiden ist es allgemein von Interesse, die Ermüdung von Brücken und ihre Belastungsgrenzen besser zu verstehen ebenso wie die abnutzenden und Belastbarkeit reduzierenden Einflüsse zu identifizieren. Insbesondere stellen sich die Fragen, welche Möglichkeiten bestehen, um das Risiko eines vorzeitigen Einsturzes zu minimieren und unter welchen Umständen ein vorzeitiger Einsturz wahrscheinlich wird. Je besser es möglich ist, die Gefahr eines Einsturzes rechtzeitig zu erkennen oder gar nicht erst aufkommen zu lassen, desto mehr Unglücke lassen sich vermeiden.

Abbildung 1 zeigt den Verlauf der Breite eines der beobachteten Risse über die Beobachtungszeitraum gemeinsam mit der Temperatur an der Unterseite der Brücke. Erkennbar ist hierbei, dass es einen Zusammenhang zwischen der Temperatur und der Breite des Risses zu geben scheint, was die Vermutung nahe legt, dass die Temperatur einen Einfluss auf die Breite des Risses besitzt. Im Rahmen dieser Arbeit wird daher nach einem Modell gesucht, welches die Breite der Risse in Abhängigkeit der gemessenen Temperaturen und dem Zeitpunkt möglichst gut erklärt. Im Idealfall wird erhofft, dass ein Zusammenhang entdeckt werden kann, der bei allen Rissen gleich und eindeutig ist.

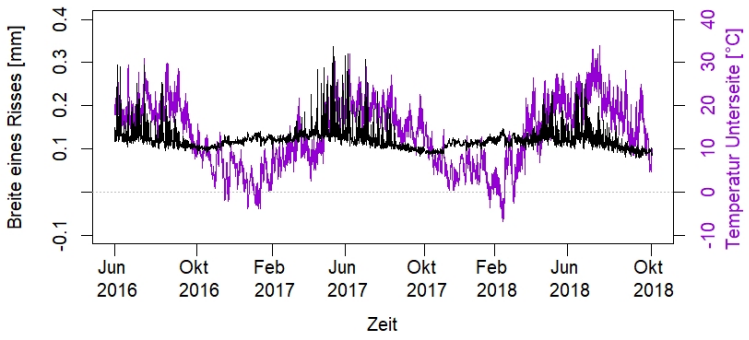

Abbildung 1: Verläufe der Breite eines Risses und der Temperatur an der Unterseite der Brücke über den Beobachtungszeitraum

Außerdem soll untersucht werden, ob die Risse im Laufe der Zeit breiter geworden sind, was ein Indiz für einen drohenden Einsturz sein könnte. So wurde z.B. am 8. April 2020 auf der Internetseite *tagesschau.de* von dem Einsturz einer Brücke in Italien berichtet, an der zuvor Risse aufgefallen waren. Die genaue URL zu diesem Bericht findet sich im Literaturverzeichnis dieser Arbeit.

Im folgenden Kapitel werden der Aufbau des Monitorings und der Brücke sowie das benutzte Datenmaterial beschrieben. In Kapitel 3 werden die angewandten statistischen Methoden und das Vorgehen erläutert. Kapitel 4 beinhaltet schließlich die Resultate dieser Arbeit, eine Diskussion und Bewertung der erhaltenen Ergebnisse.

2 Datenmaterial und Brückenmonitoring

2.1 Beschreibung der Brücke und des Monitorings

Das Brückenmonitoring wurde in Folge einer Routineuntersuchung durchgeführt und im Juni 2016 begonnen (Abbas et al 2019). Bei der Untersuchung wurden Risse entdeckt, welche mehr als 0,5mm breit waren. Es wurden an 16 Stellen der Brücke daher Wegaufnehmer angebracht, welche die Breite der Risse alle 2 Sekunden gemessen haben.

Die Brücke bestand aus zwei Überbauten, einem südlichen und einem nördlichen. Je Überbau wurden 8 Wegaufnehmer eingesetzt, davon jeweils vier westlich und vier östlich. Bezeichnet wurden diese entsprechend ihrer Position mit WON1 bis WON4, WWN1 bis WWN4, WOS1 bis WOS4 und WWS1 bis WWS4 (Abbas et al 2019). Hierbei steht WWN2 z.B. für „**W**egaufnehmer im **W**esten des **s**üdlichen Überbaus Nummer **2**". Analog sind die Bezeichnungen der übrigen Wegaufnehmer zu interpretieren.

2.2 Bedingungen des Monitorings

In der Mitte der Brücke fuhr eine Straßenbahn, die äußeren Spuren wurden für den öffentlichen Straßenverkehr genutzt. Der südliche Überbau wurde im Oktober 2017 abgerissen, der Rest der Brücke im Oktober 2018. Während des Monitorings wurden Einschränkungen im Straßenverkehr getroffen um die Belastung auf die Brücke zu reduzieren. Unter anderem wurde ein Maximalgewicht von 24 Tonnen pro Fahrzeug und das einspurige Befahren pro Fahrtrichtung beschlossen. Somit wurde die Brücke während des Monitorings seitens des Straßenverkehrs vermutlich weniger belastet als vorher.

2.3 Datenmaterial

Es liegen für diese Arbeit die Messungen der Wegaufnehmer vor, sowie zwei verschiedene Temperaturmessungen. Es wird angemerkt, dass zwar eine dritte Temperaturmessung stattgefunden hat, diese jedoch im Rahmen dieser Arbeit keine Beachtung findet. Die Rohdaten sind in regelmäßigen Abständen von 2 Sekunden erhoben worden, betrachtet werden hier jedoch ausschließlich auf die Stunde mittels Median gemittelte Werte. Die technische Verarbeitung der Daten vor ihrer Anwendung im Rahmen dieser Arbeit wird in Abbas et al (2019) in Abschnitt 2.2 beschrieben.

Alle Risse, die untersucht worden sind, besitzen eine Breite von mindestens 0,5 mm. Betrachtet werden daher die Differenzen zwischen den gemessenen Breiten und 0,5mm. Da sich die Risse abhängig von den Einflüssen auch verengt haben, kommen daher bei einigen Wegaufnehmern selten Datenpunkte mit weniger als 0 mm vor. Diese Werte bedeuten, dass der entsprechende Riss zu dem Zeitpunkt geringer als 0,5mm breit gewesen ist.

Zur Qualität der Daten lässt sich sagen, dass diese bei den Rissbreiten als sehr hoch angenommen werden kann. Gemäß der Bachlorarbeit von Thunich (2017, S. 8), welcher seine Informationen auf zwei Berichte der „König und Heunisch Planungsgesellschaft (KHP)" stützt (Heinrich 2016; König und Heunisch Planungsgesellschaft 2016), erfassen die Wegaufnehmer Änderungen der Rissbreiten bis auf 10^{-5}mm genau. Die Temperaturen wurden in °C gemessen, für diese Arbeit aber in Kelvin umgerechnet, da dies als die natürliche Einheit angesehen und daher als sinnvoller angenommen wird. Die Genauigkeit der Temperaturen hängt von der Messgenauigkeit der benutzten Geräte ab und es wird davon ausgegangen, dass sie mindestens auf 1 Kelvin [K] genau sind.

Es sei hierbei angemerkt, dass in dem für diese Arbeit benutzten Datensatz zwar auch die Daten der südlichen Wegaufnehmer zur Verfügung standen, diese aber dennoch aufgrund des vorzeitigen Abrisses des südlichen Überbaus vollständig vernachlässigt werden. Außerdem werden auch die Daten vom 23. und 24. Oktober 2017 bei dieser Arbeit nicht mit berücksichtigt, da an diesen Tagen der südliche Überbau entfernt wurde und dementsprechend angenommen wird, dass dies auch den nördlichen Überbau beeinflusst hat, weshalb den betroffenen Datenpunkten ein verzerrender Einfluss auf die geplante Modellbildung unterstellt wird.

Insgesamt liegen je gemessener Variable 21144 Werte vor. Unter diesen sind 50 als „fehlende Werte" im Datensatz gekennzeichnet.

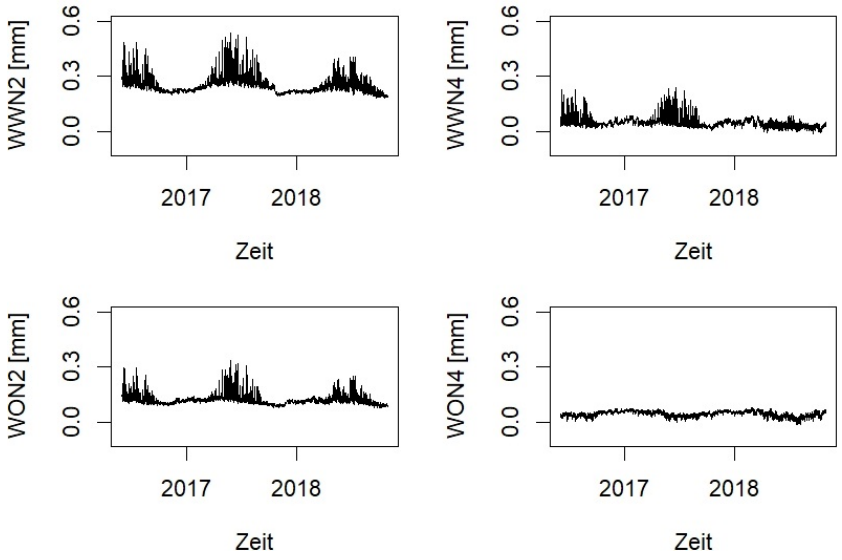

Abbildung 2: 4 Verläufe von Rissbreiten (des nördlichen Überbaus) über die Zeit

2.4 Problemstellung

Ziel dieser Arbeit ist es, ein Modell zu finden, dass die Breite wenigstens eines Risses möglichst gut mithilfe der gegebenen Temperaturen erklärt. Hierbei ist optimal, wenn das Modell eine Erkenntnis liefert, die sich auf alle Risse übertragen lässt. Abbildung 1 hat bereits motiviert, warum ein Modell ausgerechnet in Abhängigkeit der Temperatur gesucht ist. Abbildung 2 verdeutlicht hierbei die Schwierigkeit. Sie zeigt, dass die Verläufe der Rissbreiten sich teilweise sehr unterschiedlich verhalten. Dabei ist zu beachten, dass die Temperaturverläufe bei allen Rissen gleich bleiben. Der Vollständigkeit halber findet sich im Anhang eine Grafik mit den Rissbreiten der übrigen 4 Wegaufnehmer des nördlichen Überbaus.

Es lässt sich augenscheinlich anhand dieser Grafiken feststellen, dass die meisten Rissverläufe in den warmen Jahreszeiten breiter werden und stärker streuen, womit die Motivation für diese Arbeit weiterhin besteht. Allerdings deutet sich hieran anhand dieser Grafiken bereits an, dass es sehr schwierig werden könnte, ein Modell zu finden, dass bei allen Rissverläufen zufriedenstellend ist. Dies wird sich in Kapitel 4 bestätigen. Es wird daher an dieser Stelle betont, dass das primäre Ziel ist, ein geeignetes Modell für wenigstens einen der Rissverläufe zu finden.

3 Statistische Methoden

In diesem Kapitel werden die benutzten statistischen Methoden beschrieben. Es sei angemerkt, dass alle Anwendungen der Methoden und weitere Verarbeitungen der Daten, ebenso wie die Erstellungen der Grafiken mithilfe der Software `R` (R Core Team 2019) durchgeführt wurde. Mehr dazu in Abschnitt 3.3 dieser Arbeit.

3.1 Statistische Modellierung

3.1.1 Notation und Voraussetzungen

Im gesamten Abschnitt 3.1 bezeichnen $Y_1, \ldots, Y_n$ untereinander unabhängige Zufallsvariablen und $y_1, \ldots, y_n$ deren Realisierungen. Hierbei ist n der Stichprobenumfang und p die Anzahl weiterer Größen, von denen die $Y_1, \ldots, Y_n$ abhängen und welche mit $x_{1,1}, \ldots, x_{n,1}, x_{1,2}, \ldots, x_{n,2}, \ldots, x_{1,p}, \ldots, x_{n,p}$ bezeichnet werden. Mit Y ist der Zufallsvektor gemeint, welcher die Komponenten $Y_1, \ldots, Y_n$ beinhaltet. Analog ist y ein Vektor, der die Realisierungen $y_1, \ldots, y_n$ beinhaltet. Mit X_i wird der Vektor bezeichnet, welcher die Einträge $x_{1,i}, \ldots, x_{n,i}$ besitzt. Bei den Erläuterungen der Modelle werden Y die abhängige Variable und $X_1, \ldots, X_p$ die unabhängigen Variablen genannt. Die sogenannte Designmatrix X besitzt die Spaltenvektoren $X_1, \ldots, X_p$, sieht also wie folgt aus:

$$X = \begin{pmatrix} X_1 & \ldots & X_p \end{pmatrix}.$$

Mit $e_1, \ldots, e_n$ werden die sogenannten Residuen bezeichnet, welche die Komponenten des Residuenvektors e sind. Diese sind gleichzeitig Realisierungen unabhängiger Zufallsvariablen $\varepsilon_1, \ldots, \varepsilon_n$, welche in dem Zufallsvektor ε zusammengefasst werden. Für alle $\varepsilon_1, \ldots, \varepsilon_n$ wird vorausgesetzt, dass sie Erwartungswert 0 und eine eventuell unbekannte, aber identische Varianz σ^2 besitzen.

Sei $v \in \mathbb{R}^m, m \geq 1$ ein beliebiger Vektor. Mit $\|v\|_2 := \sqrt{\sum_{j=1}^{m} v_j^2}$ wird die euklidische Norm und mit $\|v\|_1 := \sum_{j=1}^{m} |v_j|$ die Summennorm von v bezeichnet.

Schließlich bezeichnet $\beta \in \mathbb{R}^p$ einen unbekannten Parametervektor, dessen Komponenten $\beta_1, \ldots, \beta_p$ sind. Falls bei X_1 $\mathbb{R}^n$ alle Komponenten exakt 1 sind, ist von einem Intercept die Rede, dessen Effekt β_1 ist. Allgemein wird eine Schätzung einer beliebigen Größe γ mit $\hat{\gamma}$ bezeichnet. Insbesondere ist also $\hat{\beta}$ eine Schätzung von β und $\hat{\beta}_i$ eine Schätzung der i-ten Komponente von β. Der Vektor $\hat{y} = (\hat{y}_1, \ldots, \hat{y}_n)^T$ bezeichnet daher

eine Schätzung für y.

In dem Abschnitt zu den statistischen Tests (3.2) werden eigene Notationen benutzt. Die hier vor gestellten Notationen sind nur gültig für den Abschnitt 3.1.

3.1.2 Lineares Modell

Beim linearen Modell (Toutenburg 2003, S. 89 - 194) wird davon ausgegangen, dass die abhängige Variable Y einer Linearkombination der unabhängigen Variablen und ε entspricht. Hierbei wird also ein Zusammenhang wie in (1) unterstellt, wobei β ein unbekannter Parametervektor ist (Toutenburg 2003, S. 90).

$$Y = X\beta + \varepsilon \tag{1}$$

Ziel beim linearen Modell ist, die Komponenten $\beta_1, \ldots, \beta_p$ des Vektors β anhand der beobachteten Realisierungen $y_1, \ldots, y_n$ und der Designmatrix X geeignet zu schätzen. Hierfür wird das Prinzip der kleinsten Quadrate betrachtet, bei dem β so zu wählen ist, dass die euklidische Distanz zwischen dem Vektor der beobachteten Realisierungen $y_1, \ldots, y_n$ und den durch den Zusammenhang aus (1) erwarteten Größen $\hat{y}_1, \ldots, \hat{y}_n$ minimiert wird. Das Kleinste Quadrate Kriterium ist in (2) gegeben.

$$\hat{\beta} := \underset{\tilde{\beta}}{\operatorname{argmin}} \left\{ (y - X\tilde{\beta})^T (y - X\tilde{\beta}) \right\} = \underset{\tilde{\beta}}{\operatorname{argmin}} \left\{ \left\| y - X\tilde{\beta} \right\|_2^2 \right\} \tag{2}$$

Somit fordert das Kleinste-Quadrate Kriterium die Minimierung der euklidischen Länge des Residuenvektors. Dieser ist wie folgt definiert:

$$e := y - X\hat{\beta}.$$

Außerdem sei

$$\hat{y} := X\hat{\beta}$$

der Vektor der durch das Modell vorhergesagten Werte für y.

