Masterarbeit

Erkennung von Anomalien beim Brückenmonitoring und Verkehrsschätzung

Verfasser: Anne Michels

Betreuer: Prof. Dr. Christine Müller

TU Dortmund

Datum: 28. Mai 2019

Inhaltsverzeichnis

1 Einleitung **1**

2 Problemstellung **3**

2.1 Beschreibung des Datenmaterials 3

2.2 Zielsetzung 5

3 Statistische Methoden **6**

3.1 Gleitender Median 7

3.2 Changepointanalyse 7

3.3 K-Means 8

3.4 K-fache Kreuzvalidierung 8

3.5 Logistische Regression 9

3.6 Integrationsbasiertes Variationsmaß 9

4 Das Changepoint Paket **11**

4.1 Optimal Partitioning 12

4.2 PELT 13

4.3 Die Kostenfunktion 15

4.4 MBIC 16

4.5 Verringerung der Laufzeit durch den Pruningschritt in PELT 18

4.6 Unterschiede zwischen dem changepoint Paket und Killick et al. (2012) [11] 18

4.6.1 Unterschiede in der Ausgabe der Changepoints 19

4.6.2 Minimale Segmentlänge 20

4.6.3 Unterschiede der Kostenfunktion 21

5 Methoden zur Erkennung von Anomalien **21**

5.1 Variance Changepoint Detection 21

5.2 Modified Variance Changepoint Detection 23

5.2.1 Gründe für die Unterschiede bei der Anwendung der VCP Methode auf die umgekehrte Zeitreihe 25

5.3 Clustering of MAD filtered Data 29

6 Vergleich der Methoden zur Erkennung von Anomalien **31**

6.1 Vergleich der Methoden bezüglich Einheitlichkeit der Ergebnisse 32

6.2 Vergleich der Methoden anhand von manuell bestimmten Anomalien in ausgewählten Wochen 38

6.2.1 Fehlerraten der einzelnen Tage an denen Anomalien auftreten . . 40

6.2.2 Fehlerraten pro Woche 45

6.3 Klassifikation bezüglich des Vorliegens einer Anomalie an einem Tag . . 47

6.3.1 Verbesserung der Zuordnung bezüglich des Vorliegens einer Anomalie an einem Tag mit Hilfe der logistischen Regression 49

6.4 Klassifikation bezüglich der Anzahl Anomalien an einem Tag 56

7 Verkehrsschätzung **60**

7.1 Vergleich der Schätzungen in Abhängigkeit der identifizierten Anomalien 64

7.2 Ansätze zur Verbesserung der Schätzung 67

7.3 Verkehrsschätzung mittels linearer Regression 70

7.4 Probleme bei der Verkehrsschätzung 72

7.4.1 Probleme durch den Einfluss der Temperatur auf die Rissbreiten . 72

7.4.2 Probleme durch das Fehlen von Verkehrszahlen zu den meisten Tageszeiten 74

8 Zusammenfassung **75**

Anhang **81**

A Tabellen **81**

B R output **88**

Literatur **89**

1 Einleitung

Ein Großteil der 39 231 Brücken in Deutschland (Stand 31.12.12) ist in den Jahren von 1965 bis 1985 erbaut worden. Zu dieser Zeit waren andere Vorschriften für die Errichtung von Brücken gegeben, als heutzutage. Daraus resultiert, dass ein hoher Anteil der Brücken konstruktive Schwachstellen aufweist, wodurch die Stabilität der Brücke unter den aktuellen Bedingungen nicht sicher ist. Problematisch ist z.B. die Erhöhung des maximalen Gesamtgewichts für LKW von 24 t im Jahr 1956 auf 44 t. Diese Gewichtsbegrenzung wird auch immer häufiger voll ausgeschöpft oder sogar überschritten. Zudem ist die Anzahl der Schwertransporte enorm gestiegen. Um das unkontrollierte Zusammenbrechen einer Brücke verhindern zu können, müssen regelmäßige Bauwerksprüfungen vorgenommen werden (siehe Bundesministerium für Verkehr, Bau und Stadtentwicklung, 2013 [3]). Zu diesem Zweck existiert eine Nachrechnungsrichtlinie für Straßenbrücken, mit deren Hilfe Tragfähigkeit und Stabilität unter dem Einfluss des höheren Verkehrsaufkommens beurteilt werden soll. Werden nach dieser Richtlinie Defizite an einer Straßenbrücke festgestellt, stehen verschiedene Kompensationsmaßnahmen, zur Erhaltung der Sicherheit der entsprechenden Brücke, zur Verfügung. Dazu zählen sowohl Verkehrsmaßnahmen, wie eine Gewichtsbeschränkung für LKW, Geschwindigkeitsbegrenzungen und die Sperrung von Fahrspuren, als auch Überwachungsmaßnahmen wie ein permanentes Rissmonitoring (siehe Bundesministerium für Verkehr, Bau und Stadtentwicklung, 2011 [2]).

In dieser Arbeit wird die Brücke, welche die Wittener Straße über den Sheffieldring in Bochum führt, betrachtet. An dieser im Jahr 1961 erbauten Spannbetonbrücke ist bei der Nachrechnung gemäß der oben erwähnten Nachrechnungsrichtlinie eine Beeinträchtigung der Tragfähigkeit festgestellt worden. Dieses Defizit macht eine Erneuerung der Brücke notwendig. Bis zum Abriss der Brücke werden Kompensationsmaßnahmen zum Erhalt der Sicherheit durchgeführt. Dazu gehören ein Rissmonitoring, die Einschränkung des Verkehrs durch eine Reduzierung der zweispurigen Fahrbahnen für jede Richtung auf jeweils eine Spur und eine Beschränkung des maximalen Gesamtgewicht pro Fahrzeug auf 24 t (siehe Heinrich, 2016 [8]).

Ziel dieser Arbeit ist eine Verkehrsschätzung anhand der Daten des Rissmonitoring. Dazu muss zunächst das Teilziel, welches in der Entfernung der Anomalien aus den Rissda-

ten besteht, erfüllt werden. Als Anomalien werden starke Schwankungen der Rissbreiten, die nicht auf den Verkehr zurückgeführt werden können, bezeichnet. Zur Identifikation der Anomalien, wurden innerhalb der Veranstaltung *Fallstudien II* bereits zwei Methoden entwickelt. Die VCP (Variance Changepoint Detection) Methode und die CMAD (Clustering of MAD filtered Data) Methode. Ein weiteres Ziel dieser Arbeit besteht darin die VCP Methode weiterzuentwickeln bzw. zu verbessern. Zudem sollen die verschieden Methoden zur Identifikation von Anomalien miteinander verglichen werden und es soll herausgefunden werden, unter welchen Umständen die Methoden gut oder weniger gut funktionieren. Die VCP Methode beruht auf der `cpt.var()` Funktion aus dem Paket `changepoint` (siehe Killick et al. 2016 [12]), welche Änderungen der Varianz auf einer Zeitreihe finden soll. Da die Dokumentation dieses Paketes sehr schwammig gehalten ist, soll die Funktionsweise der `cpt.var()` Funktion genauer erläutert werden.

Zum erreichen, der für diese Arbeit gesetzten Ziele, werden die Methoden zur Identifikation der Anomalien, bezüglich der Einheitlichkeit der als Anomalie klassifizierten Bereiche und der Übereinstimmung mit manuell identifizierten Anomalien, verglichen. Zudem wird überprüft ob die Methoden richtig erkennen, ob an einem Tag eine Anomalie vorliegt oder nicht und ob die richtige Anzahl an Anomalien identifiziert wird. Um die Methoden anhand dieser Punkte vergleichen zu können, werden entsprechende Fehlklassifikationsraten bestimmt.
Ein Ansatz zur Verkehrsschätzung aus der Veranstaltung *Fallstudien II* beruht auf einem integrationsbasierten Variationsmaß. Es wird versucht die daraus resultierende Schätzung durch die Berücksichtigung der Temperatur zu verbessern. Zudem soll die Fahrzeuganzahl pro Stunde mit Hilfe einer linearen Regression vorhergesagt werden. Außerdem wird auf eventuelle Probleme bei der Verkehrsschätzung auf Grundlage der zur Verfügung stehenden Daten eingegangen.

In Kapitel 2 werden die im Rahmen des Rissmonitoring erhobenen Daten und die genauen Ziele der Arbeit detailliert beschrieben. Die zur Identifikation der Anomalien und zur Verkehrsschätzung verwendeten statistischen Methoden werden in Kapitel 3 vorgestellt. Eine Erläuterung der `cpt.var()` Funktion aus dem `changepoint` Paket von Killick et al. (2016) [12] erfolgt in Kapitel 4. Dabei wird sowohl auf die zugrundeliegenden Me-

thoden, als auch auf die Wahl der Kostenfunktion und des Strafparameters eingegangen. Außerdem wird auf Unterschiede zwischen der Definition der für das Paket verwendeten Methoden aus Killick et al. (2012) [11] und der tatsächlichen Implementierung der Funktion eingegangen. In Kapitel 5 werden die zur Identifikation der Anomalien innerhalb der Rissdaten entwickelten Methoden vorgestellt. Der Vergleich dieser Methoden miteinander und die Beurteilung der Qualität dieser Methoden wird in Kapitel 6 vorgenommen. Die Verkehrsschätzung erfolgt in Kapitel 7. Zum Schluss werden die Ergebnisse aus dem Vergleich der Methoden und der Verkehrsschätzung in Kapitel 8 zusammengefasst und beurteilt.

2 Problemstellung

Die Grundlage für diese Arbeit bilden die im Rahmen des Rissmonitoring an der 1961 erbauten Brücke, welche die Wittener Straße in Bochum über den Sheffieldring führt, erhobenen Daten. Dabei handelt es sich um eine dreifeldrige Spannbetonbrücke mit einer Gesamtlänge von 63.50 m. Die Brücke besteht aus zwei getrennten Überbauten, so dass für jede Fahrtrichtung ein eigener Überbau vorhanden ist. Auf jedem Überbau sind zwei Fahrspuren, sowie eine Spur für die Straßenbahn, vorhanden.
Da bei der Nachrechnung im Rahmen der „Nachrechnungsrichtlinie für Straßenbrücken im Bestand“, Tragfähigkeitsdefizite für diese Brücke ermittelt worden sind, darf auf jedem Überbau lediglich eine der beiden Fahrspuren genutzt werden. Zudem ist das maximale Gewicht pro Fahrzeug auf 24 t beschränkt. Diese Informationen bezüglich der Brücke sind aus Heinrich (2016) [8] entnommen.
In Abschnitt 2.1 wird das, aufgrund des Rissmonitorings, vorliegende Datenmaterial beschrieben. Die Ziele dieser Arbeit werden in Abschnitt 2.2 vorgestellt.

2.1 Beschreibung des Datenmaterials

Es liegen Rissdaten für 16 Messstellen vor. Die Variablenbezeichnung der einzelnen Messstellen beginnt dabei immer mit einem **W** für Wegaufnehmer, anschließend steht **O** oder **W** für östliches bzw. westliches Ende, gefolgt von **N** oder **S** für nördlicher bzw. südlicher Überbau. Zum Schluss folgt eine Ziffer von 1 bis 4, da an jedem Ende jeden Überbaus 4 Wegaufnehmer angebracht sind. **WOS4** steht zum Beispiel für Wegaufnehmer am öst-

lichen Ende des südlichen Überbaus 4. Die Rissbreiten an den Messpunkten lagen zu Beginn des Rissmonitorings zwischen 0.1 und 0.3 mm. Der Messbereich der Wegaufnehmer liegt am nördlichen Überbau bei ± 5 mm und am südlichen Überbau bei ± 2 mm, jeweils ausgehend von der Rissbreite bei Installation der Wegaufnehmer.

Es gibt 3 Messstellen, an denen die Temperatur in Grad Celsius gemessen wird. Unter **TBruecke** wird die an der Unterseite der Brücke gemessene Temperatur gespeichert, unter **TSonne** die an der Oberseite der Brücke gemessene Temperatur und unter **TSchalt** wird die im Schaltschrank für die Messgeräte gemessene Temperatur gespeichert.

Für jede Variable wird alle 2 Sekunden ein Messwert gespeichert. Das ergibt täglich $43\,200$ Werte pro Messstelle. An einigen Tagen liegen weniger Beobachtungen vor, was auf die Synchronisation der Messgeräte zurückzuführen ist. Die Messdaten liegen tageweise als MDT-Dateien (Microsoft Access Add-in Data) vor. Die Informationen zum Rissmonitoring sind aus Heinrich (2016) [8] entnommen. In dieser Arbeit werden Messdaten vom 01.06.2016 bis zum 20.10.2017 betrachtet.

Abbildung 1 zeigt die Rissbreiten, gemessen an Wegaufnehmer **WOS4**, sowie die Temperaturmessungen an den Messstellen **TBruecke** und **TSonne** vom 12.06.2016.

Es ist zu erkennen, dass der Verlauf der Rissbreiten in etwa dem zeitlich verzögerten Verlauf der Temperaturkurve entspricht. Zwischen 09:00 Uhr und 11:00 Uhr ist innerhalb der Rissdaten ein „Knubbel“ erkennbar. Solche „Knubbel“ werden als Anomalien bezeichnet, da sie keine erkennbare Ursache haben. In Abbildung 2 ist ein kleinerer Ausschnitt der Zeitreihe zu sehen, wodurch Anomalien besser erkennbar sind.

Hier sind auch weitere Anomalien sichtbar, beispielsweise gegen 08:00 Uhr und zwischen 12:00 Uhr und 13:00 Uhr. Die Anomalien treten nicht täglich auf und sind unabhängig von der Tageszeit.

Sowohl in Abbildung 1 als auch in Abbildung 2 sind einzelne längere Ausschläge (Peaks) zu erkennen. Diese treten sowohl außerhalb als auch innerhalb von Anomalien auf. Peaks, die nicht Teil einer Anomalie sind, können auf den Verkehr zurückgeführt werden. Sie entstehen, wenn z.B. ein schweres Fahrzeug oder sehr viele Fahrzeuge hintereinander über den Riss fahren. Somit sollten Peaks nur als Anomalie gewertet werden, wenn sie innerhalb einer solchen auftreten.

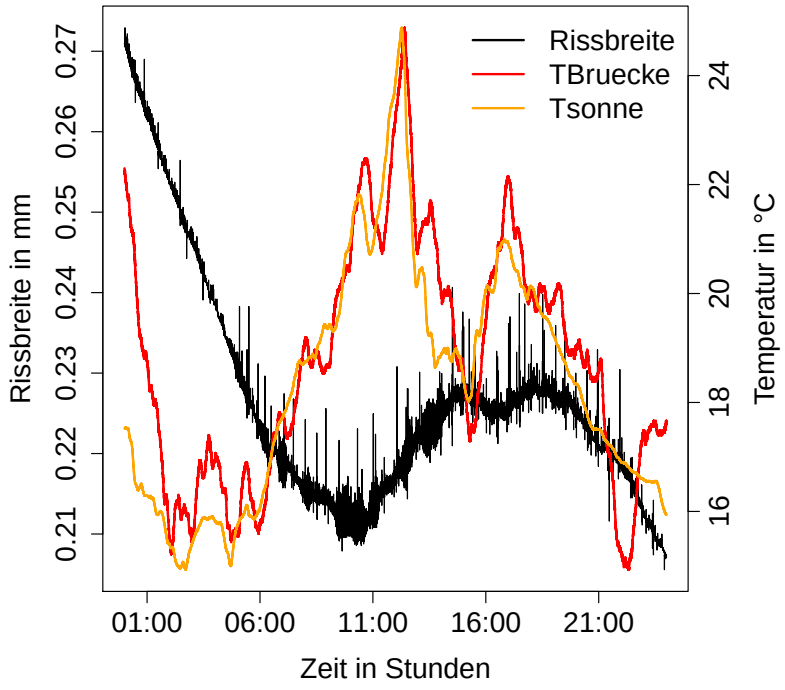


Abbildung 1: Rissbreite, gemessen an Wegaufnehmer **WOS4**, sowie die Temperaturmessungen der Messstellen **TBruecke** und **TSonne** vom 12.06.2016.

2.2 Zielsetzung

Im Rahmen der Veranstaltung *Fallstudien II* sollten Methoden zur Erkennung von Anomalien innerhalb des Rissmonitoring entwickelt werden. Dabei sind die Methoden *Clustering of MAD filtered data* (CMAD) und *Variance Change Point Detection* (VCP) entstanden. Ziel dieser Arbeit ist eine Verbesserung der VCP Methode, sowie der Vergleich der Methoden zur Erkennung von Anomalien in den Rissdaten. Anschließend soll ein Modell zur Verkehrsschätzung aufgestellt werden.

Da die, der VCP Methode zugrundeliegende, `cpt.var()` Funktion aus dem Paket `changepoint` (siehe Killick et al., 2016 [12]) nicht sehr ausführlich dokumentiert ist, besteht ein weiteres Ziel dieser Arbeit in der Aufarbeitung der Verhaltensweise der Funktion. Dabei sollen sowohl der Ablauf der zugrundeliegenden Methoden, als auch die tatsächliche Implementierung der Funktion berücksichtigt werden.

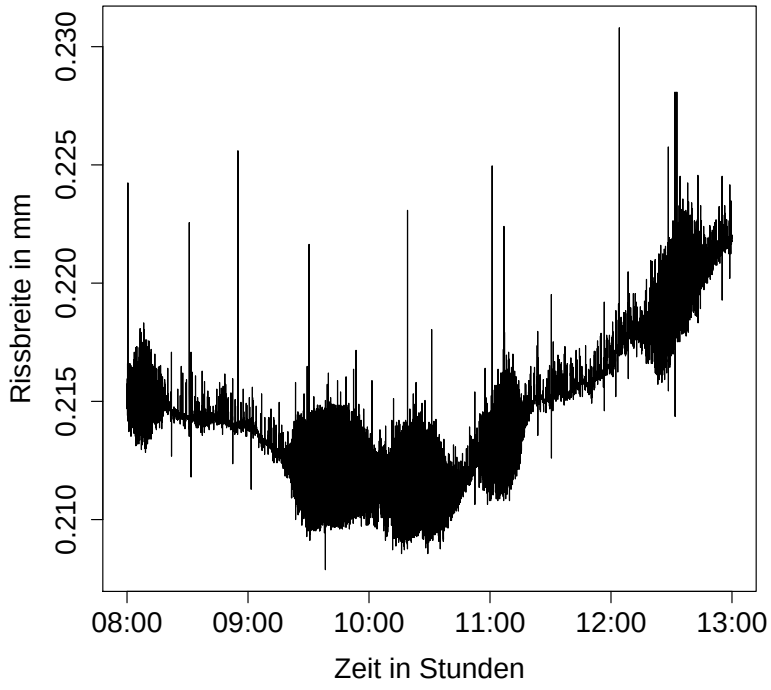


Abbildung 2: Rissbreite, gemessen an Wegaufnehmer **WOS4** am 12.06.2016 zwischen 08:00 Uhr und 13:00 Uhr.

3 Statistische Methoden

In dieser Arbeit werden verschiedene statistische Methoden verwendet. Der **Running Median** (Abschnitt 3.1) und die K**-Means** Methode (Abschnitt 3.3) werden von allen vorgestellten Methoden zur Identifikation von Anomalien benötigt. Die **Changepoint-analyse** (Abschnitt 3.2) wird von der VCP und der MVCP zur Identifikation von Anomalien verwendet. Die **logistische Regression** (Abschnitt 3.5) dient bei der MVCP und VCP Methode zur Vorentscheidung, ob an dem entsprechenden Tag eine Anomalie vorliegt und die K**-fache Kreuzvalidierung** (Abschnitt 3.4) approximiert die zugehörige wahre Fehlerrate. Das **Integrationsbasierte Variationsmaß** (Abschnitt 3.6) wird für die Verkehrsschätzung in Kapitel 7 genutzt.

Für sämtliche Berechnungen und die Erstellung von Grafiken wird die Statistiksoftware R (R Core Team, 2017 [15]) verwendet. In dieser Arbeit wird die folgende Notation verwendet: X_{t_n}, mit $n = 1, \ldots, N$ sind Zufallsvariablen einer Zeitreihe, deren zugehörige Realisierungen als x_{t_n} bezeichnet werden. Dabei bezeichnen $t_1 < t_2 < \cdots < t_{N-1} < t_N$ den Zeitpunkt der Beobachtung, wobei N der Anzahl der Beobachtungen entspricht. Bei

den für diese Arbeit verwendeten Daten, gilt immer $t_{n+1} - t_n = 2$, für $n = 1, \ldots, N-1$.

3.1 Gleitender Median

Der gleitende Median ist eine Methode zur Glättung einer Zeitreihe $X_{t_1}, \ldots, X_{t_N}$. Dabei werden die Beobachtungen $X_{t_n}, n = k+1, \ldots, N-k$ durch

$$X^*_{t_n} = \text{median}(X_{t_{n-k}}, \ldots, X_{t_{n+k}}) \tag{1}$$

ersetzt. Der Bereich $X_{t_{n-k}}, \ldots, X_{t_{n+k}}$ wird dabei als Fenster um X_{t_n} mit Fensterbreite $K = 2k+1$ bezeichnet (Härdle und Steiger [6] 1995). Für $1 < n \leq k$ bzw. $N-k+1 \leq n < N$ wird die Fensterbreite mit jedem Schritt Richtung Rand um 2 Beobachtungen verringert. Bei Beobachtung X_{t_2} wird somit eine Fensterbreite von 3 genutzt um $X^*_{t_2}$ zu bestimmen. Die Berechnung des Anfangspunktes ($X^*_{t_1}$) und des Endpunktes ($X^*_{t_N}$) erfolgt mit der folgenden Formel:

$$X^*_{t_1} = \text{median}(X_{t_1}, X^*_{t_2}, 3X^*_{t_2} - 2X^*_{t_3}) \tag{2}$$

$$X^*_{t_N} = \text{median}(X_{t_N}, X^*_{t_{N-1}}, 3X^*_{t_{N-1}} - 2X^*_{t_{N-2}}) \tag{3}$$

berechnet (Tukey 1977 [16], S. 221 - 222).

3.2 Changepointanalyse

Seien $x_{t_1}, x_{t_2}, \ldots, x_{t_N}$ Realisationen von Zufallsvariablen $X_{t_n} \sim \mathcal{N}(\mu_{t_n}, \sigma^2_{t_n})$ einer Zeitreihe, mit $n = 1, \ldots, N$. Die drei verschiedenen Arten von Changepointproblemen sind die Suche nach Änderungen des Mittelwerts, die Suche nach Änderungen der Varianz und die Suche nach Änderungen des Mittelwerts und der Varianz (siehe Chen und Gupta, 2000 [5], S.5).

In dieser Arbeit wird die Changepointanalyse zur Suche nach Änderungen der Varianz verwendet, wobei von einem gleichbleibenden, bekannten Erwartungswert μ ausgegangen wird. Dazu wird die Funktion `cpt.var()` aus dem `changepoint` Paket (Killick et al., 2016 [12]) verwendet. Das Paket wird in Kapitel 4 genauer betrachtet. Innerhalb der `cpt.var()` Funktion wird die PELT (*Pruned Exact Linear Time*) Methode zur Bestimmung der Changepoints bezüglich der Varianz verwendet. Diese Methode wird in Kapitel 4.2 genauer erläutert.

3.3 K-Means

Bei K-Means handelt es sich um eine iterative Clustermethode die gegebene Beobachtungen $x_{t_1}, \ldots, x_{t_N}$ so auf K Cluster aufteilt, dass die Varianz innerhalb der Cluster minimiert wird. Zu Beginn des Verfahrens werden die Beobachtungen zufällig in K Cluster eingeteilt, wobei alle Cluster die gleiche Anzahl Beobachtungen enthalten. Die Schreibweise $C(t_n) = k$ besagt dabei, dass Beobachtung x_{t_n} Cluster k zugeordnet ist. Zur Verbesserung der zufälligen Clusterung werden die folgenden Schritte durchgeführt:

1. Die Mittelwerte der aktuellen Clusteraufteilung $\{m_1, \ldots, m_K\}$ werden ermittelt. Dabei bezeichnet m_k den Mittelwert des k-ten Clusters.

2. Die Beobachtungen werden dem Cluster zugeordnet, zu dessen Mittelpunkt der geringste Abstand besteht. Das hierfür verwendete Abstandsmaß ist die Euklidische Distanz. Daraus ergibt sich für eine Beobachtung x_{t_n}:

$$C(t_n) = \text{argmin}_{1 \leq k \leq K} \parallel x_{t_n} - m_k \parallel^2 . \tag{4}$$

Die beiden Schritte werden solange wiederholt bis die Clusteraufteilung unverändert bleibt (Hastie et al. 2009 [7], S. 509 - 510).

3.4 K-fache Kreuzvalidierung

K-fache Kreuvalidierung ist eine Methode zur Approximation der wahren Fehlerrate einer Klassifikation. Diese wird genutzt, wenn nicht genügend Beobachtungen vorhanden sind, um den Datensatz in einen Trainings- und in einen Testdatensatz geeigneter Größe einzuteilen. Dazu wird der gegebene Datensatz zufällig in K möglichst gleich große Teile aufgeteilt. Jeder Teil wird einmal als Testdatensatz genutzt, auf den übrigen $K - 1$ Teilen wird dann jeweils das Modell angepasst. Somit können insgesamt K Vorhersagefehlerraten berechnet werden. Die kreuzvalidierte Fehlerrate lässt sich als Mittelwert der K Vorhersagefehlerraten berechnen.
Ein Spezialfall der K-fachen Kreuzvalidierung ist die Leave-one-out Kreuzvalidierung. Dabei wird $K = N$ gewählt. Das Modell wird somit N-mal auf $N - 1$ Beobachtungen angepasst und auf einer Beobachtung getestet. Die sich daraus ergebende Kreuzvalidierte Fehlerrate wird als Leave-one-out Fehlerrate bezeichnet (Hastie et al. 2009 [7], S. 241 - 242).

K-fache Kreuzvalidierung lässt sich in `R` mit der Funktion `cv.glm()` aus dem Paket `boot` (Canty, A. und Ripley, B., 2017 [4]) durchführen.

3.5 Logistische Regression

Das Ziel der logistischen Regression ist es, das beste Modell zur Erklärung der Zielvariablen Y durch die erklärenden Variablen $X = (X_1, \ldots, X_m)$ zu finden. Die Zielvariable Y ist dabei binär, d.h. sie kann nur die Werte 0 und 1 annehmen. Aus diesem Grund modelliert die logistische Regression die Wahrscheinlichkeit $\pi(x)$, dass ein Ereignis eintritt (also das $y = 1$ gilt), in Abhängigkeit von $X = x$. Dabei basiert das zugehörige Modell auf der logistischen Verteilung und hat die Form

$$\pi(x) = \frac{\exp(\alpha + \beta_1 x_1 + \cdots + \beta_m x_m)}{1 + \exp(\alpha + \beta_1 x_1 + \cdots + \beta_m x_m)}. \tag{5}$$

Die Modellparameter β_j mit $j = 1, \ldots, m$, sowie der Achsenabschnitt α, können mit der Maximum-Likelihood Methode geschätzt werden (Hosmer et al. 2013 [9], S. 1 - 8). Durch eine Logarithmus Transformation der sogenannten Chance, welche sich durch

$$\frac{\pi(x)}{1 - \pi(x)}$$

berechnen lässt, ergibt sich der $\mathrm{logit}(\pi)$. Dieser hat die Form

$$\mathrm{logit}(\pi) = \log\left(\frac{\pi(x)}{1 - \pi(x)}\right) \overset{(5)}{=} \alpha + \beta_1 x_1 + \cdots + \beta_m x_m \tag{6}$$

und weißt einige nützliche Eigenschaften einer linearen Regressionsgleichung auf (Bender et al. (2007) [1]).

3.6 Integrationsbasiertes Variationsmaß

Sei X_T eine zentrierte univariate Zeitreihe mit $T = \{t_1, \ldots t_N\} \subseteq \mathbb{Z}$. Die Beobachtungen x_{t_n} werden durch linear interpolierte Pfade verbunden. Das für die Variabilität verwendete Maß ist der Flächeninhalt der von den oben beschriebenen Pfaden und der x-Achse eingeschlossenen Fläche. Die Berechnung des dazugehörigen Integrals erfolgt stückweise. Zu diesem Zweck werden die folgenden beiden Fälle unterschieden:

1. $x_{t_n} x_{t_{n+1}} \geq 0$: In diesem Fall sind die Vorzeichen aufeinanderfolgender Realisierungen gleich, was zu einer Trapez Form der betrachteten Fläche führt. Der Flächenin-

halt wird durch

$$F_{t_n,t_{n+1}} = (t_{n+1} - t_n)\frac{\mid x_{t_n} \mid + \mid x_{t_{n+1}} \mid}{2} = \mid x_{t_n} + x_{t_{n+1}} \mid$$

berechnet. Die Gleichung gilt, da für die in dieser Arbeit betrachteten Daten immer $t_{n+1} - t_n = 2$ gilt und x_{t_n} und $x_{t_{n+1}}$ das gleiche Vorzeichen aufweisen.

2. $x_{t_n} x_{t_{n+1}} < 0$: In diesem Fall sind die Vorzeichen aufeinanderfolgender Realisierungen unterschiedlich, wodurch die betrachtete Fläche aus zwei Dreiecken besteht. Der Schnittpunkt mit der x-Achse wird als

$$s_{t_n,t_{n+1}} = t_n + \frac{2x_{t_n}}{x_{t_n} - x_{t_{n+1}}}$$

bezeichnet. Dies gilt, da für die in dieser Arbeit betrachteten Daten immer $t_{n+1} - t_n = 2$ gilt. Damit lässt sich der Flächeninhalt berechnen, welcher aus den aufsummierten Flächeninhalten der beiden Dreiecke besteht:

$$\begin{aligned}
F_{t_n,t_{n+1}} &= \frac{\mid x_{t_n} \mid}{2}(s_{t_n,t_{n+1}} - t_n) + \frac{\mid x_{t_{n+1}} \mid}{2}(t_{n+1} - s_{t_n,t_{n+1}}) \\
&= \frac{\mid x_{t_n} \mid}{2}\frac{2x_{t_n}}{x_{t_n} - x_{t_{n+1}}} + \frac{\mid x_{t_{n+1}} \mid}{2}\left(2 - \frac{2x_{t_n}}{x_{t_n} - x_{t_{n+1}}}\right) \\
&= \frac{\mid x_{t_n} \mid x_{t_n} + \mid x_{t_{n+1}} \mid (x_{t_n} - x_{t_{n+1}}) - \mid x_{t_{n+1}} \mid x_{t_n}}{x_{t_n} - x_{t_{n+1}}} \\
&= \frac{\mid x_{t_n} \mid x_{t_n} - \mid x_{t_{n+1}} \mid x_{t_{n+1}}}{x_{t_n} - x_{t_{n+1}}} \\
&= \frac{x_{t_{n+1}}^2 + x_{t_n}^2}{\mid x_{t_{n+1}} - x_{t_n} \mid}.
\end{aligned}$$

Diese Gleichung gilt, da für die in dieser Arbeit betrachteten Daten immer $t_{n+1} - t_n = 2$ gilt. Die letzte Gleichheit ergibt sich wieder aus einer Fallunterscheidung:

(a) $x_{t_n} < 0, x_{t_{n+1}} > 0$:

$$\frac{-x_{t_n}^2 - x_{t_{n+1}}^2}{x_{t_n} - x_{t_{n+1}}} = \frac{-(x_{t_n}^2 + x_{t_{n+1}}^2)}{-(x_{t_{n+1}} - x_{t_n})} = \frac{x_{t_n}^2 + x_{t_{n+1}}^2}{\mid x_{t_{n+1}} - x_{t_n} \mid}.$$

Dies gilt, da $x_{t_{n+1}} - x_{t_n} > 0$ gilt, da $x_{t_{n+1}} > 0$ und $x_{t_n} < 0$ nach Voraussetzung.

(b) $x_{t_n} > 0, x_{t_{n+1}} < 0$:

$$\frac{x_{t_n}^2 + x_{t_{n+1}}^2}{x_{t_n} - x_{t_{n+1}}} = \frac{x_{t_n}^2 + x_{t_{n+1}}^2}{\mid x_{t_{n+1}} - x_{t_n} \mid}.$$

Dies gilt, da $x_{t_n} - x_{t_{n+1}} = x_{t_n} + \mid x_{t_{n+1}} \mid = \mid x_{t_{n+1}} - x_{t_n} \mid$ gilt, da $x_{t_{n+1}} < 0$ und $x_{t_n} > 0$ nach Voraussetzung.

Der Gesamtflächeninhalt F ist die Summe der Teilflächen $F_{t_n,t_{n+1}}$ für $n = 1, \dots N-1$:

$$F = \sum_{n=1}^{N-1} F_{t_n,t_{n+1}}. \tag{7}$$

Die Beschreibung des Integrationsbasierten Variationsmaßes ist aus Kohlenbach (2017) [13] entnommen.

4 Das Changepoint Paket

Für die Berechnung der Changepoints innerhalb der VCP-Methode wird das `R`-Paket `changepoint()` [12] verwendet. Das Paket bietet die Möglichkeit nach Änderungen in der Varianz, im Mittelwert und in Varianz und Mittelwert zu suchen. Da die VCP-Methode nach Änderungen in der Varianz sucht, wird die Funktion `cpt.var()` genutzt. Als Methode wird PELT (*Pruned Exact Linear Time*) gewählt und für alle anderen Variablen werden die Standardeinstellungen verwendet. Die PELT Methode basiert auf der Optimal Partitioning Methode, welche nach einer optimalen Partition der Zeitreihe sucht, ohne die Reihenfolge der Beobachtungen zu verändern. In Abschnitt 4.1 wird zunächst Optimal Partitioning vorgestellt. Die Funktionsweise der PELT Methode folgt in Abschnitt 4.2. Auf die verwendete Kostenfunktion $\mathcal{C}$ und den Strafparameter β wird in den Kapiteln 4.3 und 4.4 eingegangen. Da das Vorgehen der `cpt.var()` Funktion an einigen Stellen von der Definition der verwendeten Methoden abweicht, werden die entsprechenden Unterschiede in Abschnitt 4.6 erläutert.

Des weiteren wird die folgende Notation verwendet, $x = (x_{t_1}, x_{t_2}, \dots, x_{t_N}) \in \mathrm{R}^N$ ist der Vektor der Realisationen von Zufallsvariablen X_{t_n}, $n \in \{1, 2, \dots N\}$ einer Zeitreihe mit Beobachtungszeitpunkten $t_1, \dots, t_N$ und $N \in \mathrm{N}$. Der Vektor der Beobachtungen von Zeitpunkt t_i bis Zeitpunkt t_j, mit $i < j$ wird als $x_{t_i:t_j} = (x_{t_i}, x_{t_{i+1}}, \dots, x_{t_{j-1}}, x_{t_j})$ bezeichnet. Die unbekannten Changepoints, deren Anzahl $M < N$ nicht bekannt ist, werden als t_{τ_m}, mit $m = 0, \dots, M$ und $0 = \tau_0 < \tau_1 < \cdots < \tau_M < \tau_{M+1} < N$, bezeichnet.