Ein Vektor γ, der folgende Gleichung (die sogenannte Normalgleichung) erfüllt, ist ein Vektor, der das Minimierungsproblem aus (2) löst:

$$X^T y = X^T X \gamma \tag{3}$$

Ist der Rang der Designmatrix X kleiner als ihre Anzahl Spalten p, so existieren mehrere Vektoren, die das Kriterium aus (2) erfüllen. Entspricht der Rang von X jedoch ihrer Anzahl Spalten p, so ist die Schätzung gemäß des Kleinste-Quadrate Kriteriums eindeutig. In diesem Fall lässt sich $\hat{\beta}$ durch (4) berechnen, wie aus der Normalgleichung (3) hergeleitet werden kann.

$$\hat{\beta} := \left(X^T X\right)^{-1} X^T y \tag{4}$$

Dieser Schätzer ist die Realisierung folgender Zufallsvariable:

$$B := \left(X^T X\right)^{-1} X^T Y$$

In (4) wird ein Vorteil dieser Schätzung erkennbar. Da y eine Realisierung eines Zufallsvektors Y ist und der Zusammenhang aus (1) angenommen wird, ist $\hat{\beta}$ aus folgendem Grund eine erwartungstreue Schätzung von β:

$$\mathrm{E}(B) = \mathrm{E}\left((X^T X)^{-1} X^T Y\right) = (X^T X)^{-1} X^T X \beta + \mathrm{E}(\varepsilon) = \beta$$

3.1.2.1 Multikollinearität

Besitzt die Matrix X nicht vollen Spaltenrang, so ist von starker Multikollinearität die Rede und (4) ist nicht berechenbar. Zu (2) existieren dann mehrere Lösungen. Tritt dieses Problem auf, bedeutet das, dass die Spalten von X nicht linear unabhängig sind, weshalb es möglich ist, eine Spalte zu finden, welche als Linearkombination der übrigen Spalten darstellbar ist. Wurde eine solche Spalte gefunden worden, kann sie schlicht aus der Designmatrix entfernt werden. Es können auf diese Art so lange Spalten von X eliminiert werden, bis die starke Multikollinearität nicht mehr länger besteht (Hedderich und Sachs 2018, S. 809 – 810).

Tritt schwache Multikollinearität auf (d.h. einige Spalten sind nahezu linear abhängig), so ist (4) zwar berechenbar, aber die Schätzung der Komponenten von $\hat{\beta}$ können stark streuen. Es gibt verschiedene Ansätze, mit diesem Problem umzugehen, unter anderem

jene, die in Abschnitt 3.1.3 vorgestellt werden.

3.1.2.2 R^2 und adjustiertes R^2

Um die Güte der Anpassung eines Modells aus (1) zu bewerten, kann das Bestimmtheitsmaß betrachtet werden. Es seien zunächst:

$$SSR := \|e\|_2^2 = \|y - \hat{y}\|_2^2 = \sum_{j=1}^{n} (y_j - \hat{y}_j)^2$$
$$SSX := \|\hat{y} - 1_n\bar{y}\|_2^2 = \sum_{j=1}^{n} (\hat{y}_j - \bar{y})^2$$
$$SSY := \|y - 1_n\bar{y}\|_2^2 = \sum_{j=1}^{n} (y_j - \bar{y})^2,$$

wobei 1_n der Vektor der Dimension n ist, bei dem alle Komponenten 1 sind, und $\bar{y}$ das arithmetische Mittel über y und schließlich $\hat{y}$ der Vektor der durch ein Modell wie in (1) prognostizierten Werte für Y bei gegebenen X und $\hat{\beta}$. Toutenburg (2003, S. 138 – 141) zeigt mathematisch, dass gilt:

$$\begin{aligned} SSY &= SSX + SSR \\ \Leftrightarrow 1 &= \frac{SSX}{SSY} + \frac{SSR}{SSY} \\ \Leftrightarrow \frac{SSX}{SSY} &= 1 - \frac{SSR}{SSY} \end{aligned}$$

Dass heißt, dass die Gesamtvariablilität der beobachteten $y_1, \ldots, y_n$, welche SSY ist, sich zerlegen lässt in eine Variabilität zur Modellprognose SSX und in die Variabilität der Residuen des Modells SSR. Darauf aufbauend wird folgendes Maß definiert, welches das Bestimmtheitsmaß R^2 genannt wird (Toutenburg 2003, S. 147 oder Pflaumer et al 2001, S. 166):

$$R^2 := \frac{SSX}{SSY} = 1 - \frac{SSR}{SSY}$$

Es gilt immer $R^2 \in [0, 1]$. Je größer R^2 ist, desto mehr von der Variabilität von y wird durch das Modell erklärt, und umso erfolgreicher ist das Modell.

Das Bestimmtheitsmaß hat jedoch den Nachteil, dass es immer größer wird, mit je

mehr unabhängigen Variablen ein Modell „gefüttert" wird, auch dann, wenn einige dieser abhängigen Variablen in der Realität keinen Einfluss auf Y haben und daher das Modell von der Realität eher entfernen.

Deshalb wird zum Vergleichen verschiedener Modelle das korrigierte Bestimmtheitsmaß $\tilde{R}^2$ betrachtet, welches hier wie folgt definiert wird (Toutenburg 2003, S. 145 – 149 oder Pflaumer et al 2001, S. 248):

$$\tilde{R}^2 := 1 - (1 - R^2)\frac{n-1}{n-p}$$

Bei einem Modell mit nur einer unabhängigen Variablen, sodass $p = 1$, gilt daher $R^2 = \tilde{R}^2$. Für $p > 1$ hingegen folgt $R^2 > \tilde{R}^2$. Dem adjustierten Bestimmtheitsmaß ist es möglich, bei einem Modell mit $p + q$ unabhängigen Variablen kleiner zu werden als bei nur p unabhängigen Variablen. Daher wird $\tilde{R}^2$ bei der Beurteilung von Modellen bevorzugt werden.

3.1.2.3 Kodierung kategorialer Merkmale

In dem Modell aus (1) wird ersichtlich, dass es nicht interpretierbar wäre, wenn für eine Spalte von X keine quantitativen Größen beinhalten würde, sondern stattdessen kategoriales Skalenniveau besäße. Im Falle einer Variable mit kategorialem Skalenniveau wird diese deshalb in geeignete „Untervariablen" aufgeteilt. Diese „Untervariablen" werden Dummys genannt und sind geeignete Kodierungen der Kategorien der ursprünglichen Variable. Allgemein sind zwei verschiedene Formen der Kodierung üblich, wobei in diesem Bericht der Einfachkeit halber ausschließlich die Dummykodierung benutzt wird (nachzulesen in Toutenburg 2003, S. 70–71).

Bei einem Modell mit Intercept wird eine kategoriale Variable mit k Kategorien hierbei in $(k-1)$ Dummys aufgeteilt, wobei für die Größe x_{ij} gilt, falls X_j eine Dummy-Variable zur Kategorie z der genannten „Übervariable" ist:

$$x_{ij} = \begin{cases} 1, & \text{die i-te Beobachtung gehört zur Kategorie z} \\ 0, & \text{sonst.} \end{cases}$$

Bei einem Modell ohne Intercept wird eine beliebige kategoriale Variable mit k Dummys versehen und der Intercept sodurch für die übrigen kategorialen Variablen ersetzt.

3.1.3 Elastisches Netz

Das elastische Netz (vorgeschlagen von Zou und Hastie 2005) ist ein Verfahren, bei der ein Modell wie in (1) angepasst wird. Anstatt des Kleinste-Quadrate Kriteriums wird der Parametervektor β auf Basis des in (5) dargestellten Kriteriums geschätzt (Friedman et al 2010, S. 3).

$$\hat{\beta} := \underset{\tilde{\beta}}{\operatorname{argmin}} \left\{ \frac{1}{2n} \left\| y - X\tilde{\beta} \right\|_2^2 + \lambda P_\alpha \left(\tilde{\beta} \right) \right\} \tag{5}$$

$$P_\alpha \left(\tilde{\beta} \right) := \left(\frac{1-\alpha}{2} \left\| \beta \right\|_2^2 + \alpha \left\| \beta \right\|_1 \right) \tag{6}$$

Die beiden Parameter λ und α sind vom Anwender geeignet zu wählen. Das Kriterium aus (5) entspricht dem Kleinste-Quadrate Kriterium aus (2) ergänzt um einen Term $P_\alpha \left(\tilde{\beta} \right)$, welcher im folgenden als „Bestrafungsterm (des elastischen Netzes)" bezeichnet wird. Motiviert wird das Ergänzen eines solchen Bestrafungsterms durch den Anspruch an das Modell, plausibel und gut interpretierbar zu sein. Hierfür sollte es nicht zu viele relevante Effekte beinhalten, da es dann vielen unabhängigen Variablen einen Einfluss auf die abhängige Variable unterstellen würde. Dies ist insofern vermeidenswert, da sich bei einem endlichen Datensatz $n < \infty$ die Erklärbarkeit der Beobachtungen $y_1, \ldots, y_n$ in einem Modell aus (1) durch das Ergänzen weiterer Variablen niemals verschlechtert, auch nicht beim Hinzufügen von Variablen, die tatsächlich keinen Einfluss besitzen.

Durch den Bestrafungsterm werden Kandidaten für β, welche sehr viele, von null (sehr) verschiedene Komponenten besitzen, disqualifiziert. Für eine sinnvolle Anwendung sollten hierbei $\lambda \in (0; \infty)$ und $\alpha \in [0; 1]$ gelten. Der Parameter λ reguliert, wie stark der Bestrafungsterm sich auf die Modellbildung auswirkt bzw. wie sehr sich das Minimierungsproblem aus (5) von dem aus (2) unterscheidet. Für den Spezialfall $\lambda = 0$ sind (5) und (2) identisch. Für $\lambda < 0$ würde der Bestrafungsterm zu einem „Belohnungsterm" werden, d.h. er würde Kandidaten für β mit vielen, von 0 verschiedenen Komponenten begünstigen. Das entspräche dem Gegenteil dessen, was er bewirken soll.

Durch die Wahl von $\alpha \in [0, 1]$ wird ebenfalls ausgeschlossen, dass der Bestrafungsterm negativ werden kann. Der Bestrafungsterm beim elastischen Netz stellt einen Kompromiss zwischen der Ridge Regression und der Lasso Regression dar, auf welche im folgenden noch eingegangen wird. Der Parameter α reguliert hierbei, ob die Bestrafung mehr nach dem Vorbild der Ridge Regression oder der Lasso Regression erfolgt. Für $\alpha = 0$ entspricht die Modellierung via elastischem Netz genau der Ridge Regression, für $\alpha = 1$ entspricht sie der Lasso Regression. Es wird darauf hingewiesen, dass die

Parameter α und λ vom Anwender bestimmt werden können. Sei ohne Beschränkung der Allgemeinheit $\lambda^* := 2n\lambda$. Dann ist das Problem aus (5) äquivalent zu folgendem Optimierungsproblem:

$$\hat{\beta} := \underset{\tilde{\beta}}{\operatorname{argmin}} \left\{ \left\| y - X\tilde{\beta} \right\|_2^2 + \lambda^* P_\alpha(\tilde{\beta}) \right\}.$$

In dieser Darstellung wird leichter ersichtlich, dass das elastische Netz in den genannten Spezialfällen einer Lasso oder Ridge Regression entspricht.

3.1.3.1 Ridge Regression

Die Ridge Regression wird durch das Problem der Multikollinearität motiviert, d.h. der beinahen oder exakten linearen Abhängigkeit der Spalten der Designmatrix X (nachzulesen in Toutenburg 2003, S. 178 – 183, sowie in Sen und Srivatava 1990, S. 256 – 261). Sei $B := \left(X^T X\right)^{-1} X^T Y$, sodass $\hat{\beta}$ aus (4) eine Realisierung von B ist. Die Varianz von B berechnet sich dann mit (1) zu:

$$\begin{aligned}
\operatorname{var}(B) &= \left(X^T X\right)^{-1} X \operatorname{var}(Y) \left(\left(X^T X\right)^{-1} X^T \right)^T \\
&= \left(X^T X\right)^{-1} X^T X \left(\left(X^T X\right)^{-1} \right)^T \operatorname{var}(\varepsilon) \\
&= \left(X^T X\right)^{-1} \sigma^2 I_n \\
&= \left(X^T X\right)^{-1} \sigma^2.
\end{aligned}$$

Dabei ist I_n die n-dimensionale Einheitsmatrix. Da $\left(X^T X\right)$ symmetrisch ist, gilt dies auch für ihre Inverse und $\operatorname{var}(B)$ lässt sich in Abhängigkeit der Eigenwerte und der Eigenvektoren von $X^T X$ darstellen (Toutenburg 2003, S. 176, 490):

$$\operatorname{var}(B) = \sigma^2 \sum_{j=1}^{p} \lambda_j^{-1} u_j u_j^T,$$

wobei $\lambda_1, \ldots, \lambda_p$ die Eigenwerte und $u_1, \ldots, u_p$ die standardisierten Eigenvektoren zu $X^T X$ sind. Sind die Spalten von X beinahe linear abhängig (schwache Multikollinearität), so ist wenigstens einer der Eigenwerte beinahe null. Werden die Diagonalelemente von $\operatorname{var}(B)$ betrachtet, wird hier sogar deutlich, dass die Varianzen einiger Komponenten

von B explodieren können.

In so einer Situation lässt sich anstelle des Schätzers aus (4) der Ridge-Schätzer anwenden (Toutenburg 2003, S.179):

$$\hat{\beta}_{Ridge} := \left(X^T X + \tilde{\lambda} I_p\right)^{-1} X^T y$$

Bei geeigneter Wahl von $\tilde{\lambda} > 0$ kann erreicht werden, dass die Varianzen der Komponenten dieses Schätzers geringer oder höchstens gleich ausfallen wie die Varianzen der Komponenten von B (Sen und Srivatava 1990, S. 258 – 261). Das liegt daran, dass der kleinste Eigenwert der Matrix $X^T X + \tilde{\lambda} I_p$ größer sein kann als der kleinste Eigenwert von $X^T X$. Der Nachteil hingegen ist, dass dieser Schätzer nicht erwartungstreu ist:

$$\begin{aligned}
\mathrm{E}(\hat{\beta}_{Ridge}) =& \mathrm{E}\left(\left(X^T X + \tilde{\lambda} I_n\right)^{-1} X^T Y\right) \\
=& \left(X^T X + \tilde{\lambda} I_n\right)^{-1} X^T \mathrm{E}(Y) \\
=& \left(X^T X + \tilde{\lambda} I_n\right)^{-1} X^T X \beta \\
\neq& \beta.
\end{aligned}$$

Der Ridge Schätzer entspricht außerdem der Lösung des folgenden Minimierungsproblems, welches einem Spezialfall aus (5) entspricht (Zou und Hastie 2005, Abschnitt 2.1):

$$\hat{\beta}_{Ridge} = \underset{\tilde{\beta}}{\operatorname{argmin}} \left\{ \left\| y - X\tilde{\beta} \right\|_2^2 + \tilde{\lambda} \left\| \tilde{\beta} \right\|_2^2 \right\}$$

Dass der Ridge-Schätzer dieses Minimierungsproblem löst, lässt sich wie folgt nachweisen:

Es seien

$$\tilde{y} := \begin{pmatrix} y \\ 0_p \end{pmatrix}, \tilde{X} := \begin{pmatrix} X \\ J_p \end{pmatrix},$$

wobei $J_p := \sqrt{\tilde{\lambda}} I_p$ mit I_p der p-dimensionalen Einheitsmatrix und 0_p dem p-dimensionalen

Vektor mit allen Einträgen null sind. Das Minimierungsproblem ist dem Modell

$$\tilde{y} = \tilde{X}\beta + \tilde{e}$$

zuzuordnen, denn es gilt:

$$\begin{aligned}
\min\{\tilde{e}^T\tilde{e}\} &= \min\{(\tilde{y} - \tilde{X}\beta)^T(\tilde{y} - \tilde{X}\beta)\} \\
&= \min\left\{\left|\left|\begin{pmatrix} y \\ 0_p \end{pmatrix} - \begin{pmatrix} X\beta \\ J_p\beta \end{pmatrix}\right|\right|_2^2\right\} \\
&= \min\left\{||y - X\beta||_2^2 + ||0_p - J_p\beta||_2^2\right\} \\
&= \min\left\{||y - X\beta||_2^2 + \lambda\,||\beta||_2^2\right\},
\end{aligned}$$

d.h. das Minimierungsproblem entspricht dem Anspruch an das Modell, die Quadratsumme der Residuen zu minimieren. Mit den Beziehungen

$$\tilde{X}^T\tilde{y} = \begin{pmatrix} X^T & J_p^T \end{pmatrix}\begin{pmatrix} y \\ 0_p \end{pmatrix} = X^T y$$

$$\tilde{X}^T\tilde{X} = \begin{pmatrix} X^T & J_p^T \end{pmatrix}\begin{pmatrix} X \\ J_p \end{pmatrix} = X^T X + \lambda I_p$$

lässt sich über die Normalgleichung aus (3) zeigen, dass der Ridge-Schätzer der Kleinste-Quadrate-Schätzer zu dem Minimierungsproblem ist:

$$\begin{aligned}
&\tilde{X}^T\tilde{y} = \tilde{X}^T\tilde{X}\hat{\beta}_{KQ} \\
\Leftrightarrow &X^T y = \tilde{X}^T\tilde{X}\hat{\beta}_{KQ} \\
\Leftrightarrow &(\tilde{X}^T\tilde{X})^{-1}X^T y = (\tilde{X}^T X)^{-1}\tilde{X}^T\tilde{X}\hat{\beta}_{KQ} \\
\Leftrightarrow &(X^T X + \lambda I_p)^{-1}X^T y = \hat{\beta}_{KQ} \\
\Leftrightarrow &\hat{\beta}_{Ridge} = \hat{\beta}_{KQ}.
\end{aligned}$$

Bei Betrachtung des Minimierungsproblems wird ersichtlich, dass die Länge dieses Schätzers nicht „unnötig" groß werden kann, da die Größe $\|\beta\|_2^2$ ihrerseits im Minimierungsproblem vorkommt. Das ist bei hoher Multikollinearität ein Vorteil gegenüber dem Schätzer aus (4), da es offensichtlich seine Streuung begrenzt. Auch lässt sich der Ridge-Schätzer anders als der Schätzer aus (4) immer berechnen. Das ist daran erkennbar, dass die Matrix $\tilde{X}$ stets vollen Rang hat, da ihre untersten p Zeilen ihrerseits eine Einheitsmatrix bilden.

Abschließend wird angemerkt, dass die Ridge Regression häufig nach Zentrierung und Standardisierung der Daten angewandt wird. Dies ist jedoch nicht zwingend notwendig (Sen und Srivatava 1990, S. 257).

3.1.3.2 Lasso Regression

Die Lasso Regression, vorgeschlagen von Tibshirani (1996), schätzt β über folgendes Minimierungsproblem:

$$\hat{\beta}_{Lasso} := \underset{\tilde{\beta}}{\operatorname{argmin}} \left\{ \left\| y - X\tilde{\beta} \right\|_2^2 + \tilde{\lambda} \left\| \tilde{\beta} \right\|_1 \right\}.$$

Hierbei ist $\tilde{\lambda}$ erneut ein vom Anwender zu wählender Parameter, der beeinflusst, wie stark der Bestrafungsterm in die Schätzung mit eingeht. Diese Schätzung tendiert im Gegensatz zu der Schätzung aus (4) dazu, mehr Komponenten von β auf null zu schätzen und somit einige der unabhängigen Variablen faktisch zu eleminieren. Die Lasso Regression betreibt damit gleichzeitig eine Variablenselektion, weshalb sie interessant wird, wenn viele Variablen in das Modell aufgenommen werden, bei denen die Möglichkeit besteht, dass sie keinen Effekt haben könnten. Gleichzeitig besitzt die Lasso Regression ebenfalls die Eigenschaft, Kandidaten für β, welche viele betragsmäßig große Komponenten besitzen, eher abzulehnen als der Schätzer aus (4).

Hingegen gibt es einige Nachteile bei der Lasso Regression, welche von Zou und Hastie (2005, Abschnitt 1) aufgezählt werden:

- die Lasso Regression kann höchstens n Variablen einen von null verschiedenen Effekt zuweisen, was bei $p > n$ Variablen mit Effekt automatisch zu einer Unterschätzung der Variablenanzahl führt.
- bei einer Gruppe korrelierter Variablen tendiert die Lasso Regression dazu, einer Variable einen Effekt zuzuordnen und den übrigen keinen
- bei $n > p$ und hoher Korrelation einiger Variablen wurde empirisch nachgewiesen, dass die Ridge Regression erfolgreichere Vorhersagen trifft (Tibshirani 1996)

3.1.3.3 Motivation des elastischen Netzes

Das elastische Netz ist ein Kompromiss zwischen der Lasso und der Ridge Regression. Es ist daher interessant, als dass es die Möglichkeit birgt, die Vorzüge beider Verfahren bis zu einem gewissen Grad aufzunehmen und die Nachteile nur in abgefederter Form.

3.1.4 k-fache Kreuzvalidierung

Die Parameter beim elastischen Netz α und λ müssen vom Anwender gewählt werden. Um eine geeignete Auswahl zu treffen, kann die k-fache Kreuzvalidierung benutzt werden (beschrieben in Likas et al 2014, S. 3 – 5).

Diese Methode ermöglicht, den Vorhersagefehler eines Modells abzuschätzen. Hierzu wird der Datensatz in k zufällige Cluster eingeteilt, welche möglichst gleichgroß sein sollten. Anschließend werden $k-1$ Cluster als Trainingsdatensatz festgelegt, während der letzte Cluster als Testdatensatz aufgehoben wird. Mithilfe des Trainingsdatensatzes wird dann das Modell angepasst. Anschließend werden mithilfe des Modells aus dem Trainingsdatensatz Vorhersagen für die Datenpunkte aus dem Testdatensatz gemacht. Diese werden schließlich mit den Datenpunkten aus dem Testdatensatz verglichen und der mittlere Fehler wird bestimmt.

Dieses Vorgehen wird wiederholt, bis jeder Datenpunkt genau einmal im Testdatensatz eingesetzt worden ist. Abschließend kann der Mittelwert über die mittleren Fehler betrachtet werden um zu beurteilen, wie erfolgreich die Modellierung gewesen ist.

Im Falle des elastischen Netzes kann die Kreuzvalidierung wie folgt zur Wahl geeigneter Parameter angewandt werden: Für verschiedene Werte von α und λ wird mittels der k-fachen Kreuzvalidierung der Fehler bestimmt. Da die Modellierung bis auf die unterschiedlichen Werte von α und λ immer dem selben Vorgehen entspricht, können schließlich jene Werte für α und λ als geeignet angesehen werden, für die der Fehler der Kreuzvalidierung minimal ausgefallen ist.

3.1.5 Varianz des Varianzschätzers bei Normalverteilung

Während der Berechnung und der Analyse einiger Modelle erwies es sich als interessant, ein lineares Modell zu suchen, welches empirische Varianzen von Residuen in Abhängigkeit anderer Variablen erklärt. Um ein lineares Modell anzupassen, gelten die Voraussetzungen aus 3.1.1, welche auch die Gleichheit der Varianzen der Residuen und im Fall eines Modells nach (1) daher auch die Gleichheit der Varianzen der abhängigen Variablen erfordern. Daher ist die Varianz der Varianzschätzfunktion für diese Arbeit von Interesse.

Es seien $W_1, \ldots, W_n$ unabhängig identisch verteilte Zufallsvariablen mit $W_i \sim N(0, \tilde{\sigma}^2)$ für alle i. Die Realisierungen der $W_1, \ldots, W_n$ könnten also Residuen eines linearen Modells sein unter zusätzlicher Normalverteilungsannahme. Im folgenden wird die Varianz der Varianzschätzfunktion der $W_1, \ldots, W_n$ bestimmt. Hierfür lohnt es sich, Eigenschaften von $\mathcal{X}^2$-verteilten Zufallsvariablen auszunutzen (nachzulesen in Pflaumer et al 2001,

S. 80 – 81).

Die Varianz einer $\mathcal{X}^2$-verteilten Zufallsvariable Z mit m Freiheitsgeraden ergibt sich zu

$$\mathrm{var}(Z) = 2m.$$

Außerdem gilt für die $W_1, \ldots, W_n$, da sie unabhängig und normalverteilt mit gleicher Varianz und Erwartungswert sind, dass folgender Größe einer $\mathcal{X}^2$-Verteilung mit $(n-1)$ Freiheitsgeraden folgt.

$$\frac{1}{\tilde{\sigma}^2} \sum_{j=1}^{n} \left(W_j - \bar{W} \right)^2 \sim \mathcal{X}^2_{(n-1)} \tag{7}$$

$\bar{W}$ ist hierbei das arithmetische Mittel der $W_1, \ldots, W_n$. Für die Varianz der Varianzschätzung, angewandt auf die $W_1, \ldots, W_n$, folgt somit:

$$\begin{aligned}
\mathrm{var}\left(\frac{1}{n-1} \sum_{j=1}^{n} \left(W_j - \bar{W} \right)^2 \right) &= \mathrm{var}\left(\frac{1}{n-1} \frac{\tilde{\sigma}^2}{\tilde{\sigma}^2} \sum_{j=1}^{n} \left(W_j - \bar{W} \right)^2 \right) \\
&= \frac{\tilde{\sigma}^4}{(n-1)^2} \mathrm{var}\left(\frac{1}{\tilde{\sigma}^2} \sum_{j=1}^{n} \left(W_j - \bar{W} \right)^2 \right) \\
&\overset{(7)}{=} \frac{2\tilde{\sigma}^4}{n-1}.
\end{aligned}$$

Zentral wird hierbei die Erkenntnis sein, dass bei wachsender, zugrundeliegender Varianz $\tilde{\sigma}^2$ auch die Varianz der Varianzschätzfunktion zunimmt.

3.2 Statistische Tests

In diesem Abschnitt werden eigene Notationen verwendet. Die Notationen und Voraussetzungen aus 3.1.1 sind ab hier nicht mehr länger gültig. Sofern nicht anders angegeben, wird das Signifikanzniveau α aller Tests auf 0.05 gesetzt.

3.2.1 Lilliefors-Test

Gegeben seien unabhängige, identisch verteilte Zufallsvariablen $Y_1, \ldots, Y_n$, mit $n \geq 1$ dem Stichprobenumfang. Von Interesse ist, ob diese Zufallsvariablen aus einer Normal-

verteilung stammen könnten. Es wird also ein Test gesucht, welcher überprüft, ob die folgende Nullhypothese unter den genannten Voraussetzungen plausibel ist:

$$H_0 : \forall i \in \{1, \ldots, n\} : Y_i \sim N(\mu, \sigma^2).$$

Hierbei sind die Parameter μ und σ^2 unbekannt.

Ein für diese Fragestellung geeignetes Test-Verfahren des Kolmogorov-Smirnov Typs wurde von Lilliefors (1967, S. 399 – 402) vorgeschlagen. Hierbei werden die Parameter μ und σ^2 durch das arithmetische Mittel und die empirische Varianz geschätzt:

$$\hat{\mu} = \frac{1}{n} \sum_{j=1}^{n} Y_j; \quad \widehat{\sigma^2} = \frac{1}{n-1} \sum_{j=1}^{n} (Y_j - \hat{\mu})^2.$$

Mithilfe der Verteilungsfunktion der entsprechenden Normalverteilung wird anschließend die Kolmogorov-Smirnov Teststatistik berechnet. Diese entspricht dem maximalen, betragsmäßigen Abstand zwischen der empirischen Verteilungsfunktion und der vermuteten Verteilungsfunktion. Wenn $\Phi_{(\hat{\mu}, \widehat{\sigma^2})}$ die Verteilungsfunktion der betrachteten Normalverteilung und $\hat{F}$ der empirischen Verteilungsfunktion der $Y_1, \ldots, Y_n$ entspricht, so ist die Kolmogorv-Smirnov Teststatistik in diesem Fall:

$$\max_{\forall y} \left| \hat{F}(y) - \Phi_{(\hat{\mu}, \widehat{\sigma^2})}(y) \right|.$$

Lilliefors (1967) gibt verbesserte Quantile an, mit denen die Teststatistik verglichen werden kann, welche durch Monte-Carlo Methoden erhalten worden sind. Diese Quantile sind gegenüber den üblichen Quantilen des Kolmogorov-Smirnov-Tests vorzuziehen, da diese sonst zu einem konservativen Testen führen, d.h. zu einem „zu häufigen" Verzicht aufs Ablehnen der Nullhypothese und damit zu einem häufigeren Auftreten des Fehlers zweiter Art.

Dallal und Wilkinson (1986, S. 294 –296) haben die verbesserten technischen Mölichkeiten genutzt um ausführlichere Simulationen und darauf aufbauend präzisere Approximationen der Quantile der Kolmogorov-Smirnov Teststatistik für den hier beschriebenen Fall zu entwickeln.

3.2.2 Runs-Test

Seien n binäre, identisch verteilte Zufallsvariablen $Y_1, \ldots, Y_n$ mit den Ausprägungen +1 und -1 gegeben, sodass $P(Y_i = 1) = 1 - P(Y_i = -1)$ für alle i gilt. $Y_1, \ldots, Y_n$ können also als Bernoulli-Variablen aufgefasst werden mit gemeinsamen Parameter p. Dass Bernoulli-Variablen üblicherweise mit den möglichen Ausprägungen 0 und 1 definiert werden, ist kein Problem, da die hier vorgestellte Methode sich auf solche Variablen übertragen lässt. Die Realisierungen von $Y_1, \ldots, Y_n$ werden mit $y_1, \ldots, y_n$ bezeichnet. Es sei $\mathcal{I}_A$ die Indikator-Funktion zur Menge A, also

$$\mathcal{I}_A(x) = \begin{cases} 1; & x \in A \\ 0; & x \notin A \end{cases} .$$

Weiterhin sei m die Anzahl vorgekommener +1 unter den Realisierungen, also

$$m := \sum_{j=1}^{n} \mathcal{I}_{\{+1\}}(y_j).$$

Definiere nun einen Run der Länge l als die Anzahl Realisierungen ab einem Index i, für die gilt: $y_i = \ldots = y_{i+l}$ und $y_{i+l+1} \not= y_i$ sowie, falls $i > 1$: $y_{i-1} \neq y_i$. Weiterhin sei r die Anzahl Runs, welche in der Stichprobe vorkommen.

Die Teststatistik des Runs-Test ist (Siegel und Castellan 1988, S. 58 – 62):

$$T_Y(r) := \frac{r - \mu_r}{\sigma_r},$$

wobei

$$\mu_r := \frac{2m(n-m)}{n} + 1; \qquad \sigma_r := \sqrt{\frac{2m(m-n)[2m(n-m)-n]}{n^2(n-1)}}.$$

Die Nullhypothese laute:

$$H_0 : Y_1, \ldots, Y_n \text{ sind stochastisch unabhängig}$$

Falls n hinreichend groß ist, kann die Verteilung von $T_Y(r)$ durch die Standardnormalverteilung approximiert werden, also

$$T_Y(r) \overset{approx.}{\sim} N(0,1).$$

Siegel und Castellan (1988) empfehlen für die Approximation mindestens $n > 20$. Der Test lehnt H_0 zum Niveau α ab, falls die Teststatistik betragsmäßig größer dem $1 - \frac{\alpha}{2}$-Quantil der Standardnormalverteilung ist, also falls $|T_Y(r)| > q_{1-\frac{\alpha}{2}}$ gilt.

Der Runs-Test lässt sich auch zum Prüfen der Unabhängigkeit nicht binärer Zufallsvariablen benutzen. Seien $X_1, \ldots, X_n$ identisch verteilte reelwertige Zufallsvariablen und k ein geeigneter Schwellenwert. Definiere anschließend für alle i:

$$Y_i := \mathcal{I}_{[k,\infty)}(X_i) - \mathcal{I}_{(-\infty,k)}(X_i).$$

Unter Gültigkeit der Nullhypothese

$$H_0^* : X_1, \ldots, X_n \text{ stochastisch unabhängig}$$

folgt die stochastische Unabhängigkeit der $Y_1, \ldots, Y_n$, welche wieder binär sind. Lehnt der Runs-Test diese ab, folgt auch die Ungültigkeit von H_0^*.

3.3 Praktische Anwendung der Methoden

Zur Anwendung aller Methoden wurde die Programmiersprache `R` benutzt (R Core Team 2019). Für das Anpassen eines (klassischen) linearen Modells existiert die Funktion `lm`. Das elastische Netz, die Lasso und die Ridge Regression lassen sich mithilfe der Funktion aus dem gleichnamigen Paket `glmnet` durchführen (Friedman et al 2010). Die Kreuzvalidierung für die Parameter des elastischen Netzes wurde mithilfe der Funktion `train` aus dem Paket `caret` durchgeführt (Kuhn 2019).