4.1 Optimal Partitioning

Optimal Partitioning (OP) ist eine Methode, die mit Hilfe von dynamischer Programmierung eine optimale Partition auf einer Zeitreihe sucht und in Jackson et al. (2005) [10] vorgestellt wird. Die Reihenfolge der Datenpunkte X_{t_n} mit $n = 1, \ldots, N$ bleibt dabei unverändert. Als Changepoint werden dabei die Grenzen zwischen den Abschnitten der optimalen Partition bezeichnet. N bezeichnet die Anzahl der gegebenen Datenpunkte. Die Einteilung der Zeitreihe wird mit Hilfe einer Kostenfunktion $\mathcal{C}$ bewertet. Ziel des Optimal Partitioning ist es

$$\sum_{m=1}^{M+1} \mathcal{C}(x_{t_{(\tau_{m-1}+1):t_{\tau_m}}}) \tag{8}$$

zu minimieren. Dabei bezeichnen die t_{τ_m}, mit $0 = \tau_0 < \tau_1 < \cdots < \tau_M < \tau_{M+1} = N$, die unbekannten Changepoints, mit unbekannter Anzahl $M \in \mathbb{N}$. Die Kostenfunktion kann beliebig gewählt werden und darf auch einen Strafparameter enthalten, welcher die Hinzunahme eines zusätzlichen Changepoints sanktioniert.

Sei $F(t_s)$ die Minimierung von Gleichung (8) für die Daten $x_{t_1:t_s}$ und sei $\mathcal{T}_s = \{\tau : 0 = \tau_0 < \tau_1 < \cdots < \tau_M < \tau_{M+1} = s\}$ die Menge der möglichen Changepointvektoren für diese Daten. Dabei kann M für verschiedene Vektoren aus $\mathcal{T}_s$, unterschiedliche Werte annehmen. Die Idee der OP Methode ist es, für jeden Zeitpunkt t_n den optimalen letzten Changepoint vor diesem Zeitpunkt zu bestimmen. Dabei wird davon ausgegangen, dass

$$\mathcal{C}(x_{t_{(\tau_{m-1}+1):t_{\tau_m}}}) = \mathcal{C}(x_{t_{(\tau_{m-1}+1):t_s}}) + \mathcal{C}(x_{t_{(s+1):t_{\tau_m}}}),$$

für alle s mit $\tau_{m-1} + 1 < s < \tau_m$ gilt, wenn zwischen $t_{\tau_{m-1}}$ und t_{τ_m} kein Changepoint liegt. Ist zum Zeitpunkt t_s ein Changepoint vorhanden, gilt hingegen

$$\mathcal{C}(x_{t_{(\tau_{m-1}+1):t_{\tau_m}}}) > \mathcal{C}(x_{t_{(\tau_{m-1}+1):t_s}}) + \mathcal{C}(x_{t_{(s+1):t_{\tau_m}}}).$$

Damit lässt sich $F(t_s)$ rekursiv berechnen, da

$$\begin{aligned}
F(t_s) &= \min_{\tau \in \mathcal{T}_s} \left\{ \sum_{m=1}^{M+1} \left[\mathcal{C}\left(x_{t_{(\tau_{m-1}+1):t_{\tau_m}}} \right) \right] \right\} \\
&= \min_{\tau \in \mathcal{T}_s} \left\{ \sum_{m=1}^{M} \left[\mathcal{C}\left(x_{t_{(\tau_{m-1}+1):t_{\tau_m}}} \right) \right] + \mathcal{C}\left(x_{t_{(\tau_M+1):t_{\tau_{(M+1)}}}} \right) \right\} \\
&= \min_{r \in \{0,\ldots,s-1\}} \left\{ \min_{\tau \in \mathcal{T}_r} \left\{ \sum_{m=1}^{M+1} \left[\mathcal{C}\left(x_{t_{(\tau_{m-1}+1):t_{\tau_m}}} \right) \right] \right\} + \mathcal{C}\left(x_{t_{(r+1):t_s}} \right) \right\} \\
&= \min_{r \in \{0,\ldots,s-1\}} \left\{ F(t_r) + \mathcal{C}\left(x_{t_{(r+1):t_s}} \right) \right\}
\end{aligned} \tag{9}$$

gilt. Mit Hilfe der Rekursion aus Gleichung (9), kann der OP-Algorithmus wie folgt ausgeführt werden:

1. $F(0) = 0$, $\mathrm{cp} = \emptyset$

2. Für $\tau^* = 1, \ldots, N$ werden die folgenden Werte berechnet:

$$
\begin{aligned}
F(t_{\tau^*}) &= \min_{\tau \in \{0,\ldots,\tau^*-1\}} \left[F(t_\tau) + \mathcal{C}\left(x_{(t_{(\tau+1)}:t_{\tau^*})} \right) \right] \\
\tau^l &= \mathrm{argmin}_{\tau \in \{0,\ldots,\tau^*-1\}} \left[F(t_\tau) + \mathcal{C}\left(x_{(t_{(\tau+1)}:t_{\tau^*})} \right) \right] \\
\mathrm{cp} &= \mathrm{cp} \cup \{\tau^l\}
\end{aligned}
$$

3. Die optimalen Changepoints, welche Gleichung (8) minimieren, werden bestimmt, indem mit dem letzten Wert im cp Vektor begonnen wird. Sei $n_1 = \mathrm{cp}(N)$, dann gilt $\tau_M = n_1$ und der letzte Changepoint liegt zum Zeitpunkt t_{n_1} vor. Der vorletzte Changepoint wird durch den Wert in $\mathrm{cp}(n_1 - 1)$ bestimmt. Nach diesem Schema wird solange vorgegangen, bis der Zeitpunkt t_0 erreicht wird.

Somit wird die Kostenfunktion $\mathcal{C}$ in Iteration i zur Berechnung von $F(t_i)$ und τ^l jeweils i Mal berechnet. Da der Algorithmus insgesamt N Iterationen durchläuft, wird $\mathcal{C}$ insgesamt

$$2 * (1 + 2 + \cdots + N - 1 + N) = \mathcal{O}(N^2)$$

Mal ausgewertet. Daraus ergibt sich eine Laufzeit für die OP Methode von $\mathcal{O}(N^2)$ (siehe Jackson et al., 2005 [10]).

4.2 PELT

Die *Pruned Exact Linear Time* (PELT) Methode basiert auf der in Abschnitt 4.1 beschriebenen OP Methode und wird in Killick et al. (2012) [11] vorgestellt. Um die Laufzeit zu verringern, wird ein zusätzlicher Pruning Schritt eingeführt der dafür sorgt, dass Beobachtungen die nicht als letzter Changepoint vor der aktuell betrachteten Beobachtung in Frage kommen, nicht in die Berechnung mit einbezogen werden. Ziel der PELT Methode ist es die Summe

$$\sum_{m=1}^{M+1} \left[\mathcal{C}\left(x_{t_{(\tau_{m-1}+1)}:t_{\tau_m}} \right) + \beta \right] \tag{10}$$

zu minimieren. Sei $F(s)$ die Minimierung der Summe aus Formel (10) bezüglich der Daten $x_{t_1:t_s}$ mit $\mathcal{T}_s = \{\tau : 0 = \tau_0 < \tau_1 < \cdots < \tau_M < \tau_{M+1} = s\}$ Menge der möglichen Changepointvektoren. $F(s)$ lässt sich durch

$$F(s) = \min_{\tau \in \{0,\ldots,s-1\}} \left[F(\tau) + \mathcal{C}\left(x_{(\tau+1):s}\right) + \beta\right] \tag{11}$$

berechnen.

Zudem wird angenommen, dass die Kosten gesenkt werden, wenn ein zusätzlicher Changepoint innerhalb der betrachteten Zeitpunkte hinzugefügt wird. Die Annahme lässt sich wie folgt als Formel ausdrücken:

$$\begin{aligned} &\exists \text{ Konstante } K : \forall t < s < T \in \mathbb{N} : \\ &\mathcal{C}(x_{(t+1):s}) + \mathcal{C}(x_{(s+1):T}) + K \leq \mathcal{C}(x_{(t+1):T}). \end{aligned} \tag{12}$$

Dabei bezeichnet s den letzten Changepoint vor T. Daraus folgt, wenn zu einem Zeitpunkt $T > s > t$

$$F(t) + \mathcal{C}(x_{(t+1):s}) + K \geq F(s) \tag{13}$$

gilt, kann t niemals optimaler letzter Changepoint vor T sein. Zur Berechnung von $F(T)$, wird nach Formel 11 für jeden Zeitpunkt $u < T$, also auch für s und t, der Wert $F(u) + \mathcal{C}(x_{(u+1):T}) + \beta$ ermittelt und der Zeitpunkt mit dem der minimale Wert erreicht wird, ist optimaler letzter Changepoint vor T. Damit ergibt sich die folgende Ungleichung:

$$\begin{aligned} F(s) + \mathcal{C}(x_{(s+1):T}) + \beta &\overset{(13)}{\leq} F(t) + \mathcal{C}(x_{(t+1):s}) + K + \mathcal{C}(x_{(s+1):T}) + \beta \\ &\overset{(12)}{\leq} F(t) + \mathcal{C}(x_{(t+1):T}) + \beta. \end{aligned} \tag{14}$$

Damit ist die obige Behauptung gezeigt, dass t niemals optimaler letzter Changepoint vor s sein kann, wenn Formel 13 gilt, da $F(t) + \mathcal{C}(x_{(t+1):T}) + \beta$ niemals kleiner als $F(s) + \mathcal{C}(x_{(s+1):T}) + \beta$ werden kann.

Um Formel (13) auszunutzen wird im zweiten Schritt des OP Algorithmus zusätzlich die Menge R_{n+1} berechnet, die nur diejenigen Beobachtungen enthält, die als letzter Changepoint vor der als nächstes zu betrachtenden Beobachtung in Frage kommen. Diese Menge lässt sich durch

$$R_{n+1} = \{t_n\} \cup \{\tau \in R_n : F(t_\tau) + \mathcal{C}(x_{t_{\tau+1}:t_n}) + K < F(t_n)\} \tag{15}$$

berechnen, wobei $R_1 = \{0\}$ gilt. An dieser Stelle liegt in Killick et al. (2012) [11] ein Fehler vor, da hier der Schnitt anstelle einer Vereinigung gebildet wird. Dies würde dazu

führen, dass für alle $n = 1, \ldots, N$, $R_{n+1} = \emptyset$ gilt, wodurch keine Changepoints gefunden werden können.

Die PELT Methode verwendet einen Strafterm β, welcher als Teil der Kostenfunktion verstanden werden kann. Dieser erhöht sich bei steigender Anzahl an Changepoints und „bestraft“ somit eine hohe Anzahl an Changepoints. Für die PELT Methode ergeben sich damit die folgenden Schritte:

1. $F(0) = -\beta$, $R_1 = \{0\}$, $\mathrm{cp} = \emptyset$

2. Für $\tau^* = 1, \ldots, N$ werden die folgenden Werte berechnet:

$$
\begin{aligned}
F(t_{\tau^*}) &= \min_{\tau \in R_n} \left[F(t_\tau) + \mathcal{C}\left(x_{t_{\tau+1}:t_{\tau^*}}\right) + \beta \right] \\
\tau^l &= \mathrm{argmin}_{\tau \in R_n} \left[F(t_\tau) + \mathcal{C}\left(x_{t_{\tau+1}:t_{\tau^*}}\right) + \beta \right] \\
\mathrm{cp} &= \mathrm{cp} \cup \{\tau^l\} \\
R_{n+1} &= \{t_n\} \cup \{\tau \in R_n : F(t_\tau) + \mathcal{C}(x_{t_{\tau+1}:t_n}) + K < F(t_n)\}
\end{aligned}
$$

3. Der cp Vektor wird als Changepointvektor ausgegeben.

Die Worstcase Laufzeit der PELT Methode liegt bei $\mathcal{O}(N^2)$, was der Laufzeit der OP Methode entspricht. Dies ist dann der Fall, wenn im Pruning Schritt von PELT keine Beobachtungen entfernt werden können. Die Informationen zur PELT Methode sind aus Killick et al. (2012) [11] entnommen.

4.3 Die Kostenfunktion

Die erfolgreiche Durchführung des PELT Algorithmus hängt von der Wahl einer geeigneten Kostenfunktion ab. Nach Killick et al. (2012) [11] ist zweimal die negative Loglikelihood eine passende Kostenfunktion, wenn nach Änderungen der Varianz σ^2 bei bekanntem Erwartungswert μ gesucht wird. Da in dieser Arbeit auf trendbereinigten Daten gearbeitet wird, gilt für den Erwartungswert $\mu = 0$. Damit werden die Kosten für einen

Abschnitt zwischen zwei Changepoints $x_{t_{\tau_{m-1}+1}:t_{\tau_m}}$ durch

$$\begin{aligned}
\mathcal{C}\left(x_{t_{\tau_{m-1}+1}:t_{\tau_m}}\right) &= \mathcal{C}\left(x_{t_{\tau_{m-1}+1}}, x_{t_{\tau_{m-1}+2}}, \ldots, x_{t_{\tau_{m-1}}}, x_{t_{\tau_m}}\right) = -2\log L_{x_{t_{\tau_{m-1}+1}},\ldots,x_{t_{\tau_m}}}\left(\hat{\sigma}^2_{\left(t_{\tau_{m-1}+1}:t_{\tau_m}\right)}\right) \\
&= -2\log\left(\prod_{i=\tau_{m-1}+1}^{\tau_m} \frac{1}{\sqrt{2\pi\hat{\sigma}^2_{\left(t_{\tau_{m-1}+1}:t_{\tau_m}\right)}}} \exp\left(\frac{-(x_{t_i}-\mu)^2}{2\hat{\sigma}^2_{\left(t_{\tau_{m-1}+1}:t_{\tau_m}\right)}}\right)\right) \\
&= -2\sum_{i=\tau_{m-1}+1}^{\tau_m}\left(\log\left(\frac{1}{\sqrt{2\pi\hat{\sigma}^2_{\left(t_{\tau_{m-1}+1}:t_m\right)}}}\right) + \log\left(\exp\left(\frac{-(x_{t_i}-\mu)^2}{2\hat{\sigma}^2_{\left(t_{\tau_{m-1}+1}:t_{\tau_m}\right)}}\right)\right)\right) \\
&= -2\left(-\frac{1}{2}\left(\tau_m-\tau_{m-1}\right)\log(2\pi) - \frac{1}{2}\left(\tau_m-\tau_{m-1}\right)\log(\hat{\sigma}^2_{\left(t_{\tau_{m-1}+1}:t_{\tau_m}\right)})\right. \\
&\quad \left. + \sum_{i=\tau_{m-1}+1}^{\tau_m}\left(\frac{-(x_{t_i}-\mu)^2}{2\hat{\sigma}^2_{\left(t_{\tau_{m-1}+1}:t_{\tau_m}\right)}}\right)\right) \qquad (16)
\end{aligned}$$

berechnet. Die Varianz $\hat{\sigma}^2_{t_{\tau_{m-1}+1}:t_{\tau_m}}$ wird durch den ML-Schätzer

$$\hat{\sigma}^2_{t_{\tau_{m-1}+1}:t_{\tau_m}} = \frac{\sum_{i=\tau_{m-1}+1}^{\tau_m}(x_{t_i}-\mu)^2}{\tau_m-\tau_{m-1}} \qquad (17)$$

geschätzt. Wird der Varianzschätzer aus Gleichung (17) in Gleichung (16) eingesetzt, ergibt sich für die Kosten $\mathcal{C}(x_{t_{\tau_{m-1}+1}:t_{\tau_m}})$

$$\begin{aligned}
\mathcal{C}(x_{t_{\tau_{m-1}+1}:t_{\tau_m}}) &= -2\left(-\frac{1}{2}(\tau_m-\tau_{m-1})\log(2\pi) - \frac{1}{2}(\tau_m-\tau_{m-1})\log\left(\frac{\sum_{i=\tau_{m-1}+1}^{\tau_m}(x_{t_i}-\mu)^2}{\tau_m-\tau_{m-1}}\right)\right. \\
&\quad \left. -\frac{1}{2}\frac{\sum_{i=\tau_{m-1}+1}^{\tau_m}(x_{t_i}-\mu)^2}{\sum_{i=\tau_{m-1}+1}^{\tau_m}(x_{t_i}-\mu)^2/(\tau_m-\tau_{m-1})}\right) \\
&= (\tau_m-\tau_{m-1})\left(\log(2\pi) + \log\left(\frac{\sum_{i=\tau_{m-1}+1}^{\tau_m}(x_{t_i}-\mu)^2}{\tau_m-\tau_{m-1}}\right) + 1\right) \qquad (18)
\end{aligned}$$

(siehe Killick et al., 2012 [11]).

4.4 MBIC

Die Standardeinstellung der `cpt.var()` Funktion für den Strafterm β ist MBIC, was für *Modified Bayes Information Criterion* steht. In Zhang und Siegmund (2007) [17] ist der MBIC als Fitnessfunktion zum auffinden von Changepoints im Erwartungswert definiert. Dabei wird angenommen, dass die zugrundeliegenden Beobachtungen X_{t_n} mit $n = 1, \ldots, N$ normalverteilte Zufallszahlen mit

$$X_{t_n} \sim \mathcal{N}(\mu_j, \sigma^2) \text{ für } n = \tau_j + 1, \ldots, \tau_{j+1}, j = 0, \ldots, M, \qquad (19)$$

sind, wobei τ_j die Changepoints innerhalb der Daten bezeichnet. Das so definierte Modell mit M Changepoints wird als $\mathcal{M}_M$ bezeichnet, während das Modell mit gleichbleibendem Erwartungswert über den gesamten Beobachtungszeitraum als $\mathcal{M}_0$ bezeichnet wird. Der MBIC ist dann definiert als

$$\log\left(\frac{P(x \mid \mathcal{M}_M)}{P(x \mid \mathcal{M}_0)}\right) = \frac{1}{2}\sum_{i=1}^{M+1}(\hat{t}_i - \hat{t}_{i-1})[\bar{x}_i(\hat{t}) - \bar{x}]^2 \tag{20}$$

$$- \frac{1}{2}\sum_{i=1}^{M+1}\log(\hat{t}_i - \hat{t}_{i-1}) + (\frac{1}{2} - M)\log(N) + O_p(1), \tag{21}$$

wobei

$$\hat{t} = (\hat{t}_1, \ldots, \hat{t}_M) = \text{argmax}_{0<t_1<\cdots<t_M<N}\sum_{i=1}^{M-1}(\hat{t}_i - \hat{t}_{i-1})[\bar{x}_i(\hat{t}) - \bar{x}]^2 \tag{22}$$

gilt und $\bar{x}_i(\hat{t})$ als

$$\bar{x}_i(\hat{t}) = \frac{1}{\hat{t}_i - \hat{t}_{i-1}}\sum_{j=\hat{t}_{i-1}}^{\hat{t}_i - 1} x_j \tag{23}$$

definiert ist, während $\bar{x}$ den Mittelwert über die gesamten Beobachtungen beschreibt. Der erste Term des MBIC (siehe Gleichung (20)) ist dabei das Maximum der logLikelihood unter $\mathcal{M}_M$. Der zweite Term (siehe Gleichung (21)) wird als Strafterm bezeichnet, wobei der Restterm $O_p(1)$ außer Acht gelassen wird. Dieser Strafterm kann als $\approx 3M\log(N)/2$ abgeschätzt werden und ändert sich nicht, wenn die Varianz σ^2 unbekannt ist (siehe Zhang und Siegmund, 2007 [17]).

Im `changepoint` Paket wird lediglich die Abschätzung des Strafterm des MBIC als Wert für den Parameter β verwendet. Die oben gegebene Definition des MBIC bezieht sich auf Änderungen im Erwartungswert. In der Funktion `cpt.var()` wird der zugehörige Strafterm jedoch auch in Bezug auf Änderungen der Varianz verwendet. Es muss beachtet werden, dass β in [11] zwar getrennt von der Kostenfunktion eingeführt wird, im `changepoint` Paket [12] jedoch als der Kostenfunktion zugehörig betrachtet wird, da β hier als $3M\log(N)$ abgeschätzt wird. Nach dem Vorgehen aus Zhang und Siegmund [17] wird die logLikelihood ($\log(L(\hat{\sigma}^2))$) abzüglich β als Fitnessfunktion verwendet. Durch die Multiplikation mit -2 ergibt sich für PELT:

$$\begin{aligned}-2(\log(L(\hat{\sigma}^2)) - \beta) &\approx -2(\log(L(\hat{\sigma}^2)) - 3M\log(N)/2) \\ &= -2\log(L(\hat{\sigma}^2)) + 3M\log(N).\end{aligned}$$

4.5 Verringerung der Laufzeit durch den Pruningschritt in PELT

Da die Worse Case Laufzeit der PELT Methode der Laufzeit der OP Methode entspricht, stellt sich die Frage, wie stark der Pruningschritt der PELT Methode die Datenmenge reduziert. Um diese Frage zu beantworten, wurden mit der Funktion `rnorm()` normalverteilte Zufallszahlen erzeugt. Dabei haben 10 Beobachtungen die Varianz 1.5, gefolgt von 5 Beobachtungen mit Varianz 2, 5 Beobachtungen mit Varianz 2.5, 10 Beobachtungen mit Varianz 2, 5 Beobachtungen mit Varianz 1.5, 5 Beobachtungen mit Varianz 1 und zum Schluss 10 Beobachtungen mit Varianz 2. Der Erwartungswert hat bei allen 50 Beobachtungen den Wert 0. Auf diesen Zufallszahlen wurde die PELT Methode, mit zweimal der negativen Loglikelihood als Kostenfunktion, angewendet. Bei dieser Kostenfunktion hat der Parameter K den Wert 0. Für den Strafparameter β wurde der Strafterm des MBIC (siehe Abschnitt 4.4) verwendet. In Abbildung 3 sind die Mengen R_τ, aus 100 unabhängigen Wiederholungen der PELT Methode angewendet auf normalverteilte Zeitreihen mit je 50 Beobachtungen, dargestellt. Die Färbung zeigt an, wie oft ein Zeitpunkt t in der jeweiligen Menge R_τ vertreten ist. Dunkelrote Punkte waren in mehr als 75 der insgesamt 100 Wiederholungen in dem entsprechenden R_τ enthalten, rote Punkte in mehr als 50 aber höchstens 75 Fällen, lila gefärbte Punkte zeigen Beobachtungen, die mehr als 25 aber weniger als 50 Mal auftauchen und blass rosafarbene Punkte sind höchstens 25 Mal in dem jeweiligen R_τ enthalten. Anhand der Grafik lässt sich erkennen, dass der größter Teil der Beobachtungen in 25 oder weniger der 100 Wiederholungen in dem entsprechenden R_τ auftaucht. Daraus lässt sich schließen, dass der Pruning Schritt in den meisten Fällen zu einer deutlichen Verkürzung der Laufzeit führt.

4.6 Unterschiede zwischen dem changepoint Paket und Killick et al. (2012) [11]

Es liegen Unterschiede zwischen der Implementierung der Funktion `cpt.var()` aus dem Paket `changepoint` [12] und der Beschreibung der PELT Methode in Killick et al. (2012) [11] vor. Diese bestehen in:

- Der Ausgabe der Changepoints
- Dem Hinzufügen einer minimalen Segmentlänge

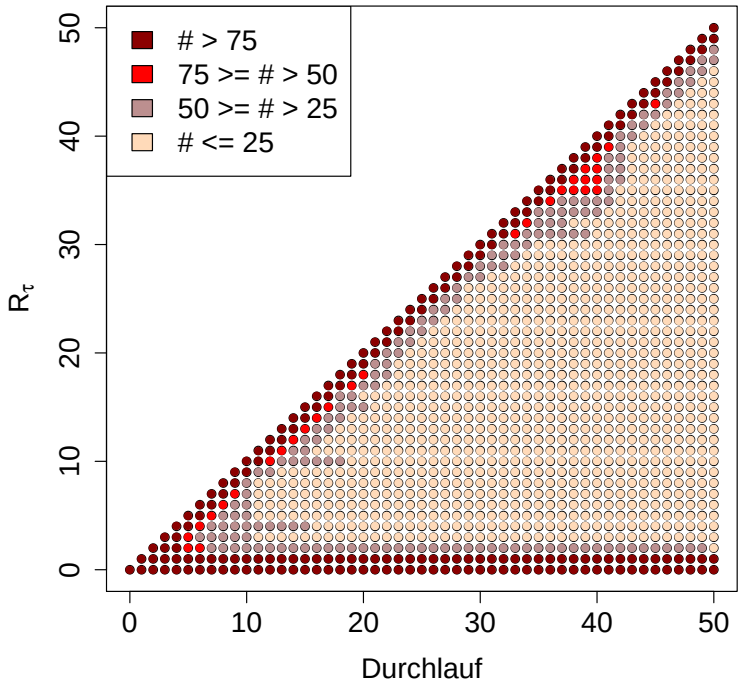


Abbildung 3: Die Menge R_τ bei Anwendung der PELT Methode auf eine normalverteilte Zeitreihe mit 50 Beobachtungen. Es wurden 100 unabhängige Wiederholungen durchgeführt. Die Färbung erfolgt je nachdem wie oft eine Beobachtung in der jeweiligen Menge R_τ auftaucht.

- Der Kostenfunktion.

4.6.1 Unterschiede in der Ausgabe der Changepoints

Nach Killick et al. (2012) [11] werden alle im cp Vektor gespeicherten Werte als Changepoints ausgegeben. In R Output B.1 ist das Ergebnis des cp Vektor für den 22.02.17 von 05 bis 09 Uhr zu sehen. Doppelte Werte wurden dabei der Übersichtlichkeit halber entfernt. Insgesamt enthält der Vektor 220 verschiedene Einträge. Mit der Funktion `cpt.var()` werden auf der gleichen Zeitreihe nur 120 Changepoints berechnet. Diese sind in R output B.2 zu sehen. Da der vollständige cp Vektor (siehe R Output B.1) 100 Punkte mehr enthält, als die tatsächlich durch die Funktion ausgegebenen Changepoints (siehe R Output B.2), ist davon auszugehen, dass die `cpt.var()` Funktion nicht den vollständigen cp Vektor ausgibt.

Die PELT Methode basiert auf der OP Methode und unterscheidet sich von dieser nur

durch den Pruning Schritt, welcher zu einer kürzeren Laufzeit führt. Daher liegt die Annahme nahe, dass das gleiche Vorgehen zur Ausgabe der finalen Changepoints verwendet wird. In Abschnitt 4.1 ist das bei OP verwendete Backtracking näher erläutert. Da dabei immer vom Ende eines Blocks ausgehend der Anfang des jeweiligen Blocks gefunden wird (der cp Vektor enthält für jede Beobachtung den optimalen Changepoint vor dieser Beobachtung), wird von dem Anfang des zuletzt gefundenen Blocks eine Beobachtung nach links gegangen um das Ende des vorherigen Blocks zu erreichen. Somit enthält jeder Abschnitt mindestens zwei Beobachtungen.
Im Quellcode des `changepoint` [12] Paket ist jedoch zu sehen, dass zur Identifikation der endgültigen Changepoints der letzte Optimale Changepoint ausgehend vom Beginn des zuletzt bestimmten Blocks der Partition gewählt wird. Durch dieses Vorgehen kann die optimale Partition auch Blöcke der Länge 1 enthalten. Im `changepoint` Paket wird dies durch die Einführung einer minimalen Segmentlänge (siehe Abschnitt 4.6.2) verhindert.

4.6.2 Minimale Segmentlänge

Bei der Suche nach Änderungen der Varianz muss ein Segment mindestens 2 Elemente enthalten (siehe Killick et al. 2012 [11]). Durch die in `cpt.var()` verwendete Methode zur Ausgabe der optimalen Partition (siehe Abschnitt 4.6.1) ist dies nicht gewährleistet. Daher beinhaltet die Funktion eine Variable `minseglen` welche die minimale Segmentlänge festlegt. Die Variable erwartet einen ganzzahligen Eingabewert zwischen 2 und $\lfloor M/2 \rfloor$ und ist standardmäßig auf 2 gesetzt.
Damit jedes Element des cp Vektor einen Wert enthält, welcher die minimale Segmentlänge einhält, werden $F(t_n)$ und $\mathrm{cp}(n)$ durch

$$F(t_n) = \min_{j \in R_n \setminus \{n-\text{minseglen}+1,\ldots,n-1\}} \{F(t_j) + \mathcal{C}(x_{t_{j+1}:t_n}) + \beta\}$$
$$\mathrm{cp}(n) = \mathrm{argmin}_{j \in R_n \setminus \{n-\text{minseglen}+1,\ldots,n-1\}} \{F(t_j) + \mathcal{C}(x_{t_{j+1}:t_n}) + \beta\},$$

anstatt wie in Kapitel 4.2 beschrieben, berechnet. Dies gewährleistet, dass der Abstand zum letzten Changepoint vor dem Zeitpunkt t_n mindestens der minimalen Segmentlänge entspricht. Für Zeitpunkte t_n, für die gilt $n \leq \mathrm{minseglen}$ werden $F(t_n)$ und $\mathrm{op}(n)$ nicht wie oben beschrieben, sondern durch

$$F(t_n) = \mathcal{C}(x_{t_0:t_n})$$
$$\mathrm{cp}(n) = 0$$

bestimmt. Somit ist t_2 der frühst mögliche Zeitpunkt, zu dem ein Changepoint auftreten kann, wenn 2 als minimale Segmentlänge gewählt wird.

4.6.3 Unterschiede der Kostenfunktion

An die Kostenfunktion aus Gleichung (18) wird im Source Code der `cpt.var()` (siehe Killick et al., 2016 [12]) Funktion, bei der Wahl von MBIC als Strafparameter, $\log(n-j)$ addiert. Damit ergeben sich die Kosten für einen Abschnitt $x_{t_{j+1}:t_n}$ einer Partition, durch

$$\mathcal{C}(x_{t_{j+1}:t_n}) = (n-j)\left(\log(2\pi) + \log\left(\frac{\sum_{i=j+1}^{n}(x_{t_i}-\mu)^2}{n-j}\right) + 1\right) + \log(n-j).$$

Durch die Addition von $\log(n-j)$ werden viele kurze Blöcke innerhalb einer Partition bestraft.

5 Methoden zur Erkennung von Anomalien

In dieser Arbeit werden drei Methoden zur Identifikation von Anomalien innerhalb der Rissdaten, betrachtet. Dabei handelt es sich um die Variance Changepoint Detection (VCP) Methode, die Modified Variance Changepoint Detection (MVCP) Methode und die Clustering of MAD filtered Data (CMAD) Methode. Alle drei Methoden haben gemeinsam, dass sie auf der trendbereinigten Zeitreihe $\tilde{X}_{t_n}$ arbeiten. Die Trendbereinigung erfolgt, indem der gleitende Median auf die Zeitreihe angewendet wird und die daraus entstandene Zeitreihe $X^*_{t_n}$ von der original Zeitreihe abgezogen wird. Dieses Vorgehen, kann durch

$$\tilde{X}_{t_n} = X_{t_n} - X^*_{t_n},$$

für $n = 1, \ldots, N$, als Formel ausgedrückt werden. Die verschiedenen Methoden verwenden dabei unterschiedliche Fensterbreiten K.

5.1 Variance Changepoint Detection

Variance Changepoint Detection (VCP) ist eine Methode zur Identifikation von Anomalien innerhalb der Daten einer Zeitreihe. Wie dabei vorgegangen wird ist in Abbildung 4 grafisch dargestellt. Die betrachte Zeitreihe wird zunächst trendbereinigt. Dabei wird eine Fensterbreite von $K = 301$ verwendet. Auf die trendbereinigte Zeitreihe wird die

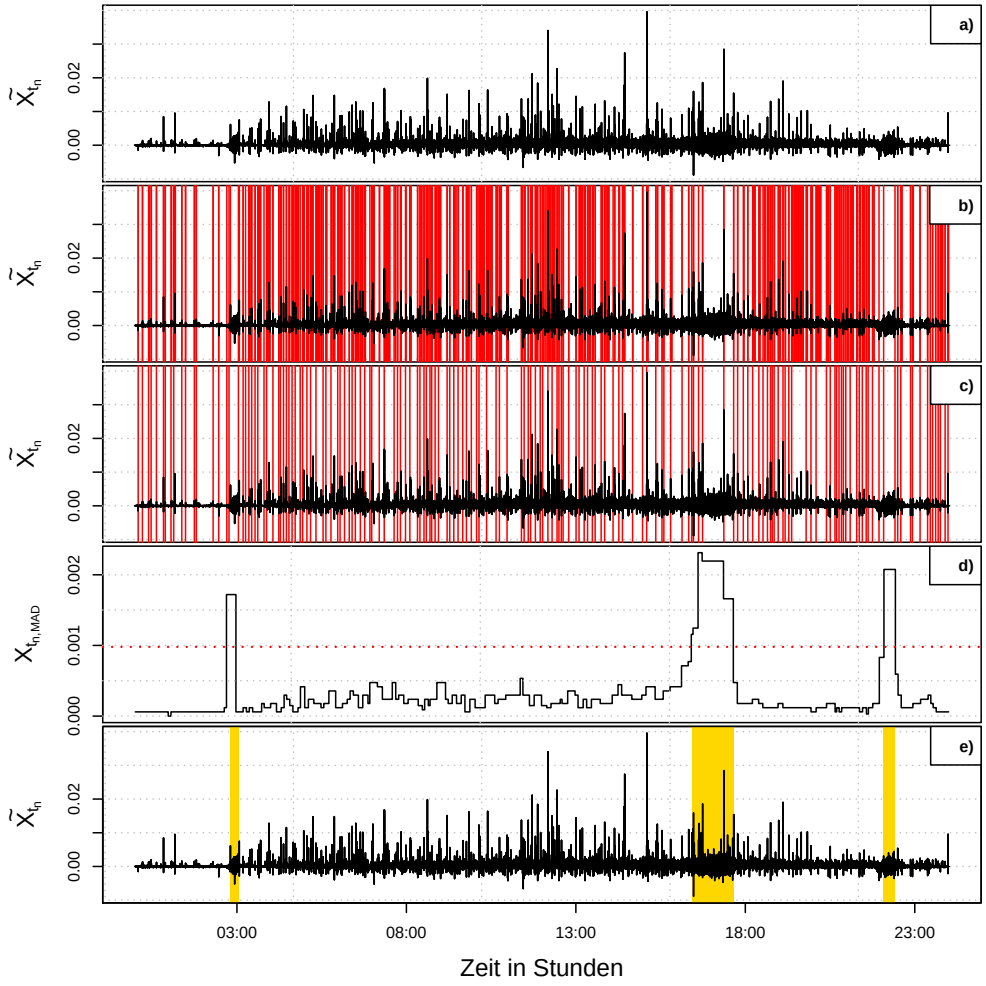


Abbildung 4: Vorgehensweise der VCP Methode am Beispiel des 09.06.2016 an Wegaufnehmer WOS4. a) Die trendbereinigte Zeitreihe $\tilde{X}_{t_n}$ mit Fensterbreite 301, b) Die mit `cpt.var()` identifizierten Strukturbrüche sind rot in der trendbereinigten Zeitreihe eingezeichnet, c) Die Changepoints mit einem Abstand von mindestens 80 Beobachtungen zum vorherigen Changepoint sind rot in der trendbereinigten Zeitreihe eingezeichnet, d) Treppenfunktion über die MADs zwischen den Strukturbrüchen aus c), die gepunktete rote Linie stellt die Grenze zwischen den beiden Clustern „Anomalie“ und „keine Anomalie“ dar, e) Die mit VCP identifizierten Anomalien sind in der trendbereinigten Zeitreihe gelb hinterlegt.