Der Lilliefors-Test wird mithilfe der Funktion `lillie.test` aus dem Paket `nortest` ausgeführt (Gross und Ligges 2015), und der Runs-Test mithilfe der Funktion `runs.test` aus dem Paket `tseries` (Trapletti und Hornik 2019).

4 Erstellung, Bewertung und Auswahl von Modellen

4.1 Vorplanung und erster Modellierungsversuch

Das primäre Ziel ist, zu wenigstens einem der Risse des nördlichen Überbaus ein geeignetes Modell zu finden. Hierzu ist zunächst abzuklären, welche Variablen in das Modell einfließen sollen und wie modelliert werden soll.

4.1.1 Auswahl der Variablen

Zu jeder Stunde sind zu jedem Wegaufnehmer die mediane Rissbreite, die mediane Temperatur an der Unterseite und die mediane Temperatur an der Oberseite der Brücke bekannt. Durch Untersuchung mittels deskriptiver Methoden wie in Abbildung 1 drängt sich die Vermutung auf, dass eine hohe Temperatur ein breiter werden des Risses begünstigt, ebenso wie das eine hohe Temperatur größere Schwankungen in der Rissbreite verursacht. Ungeklärt bleibt jedoch, wie genau die Einflussnahme der Temperatur auf die Rissbreite aussieht.

Neben den gegebenen Variablen kommen daher viele weitere Variablen in Frage, die durch nichtlineare Transformation der Temperaturen erhalten werden. Diese Variablen sind im Vorfeld durch die Messungen nicht gegeben und müssen explizit berechnet werden. Da sich durch solche Transformationen unendlich viele Variablen bestimmen lassen, lässt sich eine gewisse Willkürlichkeit nicht vermeiden, d.h. es werden einige Variablen erstellt und berücksichtigt, während andere nicht erwogen werden.

Somit wird neben den gegebenen Variablen noch eine Auswahl weiterer Variablen in Betracht gezogen. Es ist bereits an dieser Stelle davon auszugehen, dass einige der betrachteten Variablen keine Rolle spielen werden. Neben den mittels der Temperaturmessungen bestimmten Variablen lassen sich auch noch einige vom Zeitpunkt der Erhebung des Datenpunktes abhängige Variablen erstellen. So z.B. ist anzunehmen, dass der Verkehr einen Einfluss auf die Rissbreiten hatte. Da der Verkehr als Variable jedoch nicht erhoben worden ist, wird erhofft, dass er durch seine Abhängigkeit vom Zeitpunkt der Erhebung berücksichtigt werden kann. Beispielsweise verkehren nachts für gewöhnlich weniger Fahrzeuge als tagsüber. Daher wurden auch verschiedene zeitliche Variablen betrachtet.

Ebenfalls Berücksichtigung findet die Rissbreite des zu betrachtenden Risses der nahen Vergangenheit. Dies soll unter anderem die Unabhängigkeit der Datenpunkte gewährleisten. Beispielsweise ist anzunehmen, dass ein Riss, der zum Zeitpunkt A bereits sehr groß ist, eine Minute später immer noch groß ist, auch wenn er „verschmälernden" Einflüssen ausgesetzt sein sollte.

Weiterhin wurden einige Sprungvariablen in Abhängigkeit der Temperatur aufgenommen. So unterscheiden diese Variablen, ob sich die Temperatur oberhalb von 0, 4, 10, $20°C$ befanden oder nicht. Die Motivationen sind hierbei, die Anomalie des Wassers ($4°C$) und der Gefriepunkt ($0°C$). Die beiden Sprünge zu den anderen Temperaturen dienen zum Vergleich um sicherzustellen, dass eventuell auffliegende Effekte bei den ersten beiden Sprungvariablen nicht einfach nur allgemein auf die bloße Existenz temperaturabhängiger Sprungvariablen zurückzuführen ist.

Schließlich spielt eventuell auch die verstrichene Zeit seit Beginn des Monitorings bis zur Aufnahme des Datenpunktes eine Rolle. Mithilfe dieser Variable soll die Frage untersucht werden, ob die Rissbreite sich allgemein im Laufe der Zeit verändert hat. Hat diese Variable einen positiven Effekt, deutet das auf eine konsequente Vergrößerung der Risse hin, die nicht auf die temperatur- oder verkehrsbedingten Schwankungen zurückzuführen sei.

Verschiedene Interaktionsterme wurden ebenfalls berücksichtigt. In Tabelle 1 werden alle in Erwägung gezogenen Variablen aufgelistet. Es sei hier erwähnt, dass unter dem Begriff „Variable" in diesem Bericht auch das verstanden wird, was in Abschnitt 3.1.2.2 als „Untervariable" bezeichnet wurde. Eine Klasse kategorialer Variablen, wie z.B. „Monat" mit 12 Kategorien, zerfällt in 11 Variablen (bzw. 11 Dummys) und eine Referenzkategorie. Somit entspricht die Anzahl aller Variablen der Anzahl Spalten der Designmatrix X. Dies sind insgesamt 254.

Bisher blieb unklar, welchen Rissverlauf zu modellieren am sinnvollsten ist. Die Risse des südlichen Überbaus werden aufgrund des vorzeitigen Abrisses nicht erwogen. Als entscheidendes Kriterium bezüglich der Risse des nördlichen Überbaus wird beschlossen, auf das arithmetische Mittel der Risse zu achten. Die Hoffnung hierbei ist, dass ein großer Riss Erkenntnisse eher preisgibt, da sich mehr erkennen lässt. Aus dem selben Grund wird auch die empirische Varianz betrachtet. Tabelle 2 zeigt die entsprechenden Ergebnisse.

Am Atraktivsten erscheint demnach der Riss zu WWN2, da er sowohl den größten Mittelwert als auch die größte Varianz besitzt. Zunächst werden jedoch Modelle zu den drei Rissen mit den größten Mittelwerten gesucht werden.

4.1.2 Modellierungsansatz

Da davon auszugehen ist, dass viele Variablen keinen Effekt besitzen werden, ist eine Lasso Regression zunächst ein naheliegender Ansatz, da diese verstärkt dazu tendiert, Variablen durch auf null setzen der entsprechenden Effekte zu eliminieren. Der Ansatz der Lasso Regression leidet jedoch unter anderem an dem Problem, dass er bei Gruppen sehr stark korrelierter Variablen, wie sie in diesem Fall vorliegen, dazu tendiert, alle

Tabelle 1: Auflistung aller in Erwägung gezogener Variablen, wobei h die Einheit „eine Stunde" ist. Bezeichnungen: katego. $\hat{=}$ kategorial; Interak.$\hat{=}$Interaktion(en)

Name der Varablen(gruppe)	**Typ**	**Anzahl Variablen**
Intercept	konstant	1
vor/nach Abriss	binär	1
verstrichene Zeit seit Beginn des Monitorings	stetig	1
Temperaturen unterhalb, fortan bezeichnet mit T_B	stetig	1
Temperaturen oberhalb, fortan bezeichnet mit T_S	stetig	1
Rissweite $24h$ vorher, fortan bezeichnet mit R_{vorher}	stetig	1
mittlere T_S der letzten $i \cdot 24h$, $i \in \{1, \ldots, 7\}$	stetig	7
mittlere T_B der letzten $i \cdot 24h$, $i \in \{1, \ldots, 7\}$	stetig	7
mittleres $\lvert T_B - T_S \rvert$ der letzten $i \cdot 24h$, $i \in \{1, \ldots, 7\}$	stetig	7
Die Potenzen T_B^x und T_S^x, $x \in \{\frac{1}{2}, 2, 3, 4\}$	stetig	8
$\exp(T_B)$ und $\exp(T_S)$	stetig	2
Potenzen und exp der mittleren T_B, letzte 24 h	stetig	5
Potenzen und exp der mittleren T_S, letzte 24 h	stetig	5
$T_B > k$, wobei $k \in \{0, 4, 10, 20\}$ in $^\circ C$	binär	4
Monat	katego.	11
Wochenendtag/Wochenarbeitstag	binär	1
wievielte Woche des Jahres	katego.	127
Uhrzeit	katego.	23
Wochentag	katego.	6
$R_{vorher}\cdot$ Temperatur Variablen	Interak.	21
Abriss $\cdot$ verstrichene Zeit	Interak.	1
$R_{vorher}\cdot$ Abriss	Interak.	1
$T_B > k; k \in \{0, 4, 10, 20\}^\circ C \cdot R_{vorher}$	Interak.	4
$T_B > k; k \in \{0, 4, 10, 20\}^\circ C\cdot$ Abriss	Interak.	4
$T_B > k; k \in \{0, 4, 10, 20\}^\circ C \cdot T_B$	Interak.	4

Tabelle 2: Die Mittelwerte und emp. Varianzen der Rissbreiten der Wegaufnehmer des nördlichen Überbaus, sortiert nach ihren Mittelwerten

Bezeichnung	**arith. Mittel**	**emp. Varianz**
WWN2	0,2502	0,0018
WON2	0,1229	0,0001
WWN1	0,1140	0,0016
WWN3	0,1051	0,0009
WON3	0,0909	<0,0001
WON1	0,0860	0,0009
WWN4	0,0504	0,0007
WON4	0,0471	0,0002

Variablen der Gruppe bis auf eine zu entfernen. Ein solch zufälliges Hinauswerfen einiger Variablen erscheint wenig sinnvoll. Die Ridge Regression hingegen tendiert dazu, in solch einem Fall allen Variablen der Gruppe einen ähnlich starken Effekt zuzuordnen.

Der Kompromiss, den das elastische Netz darstellt, erscheint daher in diesem Fall als am Sinnvollsten, weshalb dieses Verfahren gewählt werden wird. Die Wahl der Parameter α und λ wird hierbei mithilfe 5-facher Kreuzvalidierungen getroffen. Hierzu werden von der Funktion `train` jeweils 10 Kandidaten für α und λ gewählt. Zu jeder Kombination von Kandidaten wird dann eine 5-fache Kreuzvalidierung durchgeführt, um die Güte der entsprechenden Modellbildung einzuschätzen. Die Kombination von Kandidaten, die zur erfolgreichsten Modellbildung geführt hat, wird gewählt.

4.1.3 Anpassung der ersten Modelle und Diskussion

4.1.3.1 Resultate der ersten Modelle

Für die Wegaufnehmer WWN2, WON2 und WWN1 werden Modelle angepasst (mit allen erwogenen Variablen). Tabelle 3 zeigt die Wahl von α, λ, dem resultierten adjustierten R^2 (also $\tilde{R}^2$) und der Anzahl der beibehaltenen Variablen. Außerdem beinhaltet sie die Testergebnisse des Runs-Tests und des Lilliefors-Tests, angewandt auf die erhaltenen Residuen der jeweiligen Modelle. Der P-Wert des Runs-Test gibt hierbei an, ob die Annahme der Unabhängigkeit der Residuen erfüllt ist. Hierzu werden die Residuen mit dem Schwellenwert null verglichen, d.h. der Runs Test wird angewandt auf $z_1, \ldots, z_n$, wobei z_i folgendermaßen definiert sei ($\mathcal{I}_A$ ist die Indikatorfunktion zur Menge A, vgl. 3.2.2):

$$z_i := \mathcal{I}_{[0;\infty)}(e_i) - \mathcal{I}_{(-\infty;0)}(e_i) = \begin{cases} +1; & fallsx_i \geq 0 \\ -1; & fallsx_i < 0. \end{cases}$$

Der Liliefors-Test wird angewandt um in Erfahrung zu bringen, ob die Residuen einer gemeinsamen Normalverteilung folgen könnten. Dies ist deshalb allgemein interessant zu wissen, da es Verfahren gibt, die in diesem Fall anwendbar werden, z.B. Varianzanalysen

Tabelle 3: Ergebnisse des ersten Durchlaufs der Modellierung: Parameter des elastischen netzes, adjustiertes R^2, Anzahl Variablen mit Effekt und P-Werte der Tests

	α	λ	$\tilde{R}^2$	Anzahl Variablen	Runs-Test	Lilliefors-Test
WWN2	1	$1,5813 \cdot 10^{-5}$	0,8839	168	$< 2,2 \cdot 10^{-16}$	$< 2,2 \cdot 10^{-16}$
WON2	0,4	$8,4737 \cdot 10^{-6}$	0,7774	203	$< 2,2 \cdot 10^{-16}$	$< 2,2 \cdot 10^{-16}$
WWN1	0,3	$1,4441 \cdot 10^{-5}$	0,8380	202	$< 2,2 \cdot 10^{-16}$	$< 2,2 \cdot 10^{-16}$

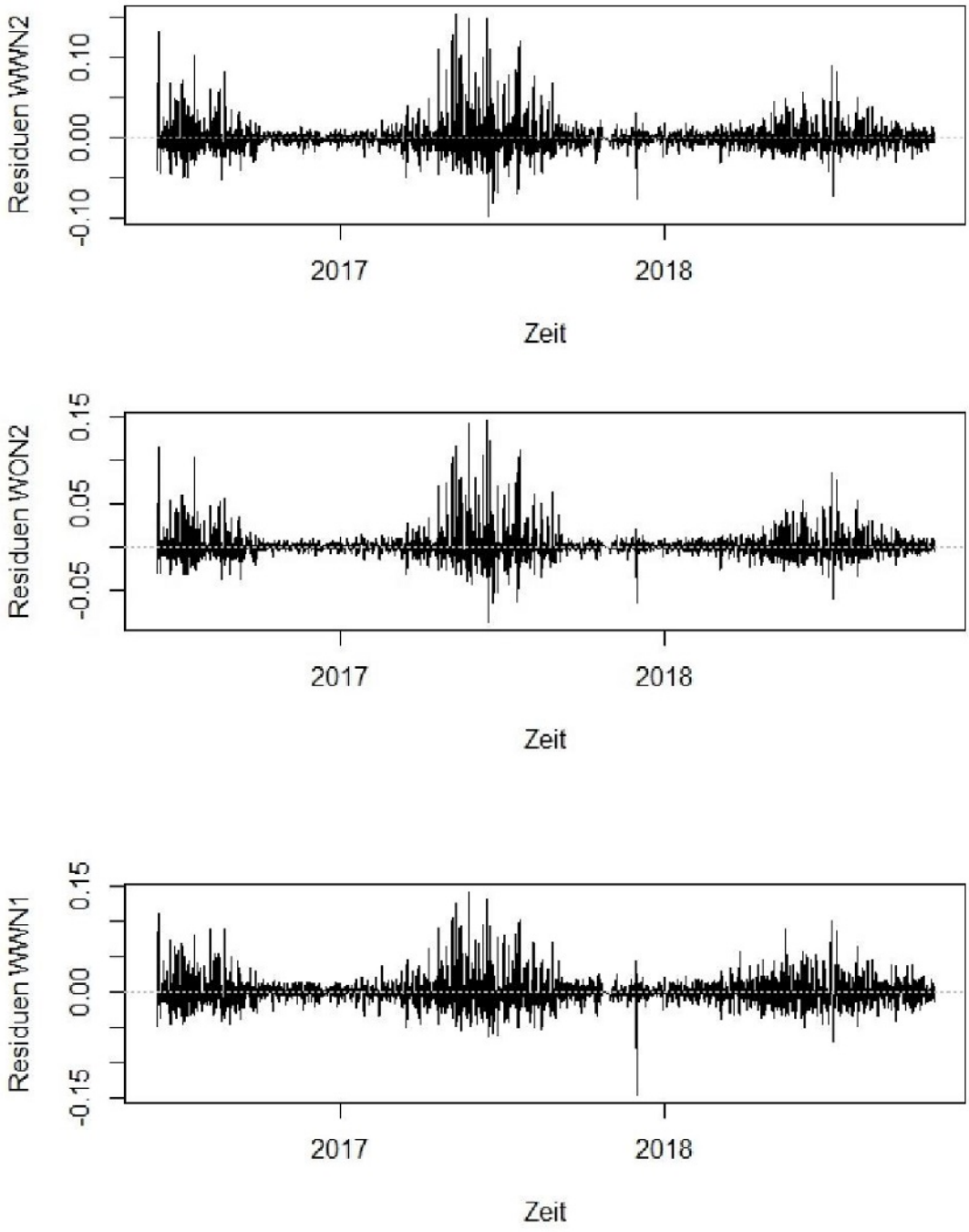

Abbildung 3: Verläufe der Residuen der vollen Modelle zu WWN2, WON2 und WWN1 gegen die Zeit

(Toutenburg 2003, S.145 –149).

Abbildung 3 zeigt die Residuen grafisch gegen die Zeit aufgetragen. Anhand dieser Grafik wird bewertet, wie es um die Anforderungen an die Residuen bezüglich homogener Varianz und Erwartungswert null steht.

4.1.3.2 Diskussion der drei Modelle

Bei allen Modellen gibt es einen schweren Einwand: Der Runs-Test lehnt die Nullhypothese, die $z_1, \ldots, z_n$ seien Realisierungen unabhängiger Zufallsvariablen, deutlich ab, was auch der Unabhängigkeitsannahme der Residuen widerspricht.

Das arith. Mittel der Residuen $\bar{e}$ erfüllt in allen drei Fällen folgendes:

$$10^{-10} > |\bar{e}| \approx 0$$

Daher wird die Anforderung an die Residuen, einen Erwartungswert von null zu besitzen, als denkbar betrachtet. An dieser Stelle wird darauf hingewiesen, dass eine alternative Möglichkeit besteht, die Ergebnisse der Runs-Tests zu interpretieren. Da es eine der Voraussetzungen des Runs-Tests ist, dass die $Y_1, \ldots, Y_n$ aus 3.2.2 einer gemeinsamen Verteilung entstammen, bzw. einer Bernoulli-Verteilung mit gleichem Parameter p, könnte das Ablehnen des Runs-Tests auch auf eine Verletzung dieser Annahme zurückgeführt werden. Dies wäre zum Beispiel der Fall, wenn die Erwartungswerte der Residuen nicht konstant identisch sind, sondern von Residuum zu Residuum variieren. Sollten für einige Residuen die Erwartungswerte kleiner als 0, für andere größer als 0 sein, würde auch das beobachtete arithmetische Mittel $\bar{e}$ der Residuen keinen Widerspruch darstellen. Aufgrund des arithmetischen Mittels wird im folgenden dennoch von einer Annahmeverletzung der Unabhängigkeit ausgegangen, da dies plausibler erscheint.

Weiterhin ist anzumerken, dass neben der Unabhängigkeitsannahme auch die Homoskedaszität der Residuen nicht erfüllt sein kann: in Abbildung 3 ist deutlich erkennbar, dass die Residuen in den warmen Jahreszeiten wesentlich stärker streuen als in den kalten. Zu den Ergebnissen der Lilliefors-Tests ist anzumerken, dass die P-Werte auch eine Folge der gerade festgestellten Verletzungen der Voraussetzungen des Tests sein könnten, wie es z.B. auch die Heteroskedaszität ist.

Es sind gleich zwei Annahmen offenkundig schwer verletzt, weshalb die Modelle als für Anwendungszwecke ungeeignet eingestuft werden. Würde von dieser Einstufung abgesehen werden, würde das adjustierte R^2 zumindest WWN1 empfehlen, dies ist durch die Annahmeverletzungen jedoch irrelevant geworden.

Eventuell könnten andere Werte in den Parametern α und λ zu besseren Resultaten

führen. Mit der Kreuzvalidierung und der Wahl von 10 Kandidaten zu jedem Parameter sind jedoch bereits 100 Kombinationen mit hinreichender Ausführlichkeit untersucht worden, sodass von einer weiteren Suche nach besseren Werten abgesehen wird. Es wird der Vollständigkeit halber darauf hingewiesen, dass bei diesem Vorgehen bei WWN2 der Wert $\alpha = 1$ zustande gekommen ist, was einer Lasso Regression entspricht.

4.2 Wiederholung des Vorgehens mit reduzierten Daten

Wie in Abschnitt 4.1.1 erwähnt wurde erhofft, den nicht erhobenen Einfluss des Verkehrs mithilfe zeitlicher Variablen auffangen und berücksichtigen zu können. Aufgrund der Unbrauchbarkeit der oben genannten Modelle wurde nun die Idee verfolgt, den Einfluss des Verkehrs stattdessen durch eine Reduzierung des Datensatzes weitestgehend zu eliminieren. Es wurde vermutet, dass die Zeit zwischen 4 und 5 Uhr morgens die Stunde am Tag ist, wo am wenigsten oder zumindest kaum Verkehr unterwegs ist. Daher wurde das Vorgehen aus 4.1 wiederholt, nachdem der Datensatz reduziert wurde und nur noch die in diesem Zeitraum vorliegenden Punkte beinhaltete (die Variable „Uhrzeit" wurde allerdings aus Gründen der Sinnhaftigkeit entfernt). Für den reduzierten Datensatz gilt mit den Bezeichnungen aus 3.1: $n = 865$.

Tabelle 4 beinhaltet analog zu Tabelle 3 die bei diesem Vorgehen erhaltenen Ergebnisse. Abbildung 4 beinhaltet erneut die Residuen aufgetragen gegen die Zeit. Auch bei diesen Modellen war das arithmetische Mittel der Residuen stets approximativ null, so dass die Annahme eines entsprechenden Erwartungswertes haltbar erscheint.

Bei allen drei Modellen lässt sich in Abbildung 4 erneut deutlich Heteroskedaszität feststllen, d.h. die Annahme gleicher Varianzen ist in allen drei Fällen verletzt.

Tabelle 4 hingegen deutet auf einen höheren Erfolg der Modellbildung hin, d.h. die Reduzierung des Datensatzes hat scheinbar tatsächlich die Erklärbarkeit der Daten durch das gewählte Vorgehen erhöht. Die $\tilde{R}^2$ sind in allen Fällen recht groß, trotz der großen Anzahl beteiligter Variablen. Der Riss zu WWN1 erscheint hier weniger erfolgsversprechend, da er sowohl das kleinste $\tilde{R}^2$ als auch die geringste Anzahl Variablen besitzt (und

Tabelle 4: Parameter α, λ sowie das adj. R^2, bezeichnet mit $\tilde{R}^2$, Anzahl beibehaltener Variablen und die P-Werte des Runs- und des Lilliefors-Tests zu den Residuen bei der erneuten Anpassung der Modelle mit reduzierten Daten

	α	λ	$\tilde{R}^2$	Anzahl Variablen	Runs-Test	Lilliefors-Test
WWN2	0,3	$1,0437 \cdot 10^{-5}$	0,9568	208	0,0228	$1,394 \cdot 10^{-8}$
WON2	0,3	$5,6029 \cdot 10^{-6}$	0,9424	215	0,7558	$7,730 \cdot 10^{-8}$
WWN1	1	$4,4226 \cdot 10^{-5}$	0,83733	146	0,0530	$1,049 \cdot 10^{-10}$

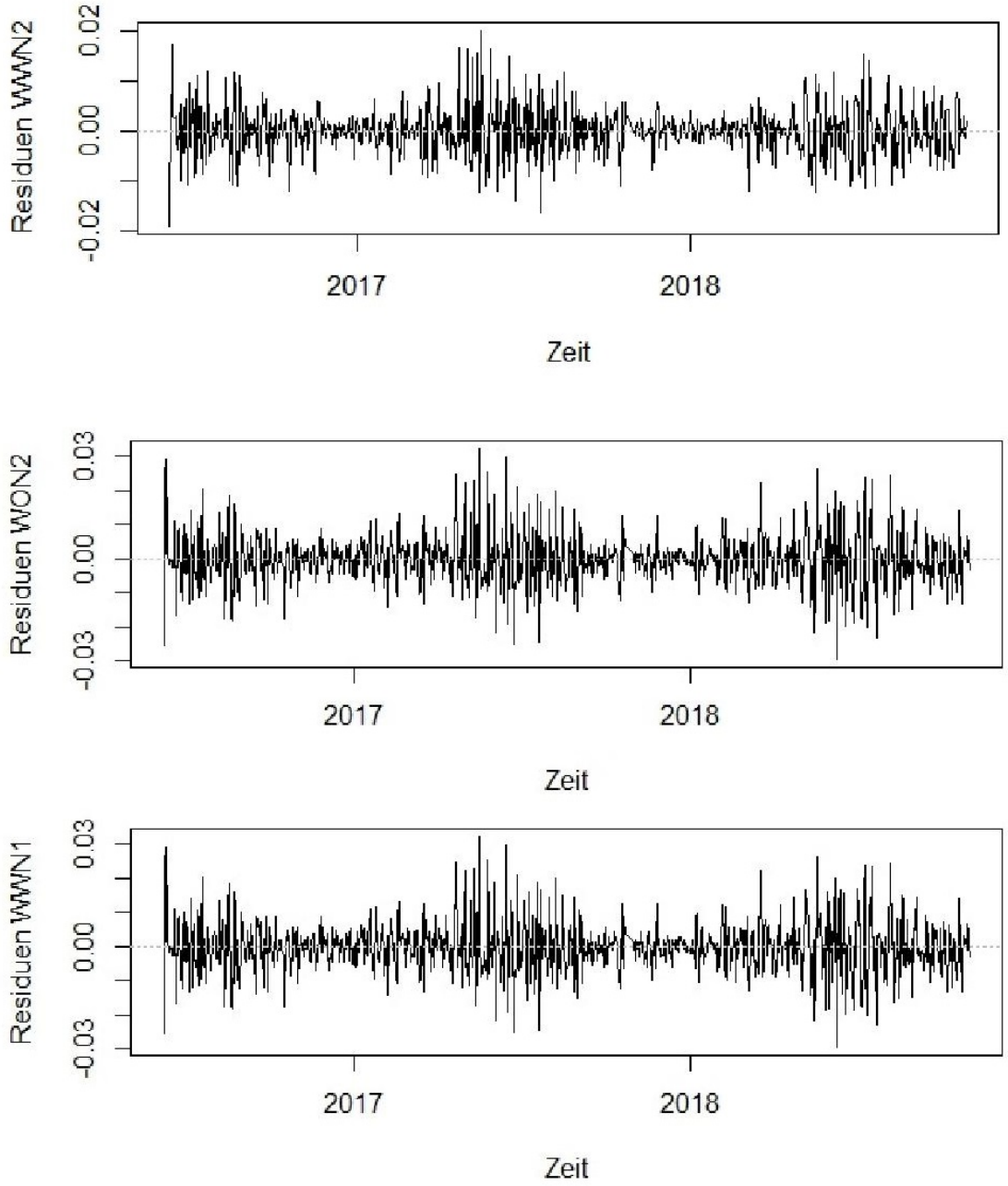

Abbildung 4: Verläufe der Residuen gegen die Zeit der 3 Modelle mit reduzierten Daten

somit den geringsten „Bestrafungseffekt" auf $\tilde{R}^2$). Es fällt auf, dass $\alpha = 1$ zustande gekommen ist, d.h. bei WWN1 wurde eine Lasso Regression durchgeführt.

Abschließend werden die Runs-Tests betrachtet. Um ein sinnvolles Modell zu finden ist es wichtig, dass die zugrunde liegenden Annahmen als (wenigstens approximativ) gültig bewertet werden können. Den Testergebnissen der Runs-Tests, welche die Residuen auf Unabhängigkeit untersuchen, kommt hierbei besondere Aufmerksamkeit zu. Als Richtwert wird hierbei das Niveau 0,05 betrachtet. Streng genommen lehnt der Runs-Test die Unabhängigkeitsannahme nur bei dem Modell zu WWN2 ab. Es ist jedoch anzumerken, dass bei WN1 der P-Wert nur sehr geringfügig größer ist als 0,05. Es kam also beinahe zu einer Ablehnung der Nullhypothese der Unabhängigkeit. Hingegen ist bei WON2 der P-Wert deutlich größer als 0,05.

Allgemein können P-Werte auch als Maß interpretiert werden, welche den Abstand zwischen dem beobachteten Verhalten und dem unter der Nullhypothese (und den weiteren Testvoraussetzungen) zu erwartendem Verhalten der Stichprobe misst. Werden die P-Werte der Runs-Tests hier so interpretiert, ermutigt dies, sich im folgenden auf WON2 zu konzentrieren, da hier die Modellannahmen besser erfüllt scheinen.

Wird von einer Interpretation des P-Wertes als Maßzahl abgesehen, so „disqualifizieren" die P-Werte der Runs-Tests das Modell zu WWN2 in gewissem Maße, da dies der einzige Wegaufnehmer ist, zu dem die Modellannahme der Unabhängigkeit der Residuen zum Niveau 0,05 abgelehnt worden ist. Ist eine Entscheidung zwischen WON2 und WWN1 erwünscht, so kann aufgrund der Argumentation bezüglich der erhaltenen $\tilde{R}^2$ auch in diesem Fall WON2 bevorzugt werden.

Im restlichen Teil dieser Bachlorarbeit wird sich daher auf WON2 konzentriert und versucht, ein Modell für diesen Riss zu finden. Entscheidend hierbei ist vor allem die Interpretation der P-Werte der Runs-Tests als „Abstandsmaß" zur Nullhypothese. Ein „gutes" Modell wie es von einem großen $\tilde{R}^2$ angedeutet wird, ist zwar ebenfalls wünschenswert, doch das Einhalten der zugrunde liegenden Annahmen wird vom Autor dieser Bachlorarbeit als wichtiger betrachtet.

4.3 Modellwahl zu WON2

Wenngleich die Annahme der Unabhängigkeit und des Erwartungswertes 0 aller Residuen im Fall von WON2 als haltbar angesehen werden kann, lässt sich in Abbildung 4 dennoch deutlich erkennen, dass die Residuen in den warmen Monaten stärker streuen, womit die Annahme der Gleichheit aller Varianzen offensichtlich nicht haltbar ist. Darüber hinaus kann es als ungeeignet angesehen werden, dass trotz des Selektionseffekts durch das elastische Netz immer noch 215 Variablen einen Effekt besitzen.

Daher wird im folgenden versucht, den Modellansatz so zu wählen, dass sowohl die Varianzen der Residuen als auch möglichst den Voraussetzungen entsprechen, als auch mehr Variablen zu selektieren. Zunächst wird zu ersterem ein Lösungsansatz vorgestellt und dessen Ausführung und Resultate.

4.3.1 Angehen des Problems der Heteroskedaszität

In den Abbildungen 3 und 4 entsteht der naheliegende Verdacht, dass die Varianzen der Residuen bei höheren Temperaturen zunehmen. Diese Vermutung wird zunächst näher untersucht. Abbildung 5 zeigt hierzu eine Grafik, welche Schätzungen der Varianzen der Residuen gegen die Temperatur aufgetragen zeigt. Zur Erstellung dieser Grafik wurden die Datenpunkte zunächst gemäß der Variable T_B (also der Temperatur der Unterseite der Brücke) sortiert. Anschließend wurde zu jeweils 7 nahe beieinanderliegenden Datenpunkten dieses sortierten Datensatzes die empirischen Varianzen bestimmt. Ebenfalls wurden zu allen 7 Datenpunkten das arithmetische Mittel von T_B gebildet.

Es sei angemerkt, dass mit den Bezeichnungen aus 3.1 für ein Modell aus (1) gilt:

$$\mathrm{var}(Y) = \mathrm{var}\left(X\beta + \varepsilon\right) = \mathrm{var}(\varepsilon)$$

D.h. die Varianzen der Rissbreite entsprächen der Varianz der Residuen, wären alle Voraussetzungen aus 3.1 erfüllt. Abbildung 5 lässt jedoch erkennen, dass die Varianzen der Rissbreite bei hohen Werten der Temperatur zuzunehmen scheinen.

Diese Erkenntnis deckt sich zunächst mit dem, was sich hat aus den Abbildungen 1 und 2 folgern lassen und legt nahe, dass die Voraussetzungen aus 3.1 zunächst nicht gelten.