Funktion `cpt.var()` aus dem `R`-Paket `changepoint` [12] angewendet. Wie diese Funktion vorgeht ist in Kapitel 4 näher erläutert. Die so gefundenen Changepoints sind in Abbildung 4 b) rot markiert. Zwischen den mit `cpt.var()` gefundenen Strukturbrüchen werden sowohl die Abstände zum vorherigen Changepoint als auch der MAD berechnet. Anhand der MADs werden die Abschnitte zwischen zwei Changepoints in mit Hilfe von K-Means in zwei Cluster eingeteilt. Abschnitte die als Anomalie klassifiziert sind und Abschnitte die nicht als Anomalie klassifiziert sind. Beim Clustern werden nur diejenigen Abschnitte verwendet, die mindestens 80 Beobachtungen enthalten, da davon ausgegangen wird, dass eine Anomalie mindestens 2.5 Minuten anhält. Eine kürzere Änderung der Varianz kann auf den Verkehr zurückgeführt werden. Die Grenzen der Abschnitte die beim Clustern berücksichtigt werden, sind in Abbildung 4 c) rot dargestellt. In Abbildung 4 d) sind die MADs zwischen den Changpoints (für Abschnitte mit mindestens 80 Beobachtungen) dargestellt. Die rot gestrichelt Linie zeigt an, wo die Grenze zwischen den beiden Clustern verläuft. Werden benachbarte Abschnitte als Anomalie klassifiziert, werden diese zu einer großen Anomalie zusammengefasst. Da Abschnitte mit weniger als 80 Beobachtungen, beim Clustern nicht berücksichtigt worden sind, und somit keine Klassifikation für diese Abschnitte vorliegt, werden sie der gleichen Gruppe wie der vorhergegangene Abschnitt zugeordnet. Enthält der erste Abschnitt $x_{t_1}, \ldots, x_{t_{\tau_1}}$ weniger als 80 Beobachtungen wird davon ausgegangen, dass es sich dabei nicht um eine Anomalie handelt. In Abbildung 4 e) sind die von der VCP Methode als Anomalien identifizierten Blöcke gelb hinterlegt.

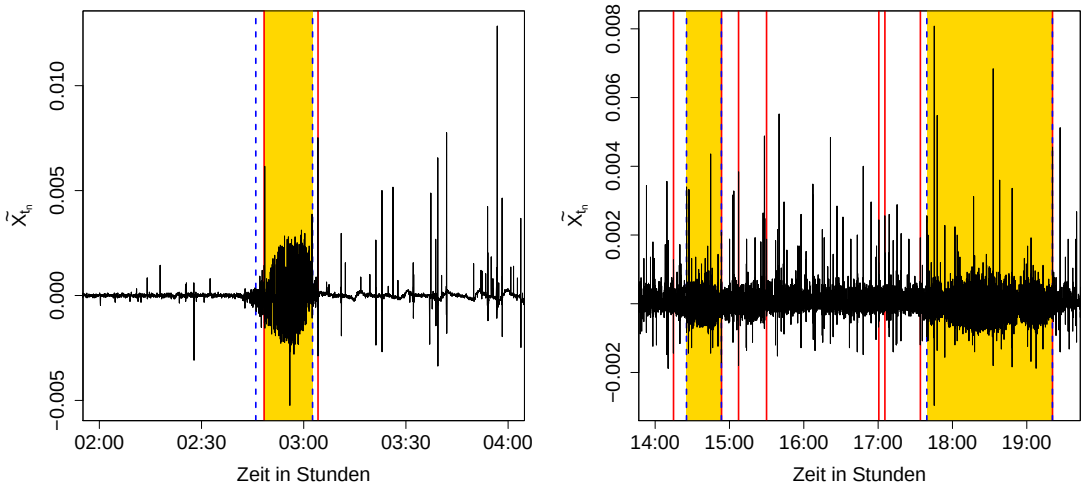
Werden Abschnitte mit weniger als 80 Beobachtungen erst nach dem Clustern entfernt, führt dies zu dem Problem, dass sehr kleine Abschnitte als Anomalie identifiziert werden, anstelle von größeren Abschnitten in denen tatsächlich eine Anomalie vorliegt. Werden Abschnitte mit weniger als 80 Beobachtungen anschließend entfernt, werden keine Anomalien identifiziert, obwohl welche vorhanden sind.

5.2 Modified Variance Changepoint Detection

Bei der Modified Variance Changepoint Detection (MVCP) handelt es sich um eine Modifikation der VCP Methode. Dabei wird die VCP Methode sowohl auf die normale als auch auf die umgekehrte Zeitreihe (d.h. die letzte Beobachtung der ursprünglichen Zeitreihe

steht an erster Stelle usw.) angewendet. Es werden nur die Beobachtungen als Teil einer Anomalie ausgegeben, die in beiden Durchläufen einer Anomalie zugeordnet werden. Die unterschiedlichen Ergebnisse der VCP Methode, bei Anwendung auf die originale Zeitreihe und auf die umgekehrte Zeitreihe, sind auf die in der `cpt.var()` Funktion des `changepoint` Pakets [12] verwendete minimale Segment Länge (siehe Abschnitt 4.6.2) zurückzuführen.

In Abbildung 5 ist am Beispiel des 09.06.16 am Wegaufnehmer WOS4 (Grafik a)), sowie des 22.02.17 am Wegaufnehmer WWS4 (Grafik b)) zu sehen, wie sich die erkannten Anomalien bei Verwendung der original Zeitreihe und der umgekehrten Zeitreihe unterscheiden. Bei höheren Temperaturen sind die Anomalien im allgemeinen leichter zu



(a) Zoom WOS4 09.06.16

(b) WWS4 22.02.17

Abbildung 5: Unterschiede wenn VCP auf die normale Zeitreihe und auf die umgekehrte Zeitreihe angewendet wird. Die roten Markierungen stellen Beginn und Ende der bei Verwendung der normalen Zeitreihe gefundenen Anomalien dar, während die blau gestrichelten Linien Beginn und Enden der Anomalien bei Verwendung der umgekehrten Zeitreihe markieren. Die gelb hinterlegten Abschnitte werden von der MVCP Methode als anomale Sequenzen ausgegeben.

erkennen. Daher liegt am 09.06.16 nur eine kleine Abweichung zwischen Beginn und Ende der ersten Anomalie zwischen ca. 02:45 Uhr und 03:10 Uhr vor. Dabei beginnt die Anomalie bei Verwendung der umgekehrten Zeitreihe früher und endet früher (blau gestrichelte Markierung), während bei Verwendung der originalen Zeitreihe die Anomalie

später beginnt und auch später endet (rote Markierung). Der gelb hinterlegte Bereich wurde in beiden Durchläufen als Anomalie erkannt und somit von der MVCP als anormale Sequenz ausgegeben.
In Teil (b) von Abbildung 5 sind größere Unterschiede zwischen der Anwendung von VCP auf die original Zeitreihe und auf die umgekehrte Zeitreihe erkennbar. Zwischen 15 und 16 Uhr und zwischen 17 und 18 Uhr gibt es Bereiche, die nur bei Anwendung der VCP Methode auf die original Zeitreihe als Anomalien klassifiziert werden. Daher werden diese Abschnitte von der MVCP Methode nicht als anomale Sequenz ausgegeben.

5.2.1 Gründe für die Unterschiede bei der Anwendung der VCP Methode auf die umgekehrte Zeitreihe

Es gibt zwei mögliche Ursachen für die Abweichungen zwischen den auf der normalen und auf der umgekehrten Zeitreihe mit der VCP Methode klassifizierten Anomalien. Diese werden anhand der am 09.06.16, an Wegaufnehmer WOS4, und der am 22.02.17, an Wegaufnehmer WWS4, gemessenen Daten erläutert. Eine Ursache ist der Umgang mit Abschnitten zwischen zwei mit `cpt.var()` gefundenen changepoints, die weniger als 80 Beobachtungen enthalten. Diese erhalten immer die gleiche Klasse wie der vorherige Abschnitt. Durch diese kurzen Abschnitte kann das Ende einer Anomalie somit weiter nach hinten verschoben werden, der Beginn einer Anomalie ändert sich aber nicht. Da auf der umgekehrten Zeitreihe von hinten begonnen wird, stellt das Ende der hier gefundenen Anomalien, bei einer Übertragung auf die originale Zeitreihe, den Anfang der Anomalie dar. Sind also auf der umgekehrten Zeitreihe kurze Abschnitte ans Ende der Anomalie angehängt worden, beginnt diese auf der originalen Zeitreihe früher als die gleiche Anomalie bei Anwendung von VCP auf die originale Zeitreihe. Umgekehrt kann somit auch der Fall auftreten, dass bei Anwendung von VCP auf die originale Zeitreihe die Anomalie, aufgrund von kurzen ans Ende angehängten Abschnitten, später endet als auf der umgekehrten Zeitreihe. Dieser Fall ist ist in Abbildung 6, am Beispiel des 09.06.16 zwischen 2 und 4 Uhr am Wegaufnehmer WOS4, dargestellt. Der Abschnitt von 02:48:22 Uhr bis 03:02:36 Uhr wird von beiden Durchläufen als Anomalie klassifiziert. Die davor und danach liegenden Abschnitte, welche mehr als 80 Beobachtungen enthalten, liegen in beiden Fällen im Zeitraum von 02:42:04 Uhr bis 02:45:56 Uhr bzw. 03:04:12 Uhr bis 03:10:56 Uhr und werden als Bereiche ohne Anomalien klassifiziert. Dazwischen lie-

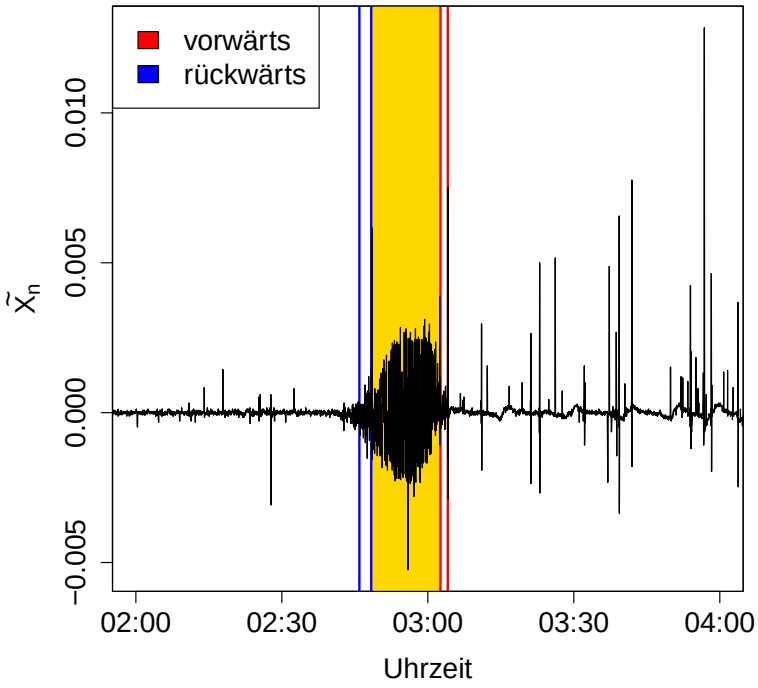
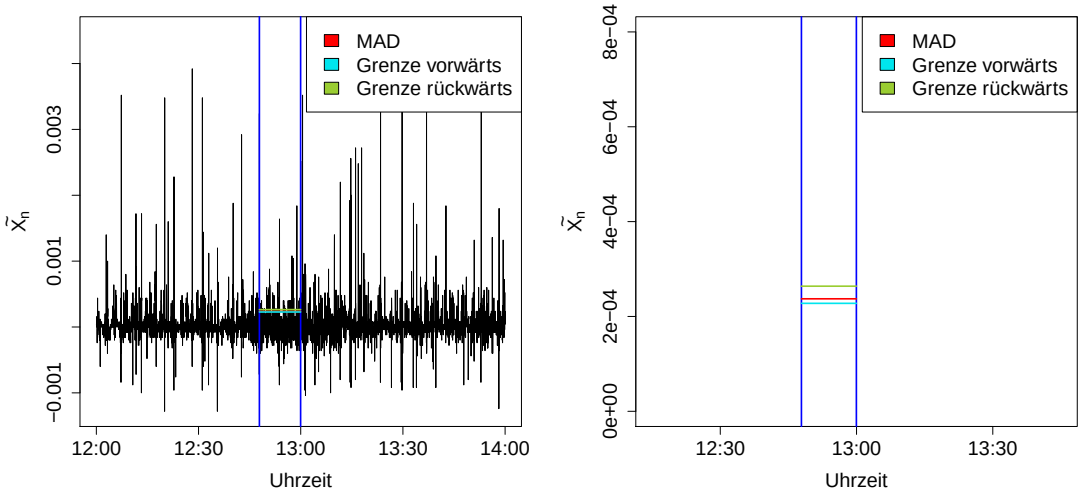


Abbildung 6: Ausschnitt der trendbereinigten Zeitreihe vom 09.06.16 an Wegaufnehmer WOS4. Der von der VCP Methode sowohl auf der normalen, als auch auf der umgekehrten Zeitreihe, als Anomalie identifizierte Bereich, ist gelb hinterlegt. Abschnitte die auf der originalen Zeitreihe direkt auf die Anomalie folgen und weniger als 80 Beobachtungen enthalten, sind von roten Linien eingefasst. Für die Anwendungen der VCP Methode auf die umgekehrte Zeitreihe sind diese Bereiche von blauen Linien eingefasst.

gen jeweils Bereiche mit weniger als 80 Beobachtungen (02:45:56 Uhr bis 02:48:22 Uhr (Abschnitt 1) bzw. 03:02:36 Uhr bis 03:04:12 Uhr (Abschnitt 2)). Da diese Abschnitte dem vorangehenden Bereich hinzugefügt werden, wird auf der originalen Zeitreihe Abschnitt 1 dem als *keine Anomalie* klassifizierten Bereich von 02:42:04 Uhr bis 02:45:56 Uhr hinzugefügt, während Abschnitt 2 an die *Anomalie* angehängt wird. Auf der umgekehrten Zeitreihe verhält es sich genau andersherum, da hier von hinten begonnen wird. Abschnitt 2 wird dem, als *keine Anomalie* klassifizierten, Bereich von 03:04:12 Uhr bis 03:10:56 Uhr hinzugefügt, während Abschnitt 1 der *Anomalie* zugeordnet wird. Der in Abbildung 6 von roten Linien eingefasste Bereich wird bei Verwendung der originalen Zeitreihe der Anomalie hinzugefügt, während die blauen Linien den Bereich markieren, der bei Verwendung der umgekehrten Zeitreihe der Anomalie zugeordnet wird. Durch

diesen Fall hervorgerufene Unterschiede können somit nur zu Beginn und am Ende einer Anomalie auftauchen. Anomalien die nur in einer Richtung identifiziert werden, kommen damit nicht zustande.

Wie dies hervorgerufen wird, wird am Beispiel der am 22.02.17, an Wegaufnehmer WWS4, gemessenen Daten erläutert. An diesem Tag identifiziert die Funktion `cpt.var()` um 03:41:34 Uhr, 03:50:10 Uhr, 08:20:34 Uhr, 08:22:06 Uhr, 14:14:44 Uhr, 14:14:50 Uhr, 17:59:22 Uhr, 19:15:16 Uhr, 21:19:08 Uhr, 22:08:24 Uhr und um 23:42:16 Uhr auf der originalen Zeitreihe Änderungen in der Varianz, die auf der umgekehrten Zeitreihe nicht entdeckt werden. Dafür findet die Funktion auf der umgekehrten Zeitreihe um 03:41:24 Uhr, 03:49:52 Uhr, 08:17:28 Uhr, 08:17:32 Uhr, 17:58:58 Uhr, 19:15:02 Uhr, 21:19:06 Uhr, 22:07:54 Uhr und um 23:41:44 Uhr Änderungen der Varianz, welche auf der originalen Zeitreihe nicht auftauchen. Für die meisten dieser Punkte gibt es ein Gegenstück, für den Fall, dass die `cpt.var()` Funktion von der anderen Seite aus angewendet wird, welches nicht mehr als 16 Beobachtungen entfernt ist. In einigen Fällen werden jedoch Paare von Beobachtungen, zwischen denen weniger als 80 Beobachtungen liegen gefunden, für die es kein Gegenstück in der anderen Richtung gibt. Die unterschiedlichen Ergebnisse der `cpt.var()` Funktion sind darauf zurückzuführen, dass die PELT Methode die Änderungen der Varianz lediglich approximiert (siehe Kapitel 4). Bei der Entfernung der Punkte, deren Abstand zum Vorgänger kleiner als 80 Beobachtung ist, fallen in beiden Richtungen jeweils 2 Punkte weg, die in der anderen Richtung nicht vorhanden sind. Dies führt dazu, dass beim Clustern über die MADs einige Unterschiede bezüglich der betrachteten Abschnitte, und somit auch über deren MADs, auftauchen. Daher ist die Grenze zwischen den beiden Clustern auf der originalen Zeitreihe und auf der umgekehrten Zeitreihe nicht identisch. Auf der originalen Zeitreihe liegt die Grenze bei 0.0002276537. Wird von hinten begonnen, wird eine Grenze, zwischen den beiden Clustern, von 0.0002638479 ermittelt. In Abbildung 7 ist der Ausschnitt der trendbereinigten Zeitreihe vom 22.02.17 zwischen 12 und 14 Uhr, gemessen an Wegaufnehmer WWS4, dargestellt. In diesen Zeitraum fällt eine Anomalie, die nur auf der unveränderten Zeitreihe von der VCP Methode als Anomalie identifiziert wird. Die mit der Funktion `cpt.var()` identifizierten changepoints, welche den Abschnitt eingrenzen, der der entsprechenden Anomalie zugrunde liegt, werden durch vertikale blaue Linien dargestellt. Diese changepoints, sowie die unmittelbar davor und danach identifizierten changepoints unterscheiden sich für die beiden



(a) 22.02.17 WWS4 12 bis 14 Uhr

(b) Zoom MAD

Abbildung 7: Ausschnitt der trendbereinigten Zeitreihe, am 22.02.17 an Wegaufnehmer WWS4. Dargestellt ist der nur auf der ursprünglichen Zeitreihe von VCP als Anomalie identifizierte Bereich. Die beiden mit `cpt.var()` identifizierten changepoints, welche den entsprechenden Abschnitt begrenzen, sind durch vertikale blaue Linien markiert. Die horizontale rote Linie stellt den MAD innerhalb dieses Abschnittes dar. *Grenze vorwärts* steht für die Grenze zwischen den beiden Clustern, bei der Anwendung der VCP Methode auf die unveränderte Zeitreihe. *Grenze rückwärts* steht für die Grenze zwischen den beiden Clustern, bei der Anwendung der VCP Methode auf die umgekehrte Zeitreihe. Ein Abschnitt wird als Anomalie klassifiziert, wenn der entsprechende MAD größer als die Grenze zwischen den beiden Clustern ist.
a) Ausschnitt der gesamten trendbereinigten Zeitreihe am 22.02.17 zwischen 12 und 14 Uhr.
b) Zoom auf den MAD und die Clustergrenzen der beiden Durchläufe zur besseren Übersicht.

Durchläufe nicht. Die Grenzen zwischen den beiden Clustern, für beide Durchläufe, sowie der MAD des Abschnittes sind als horizontale Linien dargestellt. In Teil b) der Grafik ist ein Zoom von Teil a) dargestellt, wobei die trendbereinigte Zeitreihe nicht zu sehen ist und der Fokus stattdessen auf dem MAD und den Clustergrenzen der beiden Durchläufe liegt. Hier lässt sich erkennen, dass der MAD für diesen Abschnitt zwischen den Grenzen der beiden Durchläufe liegt. Da ein Abschnitt, bei dem die entsprechende Grenze über-

schritten wird, als Anomalie klassifiziert wird, wird hier auf der originalen Zeitreihe von einer Anomalie ausgegangen. Für den Fall, dass von hinten begonnen wird, liegt keine Anomalie vor. Es können somit immer nur in der Richtung mehr Anomalien identifiziert werden, in der die Grenze zwischen den beiden Clustern den geringeren Wert annimmt. Dieser Fall kann auch direkt vor oder nach einem Abschnitt, welcher von beiden Durchläufen als Anomalie identifiziert wird, auftreten. Somit kann auch eine auf der originalen Zeitreihe, mit VCP identifizierte, Anomalie vor der entsprechenden, auf der umgekehrten Zeitreihe identifizierten, Anomalie beginnen. Eine auf umgekehrten Zeitreihe identifizierte Anomalie kann auch nach der entsprechenden, auf der originalen Zeitreihe identifizierten Anomalie enden. Das Auftreten einer Kombination der verschiedenen Ursachen für den Unterschied zwischen Anomalien ist ebenfalls möglich.

Am 01.12.16 tritt am Wegaufnehmer WWS4 ein weiterer Fall auf, bei dem große Unterschiede zwischen der Anwendung von VCP auf die originale und auf die umgekehrte Zeitreihe vorhanden sind. Der Grund für die starke Abweichung ist die Annahme der VCP Methode, dass die Gruppe, der beim Clustern mehr Abschnitte zugewiesen werden, die Gruppe ohne Anomalien ist. Am 01.12.16 ist dies für die umgekehrte Zeitreihe jedoch nicht der Fall, wodurch nur Abschnitte als Anomalie klassifiziert werden, in denen keine Anomalie vorliegt. Diese Anomalie wird bei Anwendung der VCP Methode auf die ursprüngliche Zeitreihe eindeutig identifiziert. Da die MVCP Methode nur die Abschnitte als Anomalie ausgibt, die von der VCP Methode auf der originalen und auf der umgekehrten Zeitreihe als Anomalie identifiziert werden, findet sie in diesem Fall die vorliegende Anomalie nicht. Dieses Problem ist durch einen starken Ausreißer bei der Verkehrsschätzung in Kapitel 7 aufgefallen. Es kann angenommen werden, dass dieses Problem nur sehr selten auftritt, da nur gegen Ende der Analyse ein solcher Extremfall aufgefallen ist. Das Problem kann behoben werden, indem davon ausgegangen wird, dass das Cluster, dessen Zentrum den höheren Wert hat, die Abschnitte mit Anomalien enthält.

5.3 Clustering of MAD filtered Data

Clustering of MAD filtered Data (CMAD) ist eine Methode zur Erkennung von Anomalien in einer Zeitreihe. Die Vorgehensweise der Methode ist in Abbildung 8 dargestellt.
Es wird eine Trendbereinigung der Zeitreihe durch den gleitenden Median mit Fenster-

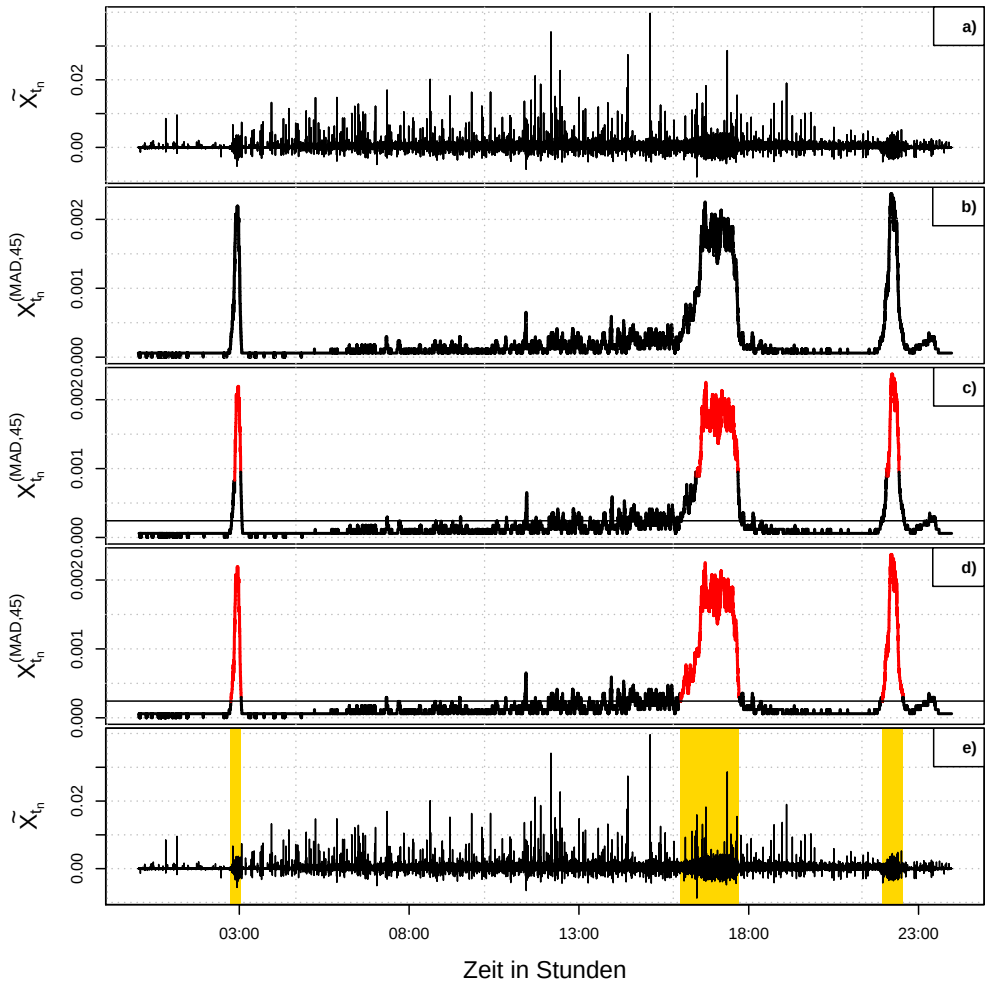


Abbildung 8: Vorgehensweise der CMAD Methode am Beispiel des 09.06.2016 an Wegaufnehmer WOS4. a) Die trendbereinigte Zeitreihe $\tilde{X}_{t_n}$ mit Fensterbreite 15. b) Die MAD transformierte Zeitreihe mit Fensterbreite 91. c) Die durch K-Means (mit $K = 2$) als Anomalie klassifizierten Beobachtungen sind rot markiert. d) Die bis zum Mittelwert erweiterten Cluster. e) Die mit CMAD identifizierten Anomalien sind in der trendbereinigten Zeitreihe gelb hinterlegt.

breite 15 vorgenommen (siehe Abbildung 8 a)). Auf die trendbereinigte Zeitreihe wird der gleitende MAD mit Fensterbreite 91 angewendet. In Abbildung 8 b) ist zu erkennen, dass Anomalien dadurch deutlicher zu sehen sind. Die MAD transformierte Zeitreihe wird als $X_{t_n}^{(\mathrm{MAD},45)}$ bezeichnet. Zur Identifikation der Anomalien werden die Beobachtungen auf $X_{t_n}^{(\mathrm{MAD},45)}$ mit K-Means in zwei Cluster eingeteilt (siehe Abbildung 8 c)). Dabei stellen die rot markierten Beobachtungen diejenigen Beobachtungen dar, die in eine anomale Sequenz fallen. Darauf wird noch einmal der gleitende Median angewendet, allerdings wird dieses Mal Fensterbreite 301 anstatt Fensterbreite 15 verwendet. Dies dient dazu, einzelne Beobachtungen, die als Anomalie klassifiziert wurden, aber verkehrsbedingte Ursachen haben, aus dem Cluster der Anomalien zu entfernen. Anschließend werden die als Anomalie klassifizierten Bereiche so lange erweitert, bis sie bis zum Mittelwert der MAD transformierten Zeitreihe reichen (Abbildung 8 d)). In Abbildung 8 e) sind die von CMAD als Anomalie ausgegebenen Bereiche auf $\tilde{X}_{t_n}$ gelb hinterlegt. Die Beschreibung der CMAD Methode wurde aus Kohlenbach (2017) [13] entnommen.

6 Vergleich der Methoden zur Erkennung von Anomalien

Die drei in dieser Arbeit betrachteten Methoden zur Identifikation von Anomalien in den, im Rahmen des Brückenmonitorings erhobenen, Rissdaten, werden anhand von verschiedenen Kriterien verglichen. In Abschnitt 6.1 erfolgt ein Vergleich bezüglich der Einheitlichkeit der Methoden. Ein Vergleich der Methoden mit manuell bestimmten Anomalien wird in Kapitel 6.2 durchgeführt. Dabei werden Fehlerraten sowohl für den Fall, dass die Methoden auf die einzelnen Tage angewendet werden, als auch für den Fall, dass die Methoden auf eine gesamte Woche angewendet werden, ermittelt. Da bei der VCP und bei der MVCP Methode Probleme auftreten, wenn in dem betrachteten Zeitraum keine Anomalien vorliegen, werden in Kapitel 6.3 untersucht wie oft die Methoden richtig erkennen, ob eine Anomalie vorliegt oder nicht. Zudem wird versucht diesen Wert mit Hilfe von Vorentscheidungen zu verbessern. In Abschnitt 6.4 wird für diese beiden Methoden zusätzlich untersucht, ob die Anzahl der an einem Tag, mit Hilfe der entsprechenden Methode, identifizierten Anomalien der Anzahl optisch identifizierter Anomalien entspricht.

6.1 Vergleich der Methoden bezüglich Einheitlichkeit der Ergebnisse

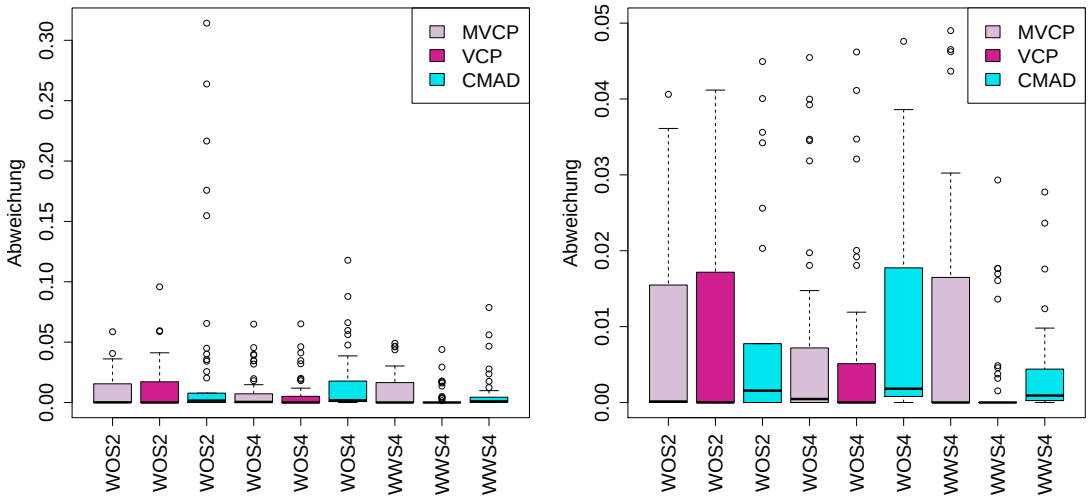
Zur Beurteilung der Einheitlichkeit der Ergebnisse der einzelnen Methoden werden jeweils drei aufeinanderfolgende Wochen betrachtet, wobei die Methoden sowohl auf die ersten beiden als auch auf die letzten beiden Wochen angewendet werden. Die mittlere Woche wird somit doppelt ausgewertet. Alle Methoden haben gemeinsam, dass sie nicht damit umgehen können wenn keine Anomalien auftreten. In diesem Fall bezeichnen sie viele kleinere Abschnitte als Anomalien. Daher wird hier davon ausgegangen, dass wenn mehr als 50 Anomalien innerhalb von 14 Tagen erkannt werden in Wirklichkeit keine Anomalien vorhanden sind. Die Grenze von 50 ergibt sich aus dem gerundeten Wert der maximal vorkommenden Anzahl an Anomalien innerhalb von 14 Tagen multipliziert mit dem Faktor 1.5. In den Daten vom 01.06.16 bis zum 31.05.17 treten innerhalb von 14 Tagen maximal 33 Anomalien auf. Dies ist der Fall im Zeitraum vom 14.01.17 bis zum 27.01.17 für den Wegaufnehmer WOS2. Daraus ergibt sich $33 * 1.5 = 49.5 \approx 50$. Der Faktor 1.5 wird gewählt um genügend Spielraum zu gewähren für den Fall, dass zusätzlich zu vorhandenen Anomalien weitere Bereiche als anomale Sequenz klassifiziert werden. Zur Berechnung der Fehlerraten werden jeweils zwei Vektoren der Länge N_d der doppelt ausgewerteten Woche erstellt. Ein Vektor für die Auswertung der ersten und zweiten Woche (y_1) und ein Vektor für die Auswertung der zweiten und dritten Woche (y_2). Für Zeitpunkte, die innerhalb einer Anomalie der jeweiligen Auswertung der mittleren Woche liegen, enthält der entsprechende Vektor an dieser Stelle den Eintrag 1, die restlichen Zeitpunkte erhalten den Eintrag 0. Die Fehlerraten werden berechnet indem die Anzahl der ungleichen Einträge der beiden Vektoren durch die Gesamtlänge der doppelt ausgewerteten Woche geteilt wird. Dieses Vorgehen kann durch

$$\frac{\sum_{n=1}^{N_d} \mathbb{1}(y_{1,n} \neq y_{2,n})}{N_d},$$

als Formel ausgedrückt werden.

In Tabelle A.1 im Anhang ist der Anteil der Abweichungen der doppelt ausgewerteten Woche für ca. ein Jahr (08.06.16 bis 23.05.17) dargestellt. Es sind zum größten Teil sehr niedrige Werte vorhanden, lediglich in den Wintermonaten fallen höhere Fehlerraten auf. Dabei muss jedoch berücksichtigt werden, dass sehr viele Beobachtungen vorhanden sind, von denen lediglich ein geringer Teil einer Anomalie angehört. Dies führt automatisch zu niedrigeren Fehlerraten. Zur Veranschaulichung der Fehlerraten aus Tabelle A.1 im An-

hang, sind diese in Abbildung 9 als Boxplots dargestellt. Teil a) der Abbildung zeigt dabei die gesamte Grafik mit allen Fehlerraten, während in Teil b), zum besseren Vergleich, ein Ausschnitt zu sehen ist, in dem extreme Ausreißer nicht betrachtet werden. Für den



(a) alle Fehlerraten

(b) ohne Berücksichtigung von Ausreißern (zoom)

Abbildung 9: Boxplots über die Anteile der Abweichungen der doppelt ausgewerteten Wochen (08.06.16 - 23.05.17).
a) Darstellung aller Fehlerraten des entsprechenden Zeitraumes.
b) Zoom (Ausreißer werden vernachlässigt).

Wegaufnehmer WWS4 lässt sich erkennen, dass die VCP Methode insgesamt die geringste Abweichung zwischen den doppelt ausgewerteten Wochen aufweist. Der Median und beide Quartile liegen bei Null. Es existieren einige Ausreißer, wobei der größte zwischen 0.04 und 0.05 liegt. Bei der MVCP Methode liegt der Median und das untere Quartil ebenfalls bei 0. Das obere Quartil und der obere Whisker nehmen mit ≈ 0.015 bzw. ≈ 0.03 für diese Methode am Wegaufnehmer WWS4 die höchsten Werte an. Für die CMAD Methode liegt das untere Quartil bei 0 und der Median knapp darüber. Der Interquartilsabstand dieser Methode liegt zwischen den Interquartilsabständen von VCP und MVCP. Einer der Ausreißer der CMAD Methode liegt bei 0.08. Dies ist der größte vorkommende Ausreißer aller drei Methoden. Die VCP Methode ist für den Wegaufnehmer WWS4 am konsequentesten was die Identifikation der Anomalien betrifft. Besonders auffällig ist hier die große Abweichung der Fehlerraten für die MVCP und die VCP Methode, da für diese Methoden im allgemeinen eher ähnliche Ergebnisse erwartet werden. Vermutlich

ist die Abweichung darauf zurück zu führen, dass bei der VCP Methode häufiger so viele Anomalien identifiziert werden, dass der Schwellenwert von 50 Anomalien überschritten wird und somit davon ausgegangen wird, dass keine Anomalien vorhanden sind. Passiert dies für beide Auswertungen der entsprechenden Woche, wird somit eine Fehlerrate von 0 erzielt. Bei der MVCP Methode können hingegen Anomalien wegfallen, wenn sie auf der original Zeitreihe und auf der umgekehrten Zeitreihe nicht in den gleichen Abschnitten auftreten. Somit kann es vorkommen, dass in einer oder beiden Auswertungen der jeweiligen Woche Anomalien entdeckt werden, bei denen zwangsläufig kleinere Abweichungen auftreten. Dies führt zu einer erhöhten Fehlerrate im Vergleich zur VCP Methode. Es muss beachtet werden, dass eine Fehlerrate von 0.03 keine dramatische Abweichung ist. Diese Fehlerrate entspricht in etwa der Lage des oberen Whiskers bei der MVCP Methode.

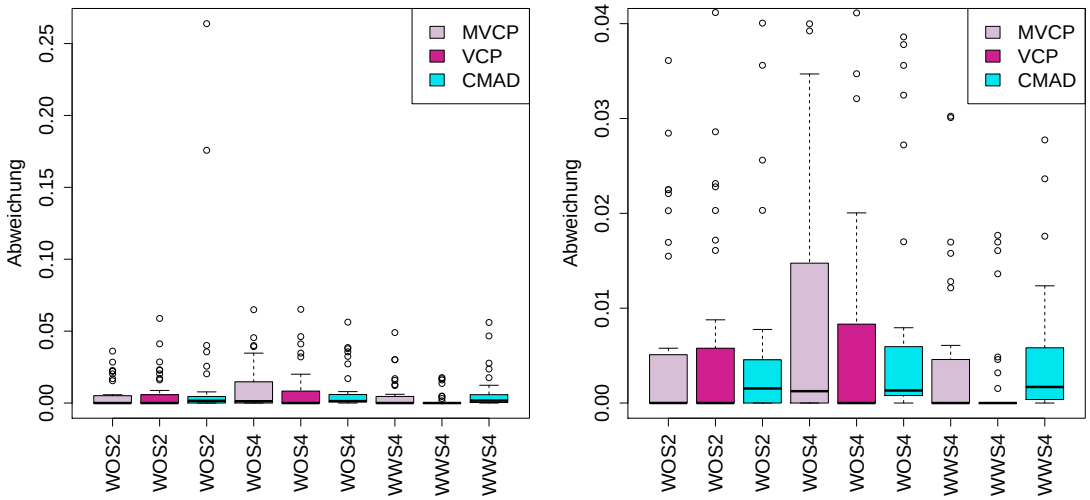
Bezüglich der Daten von Wegaufnehmer WOS4 sind die Fehlerraten der VCP und der MVCP Methode sehr ähnlich. Für beide Methoden liegen das untere Quartil und der Median bei Null. Das obere Quartil und der obere Whisker erreichen für die MVCP Methode einen geringfügig höheren Wert als für die VCP Methode. Bei beiden Methoden bleibt der obere Whisker unter dem Wert 0.02. Der größte Ausreißer liegt sowohl bei VCP als auch bei MVCP zwischen 0.06 und 0.07. Die Fehlerraten der CMAD Methode haben für den Wegaufnehmer WOS4 einen deutlich höheren Interquartilsabstand als für die anderen beiden Methoden. Der untere Whisker liegt ebenfalls bei 0, das untere Quartil und der Median liegen knapp darüber aber noch unter 0.01. Das obere Quartil der CMAD Methode liegt bei ca. 0.02 und der obere Whisker bei ca. 0.04. Der höchste Ausreißer hat eine Fehlerrate von in etwa 0.12.

In den Fehlerraten bezüglich der Daten von Wegaufnehmer WOS2 kommen insgesamt höhere Werte vor als bei den anderen beiden betrachteten Wegaufnehmern. Das untere Quartil und der Median liegt für alle drei Methoden bei Null. Bei der VCP und der MVCP Methode nimmt das obere Quartil einen höheren Wert an als bei der CMAD Methode. Das obere Quartil aller drei Methoden liegt aber unter 0.025. Der obere Whisker der VCP und MVCP Methode liegt knapp unter 0.05 und diese Methoden weisen lediglich wenige geringfügige Ausreißer auf. Für die CMAD Methode liegen viele Ausreißer vor, die teilweise auch sehr hohe Werte annehmen. Die Fehlerrate des größten Ausreißers der CMAD Methode liegt bei ca. 0.3.

Aus den Grafiken geht somit hervor, dass es je nach Wegaufnehmer unterschiedlich ist,

wie konsequent die Methoden bei der Findung von Anomalien sind.

In Abbildung 10 sind die Boxplots der Anteile der Abweichungen der doppelt ausgewerteten Wochen ohne Berücksichtigung der Wintermonate (30.11.16 - 28.02.17) dargestellt. Teil a) der Abbildung zeigt dabei die Boxplots über sämtliche Fehlerraten, während Teil b) zum besseren Vergleich einen kleineren Ausschnitt zeigt, bei dem extreme Ausreißer unberücksichtigt bleiben.



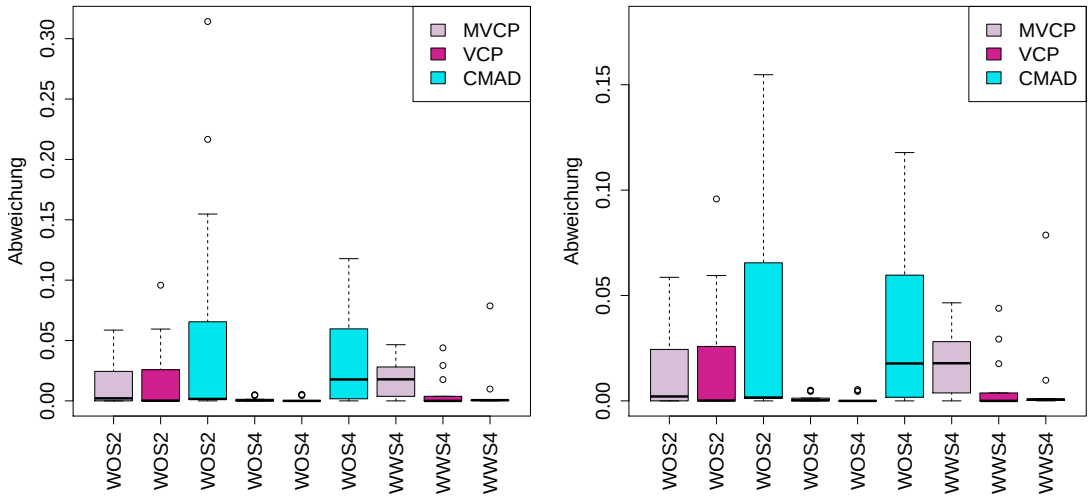
(a) alle Fehlerraten

(b) ohne Berücksichtigung von Ausreißern (zoom)

Abbildung 10: Boxplots über die Anteile der Abweichungen der doppelt ausgewerteten Wochen im Zeitraum vom 08.06.16 - 23.05.17, wenn die Wintermonate (30.11.16 - 28.02.17) nicht berücksichtigt werden.
a) Darstellung aller Fehlerraten des entsprechenden Zeitraumes.
b) Zoom (Ausreißer werden vernachlässigt).

Bleiben die Wintermonate unberücksichtigt, liegt das obere Quartil und der obere Whisker der MVCP Methode für die Daten des Wegaufnehmers WWS4 unter 0.01. Somit sind die Werte für diesen Wegaufnehmer geringer, als wenn der gesamte Zeitraum berücksichtigt wird. Für die VCP Methode liegen beide Quartile sowie der Median weiterhin bei Null. Auch der Boxplot für die CMAD Methode weist für diesen Wegaufnehmer keine größeren Veränderungen, im Vergleich zur Berücksichtigung der gesamten Daten, auf.
Für den Wegaufnehmer WOS4 weisen das obere Quartil und der obere Whisker der MVCP Methode höhere Werte auf, wenn der Winter nicht berücksichtigt wird. Die entspre-

chenden Werte haben sich im Vergleich zu den Boxplots der gesamten Fehlerraten in etwa verdoppelt. Das Verhalten der VCP Methode ist ähnlich, wobei die entsprechenden Werte mit 0.01 bzw. 0.02 unter dem oberen Quartil bzw. oberen Whisker der MVCP Methode für diesen Wegaufnehmer liegen. Bei der CMAD Methode liegt das umgekehrte Verhalten vor. Das obere Quartil und der obere Whisker weisen hier deutlich geringere Werte auf als wenn die Fehlerraten für die gesamten Daten in die Boxplots einfließen.
Für den Wegaufnehmer WOS2 nimmt das obere Quartil für alle drei Methoden einen geringeren Wert an, wenn die Wintermonate nicht berücksichtigt werden. Im Fall der VCP und der MVCP Methode hat auch der Whisker einen kleineren Wert. Bei der CMAD Methode bleibt dieser hingegen in etwa unverändert.
In Abbildung 11 sind Boxplots über die Anteile der Abweichungen der doppelt ausgewerteten Wochen für den Fall, dass ausschließlich die Wintermonate betrachtet werden, dargestellt.



(a) alle Fehlerraten

(b) ohne Berücksichtigung von Ausreißern (zoom)

Abbildung 11: Boxplots über die Anteile der Abweichungen der doppelt ausgewerteten Wochen, wenn ausschließlich die Wintermonate (30.11.16 - 28.02.17) betrachtet werden.
a) Darstellung aller Fehlerraten des entsprechenden Zeitraumes.
b) Zoom (Ausreißer werden vernachlässigt).

Für den Wegaufnehmer WWS4 fällt auf, dass bei der MVCP Methode das untere Quartil und der Median nicht mehr bei Null liegen. Der Median nimmt einen Wert von ≈ 0.018 an, das obere Quartil liegt bei ≈ 0.028 und der obere Whisker hat einen Wert von ≈ 0.05.

Dabei handelt es sich um deutlich höhere Werte als bei Betrachtung der Fehlerraten der gesamten Daten. Daraus lässt sich schließen, dass die Ursache für die relativ hohe Abweichung zwischen den Fehlerraten des gesamten Zeitraums für VCP und MVCP in den Wintermonaten zu finden ist. Im Fall der VCP Methode ist lediglich das obere Quartil geringfügig größer als Null. Der Median und das untere Quartil haben weiterhin einen Wert von Null. Bei der CMAD Methode liegen hingegen beide Quartile und der Median bei Null. Hier ist die Fehlerrate also geringer, als wenn das gesamte Jahr oder nur die Jahreszeiten Frühling, Sommer und Herbst betrachtet werden.

Für den Wegaufnehmer WOS4 liegen beide Quartile und der Median für die VCP und MVCP Methode bei Null, wenn nur die Wintermonate betrachtet werden. Zudem sind nur wenige geringfügige Ausreißer vorhanden. Im Fall der CMAD Methode liegt das untere Quartil leicht über Null, der Median hat einen Wert von 0.018, das obere Quartil liegt bei 0.06 und der obere Whisker nimmt einen Wert zwischen 0.10 und 0.15 an. Die Fehlerraten sind in diesem Fall für die Wintermonate also deutlich höher als die Fehlerraten für die übrigen drei Jahreszeiten.

Für den Wegaufnehmer WOS2 nehmen für alle drei Methoden das obere Quartil und der obere Whisker im Winter deutlich höhere Werte an, als wenn das gesamte Jahr oder die Jahreszeiten Frühling, Sommer und Herbst betrachtet werden. Dies fällt besonders bei der CMAD Methode auf. Der Median liegt trotzdem für alle drei Methoden nahe bei Null oder knapp darüber.

Liegen der Median und die Quartile im Winter bei Null und haben für das gesamte Jahr zum Teil etwas höhere Werte, kann daraus nicht automatisch geschlossen werden, dass die jeweilige Methode im Winter besser funktioniert. Die einheitlichen Ergebnisse können auch daraus entstehen, dass so viele Anomalien gefunden werden, obwohl eigentlich keine vorhanden sind, dass der Schwellenwert überschritten und somit davon ausgegangen wird, dass keine Anomalien vorliegen.

Da die Werte der Fehlerraten für die drei betrachteten Wegaufnehmer so unterschiedlich ausfallen, ist nicht eindeutig feststellbar welche Methode die einheitlichsten Ergebnisse liefert.

6.2 Vergleich der Methoden anhand von manuell bestimmten Anomalien in ausgewählten Wochen

Zur Beurteilung der Qualität der Methoden zur Erkennung von Anomalien ist es notwendig zu wissen, wie viele Anomalien tatsächlich vorhanden sind und zu welchen Zeiten diese genau auftreten. Zu diesem Zweck wurden die trendbereinigten Zeitreihen aus drei Wochen tageweise von zwei unabhängigen Personen betrachtet und die Anfangs- bzw. Endpunkte der Anomalien durch hinsehen bestimmt. Die betrachteten Zeiträume sind der 03.08. bis 09.08.16, für den Wegaufnehmer WOS2, der 01.01. bis 07.01.17 für den Wegaufnehmer WOS2 und der 22.02. bis 28.02.17 für den Wegaufnehmer WWS4.
Zur Berechnung der Fehlerraten wird für jede Methode und jede Zählung ein binärer Vektor, dessen Länge N dem betrachteten Zeitraum entspricht, erstellt. Liegt ein Zeitpunkt innerhalb einer, durch die jeweilige Methode/Zählung identifizierten, Anomalie, erhält der entsprechende Eintrag des Vektors eine 1. Ansonsten enthält er eine 0. Die Vektoren der Methoden (y_m) werden anschließend mit den Vektoren der Zählungen (y_z) verglichen. Indem die Anzahl der Abweichungen zwischen den Vektoren der Methoden und den Vektoren der Zählungen durch die Gesamtanzahl der Beobachtungen geteilt wird, ergibt sich die Fehlerrate. Dieses Vorgehen kann durch

$$\frac{\sum_{n=1}^{N} \mathbb{1}(y_{m,n} \neq y_{z,n})}{N},$$

mit $m \in \{\mathrm{VCP}, \mathrm{MVCP}, \mathrm{CMAD}\}$ und $z \in \{\mathrm{Anne}, \mathrm{Johanna}\}$, als Formel ausgedrückt werden. Für jede Methode und jeden betrachteten Zeitraum lassen sich zwei Fehlerraten berechnen. Zum einen die Fehlerrate im Vergleich zu der Zählung von Anne und zum anderen die Fehlerrate im Vergleich zu der Zählung von Johanna. Zudem wird die Fehlerrate zwischen den beiden Zählungen bestimmt. Dies dient dazu, die Qualität der Methoden besser einschätzen zu können. Sind die Unterschiede zwischen einer Methode und den Zählungen genauso groß oder geringer als die Unterschiede zwischen den Zählungen, so hat die Methode ein gutes Ergebnis erzielt. Somit kann die Fehlerrate zwischen den Zählungen als Vergleichswert bezeichnet werden.
In Kapitel 6.2.1 werden die Fehlerraten für die einzelnen Tage, an denen Anomalien auftreten, genauer betrachtet. Für die Bestimmung der tageweisen Fehlerraten wird ein Schwellenwert genutzt. Wird dieser Schwellenwert durch die Anzahl der gefundenen Anomalien überschritten, wird davon ausgegangen, dass an dem entsprechenden Tag keine

Anomalien vorliegen. Alle Einträge des Vektors zur Bestimmung der Fehlerrate erhalten dann den Wert 0. Dieses Vorgehen ist sinnvoll, wenn nicht bekannt ist, ob an einem Tag eine Anomalie vorliegt, da die Methoden, insbesondere VCP und MVCP, nicht damit umgehen können, wenn innerhalb des betrachteten Zeitraums keine Anomalie vorliegt. Der Schwellenwert ergibt sich aus der maximalen, an einem Tag vorkommenden Anzahl an Anomalien multipliziert mit dem Faktor 1.5. Dieser Faktor wird genutzt um Spielraum nach oben zu schaffen, falls zusätzlich zu den tatsächlichen Anomalien noch weitere Bereiche als Anomalie klassifiziert werden. In dem betrachteten Zeitraum vom 01.06.16 bis zum 31.05.17 treten maximal 8 Anomalien an einem Tag auf. Dies ist der Fall am 15.06.16 am Wegaufnehmer WOS2. Somit ergibt sich aus $8 * 1.5$ der Schwellenwert 12.
Um die Fehlerraten pro Woche geht es in Kapitel 6.2.2. Dabei wird kein Schwellenwert verwendet, da für die betrachten Wochen bekannt ist, dass sie Anomalien enthalten.

In Tabelle A.2 im Anhang sind die Zeiträume der Anomalien des Wegaufnehmers WOS2 in der Woche vom 03.08. bis zum 09.08.16 festgehalten. Die Bereiche der Anomalien wurden von beiden Personen sehr ähnlich bestimmt. Beide Supervisoren haben innerhalb dieser Woche 4 Anomalien an 2 Tagen ermittelt. Der größte Unterschied besteht in einer Abweichung von 8 Minuten am 05.08.2016.
Da sich die Rissbreiten im Sommer und im Winter aufgrund der unterschiedlichen Temperatur stark unterscheiden, sind in Tabelle A.3 im Anhang die Zeiträume der Anomalien des Wegaufnehmers WOS2 in der Woche vom 01.01. bis zum 07.01.17 dargestellt. Auch hier sind keine großen Abweichungen zwischen den identifizierten Anomalien der beiden Personen vorhanden. Beide Supervisoren haben innerhalb dieser Woche 5 Anomalien an 5 Tagen festgestellt. Der größte Unterschied ist eine Differenz von 10 Minuten am 05.01.17. In beiden Tabellen wurde der Wegaufnehmer WOS2 betrachtet, welcher an der Straßenbahnspur angebracht ist. Bei diesem Wegaufnehmer sind die Ausschläge deutlich gleichmäßiger als bei Wegaufnehmern an Fahrbahnen die von Autos befahren werden, wodurch Anomalien optisch leichter erkennbar sind.
Aus diesem Grund sind in Tabelle A.4 im Anhang die Zeiträume der Anomalien des Wegaufnehmers WWS4 in der Woche vom 22.02. bis zum 28.02.17 aufgeführt. Hier treten deutlich größere Unterschiede zwischen den Zählungen der beiden Personen auf. Von Anne wurden innerhalb dieser Woche 9 Anomalien an 4 Tagen und von Johanna wurden

11 Anomalien an 4 Tagen festgestellt. Die größten Abweichungen zwischen den Ergebnissen der beiden Personen treten am 22.02.17 auf. Dort gibt es 2 Zeiträume, die von der einen Person als Anomalie gewertet werden, während für die andere Person an diesen Stellen keine Anomalie erkennbar ist. Auch am 27.02.17 tritt eine Abweichung von mehr als einer halben Stunde zwischen den beiden Auswertungen auf. Daraus lässt sich schließen, dass die genauen Anfangs- und Endpunkte anomaler Sequenzen auch durch menschliche Supervisoren nicht eindeutig bestimmt werden können. Dies gilt insbesondere für Beobachtungszeiträume zu denen kalte Temperaturen herrschen.

6.2.1 Fehlerraten der einzelnen Tage an denen Anomalien auftreten

In Tabelle A.5 im Anhang sind die Fehlerraten der drei Methoden im Vergleich zur optischen Identifikation der Anomalien aus den Tabellen A.2 und A.3 dargestellt. Tabelle A.6 im Anhang zeigt die entsprechenden Fehlerraten in der Woche vom 22.02. bis zum 28.02.17 für den Wegaufnehmer WWS4. Dabei sind nur die Tage enthalten an denen optisch eine Anomalie festgestellt werden konnten. Abbildung 12 stellt die Ergebnisse aus den Tabellen A.5 und A.6, zur besseren Übersichtlichkeit, als Barplot dar.
Bei den betrachteten Tagen im August ist auffällig, dass die VCP und MVCP Methode bei der Zählung durch Anne besser abschneidet, während die CMAD Methode bei der Zählung durch Johanna einen deutlich geringeren Wert erzielt. Der jeweils niedrigere Wert liegt dabei sehr nah an, und im Fall von CMAD sogar deutlich unter, der Rate für den Unterschied zwischen den beiden Zählungen. An den Tagen im Winter an denen Anomalien auftreten, bis auf den 04.01.17, haben die Fehlerraten für CMAD einen höheren Wert als für die anderen Methoden und für den Vergleichswert. Zudem sind die folgenden Auffälligkeiten in den Fehlerraten erkennbar:

- Am 01.01.17,am 04.01.17 und am 05.01.17 hat der Vergleichswert einen deutlich geringeren Wert als die Fehlerraten der Methoden.
- Am 04.01.17 treten für alle Methoden die gleichen Fehlerraten auf.
- Am 05.01.17 ist die außergewöhnlich hohe Fehlerrate der CMAD Methode auffällig.
- Am 07.01.17 haben VCP und MVCP im Vergleich zur Zählung von Anne eine sehr geringe Fehlerrate und im Vergleich zur Zählung von Johanna, eine Fehlerrate

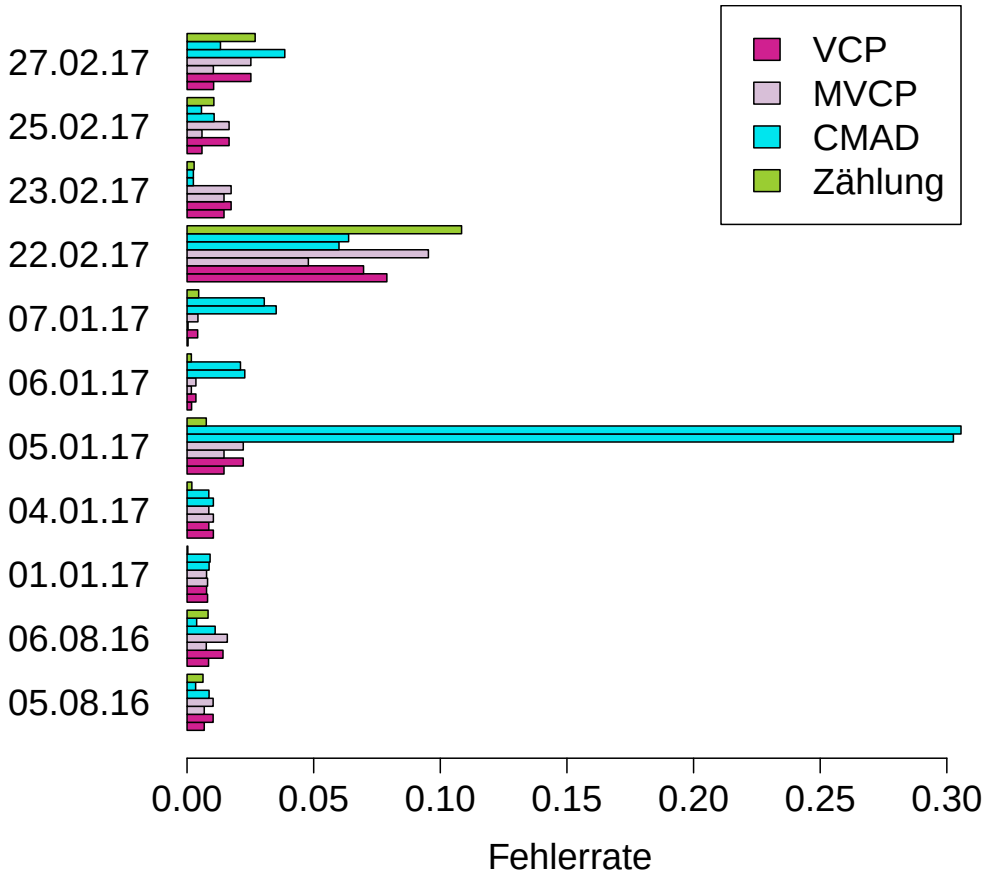


Abbildung 12: Barplot über die Fehlerraten pro Tag, beim Vergleich der Methoden mit der optischen Identifikation der Anomalien. Zudem ist der Vergleich der beiden Zählungen der Fehlerraten untereinander dargestellt. Für jede Methode sind pro Tag jeweils zwei Balken abgebildet. Der untere Balken zeigt dabei die Fehlerrate für den Vergleich mit der Zählung von Anne und der obere Balken die Fehlerrate für den Vergleich mit der Zählung von Johanna. Es sind nur Tage abgebildet an denen Anomalien vorliegen.

die in etwa dem Vergleichswert entspricht. CMAD weist im Gegensatz dazu einen deutlich höheren Wert auf.

- Am 22.02.17 ist der Vergleichswert deutlich höher als die Fehlerraten der Methoden.

- Am 23.02.17 erzielt die CMAD Methode einen deutlich geringeren Wert als VCP und MVCP.

- Am 25. und 27.02.17 schneiden VCP und MVCP im Vergleich zu Annes Zählung besser ab, während CMAD näher an Johannas Zählung liegt.

Aufgrund dieser Auffälligkeiten werden der 04.01.17, der 05.01.17 und der 07.01.17 für den Wegaufnehmer WOS2 und der 23.02.17 für den Wegaufnehmer WWS4 genauer betrachtet. Der Unterschied zwischen den Zählungen kommt dadurch Zustande, dass Johanna Bereiche als Anomalien wahrgenommen hat, die für Anne keine Anomalien darstellen (siehe Tabelle A.4).
In Abbildung 13 ist ein Ausschnitt der trendbereinigten Zeitreihe am 04.01.17, gemessen an Wegaufnehmer WOS2, zu sehen. Die einzige optisch identifizierte Anomalie an diesem

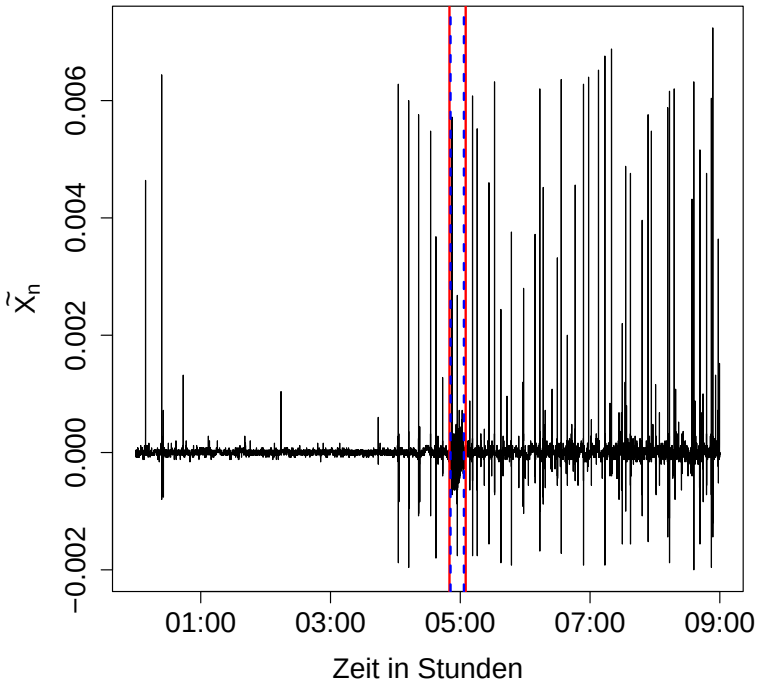


Abbildung 13: Ausschnitt vom 04.01.2017 für den Wegaufnehmer WOS2. Der Verlauf des restlichen Tages ähnelt dem Abschnitt von 07:00 bis 09:00 Uhr. Der von Anne optisch als Anomalie identifizierte Bereich ist von roten Linien eingefasst, während der von Johanna optisch als Anomalie identifizierte Bereich innerhalb der blau gestrichelten Linien liegt.

Tag ist in dem Ausschnitt enthalten. Dabei markieren die roten Linien den von Anne als Anomalie wahrgenommenen Bereich und die blau gestrichelten Linien umschließen den von Johanna als Anomalie erkannten Bereich. Alle drei Methoden weisen hier die gleiche

Fehlerrate auf, da sie insgesamt mehr als 12 Bereiche als Anomalien klassifizieren und somit davon ausgegangen wird, dass an diesem Tag keine Anomalien vorliegen.
In Abbildung 14 ist die trendbereinigte Zeitreihe, gemessen an Wegaufnehmer WOS2 am 05.01.17, dargestellt. Die optisch identifizierten Anomalien sind auch hier wie oben be-

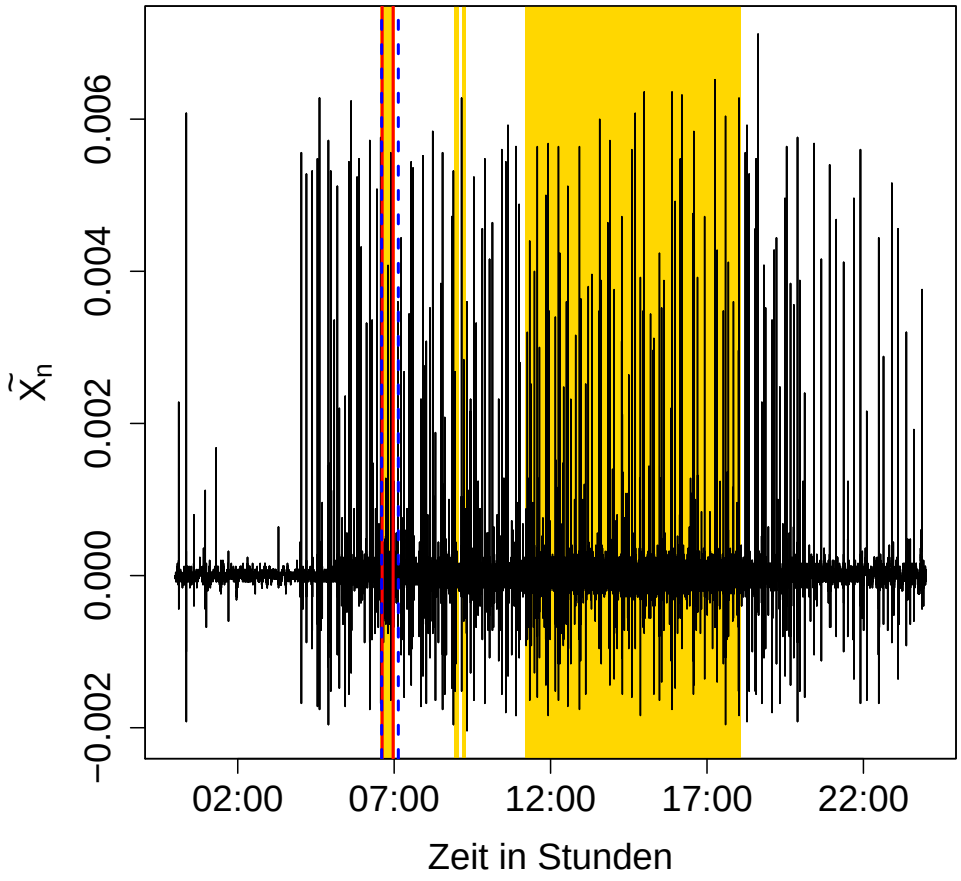


Abbildung 14: Trendbereinigte Zeitreihe am 05.01.2017 für den Wegaufnehmer WOS2. Der von Anne optisch als Anomalie identifizierte Bereich ist von roten Linien eingefasst, während der von Johanna optisch als Anomalie identifizierte Bereich innerhalb der blau gestrichelten Linien liegt. Die gelb hinterlegten Abschnitte wurden von der CMAD Methode als Anomalie klassifiziert.

schrieben markiert. Die gelb hinterlegten Bereiche wurden von der CMAD Methode als Anomalien ausgegeben. VCP und MVCP haben jeweils mehr als 12 Anomalien identifiziert, weshalb davon ausgegangen wird, dass an diesem Tag keine Anomalien vorliegen. Die sehr hohe Fehlerrate von CMAD ist vor allem darauf zurück zu führen, dass die

Methode zwischen 11 und 18 Uhr mehrere Stunden als Anomalie bezeichnet, in denen optisch keine Auffälligkeiten erkennbar sind.