4.3.1.1 Lösungsansatz durch eine Hilfsfunktion

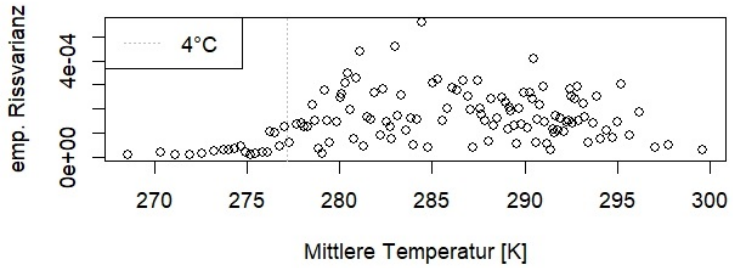

Abbildung 5: empirische Varianzen der Rissbreiten von WON2 von jeweils 7 Datenpunkten, welche möglichst ähnliche zugehörige Temperaturwerte besitzen

Um das Problem der Heteroskedaszität in diesem Fall zu lösen, wird auf die Idee einer Hilfsfunktion $s(T_B)$ zurückgegriffen. Es wird hierbei davon ausgegangen, dass alle Annahmen aus 3.1.1 erfüllt sind, jedoch der Zusammenhang zwischen Y und X nicht dem aus (1) entspricht, sondern einer leicht veränderten Version:

$$\begin{aligned} & Y = X\beta + \varepsilon \cdot s(T_B) \\ \Leftrightarrow & \frac{Y}{s(T_B)} = \frac{X}{s(T_B)}\beta + \varepsilon. \end{aligned} \tag{8}$$

Die Heteroskedaszität wäre dann darauf zurückzuführen, dass bisher versucht worden ist, ein Modell wie in (1) anzupassen, obwohl ein Zusammenhang wie in (8) der Wahrheit entspricht, wobei die Funktion s als positiv definit und monoton wachsend angenommen wird. Gilt dieser Zusammenhang, so ist $\text{var}(Y) = (s(T_B))^2 \sigma^2$ und es wäre erklärbar, wie Abbildung 5 zustande kommt. Außerdem sei in diesem Fall o.B.d.A. $\sigma = 1$ angenommen, da die Variabilität der Residuen nun durch s beschrieben werden kann, bzw. (8) auch mithilfe einer Funktion $\tilde{s}(T) := s(T)\sigma$ hätte formuliert werden können.

Es stellt sich an dieser Stelle jedoch auch die Frage, warum die Punkte in Abbildung 5 bei größeren Temperaturen stärker zu streuen scheinen. Dies drängt die Vermutung auf, dass die Varianzen nicht monoton mit der Temperatur steigen, sondern bei hoher Temperatur viel mehr einfach nur variieren, also eventuell auch wieder schrumpfen.

An dieser Stelle wird erstens angemerkt, dass in Abbildung 5 Varianzschätzungen vorliegen, und zweitens auf Abschnitt 3.1.5 verwiesen. Dort wurde die Varianz der Varianzschätzfunktion bei Annahme zugrundeliegender, normalverteilter, unabhängiger Daten berechnet. Das erhaltene Ergebnis lieferte, dass bei größerer, zugrunde liegender Varianz auch die Varianz der Schätzungen zunimmt. Somit ließe sich plausibel erklären, weshalb die Punkte aus Abbildung 5 stärker streuen. Es muss an dieser Stelle aber auch darauf hingewiesen werden, dass für die Rissbreiten von WON2 weder Unabhängigkeit noch Normalverteilung angenommen werden können (der Lilliefors-Test liefert einen P-Wert von $< 2,2 \cdot 10^{-16}$). Dass die Varianz des Varianzschätzers auch bei den hiesigen Daten zunimmt kann jedoch, nachdem ein theoretischer Fall, in dem das vorkommt, gefunden wurde, nicht als allgemein unplausibel angesehen werden.

4.3.1.2 Wahl der Streufunktion

An die Funktion s, welche im folgenden Streufunktion genannt wird (weil sie die Streuung der Residuen beeinflusst), wurden die Anforderungen gestellt, dass sie

1. positiv definit sei: $\forall T : s(T) > 0$

2. monoton wachsend sei: $\forall \delta > 0 : s(T + \delta) \geq s(T)$

3. von der Temperatur an der Unterseite der Brücke T_B abhängt.

Die Anforderung 3 ist durch Abbildung 5 motiviert. Die 1. Eigenschaft wird gefordert, da unter anderem der Spezialfall $s(T) = 0$ ausgeschlossen werden muss, aber auch, weil sie mit der Eigenschaft 2 zusammen bewirkt, dass die Funktion s für steigendes T eine Erhöhung der Rissvarianzen erzielt. Außerdem gilt somit immer: $s(T) = \sqrt{(s(T))^2}$, was sich im folgenden als nützlich erweisen wird.

Das genaue Aussehen der Funktion ist natürlich nicht bekannt. Es ist nicht ausgeschlossen, dass sie eventuell noch von weiteren Variablen abhängt. Um eine sinnvolle Streufunktion zu erhalten ist ein Modell gesucht, welches die Varianzschätzungen der Rissbreiten in Abhängigkeit der Temperatur unterhalb der Brücke darstellt. Da die Funktion unkompliziert sein sollte und fürs erste nur vom Intercept und der Temperaturvariablen abhängen sollte, erscheinen Schätzverfahren wie Lasso, Ridge oder das elastische Netz als unangemessen, sodass der naheliegendste Ansatz ein einfaches, lineares Modell (wie in Abschnitt 3.1.2 beschrieben) ist.

Ansatz 1:

$$s_1(T) = \gamma_0 + \gamma_1 T + \tilde{\varepsilon}.$$

Hierbei sei mit γ_0 der Intercept, mit γ_1 der Effekt der Temperatur und mit $\tilde{\varepsilon}$ der Fehler dieses Modells bezeichnet. Somit ist $\gamma := (\gamma_0, \gamma_1)^T$ der unbekannte Parametervektor. Ansatz 1 scheint aufgrund seiner Schlichtheit geeignet, birgt jedoch den Nachteil, dass es nicht sicher positiv definit ist (da aber der Physik zufolge $T \in [0, \infty)$ gilt, könnte s_1 theoretisch positiv definit sein).

Ein Ansatz, der die positive Definitheit sicher gewährleistet, ist der folgende:

Ansatz 2:

$$s_2(T) = \exp\{\gamma_0 + \gamma_1 T + \tilde{\varepsilon}\} \quad \Leftrightarrow \quad \log(s(T)) = \gamma_0 + \gamma_1 T + \tilde{\varepsilon}$$

Anzumerken zu beiden Ansätzen ist, dass bereits angenommen wird, dass die Varianz des Varianzschätzers zunimmt. Die Annahme gleicher Varianzen des linearen Modells kann bei der Berechnung von s_1 und s_2 nicht erfüllt sein. Da hierbei jedoch nicht ein gutes Modell, sondern einfach nur möglichst geeignete Streufunktionen gesucht sind,

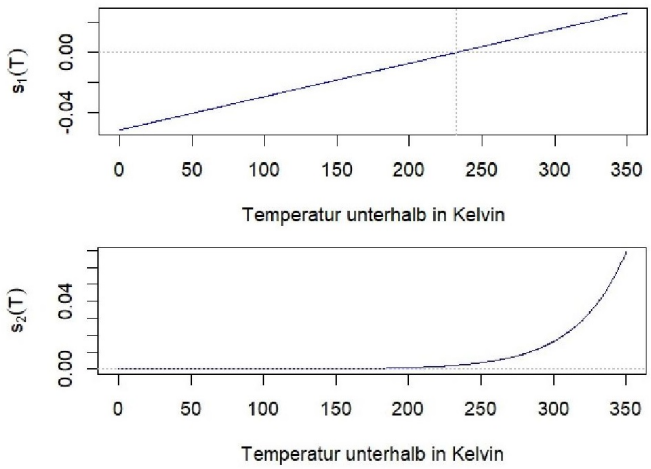

Abbildung 6: Die Verläufe von s_1 aus Ansatz 1 (oben) und s_2 aus Ansatz 2 (unten)

werden diese Annahmeverletzungen in Kauf genommen in der Hoffnung, dass sie sich im Rahmen halten.

Tabelle 5 zeigt die erhaltenen Schätzungen für die Parameter γ_0 und γ_1 und Abbildung 6 zeigt die Verläufe der Funktionen s_1 und s_2 der beiden Ansätze. Es ist zu erkennen, dass s_1 bei bestimmten Temperaturen negativ werden kann. Die Nullstelle liegt bei etwa $232K$ (umgerechnet sind das etwa $-41°C$). Derart kalte Temperaturen erscheinen im Datensatz nicht, d.h $s_1(T) > 0$ für alle betrachteten T.

Mithilfe dieser beiden Funktionen werden nun erneut zwei Modelle zu WON2 angepasst. Das Vorgehen entspricht hierbei dem aus 4.2, wobei allerdings die Designmatrix und die abhängige Variable wie in (8) transformiert werden.

Die Tabelle 6 enthält die selben Größen dieser beiden Modelle, welche auch in den Tabellen 3 und 4 für die dort thematisierten Modelle aufgelistet worden sind. Abbildung 7 enthält die Residuen dieser Modelle, aufgetragen gegen die Zeit.

Beide Modelle weisen bezüglich des Runs-Tests einen zufriedenstellend hohen P-Wert auf. Die Annahme der Unabhängigkeit der Residuen ist somit haltbar. Selbiges gilt für die Annahme, die Residuen besäßen den Erwartungswert null. Sowohl das arithmetische Mittel der Residuen verschwindet in beiden Fällen, als auch spricht Abbildung 7 dafür,

Tabelle 5: Schätzungen der Parameter für die beiden Ansätze zur Modellierung von s

	$\hat{\gamma}_0$	$\hat{\gamma}_1$
Ansatz 1	-0,05136	0,00022
Ansatz 2	-12,75726	0,02881

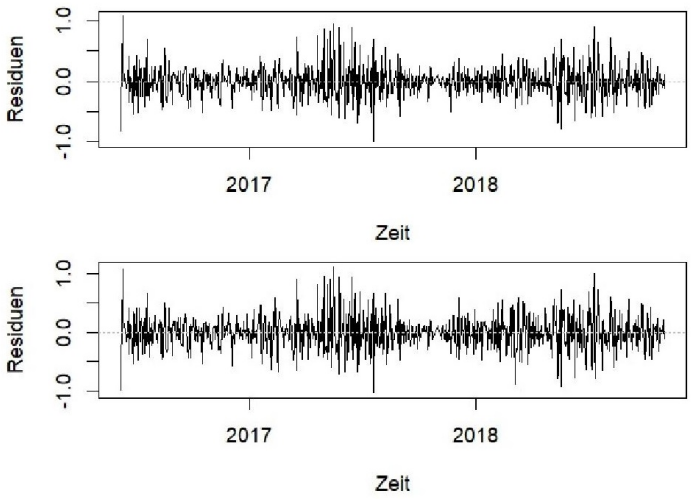

Abbildung 7: Residuen gegen die Zeit. Ansatz 1 oben und Ansatz 2 unten

dass die Residuen zumindest augenscheinlich um die null streuen.

Es fällt in Abbildung 7 erneut auf, dass Heteroskedaszität vorzuliegen scheint, wenngleich sie weniger deutlich zu Tage tritt. Beide Ansätze waren nicht fähig, dieses Problem vollständig zu beheben, aber es abzumildern. Hierbei fällt auf, dass es nach wie vor die warmen Jahrszeiten sind, an denen die Streuung der Residuen stärker zu sein scheint als zu den übrigen Zeiten. Dies stellt eine Motivation dar, eventuell die Streufunktion s anders zu modellieren, vielleicht unter Berücksichtigung weiterer Variablen. Hierfür böten sich beispielsweise die Monate an.

Es fällt auf, dass beide Modelle sehr hohe adj. R^2 besitzen, aber auch viele Variablen beibehalten haben, vor allem Ansatz 2. Dies könnte auch daran liegen, dass der Parameter α bei Ansatz 1 wesentlich größer ausgefallen ist als bei Ansatz 2, womit das elastische Netz bei Ansatz 1 deutlich „Lasso-lastiger" gerechnet wurde.

Alles in allem scheinen die beiden Modelle vergleichbar erfolgreich. Nimmt man das $\tilde{R}^2$ als Kriterium zur Modellwahl, so würde das Modell zu Ansatz 2 bevorzugt werden. Die deutlich höhere Anzahl Variablen - wenngleich $\tilde{R}^2$ bereits ein Maß ist, welches das

Tabelle 6: Parameter α, λ, sowie adj. R^2 ($\tilde{R}^2$), die Anzahl Variablen mit Effekt und die P-Werrte des Runs-Tests und des Lilliefors-Tests, angewandt auf die Residuen

	α	λ	$\tilde{R}^2$	Anzahl Variablen	Runs-Test	Lilliefors-Test
Ansatz 1	0,9	$8,2927 \cdot 10^{-4}$	0,9785	165	0,8186	$9,034 \cdot 10^{-9}$
Ansatz 2	0,3	0,00121	0,9857	204	0,8118	$4,6969 \cdot 10^{-4}$

bestraft - könnte hier die Versuchung regen, Ansatz 1 dennoch den Vortritt zu geben.

Der Lilliefors-Test lehnt in beiden Fällen die Annahme der Normalverteilung ab. Eine Normalverteilungsannahme ist auch bei diesen beiden Modellen also nicht haltbar.

Schlussendlich wird Ansatz 2 in dieser Arbeit gegenüber Ansatz 1 bevorzugt. Einerseits wird so entschieden, da Ansatz 2 sich besser verallgemeinern lässt als Ansatz 1, welcher bei extrem tiefen Temperaturen anzunehmender Weise Unsinn liefern würde (und den streuungsausschließenden Fall $s(T) = 0$ zulässt). Darüber hinaus scheinen die Ansätze nahezu gleichwertig. Den Vorteil von Ansatz 1, zu weniger Variablen geführt zu haben, könnte sich auch auf Ansatz 2 übertragen, wenn noch weitere Werte für die Parameter α und λ ausprobiert werden. Dies wird ohnehin nötig sein, da sowohl Ansatz 1 als auch Ansatz 2 keine gut interpretierbaren Modelle liefern, da sie den Großteil der Variablen beibehalten.

4.3.1.3 Diskussion zu Ansatz 2 mit temperaturunabhängigen, saisonalen Variablen

Wie bereits erwähnt könnte es von Interesse sein, eine weitere Variable einzuführen, welche einen saisonalen Effekt widerspiegelt, der nicht durch die Temperatur aufgefangen werden kann. Hierfür kommen unter den bereits betrachteten Variablen aus Tabelle 1 vor allem die Monate in Frage. Interessant ist das Einführen solcher zeitlichen Variablen auch deshalb, da sich so durch eventuell auch relevante Einflüsse, welche gar nicht erhoben worden sind, berücksichtigen lassen. Beispiele hierfür sind Urlaubsverkehr oder Feuchtigkeit.

Es ist daher nicht völlig uninteressant, eventuell einen Blick auf eine solche Erweiterung von Ansatz 2 zu werfen. In dieser Arbeit wird aber dennoch davon abgesehen. Ein Grund hierfür ist, dass beispielsweise ein Monat an sich naturwissenschaftlich betrachtet offensichtlich keinen Einfluss auf die Rissbreite haben kann, da Monate eine menschliche, nicht physikalische Erfindung sind. Wenn solche Variablen beim Modellieren Effekte erhalten, liegt das daran, dass sich die Effekte anderer, nicht erfasster Einflüsse in diese Variablen „retten" können. Die Interpretierbarkeit solcher Variablen ist daher nur begrenzt möglich.

Weiterhin wird eingewandt, dass die Einführung der Streufunktion s bereits eine Modifikation des eigentlich gesuchten Modells ist. Eine perfekte Funktion s zu finden ist nicht das ursprüngliche Ziel dieser Arbeit, welche anstrebt, das Verhalten eines Risses zu modellieren, und nicht dessen Varianz. Durch die Modellierung der Rissvarianzen sollte lediglich das Problem der Heteroskedaszität angegangen werden. Je erfolgreicher s modelliert wird, desto eher gelingt es natürlich, die Heteroskedaszität zu überwinden. Dennoch liegt auf die Perfektionierung von der Streufunktion nicht der Hauptfokus dieser

Tabelle 7: Ergebnisse der Modelle nach Ansatz 2 mit 20 und 40 Kandidaten

	α	λ	$\tilde{R}^2$	Variablenanzahl	Runs-Test	Lilliefors-Test
20 K.	0,28947	$8,1530 \cdot 10^{-4}$	0,9856,	206	0,924 6	$1,25 \cdot 10^{-4}$
40 K.	0,6538	$8,2407 \cdot 10^{-4}$	0,9858	197	0,6569	$5,069 \cdot 10^{-4}$

Arbeit.

Abschließend ist zu erwähnen, dass eine einfache Form der Streufunktion besonders erstrebenswert ist. Die Funktionen s_1 und s_2 in der Form in der sie oben vorgestellt wurden, sind stetig und unkompliziert. Hingegen würde die Einführung kategorialer Variablen wie die Monate es wären, die s_1 und s_2 vermutlich zu weniger leicht handhabbaren Treppenfunktionen werden lassen - außer alle Monatseffekte werden exakt null, dann haben sie aber auch nichts geändert.