In Abbildung 15 ist die trendbereinigte Zeitreihe, gemessen an Wegaufnehmer WOS2 am 07.01.17, dargestellt. Die optisch identifizierten Anomalien sind wie in den vorheri-

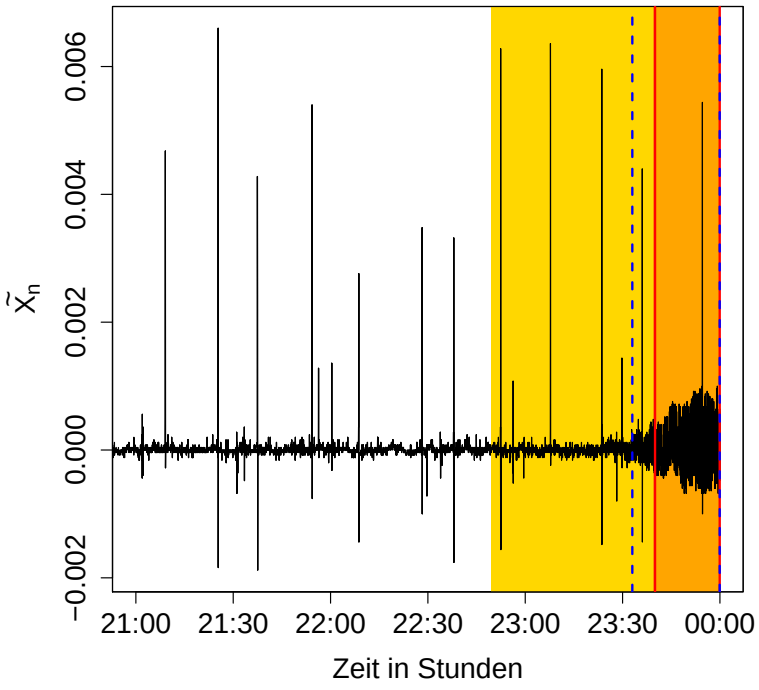


Abbildung 15: Ausschnitt vom 07.01.2017 für den Wegaufnehmer WOS2. Der Verlauf des restlichen Tages ähnelt dem Abschnitt von 21:00 bis 23:00 Uhr. Der von Anne optisch als Anomalie identifizierte Bereich ist von roten Linien eingefasst, während der von Johanna optisch als Anomalie identifizierte Bereich innerhalb der blau gestrichelten Linien liegt. Der orange hinterlegte Abschnitt wurde von allen drei Methoden als Anomalie erkannt, während der gelb hinterlegte Abschnitt nur von der CMAD Methode als Anomalie bezeichnet wird.

gen beiden Beispielen gekennzeichnet. Der orange hinterlegte Bereich wurde von allen drei Methoden als Anomalie erkannt, während der gelb hinterlegte Bereich nur von der CMAD Methode als Anomalie wahrgenommen wird. Dies führt zu einer höheren Fehlerrate bei der CMAD Methode, da der von dieser Methode als Anomalie bezeichnete Bereich deutlich vor der optisch ersichtlichen Auffälligkeit beginnt.

In Abbildung 16 ist ein Ausschnitt der trendbereinigten Zeitreihe am 23.02.17, gemessen

an Wegaufnehmer WWS4, zu sehen. Die farbigen Markierungen entsprechen dabei denen

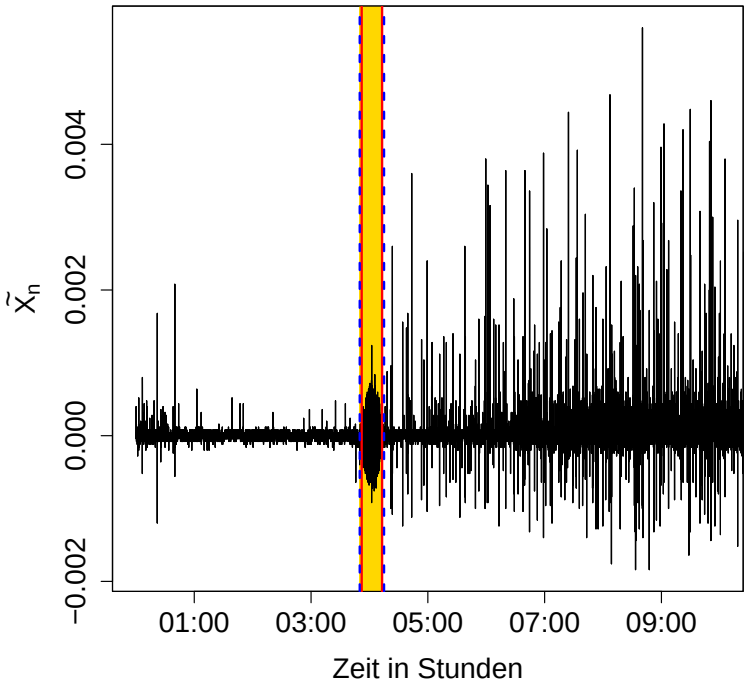


Abbildung 16: Ausschnitt vom 23.02.2017 für den Wegaufnehmer WWS4. Der Verlauf des restlichen Tages ähnelt dem Abschnitt von 07:00 bis 10:00 Uhr. Der von Anne optisch als Anomalie identifizierte Bereich ist von roten Linien eingefasst, während der von Johanna optisch als Anomalie identifizierte Bereich innerhalb der blau gestrichelten Linien liegt. Der gelb hinterlegte Abschnitt wurde von der CMAD Methode als Anomalie erkannt.

aus Grafik 14. Auch hier haben die VCP und die MVCP Methode mehr als 12 Bereiche als Anomalie klassifiziert, was zu der Annahme führt, dass an diesem Tag keine Anomalien vorliegen. Die CMAD Methode findet dagegen die einzige optisch erkennbare Anomalie an diesem Tag.

6.2.2 Fehlerraten pro Woche

Tabelle A.7 im Anhang zeigt die Fehlerraten der drei Methoden, wenn diese jeweils auf eine gesamte Woche angewendet werden. Zur besseren Übersicht sind die Fehlerraten in Abbildung 17 als Barplot dargestellt.
In der Woche vom 03.08. bis zum 09.08.16 haben VCP und MVCP im Vergleich zur

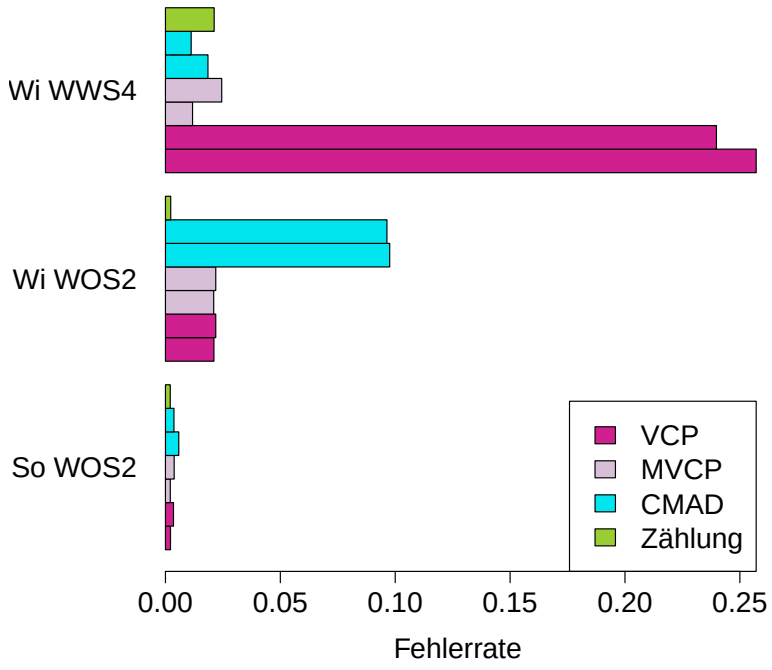


Abbildung 17: Barplot über die Fehlerraten pro Woche, beim Vergleich der Methoden mit der optischen Identifikation der Anomalien. Zudem ist der Vergleich der beiden Zählungen der Fehlerraten untereinander dargestellt. Für jede Methode sind pro Tag jeweils zwei Balken abgebildet. Der untere Balken zeigt dabei die Fehlerrate für den Vergleich mit der Zählung von Anne und der obere Balken die Fehlerrate für den Vergleich mit der Zählung von Johanna. **So WOS2** steht dabei für *Sommer WOS2* und bezeichnet die Woche vom 03.08. bis zum 09.08.16, **Wi WOS2** steht für *Winter WOS2* und bezeichnet die Woche vom 01.01. bis zum 07.01.17 und **Wi WWS4** steht für *Winter WWS4* und bezeichnet die Woche vom 22.02. bis zum 28.02.17.

Zählung von Anne deutlich niedrigere Fehlerraten als CMAD. Die Fehlerraten von VCP und MVCP entsprechen dabei in etwa dem Vergleichswert. Im Vergleich zur Zählung von Johanna haben alle drei Methoden ähnliche Fehlerraten, die über dem Vergleichswert liegen. In der Woche vom 01.01. bis zum 07.01.17 liegen die Fehlerraten aller drei Methoden deutlich über dem Vergleichswert. CMAD weißt im Vergleich zu beiden Zählungen deutlich höhere Fehlerraten auf als VCP und MVCP. Für die Woche vom 22.02. bis zum 28.02.17 sind die geringsten Fehlerraten bei CMAD zu finden. Für beide Zählungen liegen die Werte unter dem Vergleichswert. Die höchsten Fehlerraten weißt hier die VCP

Methode auf. Dies ist darauf zurückzuführen, dass diese Methode in dieser Woche sehr viele Bereiche als Anomalie klassifiziert. Darunter fallen auch Bereiche in denen optisch keine Anomalie erkennbar ist. Die Fehlerraten der MVCP Methode wurden im Vergleich zur VCP Methode deutlich verbessert. Im Vergleich zur Zählung von Anne liegt die Fehlerrate deutlich unter dem Vergleichswert, im Vergleich zur Zählung von Johanna nur knapp darüber. Daraus lässt sich schließen, dass MVCP gegenüber VCP im Vorteil ist, wenn die Anomalien schwerer zu erkennen sind, da durch VCP falsch als Anomalie klassifizierte Bereiche oftmals wegfallen können.

6.3 Klassifikation bezüglich des Vorliegens einer Anomalie an einem Tag

Wie bereits erwähnt, haben die VCP und die MVCP Methode häufig Probleme, wenn innerhalb der betrachteten Zeitreihe keine Anomalien auftauchen. In diesem Fall werden dann oftmals sehr viele kurze Bereiche als Anomalie klassifiziert. Um zu bestimmen, wie oft richtig erkannt wird, ob an einem Tag eine bzw. keine Anomalie vorliegt, werden für den Zeitraum vom 01.06.16 bis zum 31.05.17 entsprechende Fehlerraten berechnet. Zur Berechnung der Fehlerraten wurde für das gesamte Jahr optisch festgestellt, ob an den entsprechenden Tagen Anomalien vorliegen. Liegt an einem Tag eine Anomalie vor, wird in einem binären Vektor der Länge 365 an der entsprechenden Stelle der Wert 1 eingetragen. Wenn keine Anomalie vorliegt erhält der entsprechende Eintrag den Wert 0. Die optische Identifikation der Tage mit Anomalien wurde im Zeitraum vom 01.06.16 bis zum 28.02.17, im Rahmen der Veranstaltung Fallstudien 2, von Johanna vorgenommen. Für den Zeitraum vom 01.03.17 bis zum 31.05.17 ist die optische Identfikation der Anomalien von Anne durchgeführt worden.
Für die Ergebnisse der VCP und der MVCP Methode werden entsprechende Vektoren auf die gleiche Art erstellt. Die Anzahl der Tage an denen der Vektor der betrachteten Methode mit dem Vektor der optischen Betrachtung nicht übereinstimmt wird aufsummiert, und anschließend durch die Gesamtanzahl der Tage dividiert. Dabei wird einmal der Fall betrachtet, dass die Methode auf eine Woche angewendet wird, wobei dann nur die an dem mittleren Tag identifizierten Anomalien in die Auswertung eingehen. Dadurch werden insgesamt nur 359 anstatt 365 Tage betrachtet, da am Anfang und am Ende jeweils 3

Tage wegfallen, da diese nicht als mittlerer Tag aus 7 Tagen verwendet werden können. Dieses Vorgehen wird in der Tabelle als *Woche* bezeichnet. Bei den, der Auswertungsmethode *Tag* zugeordneten, Fehlerraten werden die Methoden jeweils auf einen einzelnen Tag angewendet. Dabei werden alle 365 Tage innerhalb des betrachteten Zeitraum verwendet. Die so berechneten Fehlerraten sind in Tabelle A.8 im Anhang aufgeführt und zur besseren Übersicht in Abbildung 18 dargestellt.

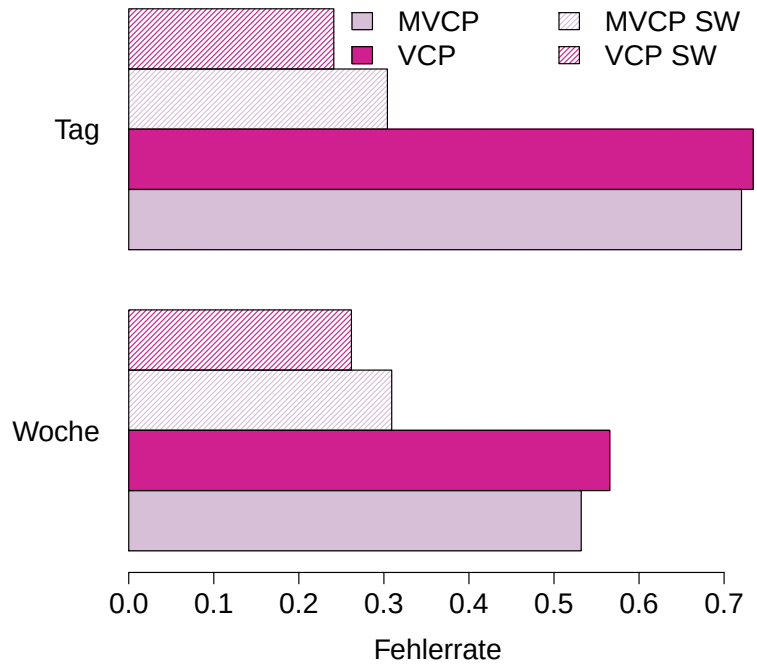


Abbildung 18: Barplot über die Fehlerraten für das Vorliegen einer Anomalie an einem Tag im Zeitraum vom 01.06.16 bis zum 31.05.17. *Tag* bedeutet, dass die Methoden jeweils tageweise angewendet wurden, bei *Woche* wurden die Methoden jeweils auf 7 Tage angewendet wobei, die Anomalien des jeweils mittleren Tages ausgewertet werden. Bei den schraffierten Balken ist ein Schwellenwert (SW) von 12 verwendet worden. An Tagen, an denen mehr Anomalien identifiziert werden, wird davon ausgegangen, dass keine Anomalie vorliegt. Bei den ausgefüllten Balken wurde kein Schwellenwert verwendet.

Wird kein Schwellenwert zur Hilfe genommen, sind die Fehlerraten insgesamt sehr hoch. Die Betrachtung des mittleren Tages eines 7-Tage Abschnittes schneidet dabei für beide Methoden deutlich besser ab, als die Betrachtung der einzelnen Tage. Für beide Betrachtungsweisen ist die Fehlerrate der MVCP Methode geringfügig niedriger als die Fehler-

rate der VCP Methode.

Wird ein Schwellenwert hinzugenommen, können die Fehlerraten für beide Methoden und beide Auswertungsweisen deutlich verbessert werden. Hierbei hat die VCP Methode deutlich geringere Fehlerraten als die MVCP Methode und die Auswertung der einzelnen Tage schneidet geringfügig besser ab, als wenn der mittlere Tag eines 7-Tage Abschnittes betrachtet wird. Die Fehlerrate mit dem geringsten Wert liegt dann bei 0.2411.

6.3.1 Verbesserung der Zuordnung bezüglich des Vorliegens einer Anomalie an einem Tag mit Hilfe der logistischen Regression

Es wird versucht, mit Hilfe der logistischen Regression eine bessere Zuordnung zu erreichen. Dazu wird ein Modell aufgestellt, in dem die Zielvariable **Anomalie** durch die Variablen **max_temp**(maximale, im betrachteten Zeitraum vorkommende Temperatur), **min_temp** (minimale, im betrachteten Zeitraum vorkommende Temperatur), **anz_schweb_MVCP_t** (Anzahl der mit der MVCP Methode gefundenen Anomalien) und **anz_schweb_VCP_t** (Anzahl der mit der VCP Methode gefundenen Anomalien), erklärt wird. Der zugehörige $\mathrm{logit}(\pi)$ hat die Form:

$$\begin{aligned}\mathrm{logit}(\pi) = \alpha + \beta_1 \,\mathrm{max\_temp} + \beta_2 \,\mathrm{min\_temp} + \beta_3 \,\mathrm{anz\_schweb\_MVCP\_t} \\ + \beta_4 \,\mathrm{anz\_schweb\_VCP\_t},\end{aligned} \tag{24}$$

wobei π der Wahrscheinlichkeit, dass eine Anomalie vorliegt, entspricht. Wird das Modell auf allen 365 Tagen angepasst, ergibt sich für die Modellkoeffizienten die folgende Schätzung:

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        0.38867    0.37614   1.033 0.301448
max_temp           0.22412    0.06801   3.295 0.000983 ***
min_temp          -0.25783    0.07697  -3.350 0.000809 ***
anz_schweb_MVCP_t -0.07607    0.03373  -2.255 0.024135 *
anz_schweb_VCP_t  -0.11033    0.03061  -3.605 0.000312 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Die wahre Fehlerrate wird dabei mit Hilfe der Leave-one-out Kreuzvalidierung approximiert um eine Überanpassung an die gegebenen Daten zu vermeiden. Die Leave-one-out Fehlerrate der logistischen Regression liegt bei 0.1781, wenn ab einer Wahrscheinlichkeit von mehr als 50 % davon ausgegangen wird, dass eine Anomalie vorliegt. Somit wird durch eine Vorentscheidung mit Hilfe eines logistischen Regressionsmodell eine Verbesserung erzielt.

Da die Anomalien nach der Identifikation aus den Daten entfernt werden sollen, um eine Verfälschung der Verkehrsschätzung zu vermeiden, ist es wichtiger, dass Tage, an denen Anomalien vorliegen richtig klassifiziert werden, als dass Tage ohne Anomalien erkannt werden. In Tabelle A.9 im Anhang sind daher die Fehlerraten der Klassifikation getrennt für den Fall, dass eine Anomalie vorliegt und, dass keine Anomalie vorliegt, aufgeführt. Bei der Vorentscheidung durch die Logistische Regression wird die wahre Fehlerrate wieder mit der Leave-One-Out Methode approximiert. Dabei wird das vorliegen einer Anomalie vorhergesagt, wenn die Wahrscheinlichkeit für das Eintreten einer solchen bei über 50 % liegt. Um die Fehlerraten für den Fall, dass keine Anomalie klassifiziert wird, obwohl eine vorhanden ist, zu berechnen, wird die Anzahl der entsprechenden Fälle durch die Gesamtanzahl der Tage, an denen eine Anomalie vorliegt geteilt. Die Berechnung der Fehlerraten für den Fall, dass eine Anomalie klassifiziert wird, obwohl keine vorhanden ist, erfolgt analog. Zur besseren Übersicht sind die Fehlerraten aus Tabelle A.9 im Anhang in Abbildung 19 als Barplot dargestellt.
Dabei lässt sich erkennen, dass die verhältnismäßig niedrige Fehlerrate der logistischen Regression hauptsächlich darauf beruht, dass hier fast immer richtig klassifiziert wird, wenn keine Anomalie vorliegt, während die Fehlerrate für den Fall, dass eine Anomalie vorliegt bei über 40 % liegt. Da dies in Bezug auf eine Verkehrsschätzung der wichtigere Wert ist, scheint eine Vorentscheidung mit Hilfe der logistischen Regression doch nicht optimal zu sein. Wird die VCP Methode ohne Schwellenwert verwendet, wird jeden Tag mindestens eine Anomalie entdeckt. Dies führt dazu, dass wenn eine Anomalie vorliegt, immer richtig, und wenn keine Anomalie vorliegt, immer falsch klassifiziert wird. Die Verwendung der MVCP Methode ohne Schwellenwert weist ein ähnliches Verhalten auf. Wird ein Schwellenwert von 12 (maximale Anzahl Anomalien pro Tag $*$ 1.5) verwendet, klassifiziert die MVCP Methode ca. 14 % und die VCP Methode ca. 20 % der Fäl-

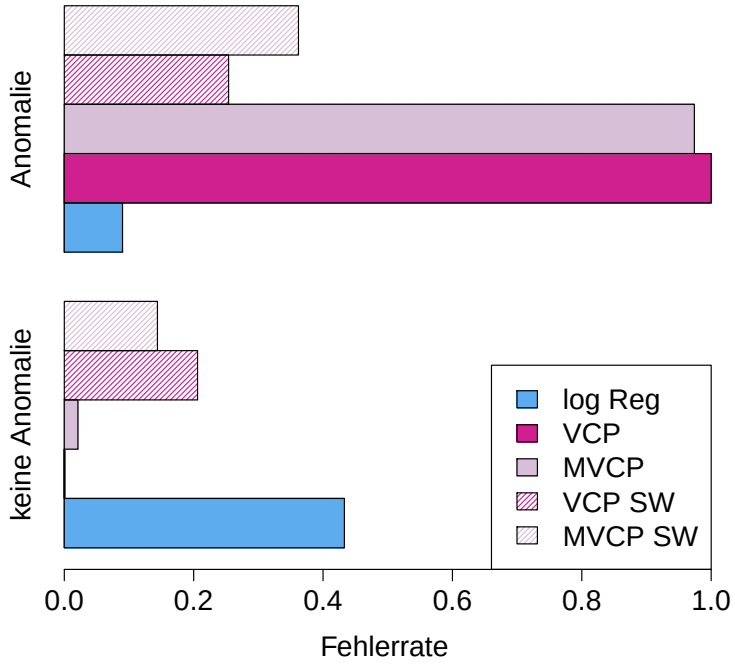
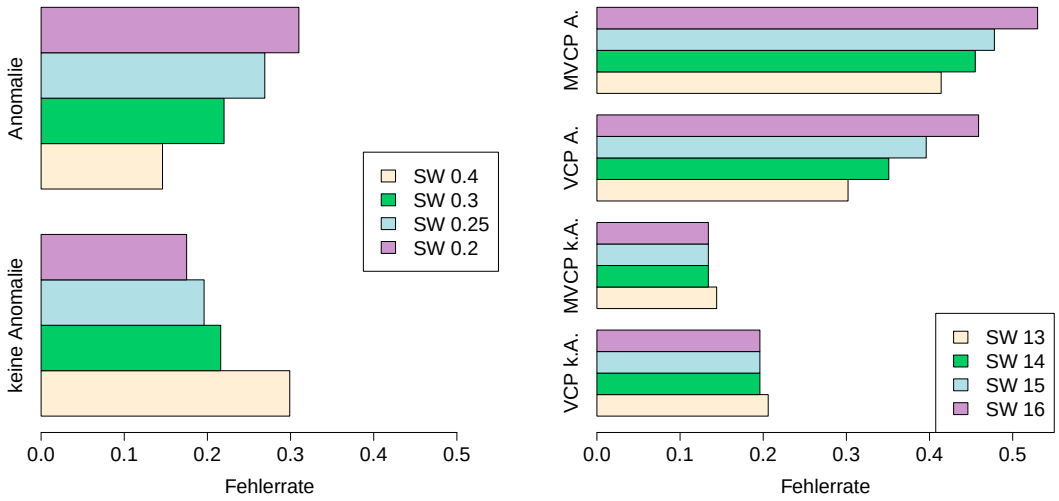


Abbildung 19: Barplot über die Fehlerraten für das Vorliegen einer Anomalie an einem Tag im Zeitraum vom 01.06.16 bis zum 31.05.17, getrennt für den Fall, dass eine Anomalie vorliegt und, dass keine Anomalie vorliegt. Dabei sind die Methoden jeweils tageweise angewendet worden. *Keine Anomalie* bedeutet, dass keine Anomalie klassifiziert wird, obwohl eine Anomalie vorliegt. *Anomalie* steht für den Fall, dass eine Anomalie klassifiziert wird, obwohl keine Anomalie vorliegt. Bei den Schraffierten Balken ist ein Schwellenwert (SW) von 12 verwendet worden. An Tagen, an denen mehr Anomalien identifiziert werden, wird davon ausgegangen, dass keine Anomalie vorliegt. Bei den ausgefüllten Balken wurde kein Schwellenwert verwendet. Die Vorentscheidung mit Logistischer Regression klassifiziert eine Anomalie ab einer Wahrscheinlichkeit von 50 %.

le, in denen eine Anomalie vorliegt, falsch. Für den Fall, dass keine Anomalie vorliegt, liegt die Fehlklassifikationsrate für die MVCP Methode, bei Nutzung eines Schwellenwertes, bei ca. 36 % und die Fehlklassifikationsrate für die VCP Methode bei ca. 25 %. Die Ergebnisse aus Tabelle A.9 legen nahe, dass sowohl die VCP Methode, als auch die MVCP Methode in einigen Fällen mehr als 12 Anomalien identifizieren, wenn tatsächlich Anomalien vorliegen. Um die Ursache dieses Problems weiter einzugrenzen, wird untersucht, ob sich die Fehlerraten zwischen den verschiedenen Jahreszeiten stark unter-

scheiden. Zudem wird untersucht, ob durch Veränderung des Schwellenwertes oder das Absenken der Wahrscheinlichkeit, bei der das Eintreten des Ereignisses auf 1 gesetzt wird, eine Verbesserung der Fehlerrate, wenn eine Anomalie vorliegt, erzielt werden kann.
In Tabelle A.10 im Anhang sind die Leave-One-Out Fehlklassifikationsraten der logistischen Regression, bei Herabsetzung des Schwellenwertes, sowie die Fehlklassifikationsraten (nur in Bezug darauf, ob an einem Tag Anomalien vorliegen oder nicht) der VCP und der MVCP Methode, bei Erhöhung des Schwellenwertes, aufgeführt. Diese wurden getrennt für den Fall, dass eine Anomalie und für den Fall, dass keine Anomalie vorliegt, berechnet. Zur besseren Übersicht sind die entsprechenden Fehlerraten in Abbildung 20 als Barplot dargestellt.
Je geringer der Schwellenwert der logistischen Regression angesetzt ist, umso niedriger ist die Fehlklassifikationsrate für den Fall, dass eine Anomalie vorliegt. Die Fehlklassifikationsrate für den Fall, dass keine Anomalie vorliegt, verhält sich gegenläufig und steigt somit bei Verringerung des Schwellenwertes. Werden die Methoden ohne Vorentscheidung, aber mit Schwellenwert verwendet, steigt die Fehlklassifikationsrate für den Fall, dass keine Anomalie vorliegt, sowohl für die VCP als auch für die MVCP Methode an, je höher der Schwellenwert ist. Liegen Anomalien vor, gibt es für beide Methoden keine Änderungen der Fehlklassifikationsrate zwischen den Schwellenwerten 12 (siehe Tabelle A.9 im Anhang) und 13, sowie zwischen den Schwellenwerten 14, 15 und 16. Zwischen den Schwellenwerten 13 und 14 sinkt die Fehlklassifikationsrate für beide Methoden um 1 %. Für die Logistische Regression mit Schwellenwert 0.25 und die VCP Methode mit Schwellenwert 14, 15 und 16 ist die Fehlklassifikationsrate für den Fall, dass eine Anomalie vorliegt identisch und liegt knapp unter 20 %. Die Fehlklassifikationsraten für den Fall, dass keine Anomalien vorhanden sind, sind dabei jedoch nicht identisch. Bei der logistischen Regression erreicht diese einen Wert von $\approx$ 27 % und für die VCP Methode $\approx$ 35 % (wenn der Schwellenwert 14 verwendet wird). Daher ist die logistische Regression mit Schwellenwert 0.25 der VCP Methode mit Schwellenwert 14 vorzuziehen. Dies legt die Vermutung nahe, dass die Vorentscheidung mit Hilfe logistischer Regression, der VCP Methode mit Schwellenwert vorzuziehen ist. Den insgesamt niedrigsten Wert innerhalb der Tabelle für den Fall, dass Anomalien vorliegen, liegt bei $\approx$ 13 % und wird von der MVCP Methode mit Schwellenwert 14 erreicht. Die Fehlklassifikationsrate für den Fall, dass keine Anomalien vorliegen, liegt in diesem Fall allerdings bei $\approx$ 46 %.



(a) Logistische Regression

(b) VCP und MVCP

Abbildung 20: Barplots über die Fehlerraten bezüglich der Entscheidung, ob an einem Tag eine Anomalie vorliegt, bei Verwendung verschiedener Schwellenwerte. Betrachtet werden die Daten des Wegaufnehmers WWS4 im Zeitraum vom 01.06.16 bis zum 31.05.17. *Keine Anomalie (k.A.)* bezeichnet den Fall, dass keine Anomalie erkannt wird obwohl eine Anomalie vorliegt. *Anomalie (A.)* bezeichnet den Fall, dass fälschlicherweise eine Anomalie identifiziert wird. Der Schwellenwert bei der logistischen Regression bezeichnet die Wahrscheinlichkeit, die überschritten werden muss, damit an einem Tag Anomalien klassifiziert werden können. Bei der VCP und der MVCP Methode gibt der Schwellenwert die Anzahl an Anomalien an, die überschritten werden muss, damit davon ausgegangen wird, dass keine Anomalien vorliegen.
a) Leave-One-Out Fehlklassifikationsraten der logistischen Regression.
b) Fehlklassifikationsraten der VCP und MVCP Methode mit Schwellenwert.

In Tabelle A.11 im Anhang sind die Fehlklassifikationsraten der logistischen Regression und der Methoden mit Schwellenwert, nach Jahreszeiten getrennt, dargestellt. Der dabei verwendete Schwellenwert ist jeweils hinter der Methode angegeben. Für die logistische Regression werden dabei die Schwellenwerte 0.2, 0.25 und 0.3 betrachtet. Mit dem Schwellenwert 0.2 ist für den Fall, dass keine Anomalie klassifiziert wird, obwohl eine solche vorliegt, bei Betrachtung des gesamten Jahres, der geringste Wert der logistischen Regression erzielt worden. Für den Schwellenwert 0.3 sind die beiden betrachteten

Fehlerraten der logistischen Regression am geringsten und die Ergebnisse für den Schwellenwert 0.25 liegen zwischen den beiden anderen Fällen (siehe Tabelle A.10 im Anhang). Bei den Methoden wird dabei sowohl für VCP als auch für MVCP der Schwellenwert 12 verwendet, da bei größeren Schwellenwerten die Fehlklassifikationsrate für den Fall, dass eine Anomalie vorliegt, nur geringfügig verbessert wird, während die Fehlklassifikationsrate für den Fall, dass keine Anomalie vorliegt, stark steigt. Zur besseren Übersicht sind die Fehlerraten aus Tabelle A.11 im Anhang in Abbildung 21 als Barplot dargestellt.
Fast alle betrachteten Vorgehensweisen erreichen im Sommer, für den Fall, dass keine Anomalie klassifiziert wird (*keine Anomalie*), obwohl eine solche vorliegt, eine Fehlklassifikationsrate von 0 %. Lediglich die logistisches Regression mit dem Schwellenwert 0.3 erreicht hier einen Wert von $\approx$ 5 %. Dabei wird hier für den Fall, dass fälschlicherweise eine Anomalie klassifiziert wird (*Anomalie*), mit einer Leave-One-Out Fehlerrate von $\approx$ 6 % der geringste Wert erreicht. Im Winter werden in fast allen Fällen die höchsten Fehlklassifikationsraten erreicht. Lediglich bei der MVCP Methode, für den Fall *Anomalie*, wird im Herbst eine höhere Fehlklassifikationsrate erreicht als im Winter. Die Fehlerrate für den Fall, *keine Anomalie*, ist bei den betrachteten Vorgehensweisen in fast allen Fällen geringer als die Fehlerrate für den Fall *Anomalie*. Eine Ausnahme stellt dabei die VCP Methode im Frühling dar. Hier liegt die Fehlerrate, für den Fall *keine Anomalie* bei 25 %, während die Fehlklassifikationsrate für den Fall *Anomalie* bei $\approx$ 17 % liegt. Je nach Jahreszeit erzielen unterschiedliche Vorgehensweise das beste Ergebnis. Im Sommer ist die Vorentscheidung mit Logistischer Regression und einem Schwellenwert von 0.25 die beste Vorgehensweise, da hier der Fall *Anomalie* eine Leave-One-Out Fehlerrate von 0 % erreicht und die Fehlerrate für den Fall *Anomalie* mit $\approx$ 8 % einen sehr geringen Wert aufweist. Im Winter ist die Vorentscheidung mit logistischer Regression und einem Schwellenwert von 0.3 vorteilhafter, da hier für fast alle Vorgehensweise für den Fall *keine Anomalie* eine Fehlerrate von ca. 31 % erreicht wird. Dabei weist die logistische Regression für den Fall *Anomalie*, mit ca. 33 %, den geringsten Wert auf. Lediglich die Verwendung der MVCP Methode mit Schwellenwert 12 weist für den Fall *keine Anomalie* mit ca. 19 % eine geringere Fehlerrate auf. Dabei ist die Fehlerrate für den Fall *Anomalie* mit ca. 42 % jedoch deutlich zu hoch. Im Frühling und im Herbst ist die Wahl der besten Vorgehensweise nicht eindeutig. Für den Frühling sollte entweder die Vorentscheidung mit logistischer Regression und Schwellenwert 0.3 oder die MVCP Methode mit Schwel-

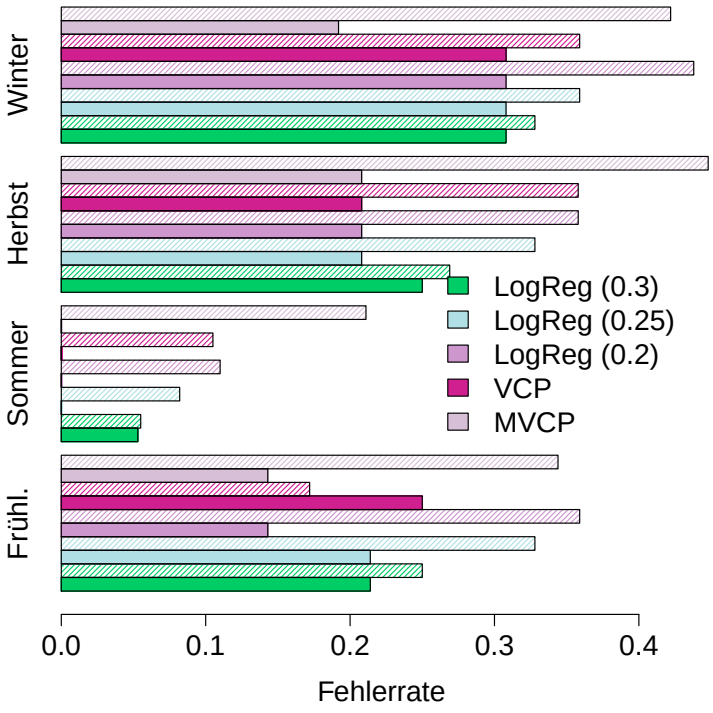


Abbildung 21: Barplot über die Fehlerraten für das Vorliegen einer Anomalie an einem Tag im Zeitraum vom 01.06.16 bis zum 31.05.17, nach Jahreszeiten getrennt. *Sommer* steht dabei für die Monate Juni, Juli und August, *Herbst* steht für September, Oktober und November, *Winter* steht für Dezember, Januar und Februar und *Frühling* steht für die Monate März, April und Mai. Die Methoden sind tageweise angewendet worden. Bei der VCP und bei der MVCP Methode wurde der Schwellenwert 12 verwendet. *LogReg* steht für logistische Regression und die Zahl in Klammern gibt die Wahrscheinlichkeit an, bei deren Überschreitung davon ausgegangen wird, dass eine Anomalie vorliegt. Ausgefüllte Balken stehen für den Fall, dass eine Anomalie vorliegt, die jeweilige Methode aber keine Anomalie klassifiziert. Schraffierte Balken stehen für den Fall, dass fälschlicherweise eine Anomalie klassifiziert wird.

lenwert 12 gewählt werden. Die erste der beiden Vorgehensweisen erzielt für die beiden Fälle mit ca. 21 % (*keine Anomalie*) und ca. 25 % (*Anomalie*) sehr ähnliche Fehlerraten. Die zweite Vorgehensweise erzielt für den Fall *keine Anomalie* mit ca. 14 % einen deutlich geringeren Wert, dafür wird für den Fall *Anomalie* mit ca. 34 % eine deutlich höhere Fehlklassifikationsrate erreicht. Für den Herbst ist entweder die Logistische Regression mit Schwellenwert 0.3 oder die Logistische Regression mit Schwellenwert 0.25 am vor-

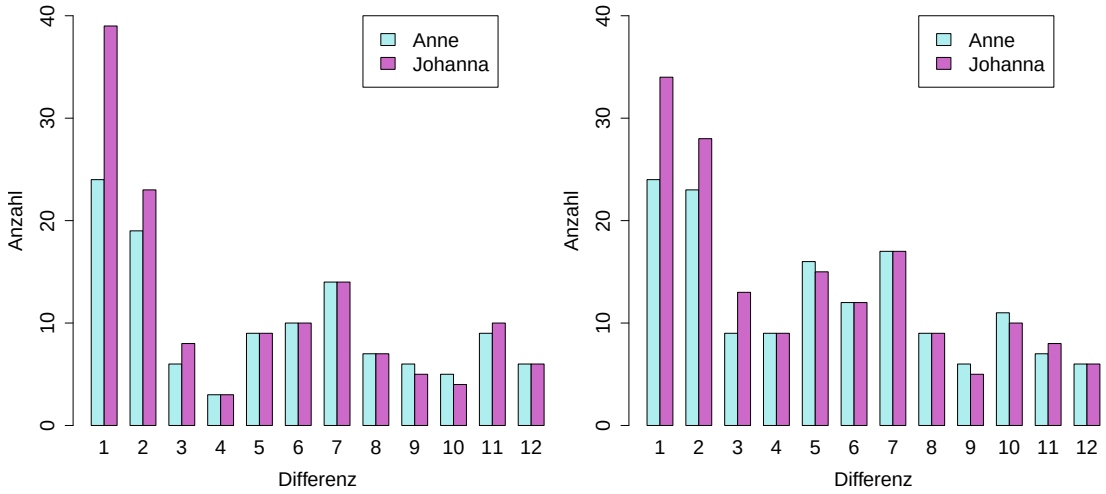
teilhaftesten. Die erste der beiden Vorgehensweisen hat für die beiden Fälle mit ca. 25 % (*keine Anomalie*) und ca. 27 % (*Anomalie*) sehr ähnliche Fehlklassifikationsraten. Für die Logistische Regression mit Schwellenwert 0.25 ist die Fehlklassifikationsrate für den Fall *keine Anomalie* mit ca. 21 % etwas geringer, während die Fehlklassifikationsrate für den Fall *Anomalie* mit ca. 33 % etwas höher ausfällt. Nur im Sommer treten bei einigen Methoden wünschenswerte Ergebnisse auf. Für die übrigen Jahreszeiten liegt deutliches Verbesserungspotential vor.

6.4 Klassifikation bezüglich der Anzahl Anomalien an einem Tag

Neben der richtigen Klassifikation von Tagen mit Anomalien ist es wichtig, dass die richtige Anzahl Anomalien klassifiziert wird. Um dies zu überprüfen werden die Methoden tageweise angewendet und die Anzahl der identifizierten Anomalien in einem Vektor gespeichert. Dabei wird der Zeitraum vom 01.06.16 bis zum 31.05.17, für den Wegaufnehmer WWS4, betrachtet. Zudem wird die Anzahl der Anomalien pro Tag, von zwei unabhängigen Supervisoren, optisch bestimmt. Anschließend wird die Differenz zwischen der optisch identifizierten Anzahl und der durch die jeweilige Methode bestimmten Anzahl ermittelt.

In Abbildung 22 sind die so berechneten Differenzen für die VCP und die MVCP Methode mit Schwellenwert 12 als Balkendiagramm dargestellt.

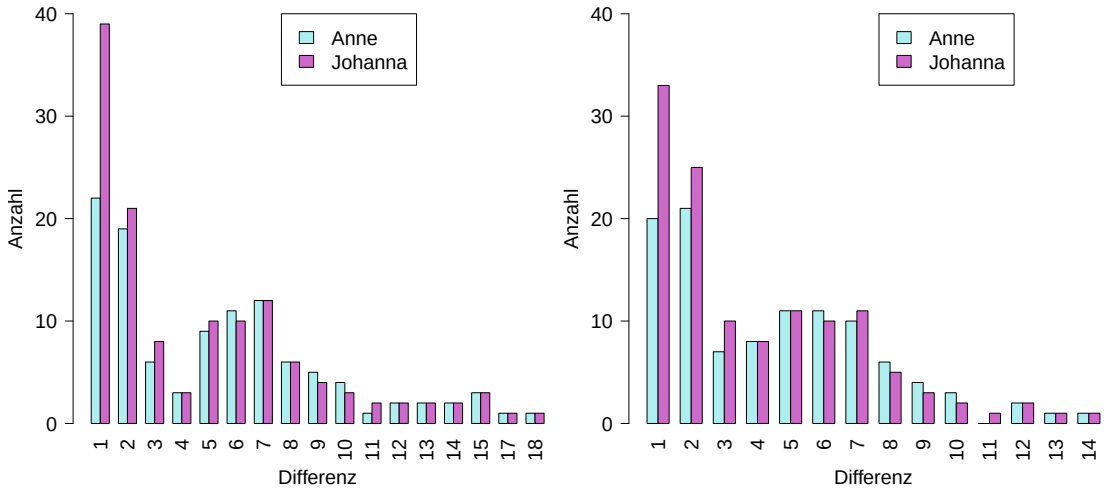
Eine Differenz von 0 wurde beim Vergleich der VCP Methode mit der Zählung von Anne 247 Mal und beim Vergleich mit der Zählung von Johanna 227 Mal erreicht. Bei der MVCP Methode wurde im Vergleich zur Zählung von Anne 216 Mal und im Vergleich zur Zählung von Johanna 199 Mal eine Differenz von 0 erreicht. Das Auftreten einer Differenz von 0 ist nicht in dem Balkendiagramm dargestellt, da dadurch die genauen Unterschiede der restlichen Balken zueinander schwerer zu erkennen sind. Es ist auffällig, dass eine Differenz von 1 im Vergleich zur Zählung von Johanna deutlich öfter auftaucht als im Vergleich zur Zählung von Anne. Dies gilt für beide Methoden, aber insbesondere für die MVCP Methode. Die VCP Methode schneidet im Vergleich zu beiden Zählungen etwas besser ab als die MVCP Methode. Dies kann darauf zurückgeführt werden, dass bei der MVCP Methode einige Anomalien wegfallen, die nur bei Anwendung auf die original Zeitreihe identifiziert werden. Wird dadurch der Schwellenwert von 12 unterschritten, werden durch die MVCP Methode an einem Tag Anomalien identifiziert, an dem optisch



(a) VCP (b) MVCP

Abbildung 22: Barplots über die Differenzen der Anzahl gefundener Anomalien der Methoden mit Schwellenwert 12 im Vergleich zur optisch bestimmten Anzahl an Anomalien durch Anne und Johanna. Die verwendeten Daten wurden an Wegaufnehmer WWS4 gemessen und liegen im Zeitraum vom 01.06.16 bis zum 31.05.17.
a) VCP Methode,
b) MVCP Methode.

keine solchen erkennbar sind. Im Gegensatz dazu überschreitet die VCP Methode den Schwellenwert, so dass richtigerweise angenommen wird, dass keine Anomalien vorliegen.
Zudem fällt auf, dass die Anzahl der jeweiligen Differenzen zwischen der VCP Methode und den Zählungen von 1 bis 4 sinkt und danach wieder deutlich ansteigt. Dies legt die Annahme nahe, dass bei dieser Methode geringe Differenzen auftreten, wenn tatsächlich Anomalien vorliegen und diese durch die Methode nicht oder nicht vollständig identifiziert werden bzw. geringfügig zu viele Anomalien identifiziert werden. Eine Differenz von 6 oder mehr taucht demnach dann auf, wenn keine Anomalie vorliegt, die VCP Methode aber dennoch weniger als 13 Anomalien identifiziert.
In Abbildung 23 sind Barplots über die Differenzen der Anzahl der Anomalien, die durch die Methoden und optisch durch Anne und Johanna identifiziert wurden, dargestellt. Dabei ist eine Vorentscheidung über das Vorliegen von Anomalien mit Hilfe der logistischen Regression mit Schwellenwert 0.3 durchgeführt worden. Das Modell der logistischen Re-



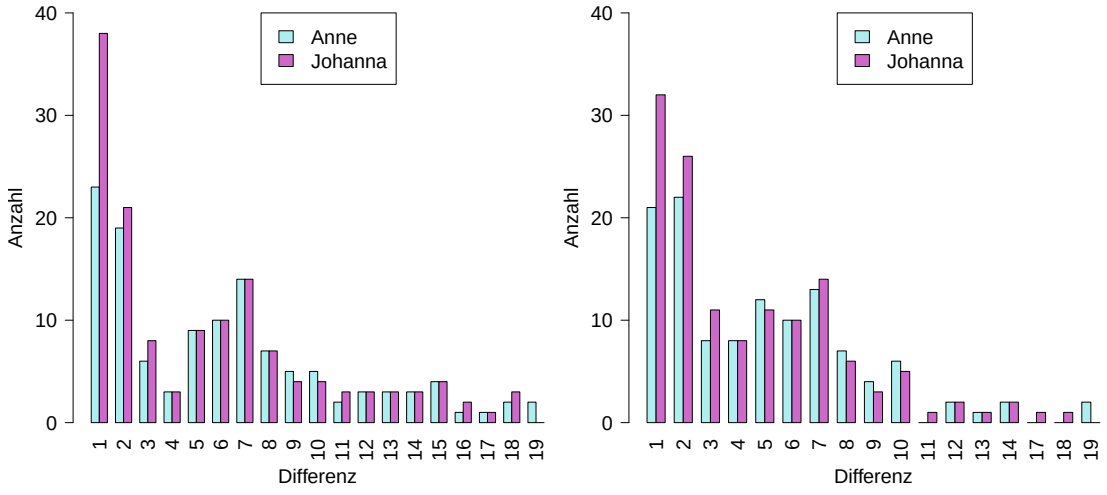
(a) VCP

(b) MVCP

Abbildung 23: Barplots über die Differenzen der Anzahl gefundener Anomalien der Methoden, bei einer Vorentscheidung, über das Vorliegen einer Anomalie an einem Tag, mit Hilfe von logistischer Regression mit Schwellenwert 0.3. Dabei wird das logistische Regressionsmodell aus Abschnitt 6.3 verwendet, welches auf den gesamten Daten des betrachten Zeitraums angepasst wurde. Die verwendeten Daten wurden an Wegaufnehmer WWS4 gemessen und liegen im Zeitraum vom 01.06.16 bis zum 31.05.17.
a) VCP Methode,
b) MVCP Methode.

gression ist dazu auf allen Daten des betrachteten Zeitraums (01.06.16 bis 31.05.17) angepasst worden. Eine Differenz von 0, in Bezug auf die Anzahl der Anomalien, liegt dabei für die VCP Methode im Vergleich zu Annes Zählung 256 Mal und im Vergleich zu Johannas Zählung 236 Mal vor. Für die MVCP Methode wurde 260 Mal im Vergleich zur Zählung von Anne und 242 Mal im Vergleich zur Zählung von Johanna eine Differenz von 0, bezüglich der Anzahl der Anomalien, erreicht. Es liegt somit für beide Methoden eine Verbesserung der Anzahl an Tagen, an denen die Anzahl der Anomalien richtig erkannt wird, im Vergleich zur Verwendung der Methoden mit Schwellenwert 12, vor. Diese Verbesserung ist lediglich auf die Anzahl richtig klassifizierter Tage, in Bezug auf das Vorliegen einer Anomalie, zurückzuführen, da die gleichen Methoden und die gleichen Daten wie in Abbildung 22 verwendet werden. Dafür gibt es, bei Verwendung der Vorentscheidung mit logistischer Regression mit Schwellenwert 0.3, Fälle in denen die

Differenz der Anzahl der durch die Methoden identifizierten Anomalien und der optisch identifizierten Anomalien, den Wert 12 übersteigt. Die größte Differenz hat einen Wert von 18 und tritt bei Verwendung der VCP Methode auf. Insgesamt tauchen bei Verwendung der VCP Methode 9 Fälle und bei Verwendung der MVCP Methode 2 Fälle auf, in denen die Differenz den Wert 12 überschreitet. Da der betrachtete Zeitraum ein gesamtes Jahr, also 365 Tage, umfasst ist dies nur ein sehr geringer Teil.
In Abbildung 24 sind die Differenzen der Anzahlen der durch die Methoden identifizierten Anomalien und der optisch identifizierten Anomalien als Barplot dargestellt. Bei Verwendung der Methoden ist eine Vorentscheidung mit Hilfe von logistischer Regression mit Schwellenwert 0.25 getroffen worden. Bei Verwendung der VCP Methode wird



(a) VCP

(b) MVCP

Abbildung 24: Barplots über die Differenzen der Anzahl gefundener Anomalien der Methoden, bei einer Vorentscheidung, über das Vorliegen einer Anomalie an einem Tag, mit Hilfe von logistischer Regression mit Schwellenwert 0.25. Dabei wird das logistische Regressionsmodell aus Abschnitt 6.3 verwendet, welches auf den gesamten Daten des betrachten Zeitraums angepasst wurde. Die verwendeten Daten wurden an Wegaufnehmer WWS4 gemessen und liegen im Zeitraum vom 01.06.16 bis zum 31.05.17.
a) VCP Methode,
b) MVCP Methode.

im Vergleich zur Zählung von Anne 243 Mal und im Vergleich zur Zählung von Johanna 225 Mal eine Differenz, bezüglich der identifizierten Anzahlen an Anomalien, von

0 erreicht. Mit der MVCP Methode wird im Vergleich zur Zählung von Anne 247 Mal und im Vergleich zur Zählung von Johanna 231 Mal eine Differenz von 0 erreicht. Somit liegt bezüglich der Tage an denen die korrekte Anzahl an Anomalien erkannt wird für die VCP Methode im Vergleich zu den Vorgehensweisen aus Abbildung 22 und 23 eine Verschlechterung vor. Die MVCP Methode verbessert sich diesbezüglich leicht im Vergleich zur Verwendung der Methode mit Schwellenwert 12 und verschlechtert sich im Vergleich zur Vorentscheidung anhand von logistischer Regression mit Schwellenwert 0.3. Zudem steigt die Anzahl der Tage an denen die Differenz der ermittelten Anzahlen den Wert 12 überschreitet, bei der VCP Methode, für beide Zählungen, auf 16 an. Bei der MVCP Methode steigt dieser Wert auf 5 an. Dies stellt für beide Methoden eine Verschlechterung, im Vergleich zur Vorentscheidung mit logistischer Regression mit Schwellenwert 0.3, dar. Die höchste Differenz zwischen der durch die Methoden identifizierten Anzahl und der optisch identifizierten Anzahl an Anomalien liegt bei 19, was ebenfalls eine Verschlechterung, im Vergleich zur Nutzung eines Schwellenwertes von 0.3 bei der logistischen Regression, darstellt.
Bei Betrachtung der richtigen Klassifikation der Anzahlen an Anomalien an einem Tag, scheint eine Vorentscheidung mit Hilfe von logistischer Regression mit Schwellenwert 0.3 und anschließender Identifikation der Anomalien mit der MVCP Methode, am vorteilhaftesten zu sein.

7 Verkehrsschätzung

In der Veranstaltung Fallstudien II ist die Idee entstanden, die Verkehrsschätzung anhand des integrationsbasierten Variationsmaßes (siehe Abschnitt 3.6) vorzunehmen. Dabei wird angenommen, dass die Fläche unter den trendbereinigten Rissbreiten einer bestimmten Anzahl an Autos entspricht. Um der Fläche innerhalb einer Stunde eine Fahrzeuganzahl zuzuordnen, wurde am 23.05.17 von 15 bis 16 Uhr am südlichen Überbau der Brücke eine Verkehrszählung durchgeführt. Dabei ist ein Verkehrsaufkommen von 835 Autos, 32 LKW und 9 Straßenbahnen ermittelt worden. Da die Straßenbahnen auf einer separaten Fahrspur fahren, wird von 867 Fahrzeugen ausgegangen. Der durch das Integrationsbasierte Variationsmaß ermittelte Flächeninhalt innerhalb dieser Stunde liegt bei ≈ 1.046 und wird als $F^{z_1}_{(15,16)}$ bezeichnet. Bei allen Schätzungen in diesem Kapitel sind die Tage

der Zeitumstellung unberücksichtigt geblieben, da diese zu Problemen bei den verwendeten Funktionen führen. Wenn davon ausgegangen wird, dass der Flächeninhalt $F^{z_1}_{(15,16)}$ - 867 Fahrzeugen entspricht, kann die Anzahl Fahrzeuge pro Stunde s, an einem Tag T, durch

$$F^{T}_{(s-1,s)} / F^{z_1}_{(15,16)} * 867, \tag{25}$$

mit $s \in \{1, 2, \ldots, 24\}$ und $T \in \{01.06.16, 02.06.16, \ldots, 31.05.17\}$, geschätzt werden. In Abbildung 25 sind Boxplots der mit Formel (25) geschätzten Fahrzeuganzahlen pro Stunde für den Wegaufnehmer WWS4 dargestellt. Vor Anwendung des Integrationsbasierten Variationsmaßes sind die Anomalien mit der MVCP Methode mit Schwellenwert 12 entfernt worden. Fehlt über die Hälfte der Beobachtungen innerhalb einer Stunde, wird die entsprechende Stunde auf `NA` gesetzt und in der Grafik nicht berücksichtigt.

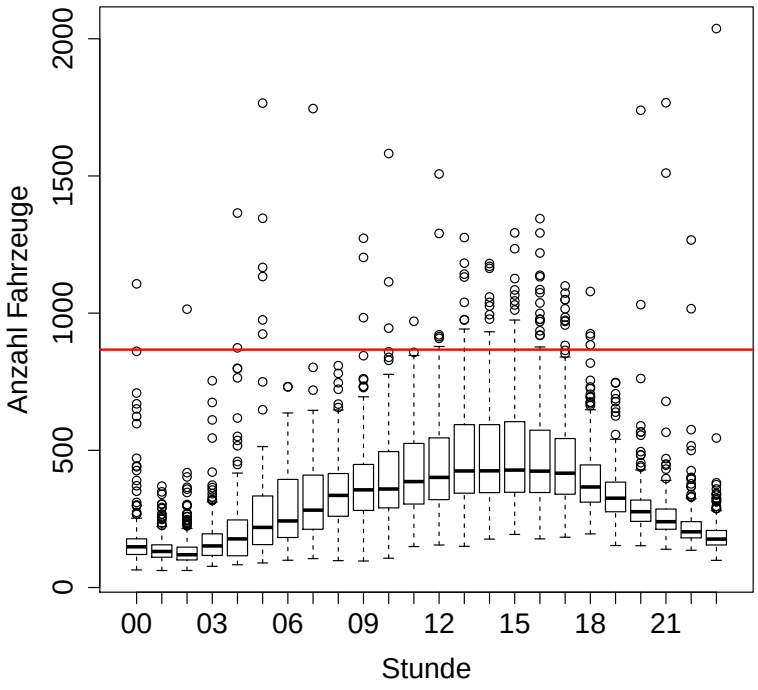


Abbildung 25: Boxplot über die mit Formel (25) geschätzten Fahrzeuganzahlen pro Stunde aus den Daten vom 01.06.16 bis zum 31.05.17 für den Wegaufnehmer WWS4, wenn davon ausgegangen wird, dass ein Flächeninhalt von 1.046 - 867 Fahrzeugen entspricht. Die rote Linie steht für die Anzahl der am 23.05.17 zwischen 15 und 16 Uhr gezählten Fahrzeuge.

Die rote Linie markiert die bei der Verkehrszählung am 23.05.17 ermittelte Fahrzeugan-

zahl. Nachts wird eine geringere Fahrzeuganzahl geschätzt als Tagsüber, was durchaus sinnvoll ist. Jedoch werden zur Berufsverkehrszeit am Morgen zwischen 7 und 9 Uhr deutlich geringere Werte als zur Berufsverkehrszeit am Nachmittag geschätzt. Dies entspricht nicht der intuitiven Annahme, dass zur Berufsverkehrszeit am Vormittag und am Nachmittag in etwa gleich viele Fahrzeuge unterwegs sind. Von 0 bis 3 Uhr fällt die geschätzte Fahrzeuganzahl, danach steigt sie bis 13 Uhr an. Zwischen 13 und 17 Uhr bleibt die geschätzte Fahrzeuganzahl in etwa konstant und fällt anschließend wieder von 17 bis 24 Uhr. Es fällt auf, dass der Median der geschätzten Fahrzeuganzahl pro Stunde zwischen 15 und 16 Uhr den maximalen Wert erreicht. Dieser Wert liegt knapp über 400 Fahrzeugen, was nur etwa der Hälfte der am 23.05.17 zu dieser Uhrzeit ermittelten Fahrzeuganzahl entspricht. Es ist davon auszugehen, dass die Anzahl der zu dieser Zeit die Brücke überquerenden Fahrzeuge keine derartig großen Abweichungen aufweist. Auch das obere Quartil liegt zu dieser Uhrzeit deutlich unter der gezählten Fahrzeuganzahl. Lediglich der obere Whisker überschreitet den entsprechenden Wert leicht. Zudem treten einige Ausreißer auf. Besonders auffällig ist eine geschätzte Fahrzeuganzahl von ungefähr 2000 Stück um 23 Uhr. Aufgrund des erheblich höheren Wertes im Vergleich zu den anderen Schätzungen zu dieser Uhrzeit, liegt die Vermutung nahe, dass an dem entsprechenden Tag zwischen 23 und 24 Uhr eine Anomalie vorliegt, die von der MVCP Methode mit Schwellenwert 12 nicht identifiziert worden ist. Der obere und der untere Whisker liegen für die meisten Stunden sehr weit auseinander. Zwischen 15 und 16 Uhr liegt der untere Whisker zum Beispiel unter 250 Fahrzeugen, während der obere Whisker bei ca. 1000 Fahrzeugen liegt. Da auch am Wochenende ermittelte Daten in die Grafik mit eingegangen sind, kann zwar davon ausgegangen werden, dass an einigen Tagen deutlich weniger als 867 Fahrzeuge die Brücke überquert haben. Ein derartig großer Unterschied entspricht jedoch nicht den Erwartungen. Insgesamt wird die Fahrzeuganzahl zwischen 15 und 16 Uhr mit dieser Methode deutlich unterschätzt.

Zum Vergleich wurde am Dienstag den 26.09.17 von 15 bis 16 Uhr eine weitere Verkehrszählung durchgeführt. Dabei sind 823 Autos, 24 LKW und 6 Straßenbahnen gezählt worden. Aufgrund der separaten Fahrspur für Straßenbahnen, wird somit von 847 Fahrzeugen ausgegangen. Der durch das integrationsbasierte Variationsmaß ermittelte Flächeninhalt zum Zeitpunkt der Verkehrszählung, liegt bei ≈ 0.335 und wird als $F^{z_2}_{(15,16)}$ bezeichnet. Da somit ein Wert von 1.046 für 867 Fahrzeuge und ein Wert von 0.335 für 847 Fahrzeuge

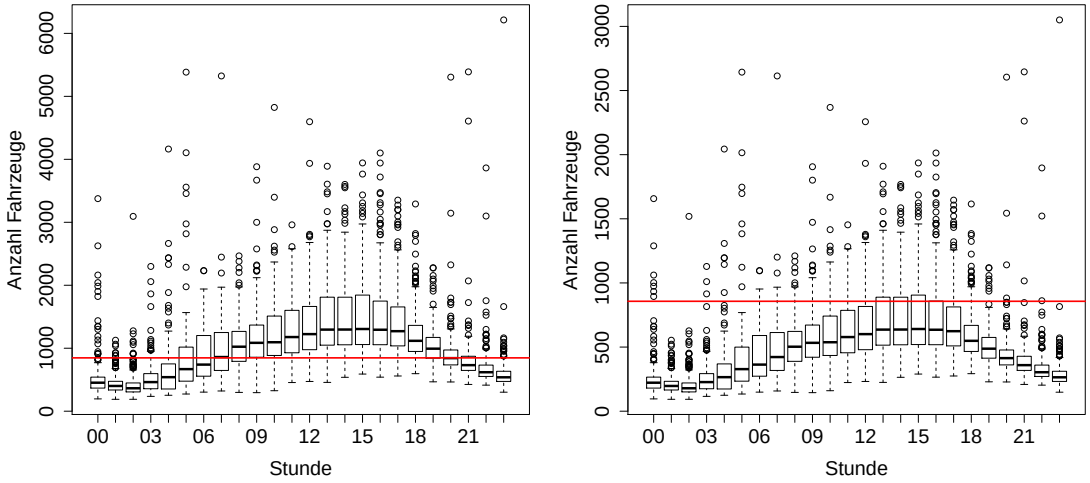
vorliegt, lässt sich ausschließen, dass der Flächeninhalt unter den trendbereinigten Rissbreiten eindeutig einer bestimmten Fahrzeuganzahl zugeordnet werden kann. Wenn davon ausgegangen wird, dass der Flächeninhalt $F^{z_2}_{(15,16)} = 0.335$ - 847 Fahrzeugen entspricht, kann die Fahrzeuganzahl pro Stunde s, an einem Tag T, durch

$$F^{T}_{(s-1,s)} / F^{z_2}_{(15,16)} * 847, \tag{26}$$

mit $s \in \{1, 2, \ldots, 24\}$ und $T \in \{01.06.16, 02.06.16, \ldots, 31.05.17\}$, geschätzt werden. Wird der Mittelwert aus den beiden Zählungen gebildet, liegt der Flächeninhalt zwischen 15 und 16 Uhr $F^{m}_{(15,16)}$ bei ≈ 0.690 und die zugehörige Fahrzeuganzahl bei 857. Die Fahrzeuganzahl pro Stunde s, an einem Tag T, wird dann durch

$$F^{T}_{(s-1,s)} / F^{m}_{(15,16)} * 857, \tag{27}$$

mit $s \in \{1, 2, \ldots, 24\}$ und $T \in \{01.06.16, 02.06.16, \ldots, 31.05.17\}$, geschätzt.
In Abbildung 26 sind Boxplots über die mit Formel (26) geschätzte Fahrzeuganzahl pro Stunde und über die mit Formel (27) geschätzte Fahrzeuganzahl pro Stunde, dargestellt.
In beiden Grafiken entspricht der Verlauf der Fahrzeuganzahl pro Stunde dem Verlauf aus Grafik 25. Das größte Fahrzeugaufkommen wird somit auch bei diesen beiden Schätzungen am Nachmittag vorhergesagt. Werden die am 26.09.17 ermittelten Daten als Grundlage für die Verkehrsschätzung verwendet, fällt die geschätzte Fahrzeuganzahl sehr hoch aus. Dabei werden zwischen 13 und 18 Uhr im Median über 1200 Fahrzeuge pro Stunde geschätzt. Es ist nicht plausibel, dass der Median der Schätzungen pro Stunde die an den beiden Tagen gezählten Fahrzeuganzahlen so stark übersteigt. Auch das untere Quartil liegt in diesem Zeitraum deutlich über der zwischen 15 und 16 Uhr ermittelten Fahrzeuganzahl. Die geschätzte Fahrzeuganzahl zwischen 7 und 8 Uhr stimmt dabei im Median in etwa, mit der am 26.09.17 zwischen 15 und 16 Uhr gezählten Anzahl Fahrzeuge überein, was Aufgrund von Berufsverkehr durchaus sinnvoll erscheint. Bei Verwendung der Mittelwerte des Flächeninhalts und der Fahrzeuganzahl der beiden Zählungen, liegt der Median zwischen 15 und 16 Uhr bei etwas über 600 Fahrzeugen. Das obere Quartil liegt dabei zwischen 13 und 17 Uhr in etwa bei der roten Linie, welche den Mittelwert der beiden Verkehrszählungen markiert. Dies ist näher an der zu dieser Zeit gezählten Fahrzeuganzahl als die anderen Vorgehensweisen. Trotzdem wird durch diese Vorgehensweise kein optimales Ergebnis erzielt. Besonders auffällig ist, dass bei allen drei Schätzungen für die meisten Uhrzeiten ein sehr großer Interquartilsabstand und somit auch sehr lange



(a) Schätzung mit Formel (26)

(b) Schätzung mit Formel (27)

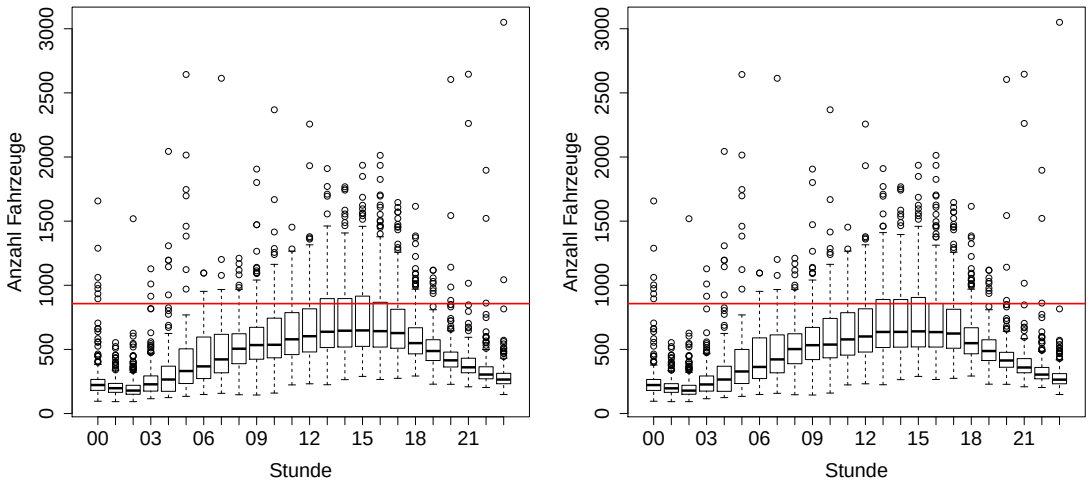
Abbildung 26: Boxplots über die mit Formel (26) bzw. Formel (27) geschätzte Fahrzeuganzahl pro Stunde aus den Daten vom 01.06.16 bis zum 31.05.17 für den Wegaufnehmer WWS4.
a) Es wird davon ausgegangen, dass ein Flächeninhalt von 0.335 - 847 Fahrzeugen entspricht (Schätzung mit Formel (26)). Die rote Linie markiert die zum Zeitpunkt der Verkehrszählung am 26.09.17 ermittelte Fahrzeuganzahl.
b) Es wird davon ausgegangen, dass ein Flächeninhalt von 0.690 - 857 Fahrzeugen entspricht (Schätzung mit Formel (27)). Die rote Linie markiert den Mittelwert der zu den beiden Verkehrszählungen ermittelten Fahrzeuganzahlen.

Whisker vorliegen, obwohl nicht davon ausgegangen werden kann, dass so große Unterschiede zwischen der Fahrzeuganzahl an verschiedenen Tagen vorliegen.