Nichtsdestotrotz könnten Verbesserungen bei der Wahl der Streufunktion allgemein zu besseren Resultaten führen. Dies könnte Stoff für weitere Forschungen sein.

4.3.2 Reduzierung der Variablen durch Variation der Parameter

Das gegenwärtig bevorzugte Modell ist jenes, welches (8) entpsricht, WON2 modelliert und mithilfe von Ansatz 2 angepasst wird. Wie bereits erwähnt hat das dort erstellte Modell den Nachteil, dass es noch sehr viele Variablen besitzt. Es wird in diesem Abschnitt daher versucht, ein geeignetes Modell mit weniger Variablen zu finden. Dies wird durch Variation der Parameter α und λ des elastischen Netzes angestrebt.

4.3.2.1 Erweiterung der Anzahl Kandidaten

Die Parameter α und λ wurden bisher durch Ausprobieren gesucht und ausgewählt: alle Kombinationen von 10 Kandidaten wurden mithilfe der 5-fachen Kreuzvalidierung getestet, die erfolgreichste Kombination schließlich benutzt. Hier wird zunächst das selbe Vorgehen nochmal durchgeführt, allerdings werden nun 20 Kandidaten je Parameter ausprobiert. Dies resultiert in 400 Kombinationen, welche zu testen sind. Da die Rechenzeit geringer ausfiel, als befürchtet, wurde anschließend das selbe Vorgehen mit 40 Kandidaten je Variable durchgeführt. Dies entspricht 1600 Kombinationen, welche untersucht werden.

Tabelle 7 zeigt wieder die Parameter, $\tilde{R}^2$, die Anzahl Variablen mit Effekt und die P-Werte der Testergebnisse. In Abbildung 8 sind die Residuen gegen die Zeit aufgetragen.

Die P-Werte der Runs-Tests sind zufriedenstellend, die arithmetischen Mittel der Residuen zu den beiden Modellen sind erneut approximativ null. Die Unabhängigkeitsannahme ist somit ebenso haltbar wie die Annahme, die Residuen hätten den Erwartungswert null.

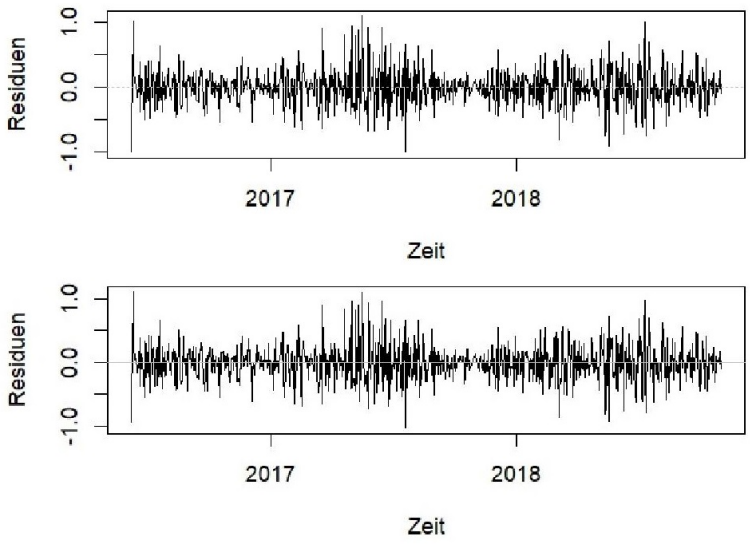

Abbildung 8: Residuen der Modelle nach Ansatz 2 für 20 kandidaten (oben) und 40 Kandidaten (unten), aufgetragen gegen die Zeit

Letzterem lässt sich auch anhand von Abbildung 8 nicht widersprechen. Die Annahme gleicher Varianzen erscheint hingegen auch hier nicht ganz erfüllt.

Der Lilliefors-Test lehnt in beiden Fällen eine Normalverteilungsannahme ab. Die beiden adj R^2 sind auch hier wieder sehr hoch.

Die Hoffnung dieser beiden Modellierungen war, Modelle zu finden, welche wenigstens vergleichbar gut sind wie die beiden aus Abschnitt 4.3.1.2 und welche aber geringere Anzahlen an Variablen besitzen. Das Modell, bei dem 20 Kandidaten betrachtet wurden, erhielt einen weniger als halb so großen Wert für α als das Modell, welches 40 Kandidaten je Parameter betrachtet hat. Bei beiden Modellen liegen die Werte für λ aber scheinbar recht nahe beieinander. Das Modell mit 40 betrachteten Kandidaten kann als geringfügig besser beurteilt werden, da es sowohl weniger Variablen besitzt als auch ein höheres $\tilde{R}^2$ aufweist. Die Anzahl Variablen ist aber in beiden Fällen unzufriedenstellend hoch geblieben.

4.3.2.2 Eliminieren von Variablen durch kontinuierliches Erhöhen von λ

Bei gemeinsamer Betrachtung von Tabelle 6 und Tabelle 7 wird das Modell mit 40 Kandidaten bevorzugt, da es sowohl den höchsten Wert bei $\tilde{R}^2$ besitzt, als auch die niedrigste Anzahl Variablen unter allen Modellen, welche Ansatz 2 umsetzen. Es ist davon auszugehen, dass sowohl ein Erhöhen des Parameters α als auch des Parameters λ zu einer geringeren Variablenanzahl führt. Auch zu befürchten ist natürlich, dass ein

Tabelle 8: $\alpha, \lambda, \tilde{R}^2$, die Variablenanzahl und die Testergebnisse des durch den in 4.3.2.2 beschriebenen Algorithmus gewonnenen Modells

α	λ	$\tilde{R}^2$	Variablenanzahl	Runs-Test	Lilliefors-Test
0,6538	0,0964	0,9698	20	0,1327	0,2111

verstärktes Eliminieren von Variablen zu einer Verringerung des $\tilde{R}^2$ führt. Sonst hätte das bisher angewandte Vorgehen zur Wahl von α und λ bereits von sich aus zu einem Modell mit weniger Variablen geführt.

Im folgenden wird abschließend versucht, die Anzahl Variablen durch kontinuierlichen Erhöhens von λ zu reduzieren. Auf ein erhöhen von α wird verzichtet, da α ohnehin die obere Schranke 1 besitzt und da es das elastische Netz immer „Lasso-lastiger" machen würde. Es ist sich jedoch von Anfang an bewusst für ein elastisches Netz und nicht für eine Lasso-Regression entschieden worden, da einige Variablen sehr korreliert sind. Es wird nun gemäß folgenden Algorithmus nach einem Modell gesucht:

1. Es werden die Parameter α und λ aus dem Modell mit 40 betrachteten Kandidaten übernommen und mit α_0 und λ_0 bezeichnet.
2. λ_0 wird um 10^{-5} erhöht.
3. das zugehörige Modell via elastischem Netz wird angepasst.
4. das adjustierte R^2 und die zugehörige Anzahl Variablen werden berechnet.
5. Wenn das adjustierte R^2 größer als 0,8 UND die Variablenanzahl größer als 20 ist, werden die Schritte 2-5 wiederholt. Sonst wird abgebrochen, und das zuletzt bestimmte Modell wird betrachtet.

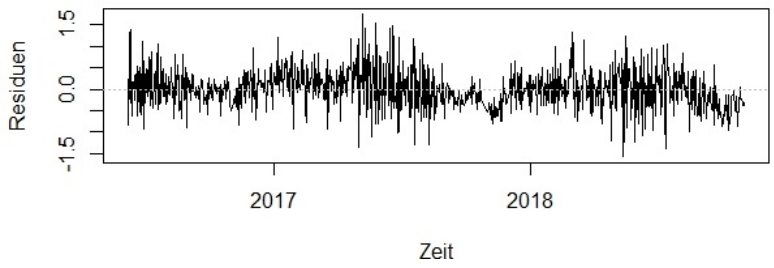

Abbildung 9: Verlauf der Residuen zu dem Modell aus 4.3.2.2 aufgetragen gegen die Zeit

Der Programmiercode zu diesem Vorgehen befindet sich im Anhang. Abbildung 9 zeigt erneut die Residuen aufgetragen gegen die Zeit. Tabelle 8 beinhaltet wieder die bei den anderen Modellen ebenfalls betrachteten Größen.

Das arithmetische Mittel der Residuen ist erneut approximativ null. Zusammen mit der Abbildung 9 und dem hinreichend großen P-Wert des Runs-Tests, lässt sich die Annahme, die Residuen hätten Erwartungswert 0, halten. Ebenso ist die Unabhängigkeitsannahme haltbar. Der P-Wert des Runs-Tests ist zwar deutlich niedriger als bei den anderen Modellen, aber dennoch deutlich größer als $0,05$.

In der Grafik zeichnet sich nach wie vor Heteroskedaszität ab, d.h. die Annahme der Gleichheit der Residuenvarianzen ist verletzt.

Anders als alle vorangegangenen Modelle lehnt der Lilliefors-Test die Annahme einer Normalverteilung diesmal nicht ab. Es kann argumentiert werden, dass die Residuen dieses Modells wenigstens einer ähnlichen Verteilung entstammen, da bei einem Stichprobenumfang von 865 eigentlich zu hoffen ist, dass der Lilliefors-Test eine Abweichung von der Normalverteilung sehr schnell aufzudecken fähig ist. Dies eröffnet die Perspektive, eventuell weitere Verfahren in Betracht zu ziehen, um dieses Modell zu analysieren, beipielsweise Varianzanalysen. Diese werden jedoch im Rahmen dieser Arbeit nicht mehr angewandt.

Neben dem Testergebnis des Lilliefors-Test ist ein weiterer zentraler Vorteil dieses Modells gegenüber allen vorangegangenen Modellen, dass das adj. R^2 sehr hoch ist, aber gleichzeitig die Anzahl Variablen erfolgreich auf 20 reduziert wurde. Dieses Modell besitzt also nicht nur sehr gutes Erklärungspotential, sondern ist im Vergleich zu allen anderen erheblich „schlanker" und somit besser interpretierbar. Daher wird dieses Modell als das erfolgreichste bzw. sinnvollste aller in dieser Arbeit betrachteten Modelle aufgefasst. Tabelle 9 zeigt eine Auflistung aller Variablen und deren Effekte, welche in diesem Modell erhalten geblieben sind.

Hierbei fallen verschiedene Sachen auf. Zunächst ist zu erwähnen, dass nur 5 der 20 Variablen keine Interaktionen sind und nur eine dieser 5 ist eine der Temperaturvariablen. Hingegen sind alle Interaktionen mit der Rissbreite des vorherigen Tages verbunden.

Auch sind einige Variablen eliminiert worden, die eigentlich sehr plausibel gewesen wären. Das betrifft unter anderem die Rissbreite des Vortages, welche außerhalb der Interaktionen nicht auftritt, aber auch die beiden nichttransformierten Temperaturmessungen und die verstrichene Zeit seit Beginn des Brückenmonitorings.

Keine der 127 Dummys der Wochen hatte einen Effekt. Von den zeitlichen Variablen sind lediglich die Dummys dreier Monate erhalten geblieben. Diese sind September, Oktober und November, also drei aufeinanderfolgende Monate. An der Stelle ist auch

Tabelle 9: Die Variablen und ihre Effekte, welche in dem Modell aus 4.3.2.2 nicht eliminiert worden sind, vgl. Tabelle 1

Variablenname	**Effektstärke ($\hat{\beta}_i$)**
Sprungvariable T_B bei $4°C$ (Anomalie des Wassers)	$4,4384 \cdot 10^{-4}$
4. Potenz der mittleren T_S, letzte $24h$	$1,5063 \cdot 10^{-11}$
November (Dummy)	$-1,1869 \cdot 10^{-3}$
Oktober (Dummy)	$-3,4387 \cdot 10^{-3}$
September (Dummy)	$-2,5732 \cdot 10^{-3}$
Interaktion: $R_{vorher}\cdot$ mittlere $\lvert T_B - T_S\rvert$, letzte $24h$	$1,3399 \cdot 10^{-2}$
alle 7 Interaktionen: $R_{vorher}\cdot$ mittlere T_S	$[1,63417 \cdot 10^{-4}; 5,1641 \cdot 10^{-4}]$
5 Interaktionen: $R_{vorher}\cdot$ mit. T_B, nicht für $i = 5, 7$	$[5,4959 \cdot 10^{-6}; 3,2847 \cdot 10^{-4}]$
Interaktion: $R_{vorher}\cdot$ Sprung T_B bei $0°C$ (Eis)	$1,0191 \cdot 10^{-2}$
Interaktion $R_{vorher}\cdot$ Sprung T_B bei $4°C$ (Anomalie)	$1,1928 \cdot 10^{-2}$

zu erwähnen, dass der Abriss des südlichen Überbaus keinen Effekt in diesem Modell hatte. Da der Abriss aber ein plausibler Effekt wäre und er sich außerdem im Oktober ereignet hat, kann vermutet werden, dass die Dummys zu Oktober und eventuell auch zu November den Effekt des Abrisses „absorbiert" haben.

Es gab vier Variablen, welche binär kodiert haben, ob die Temperatur der Unterseite der Brücke T_B sich oberhalb oder unterhalb bestimmter Schwellwerte befand. Diese waren $0, 4, 10$ und $20°C$. Die letzten beiden hierbei sind nur zum Vergleich mit den ersten beiden erhoben worden, wohingegen die ersten beiden dem Gefrierpunkt und der Anomalie des Wassers entsprechen. Die beiden Vergleichssprünge bei 10 und $20°C$ sind nicht in das Modell aufgenommen worden, wohingegen der Gefrierpunkt und die Anomalie Berücksichtigung fanden, letztere sogar als Hauptvariable und Interaktion gleichzeitig.

Es ist anzumerken, dass sowohl die Rissbreite des Vortages als auch die Temperaturvariablen (da in Kelvin) positiv sind. Somit sind es auch die zahlreichen Interaktionsterme mit Beteiligung von Temperaturvariablen. Alle zu diesen Variablen gehörigen Effekte sind ebenfalls positiv. Somit legt das Modell nahe, dass die Temperatur einen erweiternden Einfluss auf die Rissbreite besitzt, was bereits aufgrund der Abbildungen 1 und 2 vermutet wurde. Auch sind nur Temperaturvariablen in das Modell eingezogen, welche sich auf die vorangegangenen Tage beziehen. Es könnte also einen zeitversetzten Einfluss der Temperaturen auf die Rissbreite geben.

4.4 Abschließende Diskussion und Ausblick

Das Modell, welches in 4.3.2.2 schließlich gefunden wurde, wird als das geeignetste unter allen in dieser Arbeit betrachteten Modellen bewertet. Die Vorzüge sind, dass weder der

Runs- noch der Lilliefors-Test ihre jeweiligen Nullhypothesen zu den Residuen ablehnen und auch die grafische Betrachtung der Residuen keinen Anlass gegeben hat, eine der Voraussetzungen bis auf die Homogenität der Varianzen als verletzt zu betrachten, ebenso wie die geringe Variablenanzahl und den dennoch hohen Wert für das adjustierte Bestimmtheitsmaß $\tilde{R}^2$.

In dieser Arbeit wurde häufig aus den P-Werten gefolgert, dass die Voraussetzung der Unabhängigkeit der Residuen und auch die Annahme der Normalverteilung im Falle des letzten Modells haltbar seien. Hierbei ist zu betonen, dass die Tests die Erfülltheit der ihnen zugrundeliegenden Nullhypothesen nicht nachweisen können. Hohe P-Werte sind lediglich so zu interpretieren, dass die Daten nicht deutlich gegen die Nullhypothese sprechen. Da sie nicht hat abgelehnt werden können, erscheint diese zunächst nicht unplausibel. Ein signifikanter Nachweis für die Unabhängigkeit der Residuen ist jedoch an keiner Stelle erbracht worden. Selbiges gilt für das Ergebnis des Lilliefors-Tests in 4.3.2.2.

Da außerdem eine der Modellannahmen nicht erfüllt war (die Homogenität der Varianzen) sollte das Modell aus 4.3.2.2 nach wie vor mit Vorsicht interpretiert werden, auch wenn es in vielerlei Hinsicht geeignet scheint.

Auch bezieht sich das Modell ausschließlich auf die Rissbreiten des Wegaufnehmers WON2. Es gibt keinen Anlass, von einer Übertragbarkeit der hier erhaltenen Ergebnisse auf die Rissbreiten der anderen Wegaufnehmer auszugehen. Vielmehr deuten die P-Werte des Runs-Test aus Tabelle 4 an, dass sich die Risse der anderen Wegaufnehmer grundsätzlich anders verhalten und ein Modell zu diesen separat gesucht werden müsste. Nichts desto trotz ist interessant, auszuprobieren, welches Ergebnis sich erhalten lässt, wenn das Vorgehen aus 4.3.2.2 einfach auf einen anderen Wegaufnehmer angewandt wird. Es wäre vielleicht auch einen Versuch wert, zu einem der anderen Risse ein Modell anzupassen, welches von vorneherein nur die Variablen aus Tabelle 9 aufnimmt.