7.1 Vergleich der Schätzungen in Abhängigkeit der identifizierten Anomalien

Bisher liegen nur Schätzungen auf Grundlage des integrationsbasierten Variationsmaßes vor, bei denen die Anomalien mit Hilfe der MVCP Methode mit Schwellenwert 12 entfernt wurden. Diese Vorgehensweise ist nach Kapitel 6.4 nicht die genaueste der in dieser Arbeit betrachteten Vorgehensweisen zur Identifikation von Anomalien. Um zu überprü-

fen wie groß der Einfluss der richtigen Identifikation der Anomalien auf die Verkehrsschätzung durch das integrationsbasierte Variationsmaß ist, wird das Verkehrsaufkommen mit Formel (27) auf Daten, auf die verschiedenen Vorgehensweisen zur Identifikation der Anomalien angewendet wurden, geschätzt. Betrachtet wird auch hier wieder der Zeitraum vom 01.06.16 bis zum 31.05.17 am Wegaufnehmer WWS4. Formel (27) wird zur Schätzung der Fahrzeuganzahl pro Stunde verwendet, da der Mittelwert der beiden zur Verkehrszählung genutzten Tage das plausibelste der drei zur Ermittelung der Fahrzeuganzahl herangezogenen Verfahren liefert. In Abbildung 27 sind die entsprechenden Fahrzeuganzahlen pro Stunde als Boxplots dargestellt. Teil a) der Abbildung zeigt dabei



(a) Logistische Regression mit Schwellenwert 0.3

(b) MVCP mit Schwellenwert 12

Abbildung 27: Boxplots über die mit Formel (27) geschätzte Fahrzeuganzahl pro Stunde aus den Daten vom 01.06.16 bis zum 31.05.17 für den Wegaufnehmer WWS4. Es wird davon ausgegangen, dass ein Flächeninhalt von 0.690 - 857 Fahrzeugen entspricht (Mittelwert der Daten vom 23.05.17 und vom 26.09.17). Die rote Linie markiert dabei die zwischen 15 und 16 Uhr angenommene Fahrzeuganzahl.
a) Die Anomalien wurden mit Hilfe der MVCP Methode nach Vorentscheidung durch die logistische Regression mit Schwellenwert 0.3 entfernt.
b) Die Anomalien wurden mit Hilfe der MVCP Methode mit Schwellenwert 12 entfernt.

den Fall, dass die Anomalien mit der MVCP Methode nach einer Vorentscheidung durch die logistische Regression mit Schwellenwert 0.3 entfernt wurden. In Teil b) ist zum besseren Vergleich noch einmal der Fall aus Abbildung 26 b) (Entfernung der Anomalien mit

der MVCP Methode mit Schwellenwert 12) dargestellt. Es lässt sich erkennen, dass sich die beiden Vorgehensweisen im Median kaum unterscheiden. Auch die Quartile Whisker und Ausreißer sind für beide Grafiken sehr ähnlich. Die Schätzungen bei einer Vorentscheidung mit Hilfe der logistischen Regression fällt im Median geringfügig höher aus. Die maximale Fahrzeuganzahl im Median wird für beide Schätzungen weiterhin in der Stunde von 15 bis 16 Uhr erreicht. Nach einer Vorentscheidung mit der logistischen Regression, beträgt die Schätzung zu dieser Zeit im Median ca. 648 und bei Verwendung der MVCP Methode mit Schwellenwert 12 ca. 641 Fahrzeuge.

Da auch die Vorentscheidung mit logistischer Regression und Schwellenwert 0.3 kein optimales Ergebnis liefert, wird überprüft wie groß der Unterschied der Schätzungen ist, wenn die MVCP Methode zur Entfernung der Anomalien nur auf die Tage angewendet wird, an denen optisch Anomalien identifiziert wurden. Dieser Fall ist in Abbildung 28 als Boxplot dargestellt. Auf den ersten Blick ist kein großer Unterschied zu den Grafiken aus Abbildung 27 zu erkennen. Auch bei dieser Art der Vorbereitung der Daten entsprechen die stündlichen Schätzungen im Verlauf des Tages dem Verlauf der Schätzungen der anderen betrachteten Vorgehensweisen zur Entfernung der Anomalien. Die im Median maximale stündliche Fahrzeuganzahl wird hier ebenfalls zwischen 15 und 16 Uhr geschätzt und liegt bei ca. 656 Fahrzeugen. Dies entspricht einer Steigerung von 8 Fahrzeugen im Vergleich zur Verwendung der MVCP Methode nach Vorentscheidung durch die logistische Regression mit Schwellenwert 0.3. Auch bei den Quartilen, Whiskern und Ausreißern sind keine deutlichen Unterschiede erkennbar. Da diese Abweichungen eher geringfügig sind, lässt sich daraus schließen, dass eine Vorentscheidung durch die logistische Regression mit Schwellenwert 0.03, ausreichend ist. Die minimale Fahrzeuganzahl wird mit ca. 179 Fahrzeugen zwischen 2 und 3 Uhr Nachts geschätzt. Intuitiv scheint diese Zahl etwas zu hoch zu sein. Der Ausreißer zwischen 23 und 0 Uhr ist weiterhin vorhanden, wenn die MVCP Methode nur auf Tage angewendet wird, an denen Anomalien vorliegen. Daraus lässt sich schließen, dass MVCP die zu dieser Zeit vorliegende Anomalie an dem entsprechenden Tag grundsätzlich nicht identifiziert. Der betroffene Tag ist der 01.12.16 und der Grund besteht in einer bei der Implementierung der VCP Methode getroffenen Annahme, die für diesen Tag nicht erfüllt ist (siehe Kapitel 5.2.1).

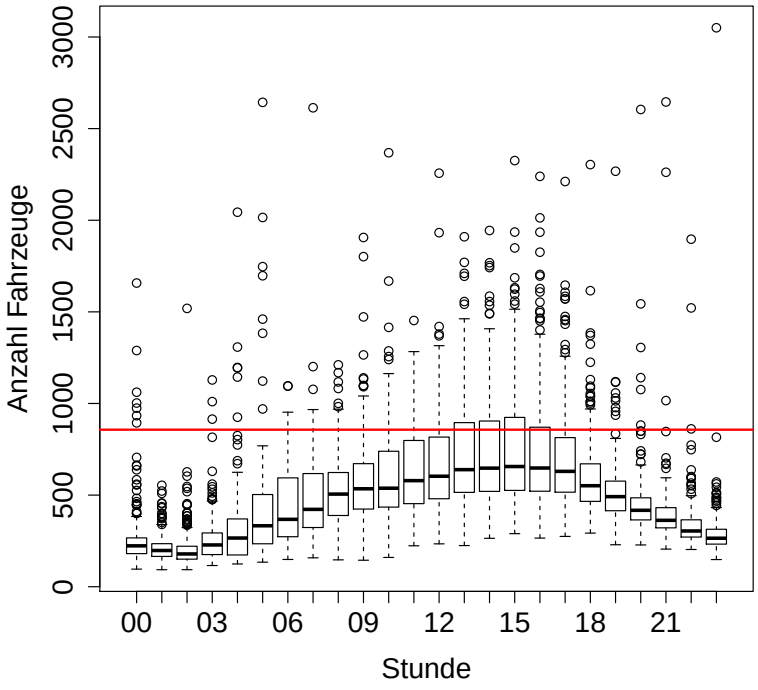


Abbildung 28: Boxplots über die mit Formel (27) geschätzte Fahrzeuganzahl pro Stunde aus den Daten vom 01.06.16 bis zum 31.05.17 für den Wegaufnehmer WWS4, wenn davon ausgegangen wird, dass ein Flächeninhalt von 0.690 - 857 Fahrzeugen entspricht (Mittelwert der Daten vom 23.05.17 und vom 26.09.17). Die rote Linie markiert die zwischen 15 und 16 Uhr angenommene Fahrzeuganzahl. Die Anomalien wurden mit Hilfe der MVCP Methode entfernt. Dabei ist die Methode nur auf die Tage angewendet worden, an denen optisch eine Anomalie erkennbar ist.

7.2 Ansätze zur Verbesserung der Schätzung

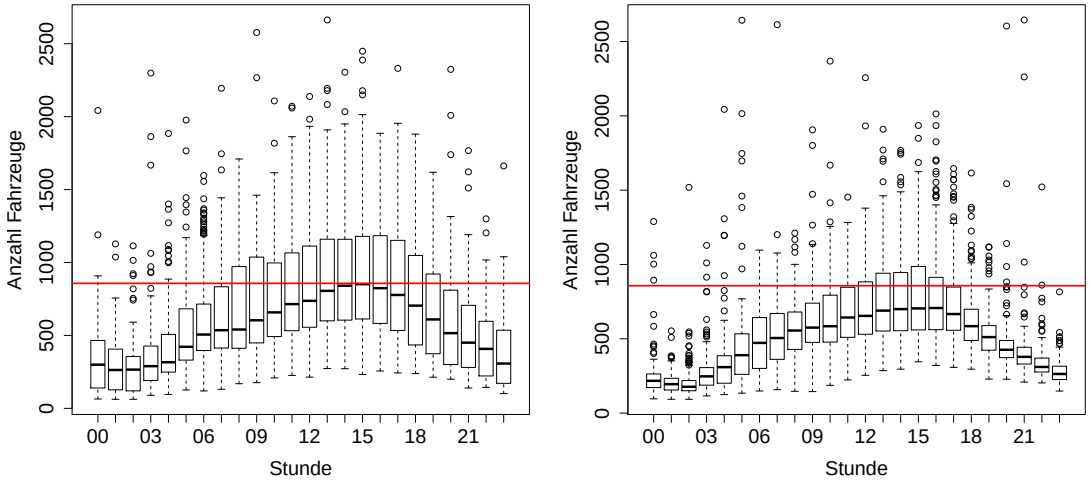
Da durch die Verwendung des integrationsbasierten Informationsmaßes kein zufriedenstellendes Ergebnis erzielt werden kann, wird nach Möglichkeiten zur Verbesserung der entsprechenden Schätzungen gesucht. Eine mögliche Ursache der großen Unterschiede im Flächeninhalt, bei einer ähnlichen Fahrzeuganzahl, zu den beiden Zeitpunkten der Verkehrszählung, kann in einem Unterschied in der Temperatur liegen. Die maximale Temperatur unterhalb der Brücke lag am 23.05.17 zwischen 15 und 16 Uhr bei 22.83°C und am 26.09.17 zwischen 15 und 16 Uhr bei 17.37°C. Dies scheint auf den ersten Blick kein besonders großer Temperaturunterschied zu sein. Um herauszufinden ob die Schätzung durch eine Einteilung nach der Temperatur verbessert werden kann, werden die Daten

vom 01.06.16 bis zum 31.05.17 anhand der maximalen Temperatur innerhalb der Stunde (tempHour) und anhand des Flächeninhalt in 3 Gruppen eingeteilt. Die Gruppen werden als *kalt*, *mittel* und *warm* bezeichnet und die durch das k-Means Verfahren ermittelten Clusterzentren sind in Tabelle 1 zu sehen. Bei Beobachtungen, die der Kategorie *warm*

Tabelle 1: Clusterzentren der drei Gruppen *kalt*, *mittel* und *warm*, bei einer Einteilung durch das k-Means Verfahren. Als Grundlage für die Clusterung haben die Daten vom 01.06.16 bis zum 31.05.17, gemessen an Wegaufnehmer WWS4, gedient.

Cluster	tempHour	Flächeninhalt
kalt	2.50	0.287
mittel	9.66	0.340
warm	19.94	0.491

zugeordnet werden, wird davon ausgegangen, dass ein Flächeninhalt von 1.046 - 867 Fahrzeugen entspricht. In der Kategorie *mittel* entspricht ein Flächeninhalt von 0.335 - 847 Fahrzeugen. Die Kategorie *kalt* wird zunächst nicht berücksichtigt, da zu diesen Temperaturen keine Verkehrszählung durchgeführt worden ist. Somit wird die Fahrzeuganzahl pro Stunde, für Stunden, die der Kategorie *warm* zugeordnet werden mit Formel (25) und für Stunden die der Kategorie *mittel* zugeordnet werden mit Formel (26), geschätzt. Abbildung 29 a) zeigt Boxplots der wie oben beschriebenen stündlichen Schätzung für den Zeitraum vom 01.06.16 bis zum 31.05.17. In Teil b) der Abbildung wird für die Kategorien *warm* und *mittel* Formel (27) verwendet. Die Stunden der Kategorie *kalt* werden dabei nicht berücksichtigt. Dies dient dazu um festzustellen, ob die Änderungen im Vergleich zu Abbildung 27 a) lediglich auf das Wegfallen von Stunden aus der Kategorie *kalt* zurückzuführen sind. Für den Fall in Abbildung 29 a) wird zwischen 15 und 16 Uhr im Median eine Fahrzeuganzahl von ca. 851 geschätzt, was sehr nahe an den tatsächlichen Zählungen zu dieser Zeit liegt. Zwischen 0 und 5 Uhr werden stündliche Werte um, im Median, 300 Fahrzeuge geschätzt. Diese Zahl erscheint intuitiv deutlich zu hoch. Zudem wird die Fahrzeuganzahl zur Berufsverkehrszeit am morgen deutlich geringer geschätzt als zur Berufsverkehrszeit am Nachmittag. Intuitiv wird aber davon ausgegangen, dass das Verkehrsaufkommen zu diesen Zeiten in etwa gleich ist. Da zu dieser Zeit jedoch keine Verkehrszählung vorliegt, kann dies nicht eindeutig beurteilt werden. Auch hier treten insgesamt große Unterschiede für die Schätzungen innerhalb einer Stunde auf. Zwischen



(a) Formel (25) für *warme* und Formel (26) für *mittlere* Tage

(b) Formel (27) (ohne *kalte* Tage)

Abbildung 29: Boxplots über die durch das integrationsbasierte Variationsmaß geschätzten Fahrzeuganzahlen pro Stunde aus den Daten vom 01.06.16 bis zum 31.05.17 für den Wegaufnehmer WWS4, wenn davon ausgegangen wird, dass sich die Daten in die Kategorien *warm*, *mittel* und *kalt* einteilen lassen. Die der Kategorie *kalt* zugeordneten Stunden werden nicht in die Grafiken einbezogen. Die Anomalien sind mit der MVCP Methode, nach einer Vorentscheidung durch die logistische Regression mit Schwellenwert 0.3, identifiziert worden. Die rote Linie markiert dabei das Mittel der beiden durch die Verkehrszählungen zwischen 15 und 16 Uhr ermittelten Fahrzeuganzahlen.
a) Es wird davon ausgegangen, dass in der Kategorie *mittel* ein Flächeninhalt von 0.335 - 847 Fahrzeugen entspricht (Verwendung von Formel (25)) und in der Kategorie *warm* entspricht ein Flächeninhalt von 1.046 - 867 Fahrzeugen (Verwendung von Formel (26)).
b) Es wird davon ausgegangen, dass ein Flächeninhalt von 0.690 - 857 Fahrzeugen entspricht (Verwendung von Formel (27)).

15 und 16 Uhr liegt der untere Whisker unter 250, während der obere Whisker 2000 Fahrzeuge überschreitet. Für den Fall in Abbildung 29 b) ist die geschätzte Fahrzeuganzahl pro Stunde, im Median, deutlich geringer als in Teil a) der Abbildung. Lediglich zwischen 8 und 9 Uhr wird eine geringfügig höhere Fahrzeuganzahl geschätzt, wenn das Mittel aus den beiden Stunden der Verkehrszählung verwendet wird. Die maximale Fahrzeuganzahl im Median wird für diesen Fall zwischen 16 und 17 Uhr geschätzt und beträgt ca. 707.

Dies ist im Vergleich zur Verkehrszählung zu wenig. Zwischen 1 und 4 Uhr sinkt die geschätzte Fahrzeuganzahl pro Stunde im Median unter 200 Fahrzeuge. Diese Werte wirken realistischer als die Werte aus Abbildung 29 a), scheinen intuitiv jedoch immer noch zu hoch zu sein. Die Interquartilsabstände und somit auch die Länge der Whisker ist geringer als in Teil a) der Abbildung.

Aus den Grafiken lässt sich schließen, dass die Verwendung von verschiedenen Werten zur Verkehrsschätzung, in Abhängigkeit der Temperatur, einen Unterschied hervorruft. Dabei wird zwischen 15 und 16 Uhr im Median ein Wert sehr nahe der zu dieser Zeit tatsächlich gezählten Fahrzeuganzahl erreicht. Die geschätzte Fahrzeuganzahl zu anderen Zeiten, insbesondere Nachts, erscheint jedoch nicht optimal. Da der extreme Ausreißer um 23 Uhr in keiner der beiden Grafiken auftaucht, kann davon ausgegangen werden, dass an dem entsprechenden Tag zu dieser Uhrzeit Temperaturen herrschten, die der Kategorie *kalt* zugeordnet werden konnten.

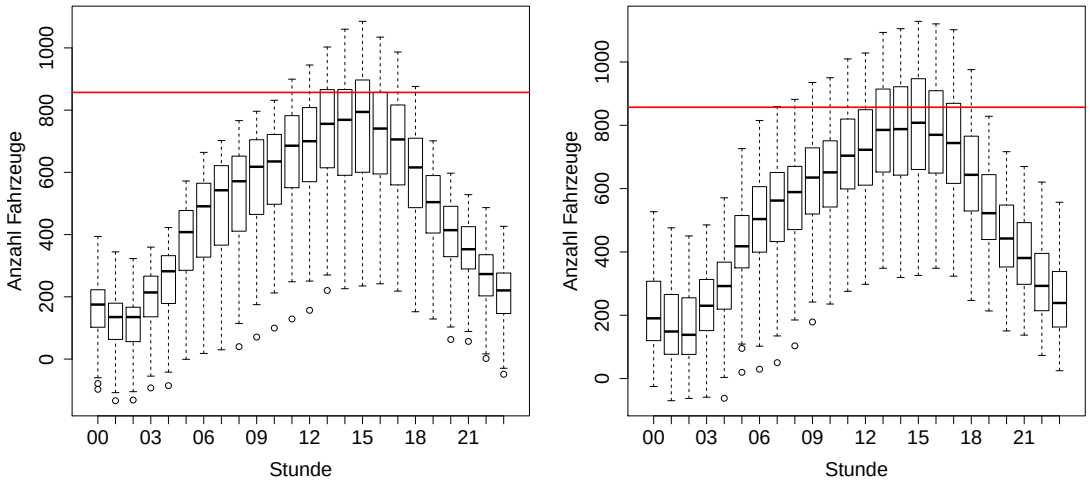
7.3 Verkehrsschätzung mittels linearer Regression

Da für Stunden, welche in die Kategorie *kalt* fallen, aufgrund der fehlenden Verkehrszählung keine Schätzung durch den Ansatz aus Kapitel 7.2 durchgeführt werden kann, wird versucht diese mit Hilfe eines linearen Modells vorherzusagen. Zur Vorhersage der Fahrzeuganzahl pro Stunde wird ein lineares Modell aufgestellt, indem die Variable **anzahl** durch die Variablen **wochentag**, **wochenende**, **ferien**, **stunde**, **tempHour** (Temperatur pro Stunde) und **wochenende**$*$**stunde**, erklärt wird. Die Modellgleichung zur Vorhersage der Fahrzeuganzahl innerhalb einer Stunde hat dann die Form:

$$\begin{aligned} \text{anzahl} = \alpha + \beta_1\,\text{wochentag} + \beta_2\,\text{wochenende} + \beta_3\,\text{ferien} + \beta_4\,\text{stunde} \\ + \beta_5\,\text{tempHour} + \beta_6\,\text{wochenende} * \text{stunde} + \varepsilon. \end{aligned} \tag{28}$$

Dabei wird der Achsenabschnitt als α und der Fehler des Modells als ε bezeichnet. Das Modell wird auf den Daten vom 01.06.16 bis zum 31.05.17 der Kategorien *warm* und *mittel* angepasst. Die Fahrzeuganzahl pro Stunde ist dabei durch das in Abschnitt 7.2 beschriebene Vorgehen geschätzt worden. In Abbildung 30 sind die durch das lineare Modell aus Gleichung (28) vorhergesagten stündlichen Fahrzeuganzahlen als Boxplots dargestellt. Teil a) der Abbildung zeigt dabei die Vorhersagen des Zeitraumes vom 01.06.16

bis zum 31.05.17 der Kategorie *kalt* und Teil b) zeigt die Vorhersagen für den Zeitraum vom 01.06.17 bis zum 20.10.17. Dabei handelt es sich um Zeiträume, die nicht zur An-



(a) Vorhersage Kategorie *kalt*

(b) Vorhersage 01.06. - 20.10.17

Abbildung 30: Boxplots über die mit dem linearen Modell vorhergesagten Fahrzeuganzahlen pro Stunde für den Wegaufnehmer WWS4. Das Modell ist auf den wie in Abbildung 29 a) geschätzten Daten angepasst worden. Die rote Linie markiert dabei den Mittelwert der in den beiden Verkehrzählungen ermittelten Fahrzeuganzahlen zwischen 15 und 16 Uhr. Die Anomalien sind mit der MVCP Methode, nach einer Vorentscheidung durch die logistische Regression mit Schwellenwert 0.3, identifiziert worden.
a) Vorhersage für die Daten der Kategorie *kalt* aus den Daten vom 01.06.16 bis zum 31.05.17.
b) Vorhersage für die Daten vom 01.06.17 bis zum 20.10.17.

passung des Modells verwendet worden sind. Wie der Name schon sagt, sind in Teil a) der Grafik nur Beobachtungen enthalten, die der Kategorie *kalt* zugeordnet werden können, während innerhalb des Zeitraumes vom 01.06. bis zum 20.10.17 (Teil b) der Grafik) hauptsächlich Beobachtungen der Kategorien *warm* und *mittel* zu finden sein werden. Die beiden Grafiken unterscheiden sich nur sehr geringfügig voneinander. In beiden Fällen wird zwischen 15 und 16 Uhr das Maximum erreicht, welches im Median knapp unter 800 Fahrzeugen liegt. Diese Zahl ist nicht all zu weit von der zu dieser Uhrzeit tatsächlich gezählten Fahrzeuganzahl entfernt. Das obere Quartil überschreitet in beiden Fällen den zu dieser Zeit gezählten Wert, bleibt jedoch unter 1000 Fahrzeugen, während das untere

Quartil in beiden Grafiken über 600 Fahrzeugen liegt. Der Interquartilsabstand und somit auch die Länge der Whisker ist weiterhin für alle Tageszeiten recht hoch. Zwischen 0 und 3 Uhr liegt die geschätzte Fahrzeuganzahl pro Stunde im Median zwischen 100 und 200 Fahrzeugen. Wie in allen bisherigen Schätzungen scheint diese Zahl intuitiv zu hoch zu sein. Die oberen Whisker überschreiten in dieser Zeit in Teil b) der Grafik 400 Fahrzeuge und liegen in Teil a) der Grafik zwischen 300 und 400 Fahrzeugen. Am Wochenende ist das Verkehrsaufkommen zu dieser Uhrzeit vermutlich zwar etwas höher als den Großteil der Woche, trotzdem erscheint diese Zahl etwas zu hoch. Dieser Umstand ist nicht weiter überraschend, da das Modell auf Daten basiert, bei denen die Fahrzeuganzahl zu dieser Uhrzeit ebenfalls zu hoch zu sein scheint. Ausreißer treten bei Verwendung des linearen Modells deutlich seltener auf, als wenn nur das integrationsbasierte Variationsmaß zur Schätzung verwendet wird. Dies kann darauf zurückgeführt werden, dass die Rissbreiten nicht direkt in das lineare Modell einfließen, sondern eine Vorhersage anhand von Kriterien wie Wochentag, Uhrzeit und Temperatur getroffen wird. Bei den Daten die nur aus der Kategorie *kalt* stammen werden etwas geringere Fahrzeuganzahlen geschätzt, als wenn der Zeitraum vom 01.06. - 20.10.17 betrachtet wird.

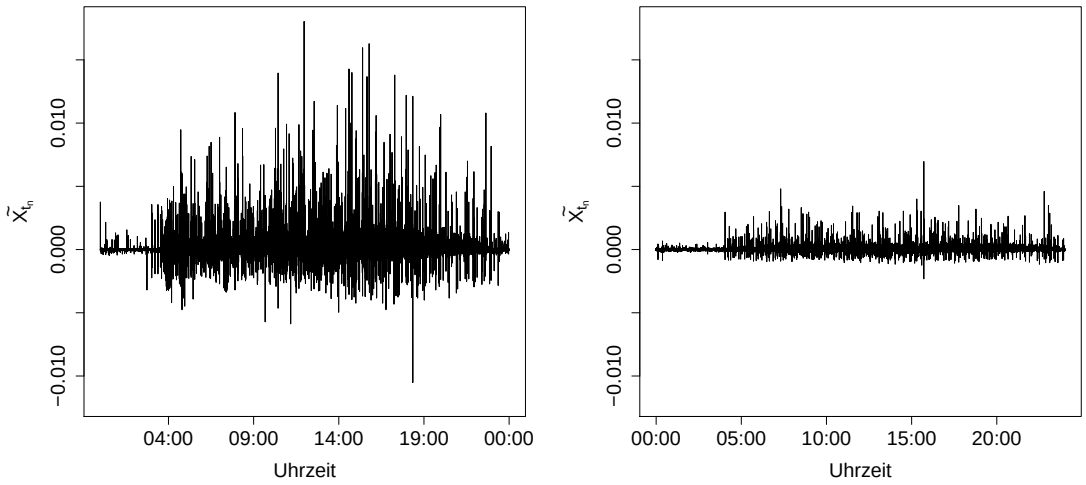
7.4 Probleme bei der Verkehrsschätzung

Bei dem Versuch der Verkehrsschätzung anhand der Rissbreiten treten verschiedene Probleme auf, welche das Erzielen von durchgehend schlüssigen Ergebnissen verhindern. Dabei handelt es sich um die folgenden Probleme:

- Einfluss der Temperatur auf die Rissbreiten
- Die genaue Fahrzeuganzahl liegt für zu wenige Zeiträume vor.

7.4.1 Probleme durch den Einfluss der Temperatur auf die Rissbreiten

Die Temperatur hat einen deutlichen Einfluss auf die Rissbreiten. Im Winter sind die Risse deutlich schmaler als im Sommer. Um dies zu veranschaulichen sind in Abbildung 31 die trendbereinigten Rissbreiten für den Wegaufnehmer WWS4 an zwei verschiedenen Tagen dargestellt. In Teil a) wird der 20.07.16, mit Temperaturen unterhalb der Brücke zwischen 22°C und 31°C, und in Teil b) wird der 23.01.17, mit Temperaturen unterhalb der Brücke zwischen -4°C und -1°C, betrachtet. Anhand der Abbildungen ist deutlich zu



(a) Rissbreite am 20.07.16

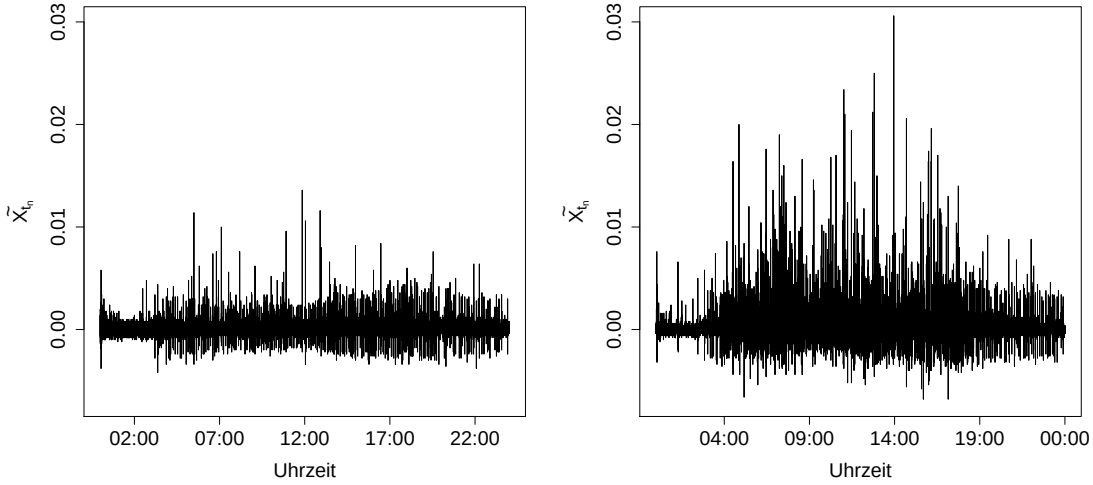
(b) Rissbreite am 23.01.17

Abbildung 31: Trendbereinigte Rissbreite $\tilde{X}_{t_n}$, gemessen am Wegaufnehmer WWS4, bei unterschiedlichen Temperaturen unterhalb der Brücke.
a) Am 20.07.16 lag die Temperatur unter der Brücke zwischen 22°C und 31°C.
b) Am 23.01.17 lag die Temperatur unter der Brücke zwischen -4°C und -1°C.

erkennen, dass die Rissbreiten am 23.01.17 schmaler sind als am 20.07.16. Soll die genaue Fahrzeuganzahl anhand der Rissbreiten geschätzt werden, führt dies dazu, dass im Winter ein geringeres Verkehrsaufkommen bestimmt wird als im Sommer. Es wird jedoch davon ausgegangen, dass das Verkehrsaufkommen im Winter nicht geringer ist, als im Sommer. Dies ist der Grund, aus dem die großen Unterschiede zwischen den Schätzungen, auf Grundlage der Daten vom 23.05.17 und auf Grundlage der Daten vom 26.09.17, entstehen (siehe Abbildung 25 und 26 a)). Die Temperatur, welche entscheidend für die Veränderung der Rissbreite ist, ist die Temperatur innerhalb der Brücke. Es liegen jedoch lediglich Zahlen für die Lufttemperatur oberhalb der Brücke und für die Temperatur an der Unterseite der Brücke vor. Es ist zwar davon auszugehen, dass diese Temperaturen in irgendeiner Weise die Temperatur innerhalb der Brücke beeinflussen. Wie dieser Einfluss aussieht, wie schnell die Veränderung der Temperatur innerhalb der Brücke vonstatten geht und welche weiteren Variablen die Brückentemperatur beeinflussen, ist jedoch nicht bekannt.

7.4.2 Probleme durch das Fehlen von Verkehrszahlen zu den meisten Tageszeiten

Es liegen insgesamt drei Verkehrszählungen für die Brücke der Wittener Straße über den Sheffieldring, in Bochum, vor. Diese haben am Dienstag den 23.05.17 von 15 bis 16 Uhr, am Dienstag den 26.09.17 von 15 bis 16 Uhr und am Sonntag den 10.01.17 von 12:40 bis 13:40 Uhr stattgefunden. Sämtliche Zählungen sind am südlichen Überbau der Brücke vorgenommen worden. Das sind insgesamt zwei Wochentage an denen für die Fahrzeuganzahl innerhalb einer bestimmten Stunde ein Vergleichswert vorliegt. Bei den Ergebnissen der Schätzungen sind Werte aufgetreten, die einer intuitiven Einschätzung widersprechen. Zum Beispiel, dass die maximale Fahrzeuganzahl zwischen 15 und 16 Uhr erreicht wird, obwohl die Hauptstoßzeit aufgrund von Berufsverkehr zwischen 16 und 18 Uhr liegen sollte. Zudem ist auffällig, dass zur Berufsverkehrszeit am Vormittag, zwischen 7 und 9 Uhr, eine deutlich geringere Fahrzeuganzahl geschätzt wird als am Nachmittag. Auch die relativ hohen Werte, welche Nachts geschätzt werden, erscheinen nicht plausibel. Um dies überprüfen und korrigieren zu können, werden Verkehrszahlen pro Stunde für sämtliche Tageszeiten benötigt. Eine Ursache für die zu hohen Werte in der Nacht könnte darin bestehen, dass die Risse auch vorhanden sind, wenn die Brücke nicht befahren wird. Dieser Umstand wird bei der Schätzung durch das integrationsbasierte Variationsmaß nicht berücksichtigt. In Abbildung 32 sind die trendbereinigten Rissbreiten $\tilde{X}_{t_n}$ vom 28.07.18 (Teil a)) und vom 05.07.18 (Teil b)), gemessen am Wegaufnehmer WWN1, dargestellt. Im Zeitraum vom 25.07.18 bis zum 20.08.18, und somit auch am 28.07.18, war die Brücke laut einem Artikel von Radio Bochum, vom 25.07.18, für LKW und PKW gesperrt. Von Straßenbahnen, Bussen und Rettungsfahrzeugen wurde die Brücke weiterhin befahren (siehe Radio Bochum, 2018 [14]). Anhand der Grafik ist zu erkennen, dass die Rissbreiten am 28.07.18 deutlich geringer sind als am 05.07.18. Trotzdem werden hier Verkehrszahlen zwischen 188 und 301 Fahrzeugen pro Stunde geschätzt, wenn davon ausgegangen wird, dass am 23.05.17 zwischen 15 und 16 Uhr auch über den nördlichen Überbau 867 Fahrzeuge gefahren sind. Diese Zahlen ähneln der geschätzten Fahrzeuganzahl, die im Median in den Nachtstunden am Wegaufnehmer WWS4 ermittelt wurde. Dies legt die Annahme nahe, dass dort zu dieser Zeit tatsächlich deutlich weniger Fahrzeuge entlang gefahren sind als der geschätzte Wert angibt. Eine genaue Schätzung durch herausrechnen der Rissbreite ohne Verkehrsaufkommen ist mit den gegebenen Daten jedoch nicht möglich, da die Rissdaten während einer Vollsperrung nur für den nörd-



(a) Rissbreite am 28.07.18

(b) Rissbreite am 05.07.18

Abbildung 32: Trendbereinigte Rissbreite $\tilde{X}_{t_n}$, gemessen am Wegaufnehmer WWN1.
a) Am 28.07.18 war die Brücke für PKW und LKW gesperrt, Straßenbahnen, Busse und Rettungsfahrzeuge konnten die Brücke weiterhin befahren.
b) Am 05.07.18 wurde die Brücke normal befahren.

lichen Überbau vorliegen, während die Verkehrszählung lediglich am südlichen Überbau stattgefunden hat. Hinzukommt, dass die Vollsperrung während einem Zeitraum, in dem verhältnismäßig hohe Temperaturen herrschten, stattgefunden hat. Dies führt dazu, dass ohne genaue Kenntnisse der Temperatur innerhalb der Brücke, oder deren genauer Berechnungsweise, keine Umrechnung auf Zeiträume mit anderen Temperaturen stattfinden kann (siehe Abschnitt 7.4.1).

8 Zusammenfassung

Im Rahmen dieser Arbeit, sind drei Methoden zur Erkennung von Anomalien in den Rissdaten des Brückenmonitorings miteinander verglichen worden. Dabei handelt es sich um die Variance Changepoint Detection (VCP) Methode, die Modified Variance Changepoint Detection (MVCP) Methode, welche eine Weiterentwicklung der VCP Methode darstellt, und die Clustering of MAD filtered Data (CMAD) Methode. Die Methoden werden bezüglich der Einheitlichkeit ihrer Ergebnisse und im Vergleich zu manuell bestimmten Anomalien, miteinander verglichen. Zum Vergleich der Einheitlichkeit der Ergebnisse wird

die entsprechende Methode auf Abschnitte von jeweils 2 Wochen angewendet, wobei sich der Beginn des betrachteten Abschnitts immer um eine Woche verschiebt. Bis auf die erste und die letzte Woche, werden somit alle Wochen doppelt ausgewertet. Anschließend werden die doppelt ausgewerteten Wochen auf die Übereinstimmung der Ergebnisse pro Methode überprüft.