Festzuhalten ist, dass das Reduzieren des Datensatzes auf die Daten einer bestimmten Uhrzeit (4 Uhr morgens) der entscheidende Schritt war, um die der Modellierung zugrundeliegende Annahme der Unabhängigkeit der Residuen nicht schwer zu verletzen. Wie sich das Vorgehen aus 4.3.2.2 bei Erweiterung auf alle Daten verhält, könnte Gegenstand weiterer Forschung sein.

Abschließend kann nochmal auf Abschnitt 4.3.1.3 verwiesen werden. Dort wurde bereits erwähnt, dass auch das Ausprobieren anderer Streufunktionen s zu einem besseren Ergebnis führen könnte und insbesondere das Potential birgt, die Forderung nach homogenen Varianzen der Residuen einzuhalten.

Anhang

A Weitere Grafiken

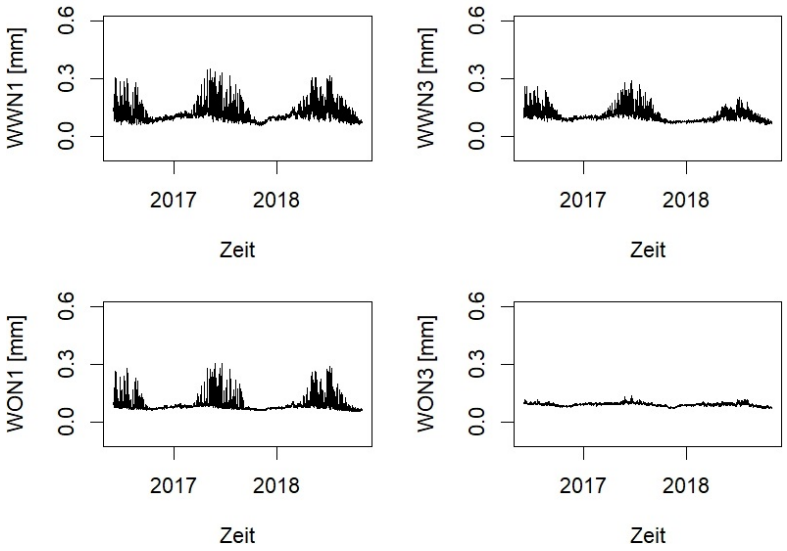

Abbildung 10: In Analogie zu Abbildung 2, die Rissverläufe der übrigen 4 nördlichen Wegaufnehmer über die Zeit

B Wichtige Ausschnitte des Programmcodes

In diesem Abschnitt des Anhangs werden nur die Passagen des benutzten R-Codes vorgestellt, welche die wichtigsten sind um die zentralen Ergebnisse dieser Arbeit zu reproduzieren.

Zunächst sind die Daten und die benötigten Pakete zu laden:

```
load("Rissdaten.RData")
library("glmnet")
library("caret")
library("nortest")
library("tseries")
```

Im folgenden wurden die Variablen zur Modellierung erstellt. Hierzu wurde öfters eine Funktion benutzt, welche dazu gedacht war, entweder die Funktion `var` oder die Funktion `mean` auf die letzten $k \cdot 24$ Einträge eines gegebenen Datenvektors zu einem gegebenem Index i zu berechnen. Es entspricht `CALC` der Funktion, welche angewandt werden soll, also entweder `mean` oder `var`. In `Index` soll der Index i eingespeist werden

und in `k` selbsterklärendeerweise k. Im Anschluss wird mithilfe von `sapply` und der hier definierten Funktion exemplarisch die durchschnittliche Temperatur T_B für die letzten k Tage, $k \in \{1, \ldots, 7\}$, berechnet

```
mTk <- function(Index, k, Temp, CALC = mean){

Zraum <- k * 24
All_Zeit <- WON_WWN$Zeit
if(Index < Abriss_indizes[1])
result <- ifelse(Index > Zraum,
                                  CALC(Temp[(Index - Zraum):(Index - 1)],
                                  na.rm = TRUE), NA)
else{

Index_kor <- Abriss_indizes[1] - 1
result <- ifelse(Index - Index_kor > Zraum,
                                                  CALC(Temp[(Index - Zraum):(Index - 1)],
                                                  na.rm = TRUE), NA)
}

return(result)
}


# Bezeichnung: *mTbi* steht fuer *mittlere Temp. Brucke der letzten i*24 Stunden*
mTb1 <- sapply(1:N_tot, mTk, Temp = Tb, k = 1); ATB_1 <- mTb1
mTb2 <- sapply(1:N_tot, mTk, Temp = Tb, k = 2); ATB_2 <- mTb2
mTb3 <- sapply(1:N_tot, mTk, Temp = Tb, k = 3); ATB_3 <- mTb3
mTb4 <- sapply(1:N_tot, mTk, Temp = Tb, k = 4); ATB_4 <- mTb4
mTb5 <- sapply(1:N_tot, mTk, Temp = Tb, k = 5); ATB_5 <- mTb5
mTb6 <- sapply(1:N_tot, mTk, Temp = Tb, k = 6); ATB_6 <- mTb6
mTb7 <- sapply(1:N_tot, mTk, Temp = Tb, k = 7); ATB_7 <- mTb7
```

Mit dem folgenden Code wird das in 4.1 beschriebene Vorgehen durchgeführt. In der Variable `Weg` ist der Riss des Wegaufnehmers gespeichert, der gerade modelliert werden soll. Dieser Code wurde mehrfach benutzt. Die Modelle zu den unterschiedlichen Wegaufnehmern wurden erhalten, indem in der Variable `Riss` vor jeder Modellierung die Daten des entsprechenden Wegaufnehmers eingespeist wurden.

```
formula_4 <- Weg4_ ~ Tb4_ + Ts4_ + dZeit4_ + Wegv4_ +
        Eis4_ + H2O_an4_ + kalt_T10_4_ + kalt_T20_4_ +
        mTs1pot24_ + mTs1pot34_ + mTs1pot44_ +
        mTs1_exp4_ + mTb1_exp4_ + mTs1pot.54_ + mTb1pot.54_ +
        mTb1pot24_ + mTb1pot34_ + mTb1pot44_ +
        Tspot24_ + Tspot34_ + Tspot44_ + Ts_exp4_ + Tspot.54_ +
        Tbpot24_ + Tbpot34_ + Tbpot44_ + Tb_exp4_ + Tbpot.54_ +
        TDiff_14_ + TDiff_24_ + TDiff_34_ + TDiff_44_ +
```

```
        TDiff_54_ + TDiff_64_ + TDiff_74_ +
        mTb14_ + mTb24_ + mTb34_ + mTb44_ + mTb54_ + mTb64_ + mTb74_ +
        mTs14_ + mTs24_ + mTs34_ + mTs44_ + mTs54_ + mTs64_ + mTs74_ +
        Monat4_ + abgerissen4_ + S_Wo4_ + weekend4_ + wochentag4_ +
        I(Wegv4_ * TDiff_14_) + I(Wegv4_ * TDiff_24_) + I(Wegv4_ * TDiff_34_) +
        I(Wegv4_ * TDiff_44_) + I(Wegv4_ * TDiff_54_) + I(Wegv4_ * TDiff_64_) +
        I(Wegv4_ * TDiff_74_) +
        I(abgerissen4_ * dZeit4_) + I(Wegv4_ * dZeit4_) +
        I(Wegv4_ * mTs14_) + I(Wegv4_ * mTs24_) + I(Wegv4_ * mTs34_) +
        I(Wegv4_ * mTs44_) + I(Wegv4_ * mTs54_) + I(Wegv4_ * mTs64_) +
        I(Wegv4_ * mTs74_) +
        I(Wegv4_ * mTb14_) + I(Wegv4_ * mTb24_) + I(Wegv4_ * mTb34_) +
        I(Wegv4_ * mTb44_) + I(Wegv4_ * mTb54_) + I(Wegv4_ * mTb64_) +
        I(Wegv4_ * mTb74_) +
        I(Eis4_ * abgerissen4_) + I(H2O_an4_ * abgerissen4_) +
        I(kalt_T10_4_ * abgerissen4_) + I(kalt_T20_4_ * abgerissen4_) +
        I(Eis4_ * Wegv4_) + I(H2O_an4_ * Wegv4_) + I(kalt_T10_4_ * Wegv4_) +
        I(kalt_T20_4_ * Wegv4_) +
        I(Eis4_ * Tb4_) + I(H2O_an4_ * Tb4_) + I(kalt_T10_4_ * Tb4_) +
        I(kalt_T20_4_ * Tb4_)

Volle_M4 <- data.frame(Weg4_, Tb4_, Ts4_, dZeit4_, Wegv4_ ,
        Eis4_, H2O_an4_, kalt_T10_4_, kalt_T20_4_,
        mTs1pot24_ , mTs1pot34_ , mTs1pot44_ ,
        mTs1_exp4_ , mTb1_exp4_ , mTs1pot.54_, mTb1pot.54_,
        mTb1pot24_ , mTb1pot34_ , mTb1pot44_ ,
        Tspot24_ , Tspot34_ , Tspot44_ , Ts_exp4_ , Tspot.54_,
        Tbpot24_ , Tbpot34_ , Tbpot44_ , Tb_exp4_ , Tbpot.54_,
        TDiff_14_ , TDiff_24_ , TDiff_34_ , TDiff_44_ ,
        TDiff_54_ , TDiff_64_ , TDiff_74_ ,
        mTb14_ , mTb24_ , mTb34_ , mTb44_ , mTb54_ , mTb64_ , mTb74_ ,
        mTs14_ , mTs24_ , mTs34_ , mTs44_ , mTs54_ , mTs64_ , mTs74_ ,
        Monat4_ , abgerissen4_ , S_Wo4_ , weekend4_, wochentag4_)

NA_Zeilen4 <- which(apply(MAR = 1, Volle_M4, FUN = function(x){any(is.na(x))}))

MM_4 <- model.matrix(formula_4)

set.seed(469)
H4_5_train <- train(form = formula_4, data = Volle_M4[-NA_Zeilen4, ],
                                        tuneLength = 10,
                                        trControl = trainControl(method = "cv", 5),
                                        method = "glmnet")

fit_H4_5 <- glmnet(x = MM_4, y = Weg4_[-NA_Zeilen4],
                                    alpha = H4_5_train$bestTune$alpha,
                                    lambda = H4_5_train$bestTune$lambda)
```

Die folgenden beiden Zeilen führen beispielhaft Runs-Tests und Lilliefors-Tests durch. In der Variable `res` müssen hierfür die Residuen eines Modells stehen.

```
runs.test(as.factor(res < 0))
```

```
lillie.test(res)
```

Folgender Aufruf trägt die Residuen gegen die Zeit auf:

```
plot(res_4 ~ Zeit4_[-NA_Zeilen4], xlab = "Zeit", ylab = "Residuen␣WON2", type = "l")
abline(lty = 3, col = "grey", 0, 0)
```

In diesen beiden Zeilen werden die Funktionen s_1 aus Ansatz 1 und s_2 aus Ansatz 2 angepasst (vgl. 4.3.1.2).

```
# Ansatz 1
s_Tb_naiv_lm <- lm(res_sd_4 ~ Tb_mean_H4_5_geordnet)
# Ansatz 2
s_Tb_exp_lm <- lm(log(res_sd_4) ~ Tb_mean_H4_5_geordnet)
```

Mit diesem Code wurde schließlich das letzte Modell angepasst (vgl. 4.3.2.2). In `H4_5_train_exp_k40` ist das Modell mit 40 Kandidaten für λ und α gespeichert (output der Funktion `train`). In `fit_red_var` ist schließlich das Resultat.

```
set.seed(99531)
a_0 <- H4_5_train_exp_k40$bestTune$alpha
l_0 <- H4_5_train_exp_k40$bestTune$lambda

calc_adj_R_sq <- function(model){

        residuals <- Weg4_exp_ - predict(model, newx = MM_4_exp)
        R_sq <- 1 - (var(residuals) / var(Weg4_exp_))
        p <- sum(model$beta != 0)
        n <- length(Weg4_exp_)
        punish <- (n - 1) / (n - p - 1)
        return(1 - ((1 - R_sq)) * punish)
}

repeat{

        l_0 <- l_0 + 1e-5
        fit_red_var <- glmnet(x = MM_4_exp, y = Weg4_exp_,
        alpha = a_0, lambda = l_0)
        adj_R_sq <- calc_adj_R_sq(fit_red_var)
        number_var_not_0 <- sum(fit_red_var$beta != 0)

        if(adj_R_sq <= 0.8)
        break()
        if(number_var_not_0 <= 20)
        break()
}
```

Literaturverzeichnis

Abbas, S., Fried, R., Heinrich, J., Horn, M., Jakubzik, M., Kohlenbach, J., Maurer, R., Michels, A. und Müller, C.H. (2019). "Detection of anomalous sequences in crack data of a bridge monitoring". In: *Applications in Statistical Computing - From Music Data Analysis to Industrial Quality Improvement*. Hrsg. von K. Ickstadt, H. Trautmann, G. Szepannek, N. Bauer, K. Lübke und M. Vichi, S. 251 –269.

Dallal, G. E. und Wilkinson, L. (1986). "An Analytic Approximation to the Distribution of Lilliefors's Test Statistic for Normality". In: *The American Statistician* 40, S. 294 –296.

Friedman, J., Hastie, T. und Tibshirani, R. (2010). "Regularization Paths for Generalized Linear Models via Coordinate Descent". In: *Journal of Statistical Software* 33.1, S. 1–22.

Gross, J. und Ligges, U. (2015). *nortest: Tests for Normality*. R package version 1.0-4. URL: `https://CRAN.R-project.org/package=nortest`.

Hedderich, J. und Sachs, L. (2018). *Angewandte Statistik. Methodensammlung mit R*. 16. Auflage. SpringerSpektrum.

Heinrich, J. (2016). "Ergebnisse aus dem Rissmonitoring. König und Heunisch Planungsgesellschaft. (unveröffentlicht)".

König und Heunisch Planungsgesellschaft (2016). "Bochum BW 179 UF d. Wittener Straße Brücke über den Sheffield Ring (L 705). (unveröffentlicht)".

Kuhn, M. (2019). *caret: Classification and Regression Training*. R package version 6.0-84. URL: `https://CRAN.R-project.org/package=caret`.

Likas, A., Blekas, K. und Kalles, D. (2014). *Artificial Intelligence: Methods and Applications*. Springer.

Lilliefors, H. W. (1967). "On the Kolmogorov-Smirnov Test for Normality With Mean and Variance Unknown". In: *Journal of the American Statistical Association* 62, S. 399 –402.

Pflaumer, P., Heiner, B. und Hartung, J. (2001). *Statistik für Wirtschafts- und Sozialwissenschaften: Induktive Statistik. Lehr- und Übungsbuch*. Oldenbourg Wissenschaftsverlag GmbH.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: `https://www.R-project.org/`.

Sen, A. und Srivatava, M. (1990). *Regression Analysis. Theory, Methods, and Applications*. Springer-Verlag.

Siegel, S. und Castellan, N. (1988). *Nonparametric statistics for the behavioral sciences*. 2. ed. McGraw-Hill.

tagesschau.de (8 April 2020). *Brücke in Italien eingestürtzt. Zugegriffen am 27.04.2020*. URL: https://www.tagesschau.de/ausland/italien-brueckeneinsturz-101.html.

Thunich, O. (2017). *Eliminierung des Temperatureffekts bei Brückenmonitoringdaten. Vergleich zwischen linearer Regression und Kriging-Modellen*. Bachlorarbeit. Technische Universität Dortmund.

Tibshirani, R. (1996). "Regression Shrinkage and Selection Via the Lasso". In: *Royal Statistical Society*. DOI: https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

Toutenburg, H. (2003). *Lineare Modelle. Theorie und Anwendungen*. zweite Auflage. Physica Verlag.

Trapletti, A. und Hornik, K. (2019). *tseries: Time Series Analysis and Computational Finance*. R package version 0.10-47. URL: https://CRAN.R-project.org/package=tseries.

Zou, H. und Hastie, T. (2005). "Regularization and variable selection via the elastic net". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.2, S. 301–320.