Dabei unterscheidet es sich je nach Wegaufnehmer welche Methode am besten abschneidet. Die Fehlerrate für die VCP Methode liegt im Median jedoch immer bei Null und die Fehlerrate der MVCP Methode liegt nur bei den an Wegaufnehmer WOS4 gemessenen Daten im Median leicht über Null. Lediglich die Fehlerrate der CMAD Methode liegt im Median in mehreren Fällen über Null. Daraus kann nicht geschlossen werden, dass VCP die beste der drei Methoden ist. Die häufigen Übereinstimmungen der doppelt ausgewerteten Woche können auch dadurch zustande kommen, dass besonders häufig der Schwellenwert überschritten wird, wodurch davon ausgegangen wird, dass keine Anomalien vorliegen. Ist dies bei beiden Auswertungen der doppelt betrachteten Woche der Fall, wird dies als Übereinstimmung gewertet, auch wenn die Anomalien hier nicht an den gleichen Stellen identifiziert wurden. Dabei wird jedoch nicht überprüft, ob in der betroffenen Woche tatsächlich keine Anomalien vorliegen.

Für den Vergleich zu manuell bestimmten Anomalien werden die Zeiträume vom 03.08. bis 09.08.16 und vom 01.01. bis 07.01.17 am Wegaufnehmer WOS2 und der Zeitraum vom 22.02. bis zum 28.02.17 am Wegaufnehmer WWS4 betrachtet. Die Ergebnisse sind dabei im Sommer deutlich besser als im Winter. Um die Ergebnisse besser beurteilen zu können, sind die Anomalien von zwei verschiedenen Personen unabhängig voneinander, optisch bestimmt worden. Die Abweichungsrate zwischen diesen beiden Zählungen wird als Vergleichswert zur besseren Beurteilung der Qualität der Methoden herangezogen. Zum einen sind die Methoden auf die einzelnen Tage angewendet worden und zum anderen auf eine ganze Woche. Bei der Betrachtung der einzelnen Tage ist ein Schwellenwert von 12 verwendet worden. Am 22.02.17 fällt der Vergleichswert mit nahezu 11 % deutlich höher aus, als die Fehlerraten der Methoden. Daraus lässt sich schließen, dass im Winter eine Fehlerrate von unter 11 % unter Umständen als gutes Ergebnis gewertet werden kann. Am 05.01.17 tritt mit ungefähr 30 % eine außergewöhnlich hohe Fehlerrate bei der CMAD Methode auf. Hier wird ein sehr großer Abschnitt, an dem optisch keine Anomalie erkennbar ist, als Anomalie klassifiziert. Bei der VCP und der MVCP Metho-

de wurden an diesem Tag so viele kleine Abschnitte als Anomalie klassifiziert, dass der Schwellenwert von 12 überschritten wurde und somit davon ausgegangen wird, dass keine Anomalie vorliegt. Am 05.01.17 wird von der CMAD Methode ebenfalls ein größerer Bereich als Anomalie klassifiziert, an dem keine solche vorliegt. Die VCP und die MVCP Methode haben die Anomalie in diesem Fall genau identifiziert.
Werden die Methoden auf eine ganze Woche angewendet, schneidet die CMAD Methode an Wegaufnehmer WOS2 am schlechtesten ab. Dies liegt daran, dass CMAD an diesem Wegaufnehmer im Winter häufig große Bereiche als Anomalie klassifiziert, an denen keine solche vorhanden ist. An Wegaufnehmer WWS4 schneidet die VCP Methode deutlich schlechter ab als die anderen beiden Methoden. Besonders auffällig ist hier der Unterschied zur MVCP Methode, welcher dadurch zustanden kommt, dass ein Teil der falsch klassifizierten Anomalien der VCP Methode bei der MVCP Methode wegfallen, da diese auf der umgekehrten Zeitreihe nicht auftauchen. Es kann davon ausgegangen werden, dass die CMAD Methode bei einer geringen Rissbreite nicht gut mit den regelmäßigen Ausschlägen der Straßenbahn umgehen kann. Daher ist die CMAD Methode für die Verwendung auf der Straßenbahnspur im Winter weniger geeignet.
Die VCP und die MVCP Methode werden zusätzlich hinsichtlich der Klassifikation bezüglich des Vorliegens einer Anomalie an einem Tag und hinsichtlich der Anzahl der klassifizierten Anomalien pro Tag verglichen. Bei Verwendung des Schwellenwertes 12 verringert sich die Fehlklassifikationsrate, bezüglich des Vorliegens einer Anomalie an einem Tag, für beide Methoden deutlich. Wird die VCP Methode ohne Schwellenwert verwendet, klassifiziert sie an jedem Tag eine Anomalie. Werden die Methoden auf 7 aufeinanderfolgende Tage angewendet, wobei nur der jeweils mittlere Tag betrachtet wird, sind die Ergebnisse ohne Verwendung eines Schwellenwertes besser als bei tageweiser Anwendung. Ein gutes Ergebnis wird mit diesem Vorgehen jedoch nicht erzielt. Bei Verwendung des Schwellenwertes 12 wird bei der tageweisen Anwendung der Methoden ein besseres Ergebnis erzielt. Das Klassifikationsergebnis soll mit Hilfe der logistischen Regression weiter verbessert werden. Wird dabei ab einer Wahrscheinlichkeit von 50 % das Vorliegen einer Anomalie klassifiziert, wird für den Fall, dass keine Anomalie vorliegt, fast immer richtig klassifiziert. Die Tage an denen Anomalien vorhanden sind werden hingegen zu 40 % falsch klassifiziert. Dies kann darauf zurück geführt werden, dass insgesamt mehr Tage ohne Anomalien vorliegen. Das Gesamtergebnis wird also besser,

wenn im Zweifelsfall für *keine Anomalie* entschieden wird. Durch Herabsetzen der Wahrscheinlichkeit ab welcher das Vorliegen einer Anomalie klassifiziert wird, kann die Fehlklassifikationsrate für den Fall, dass eine Anomalie vorliegt, verringert werden. Wird eine Wahrscheinlichkeit von 30 % als Grenzwert verwendet, ist die Fehlklassifikationsrate für den Fall, dass Anomalien auftreten und für den Fall, dass keine Anomalien auftreten, in etwa gleich hoch. Werden die Klassifikationsergebnisse nach Jahreszeiten getrennt betrachtet fällt auf, dass im Sommer die besten und im Winter die schlechtesten Ergebnisse erzielt werden. Beim Vergleich der Anzahl der von den Methoden klassifizierten Anomalien mit der Zahl der tatsächlich vorliegenden Anomalien, kann bis zu einer absoluten Differenz von 4 ein Abfallen der Häufigkeiten festgestellt werden. Danach steigt die Häufigkeit des Vorkommens mit dem Anstieg des Wertes der absoluten Differenz wieder an. Dies legt die Vermutung nahe, dass geringe Differenzen auftreten, wenn tatsächlich Anomalien vorliegen und große Differenzen auftreten, wenn keine Anomalien vorliegen, die Methoden aber welche identifizieren. Aufgrund der Ergebnisse des Vergleichs wird eine Vorentscheidung mit Hilfe der logistischen Regression mit Grenzwert 30 % und anschließende Verwendung der MVCP Methode empfohlen.

Auf den von Anomalien befreiten Rissdaten wird eine Verkehrsschätzung mit Hilfe des integrationsbasierten Variationsmaßes durchgeführt. Werden die, zur Zeit der Verkehrszählung am 23.05.17 von 15 bis 16 Uhr, gegebenen Daten als Referenzwert verwendet, wird die Fahrzeuganzahl zu dieser Uhrzeit im Median deutlich unterschätzt. Bei Verwendung der, zur Zeit der Verkehrszählung am 26.09.17 von 15 bis 16 Uhr, gegebenen Daten als Referenzwert, wird die Fahrzeuganzahl zu dieser Uhrzeit im Median deutlich überschätzt. Bei Verwendung des Mittelwerts aus den beiden Zählungen und den entsprechenden Flächeninhalten, ist die Schätzung zwischen 15 und 16 Uhr im Median näher an den tatsächlich zu dieser Uhrzeit gezählten Fahrzeugzahlen. Eine Verbesserung kann durch eine Aufteilung in drei Gruppen (*warm*, *mittel*, *kalt*), anhand der Temperatur und der Rissbreite, erzielt werden. Dabei wird die Fahrzeuganzahl innerhalb einer Stunde, welche der Kategorie *warm* zugeordnet ist, mit den Daten der Zählung vom 23.05.17 geschätzt. Die Fahrzeuganzahl innerhalb einer Stunde welche der Kategorie *mittel* angehört, wird mit den Daten der Zählung vom 26.09.17 geschätzt. Stunden die der Kategorie *kalt* angehören, können in diesem Fall nicht mit Hilfe des integrationsbasierten Varia-

tionsmaßes geschätzt werden, da für diese Gruppe keine Verkehrszählung vorliegt. Die Schätzung zwischen 15 und 16 Uhr der Kategorien *mittel* und *warm* entspricht dann im Median in etwa der bei den zu dieser Zeit vorgenommenen Verkehrszählungen ermittelten Anzahl. Um eine Verkehrsschätzung innerhalb von Stunden, die der Kategorie *kalt* zugeordnet sind durchzuführen, wird ein lineares Modell verwendet. Das Modell wird auf den Daten der Kategorien *mittel* und *warm* angepasst, wobei die zugehörigen Fahrzeuganzahlen durch eine Schätzung mit Hilfe des integrationsbasierten Variationsmaßes ermittelt worden sind. Damit werden für die Kategorie *kalt* zwischen 15 und 16 Uhr Werte geschätzt, die im Median leicht unter dem zu dieser Zeit gezählten Wert liegen. Der Interquartilsabstand ist hier jedoch sehr hoch. Es ist dabei zu beachten, dass ein lineares Modell nur so gut sein kann, wie die Daten auf denen es basiert. Nur weil die Schätzungen mit dem integrationsbasierten Variationsmaß für die Kategorien *warm* und *kalt* zwischen 15 und 16 Uhr im Median in etwa den Werten der Verkehrszählungen entsprechen, bedeutet dies nicht, dass die Schätzungen zu den anderen Uhrzeiten ebenfalls in etwa dem wahren Verkehrsaufkommen entsprechen. Vor allem Nachts geschätzte Werte erscheinen intuitiv etwas zu hoch. Außerdem fällt auf, dass zu den typischen Stoßzeiten des Berufsverkehrs nicht das größte Verkehrsaufkommen geschätzt wird. Zudem fallen die während der Berufsverkehrszeit am Morgen geschätzten Werte im Median deutlich geringer aus, als die zur Berufsverkehrszeit am Nachmittag geschätzten Werte. Die Nachts geschätzten Werte entsprechen in etwa dem Wert, welcher bei der für PKW und LKW gesperrten Brücke geschätzt wird, wenn davon ausgegangen wird, dass am 23.05.17 zwischen 15 und 16 Uhr über den nördlichen Überbau ebenfalls 867 Fahrzeuge gefahren sind. Dieser Umstand erhärtet den Verdacht, dass die in der Nacht ermittelten Werte den wahren Wert stark überschätzen. Ein weiteres Problem sind die großen Unterschiede bezüglich der Rissbreiten im Sommer und im Winter. Im Winter ist die Rissbreite deutlich geringer als im Sommer. Diese Veränderung der Rissbreite wird von der Temperatur innerhalb der Brücke beeinflusst. Die Temperatur im Inneren der Brücke wird jedoch nicht gemessen und es ist nicht bekannt, ob und wenn wie sich diese Temperatur aus den gegebenen Daten zusammensetzt. Zudem liegen zu wenige Verkehrszählungen vor um die Temperatur richtig berücksichtigen zu können. Für eine Zuverlässige Schätzung werden Verkehrszählungen zu allen Jahres- und Tageszeiten, sowie Messungen der Temperatur innerhalb der Brücke, benötigt. Mit den gegeben Daten ist somit keine zuverlässige Verkehrsschätzung

möglich. Es kann somit auch keine Aussage getroffen werden, ob sich das Verkehrsaufkommen in Abhängigkeit von der Jahreszeit ändert.
Bei der Verkehrsschätzung ist eine Vorentscheidung mit Logistischer Regression, wobei ab einer Wahrscheinlichkeit von 30 % vom Vorliegen von Anomalien ausgegangen wird, und die anschließende Verwendung der MVCP Methode, zur Entfernung der Anomalien ausreichend. Das Ergebnis der Verkehrsschätzung unterscheidet sich dabei nur geringfügig von den Schätzwerten nach Entfernung der Anomalien durch die MVCP Methode, wobei diese nur auf die Tage angewendet wird, an denen optisch eine Anomalie erkennbar ist.
Bei der Verkehrsschätzung, auf den Daten vom 01.06.16 bis zum 31.05.17, ist zwischen 23 und 0 Uhr ein extremer Ausreißer aufgefallen. Der Grund dafür ist die bei der VCP Methode getroffene Annahme, dass das Cluster, welches mehr Abschnitte enthält, die Gruppe ohne Anomalien enthält. Da dieses Problem nur für einen Fall, und erst am Ende der Analyse aufgefallen ist, kann davon ausgegangen werden, dass die entsprechende Annahme an den meisten Tagen erfüllt wird. Wird diese Annahme durch die Annahme, dass das Clusterzentrum der Gruppe mit Anomalien den größeren Wert hat, ersetzt, tritt dieser Fehler nicht mehr auf. Eine erneute Berechnung der Fehlerraten der VCP und der MVCP Methode kann Aufschluss darüber geben wie groß der Einfluss der veränderten Annahme tatsächlich ist.

Ein weiteres Ziel ist die Erläuterung der `cpt.var()` Funktion aus dem Paket `changepoint`. Dabei hat sich herausgestellt, dass die `cpt.var()` Funktion sich in der Art und Weise wie die auszugebenden Changepoints ermittelt werden von den zugrundeliegenden Methoden unterscheidet. Eine erneute Implementierung der PELT Methode, welche sich an die Definition der Methode hält, kann Aufschluss über die Auswirkungen der vorgenommenen Änderungen in Bezug auf die VCP Methode geben.

Um zukünftig eine zuverlässige Verkehrsschätzung durchführen zu können, sollte bei einem erneuten Brückenmonitoring die Temperatur im Inneren der Brücke gemessen werden. Dies kann auch hilfreich für die Modellierung des Temperatureinflusses auf die Rissbreiten sein, was wiederum für die Ermittelung von plötzlichen gravierenden Änderungen der Rissbreite von Vorteil sein kann.

Anhang

A Tabellen

Tabelle A.1: Anteil der Abweichung einer Methode bei Bestimmung der Anomalien innerhalb von 3 Wochen, wobei die Methoden zuerst auf die ersten beiden Wochen und anschließend auf die letzten beiden Wochen angewendet werden. Die doppelt berechneten Ergebnisse der mittleren Woche werden miteinander verglichen.

Woche	CMAD			VCP			MVCP		
	WOS2	WOS4	WWS4	WOS2	WOS4	WWS4	WOS2	WOS4	WWS4
08.06.16.-14.06.16	0.000483	0.005946	0.000000	0.000000	0.000605	0.000000	0.000000	0.001382	0.000000
15.06.16-21.06.16	0.004560	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
22.06.16-28.06.16	0.002814	0.007417	0.000661	0.000000	0.020047	0.016072	0.000000	0.019739	0.000020
29.06.16-05.07.16	0.001058	0.000463	0.000000	0.000003	0.000010	0.000000	0.000003	0.000000	0.000000
06.07.16-12.07.16	0.002761	0.000847	0.000529	0.002573	0.000000	0.016968	0.002573	0.000000	0.016968
13.07.16-19.07.16	0.004669	0.001250	0.003598	0.001713	0.034720	0.000000	0.001713	0.034697	0.000000
20.07.16-26.07.16	0.007497	0.001779	0.008108	0.008773	0.000304	0.004855	0.036119	0.000298	0.004851
27.07.16-02.08.16	0.000000	0.001323	0.000000	0.000000	0.002398	0.000000	0.000000	0.002404	0.000000
03.08.16-09.08.16	0.002672	0.000794	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
10.08.16-16.08.16	0.001538	0.000833	0.004401	0.000000	0.001637	0.000000	0.000000	0.001637	0.000000
17.08.16-23.08.16	0.000132	0.001852	0.023641	0.005093	0.000000	0.000000	0.005093	0.001518	0.000000
24.08.16-30.08.16	0.035608	0.000794	0.000529	0.017173	0.000000	0.017689	0.016941	0.000000	0.000549
31.08.16-06.09.16	0.000000	0.002083	0.001700	0.016088	0.006306	0.000000	0.015483	0.001250	0.000000
07.09.16-13.09.16	0.000556	0.004041	0.000833	0.023142	0.011905	0.000000	0.022504	0.011885	0.000840
14.09.16-20.09.16	0.007758	0.000417	0.002646	0.000000	0.000000	0.013625	0.000000	0.000000	0.012808
21.09.16-27.09.16	0.002490	0.000476	0.017589	0.002550	0.018069	0.000000	0.002550	0.018069	0.000000
28.09.16-04.10.16	0.003158	0.000251	0.002738	0.028602	0.008317	0.000000	0.028456	0.008317	0.000000
05.10.16-11.10.16	0.000000	0.000635	0.012358	0.000000	0.046197	0.000000	0.000222	0.045467	0.000000
12.10.16-18.10.16	0.000000	0.001224	0.001032	0.000000	0.000000	0.000000	0.000030	0.000000	0.000000
19.10.16-25.10.16	0.020314	0.000698	0.003009	0.000000	0.032093	0.000000	0.000000	0.031845	0.049011
26.10.16-01.11.16	0.040060	0.001713	0.003307	0.000000	0.065157	0.000000	0.000000	0.064921	0.012152
02.11.16-08.11.16	0.263846	0.037794	0.008654	0.022798	0.005119	0.000000	0.022490	0.000503	0.000000
09.11.16-15.11.16	0.175784	0.056257	0.000370	0.041174	0.000000	0.001534	0.003416	0.002355	0.000000
16.11.16-22.11.16	0.000000	0.000635	0.005827	0.001653	0.019197	0.003188	0.022087	0.000060	0.006078
23.11.16-29.11.16	0.025612	0.007940	0.000794	0.000010	0.000542	0.000000	0.000529	0.000179	0.030096
30.11.16-06.12.16	0.001587	0.117804	0.000198	0.059468	0.004828	0.000000	0.058621	0.004828	0.046234
07.12.16-13.12.16	0.001472	0.059656	0.000099	0.021478	0.000000	0.000000	0.021472	0.000000	0.027880
14.12.16-20.12.16	0.154755	0.000000	0.000595	0.030509	0.000000	0.000000	0.030509	0.000046	0.016485
21.12.16-27.12.16	0.001190	0.066131	0.000694	0.000000	0.000000	0.000000	0.000000	0.000413	0.027682
28.12.16-03.01.17	0.034223	0.017748	0.000278	0.025827	0.004514	0.003770	0.000000	0.004507	0.003770
04.01.17-10.01.17	0.044940	0.000000	0.000000	0.000152	0.005370	0.000000	0.040618	0.005086	0.000000
11.01.17-17.01.17	0.065509	0.087887	0.000648	0.024865	0.000000	0.029322	0.024421	0.000281	0.028102
18.01.17-24.01.17	0.001538	0.015966	0.000000	0.095814	0.000000	0.000000	0.000936	0.001310	0.000000
25.01.17-31.01.17	0.000496	0.001726	0.001091	0.000000	0.000000	0.043932	0.004544	0.001028	0.043671
01.02.17-07.02.17	0.314226	0.047603	0.000496	0.000000	0.000000	0.017649	0.002093	0.000268	0.017649
08.02.17-14.02.17	0.216577	0.029269	0.078677	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
15.02.17-21.02.17	0.000000	0.005466	0.009808	0.000000	0.000000	0.000000	0.000000	0.000000	0.017854
22.02.17-28.02.17	0.000000	0.000694	0.000992	0.000000	0.000000	0.000000	0.000000	0.000007	0.046521
01.03.17-07.03.17	0.000000	0.038604	0.009461	0.000000	0.000000	0.000000	0.000000	0.039233	0.000000
08.03.17-14.03.17	0.000000	0.032464	0.001687	0.000000	0.000000	0.000000	0.000000	0.034511	0.000000
15.03.17-21.03.17	0.000000	0.004296	0.000787	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
22.03.17-28.03.17	0.001204	0.001190	0.046653	0.005779	0.000000	0.000000	0.005779	0.000000	0.000000
29.03.17-04.04.17	0.000899	0.001104	0.000000	0.058843	0.000000	0.000000	0.000000	0.000000	0.000000
05.04.17-11.04.17	0.003896	0.001237	0.056045	0.004577	0.000000	0.000000	0.004570	0.000000	0.000000
12.04.17-18.04.17	0.000000	0.002050	0.004015	0.000000	0.000000	0.000000	0.000000	0.000000	0.015784
19.04.17-25.04.17	0.000000	0.035599	0.001693	0.000000	0.000000	0.000000	0.000000	0.000000	0.030242
26.04.17-02.05.17	0.000000	0.027216	0.027741	0.000000	0.000000	0.000000	0.000000	0.039984	0.000000
03.05.17-09.05.17	0.000000	0.017004	0.000000	0.000000	0.000000	0.000000	0.000000	0.014749	0.000000
10.05.17-16.05.17	0.002341	0.002907	0.000000	0.020301	0.041132	0.000000	0.020281	0.007193	0.000000
17.05.17-23.05.17	0.004236	0.000628	0.000251	0.000000	0.000000	0.004590	0.000000	0.000003	0.004590

Tabelle A.2: Uhrzeiten von entdeckten Anomalien bei Auszählung in der Woche vom 03.08.2016 bis 09.08.2016, am Wegaufnehmer WOS2. Tage aus diesem Zeitraum, die nicht in der Tabelle aufgeführt sind, weisen keine Anomalien auf.

	Anne	Johanna
05.08.2016	05:30-08:22	05:30-08:30
	10:05-10:30	10:05-10:31
06.08.2016	05:20-05:47	05:18-05:49
	06:45-08:17	06:40-08:20

Tabelle A.3: Uhrzeiten von entdeckten Schwebungen bei Auszählung in der Woche vom 01.01.2017 bis 07.01.2017, für den Wegaufnehmer WOS2. Tage aus diesem Zeitraum, die nicht in der Tabelle aufgeführt sind, weisen keine Anomalien auf.

	Anne	Johanna
01.01.2017	01:05:00-01:28:00	01:05:00-01:27:30
04.01.2017	04:50:00-05:05:00	04:51:00-05:03:20
05.01.2017	06:37:00-06:58:00	06:36:00-07:08:00
06.01.2017	04:45:00-05:00:00	04:42:20-05:00:00
07.01.2017	23:40:00-23:59:58	23:33:20-23:59:58

Tabelle A.4: Uhrzeiten von entdeckten Schwebungen bei Auszählung in der Woche vom 22.02.2017 bis 28.02.2017, für den Wegaufnehmer WWS4. Tage aus diesem Zeitraum, die nicht in der Tabelle aufgeführt sind, weisen keine Anomalien auf.

	Anne	Johanna
22.02.2017	03:50:00-04:38:00	03:43:20-04:50:00
		05:14:20-05:40:00
	06:02:00-07:01:30	06:00:00-07:03:20
		14:13:20-15:28:20
	17:55:00-19:22:00	17:28:20-19:23:20
	22:15:00-23:48:30	22:11:40-23:50:00
23.02.2017	03:50:02-04:13:00	03:50:00-04:15:00
25.02.2017	03:52:00-04:25:00	03:46:40-04:33:20
	10:20:00-10:35:00	10:18:20-10:35:00
27.02.2017	05:42:00-06:05:00	05:40:00-06:05:00
	08:10:00-08:50:00	07:38:20-08:55:00

Tabelle A.5: Fehlerraten der Methoden im Vergleich zur optischen Identifikation der Anomalien. Die Werte wurden anhand von an Wegaufnehmer WOS2 gemessenen Daten berechnet.

	VCP		MVCP		CMAD		Vergleich
Datum	Anne	Johanna	Anne	Johanna	Anne	Johanna	Zählungen
05.08.16	0.0068	0.0103	0.0068	0.0103	0.0087	0.0034	0.0063
06.08.16	0.0085	0.0142	0.0076	0.0159	0.0111	0.0038	0.0083
01.01.17	0.0081	0.0077	0.0081	0.0077	0.0087	0.0091	0.0003
04.01.17	0.0104	0.0086	0.0104	0.0086	0.0104	0.0086	0.0019
05.01.17	0.0146	0.0222	0.0146	0.0222	0.3026	0.3056	0.0076
06.01.17	0.0018	0.0035	0.0017	0.0035	0.0228	0.0211	0.0017
07.01.17	0.0004	0.0042	0.0004	0.0043	0.0352	0.0305	0.0046

Tabelle A.6: Fehlerraten der Methoden im Vergleich zur optischen Identifikation der Anomalien. Die Werte wurden anhand von an Wegaufnehmer WWS4 gemessenen Daten berechnet.

	VCP		MVCP		CMAD		Vergleich
Datum	Anne	Johanna	Anne	Johanna	Anne	Johanna	Zählungen
22.02.17	0.0789	0.0697	0.0479	0.0953	0.0600	0.0638	0.1084
23.02.17	0.0146	0.0174	0.0146	0.0174	0.0025	0.0025	0.0028
25.02.17	0.0059	0.0166	0.0059	0.0166	0.0107	0.0057	0.0106
27.02.17	0.0105	0.0252	0.0104	0.0252	0.0386	0.0132	0.0269

Tabelle A.7: Fehlerraten zur Identifikation der Schwebungen. In den Wochen vom 03.08.16 - 09.08.16 (Sommer, WOS2), 01.01.17 - 07.01.17 (Winter, WOS2) und 22.02.17 - 28.02.17 (Winter, WWS4).

	CMAD		VCP		MVCP		Vergleich
	Anne	Johanna	Anne	Johanna	Anne	Johanna	Zählungen
Sommer WOS2	0.0058	0.0037	0.0022	0.0035	0.0021	0.0038	0.0021
Winter WOS2	0.0976	0.0964	0.0211	0.0219	0.0210	0.0219	0.0023
Winter WWS4	0.0185	0.0112	0.2571	0.2398	0.0118	0.0245	0.0212

Tabelle A.8: Fehlerraten bezüglich des Auftretens von Anomalien, für den Zeitraum vom 01.06.16 bis zum 31.05.17. *Woche* bezeichnet die Fehlerraten für die Auswertung einer ganzen Woche, wobei immer nur der mittlere Tag betrachtet wird. *Tag*, steht für die getrennte Auswertung der einzelnen Tage. Erfolgt die Auswertung mit Schwellenwert wird davon ausgegangen, dass an Tagen an denen mehr als 12 Anomalien durch die jeweilige Methode aufgedeckt werden, in Wahrheit keine Anomalien vorliegen.

	ohne Schwellenwert		mit Schwellenwert	
Auswertung	MVCP	VCP	MVCP	VCP
Woche	0.5320	0.5655	0.3092	0.2618
Tag	0.7205	0.7342	0.3041	0.2411

Tabelle A.9: Fehlerraten getrennt je nachdem ob eine Anomalie vorliegt oder ob keine Anomalie vorliegt. *Anomalie* bezeichnet dabei den Anteil der Fälle in dem klassifiziert wird, dass eine Anomalie vorliegt obwohl dies nicht der Fall ist und *keine Anomalie* bezeichnet den Fall, dass keine Anomalie klassifiziert wird obwohl an dem entsprechenden Tag eine vorhanden ist. Als Schwellenwert wird auch hier der Wert 12 gewählt, was sich aus der maximalen Anzahl an Anomalien pro Tag multipliziert mit 1.5 zusammensetzt. Die Fehlerrate der logistischen Regression ist mit der Leave-One-Out Methode approximiert worden, wobei eine Anomalie klassifiziert wird, wenn die Wahrscheinlichkeit für das auftreten einer Anomalie mit größer als 50 % vorhergesagt wird.

falsches Klassifikationsergebnis	keine Anomalie	Anomalie
logistische Regression	0.443	0.082
VCP ohne Schwellenwert	0.000	1.000
MVCP ohne Schwellenwert	0.021	0.974
VCP mit Schwellenwert	0.206	0.254
MVCP mit Schwellenwert	0.144	0.362

Tabelle A.10: Fehlklassifikationsraten der logistischen Regression, bei Herabsetzung des Schwellenwertes, sowie die Fehlklassifikationsraten der Methoden, bei Erhöhung des Schwellenwertes. Dabei wird unterschieden ob an den entsprechenden Tagen eine Anomalie oder keine Anomalie vorliegt. Die Abkürzung *SW* steht dabei für Schwellenwert und *f.K.* steht für falsches Klassifikationsergebnis.

Logistische Regression			Methoden mit Schwellenwert				
f.K.	keine Anomalie	Anomalie	f.K.	keine Anomalie		Anomalie	
SW			SW	VCP	MVCP	VCP	MVCP
0.4	0.299	0.146	13	0.206	0.144	0.302	0.414
0.3	0.216	0.220	14	0.196	0.134	0.351	0.455
0.25	0.196	0.269	15	0.196	0.134	0.396	0.478
0.2	0.175	0.310	16	0.196	0.134	0.459	0.530

Tabelle A.11: Fehlklassifikationsraten der logistischen Regression und der Methoden mit Schwellenwert, getrennt nach Jahreszeiten. Der hinter der Methode in Klammern angegebene Wert gibt den verwendeten Schwellenwert an. Für jede Jahreszeit wird die Fehlklassifikationsrate für den Fall, dass keine Anomalie klassifiziert wird (*k.A.*) obwohl eine solche vorliegt und für den Fall, dass fälschlicherweise eine Anomalie klassifiziert wird (*A.*), angegeben.

	Frühling		Sommer		Herbst		Winter	
falsches Klassifikationsergebnis	k.A.	A.	k.A.	A.	k.A.	A.	k.A.	A.
Logistische Regression (0.3)	0.214	0.250	0.053	0.055	0.250	0.269	0.308	0.328
Logistische Regression (0.25)	0.214	0.328	0	0.082	0.208	0.328	0.308	0.359
Logistische Regression (0.2)	0.143	0.359	0	0.110	0.208	0.358	0.308	0.438
VCP (12)	0.250	0.172	0	0.105	0.208	0.358	0.308	0.359
MVCP (12)	0.143	0.344	0	0.211	0.208	0.448	0.192	0.422

B R output

R output B.1: lastchange Vektor 22.02.17 05 bis 09 Uhr WWS4

```
> lastchange_220217_5bis9_WWS4
  [1]    0  167  171  170  200  344  345  348  434  437  436  508  512  568  569  572  753  755  770  852
 [21]  863  961  962  965 1058 1062 1060 1093 1094 1097 1096 1114 1122 1309 1310 1312 1313 1335 1372 1373
 [41] 1375 1397 1399 1401 1606 1613 1617 1718 1719 1724 1723 1810 1811 1816 1912 1928 2009 2010 2194 2195
 [61] 2176 2398 2983 2984 3004 3133 3135 3377 3572 3630 3634 3632 3728 3730 3803 3820 3859 3860 3863 3862
 [81] 4004 4007 4032 4172 4173 4180 4184 4203 4206 4231 4229 4266 4306 4354 4389 4407 4403 4494 4497 4479
[101] 4506 4565 4587 4613 4614 4637 4657 4691 4762 4764 4792 4849 4903 4992 5084 5062 5095 5128 5159 5168
[121] 5169 5167 5189 5256 5257 5260 5296 5298 5310 5308 5334 5353 5363 5397 5481 5508 5511 5560 5589 5590
[141] 5595 5597 5616 5662 5668 5667 5706 5707 5710 5724 5725 5729 5728 5755 5802 5803 5805 5842 5885 5886
[161] 5888 5925 5926 5934 5927 6018 6085 6097 6064 6119 6165 6206 6278 6280 6282 6322 6324 6359 6362 6399
[181] 6401 6403 6434 6432 6529 6560 6602 6604 6654 6683 6690 6697 6689 6782 6785 6788 6816 6875 6882 6898
[201] 6903 6926 6928 6924 7004 7005 7008 7042 7062 7073 7075 7106 7127 7126 7168 7169 7171 7186 7187 7189
```

R output B.2: cpt.var 22.02.17 05 bis 09 Uhr WWS4

```
> erg_cptvar_220217_5bis9_WWS4
  [1]  167  171  345  348  434  436  508  512  569  572  753  770  852  863  962  965 1058 1060 1094 1096
 [21] 1122 1310 1335 1373 1375 1399 1401 1613 1617 1719 1723 1811 1816 1912 1928 2176 2984 3004 3133 3135
 [41] 3572 3630 3632 3728 3730 3860 3862 4004 4007 4173 4184 4203 4231 4354 4389 4494 4497 4565 4637 4657
 [61] 4762 4764 4792 5084 5095 5128 5159 5169 5189 5256 5310 5334 5397 5481 5508 5511 5560 5590 5595 5662
 [81] 5667 5707 5710 5725 5728 5755 5803 5805 5886 5888 6018 6064 6280 6282 6322 6324 6359 6362 6401 6403
[101] 6602 6604 6654 6683 6689 6782 6785 6875 6882 6898 7005 7008 7042 7106 7126 7169 7171 7187 7189 7200
```

Literatur

[1] Bender, R., Ziegler, A. und Lange, S. (2007): Logistische Regression, *DMW - Deutsche Medizinische Wochenschrift* **S01**, S.e33 - e35.

[2] Bundesministerium für Verkehr, Bau und Stadtentwicklung (2011): *Richtlinie zur Nachrechnung von Straßenbrücken im Bestand (Nachrechnungsrichtlinie)*, `https://www.bast.de/BASt_2017/DE/Ingenieurbau/Publikationen/Regelwerke/Entwurf/Nachrechnungsrichtlinie-Ausgabe-5_2011.html` (Abfrage am 21.01.2019)

[3] Bundesministerium für Verkehr, Bau und Stadtentwicklung (2013): *Bauwerksprüfung nach DIN 1076 Bedeutung, Organisation, Kosten*, `https://www.bmvi.de/SharedDocs/DE/Anlage/VerkehrUndMobilitaet/Strasse/dokumentation-bauwerkspruefung-nach-din-1076.pdf?__blob=publicationFile` (Abfrage am 21.01.2019)

[4] Canty, A. und Ripley, B. (2017): boot: Bootstrap R (S-Plus) Functions, R package version 1.3-20.

[5] Chen, J. und Gupta, A.K. (2000): *Parametric Statistical Change Point Analysis*, Birkhäuser, Boston, ISBN 0-8176-4169-6.

[6] Härdle, W. und Steiger, W. (1995): Optimal Median Smoothing, *Applied Statistics* **44**(2), S.258-268.

[7] Hastie, T., Tibshirani, R. und Friedman, J. (2009): *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2. Auflage, Springer, New York, ISBN 978-0-387-84858-7.

[8] Heinrich, J. (2016): Ergebnisse aus dem Rissmonitoring. Bericht-Nr. 001 der König und Heunisch Planungsgesellschaft. (unveröffentlicht)

[9] Hosmer, D. W., Lemeshow, S. und Sturdivant, R., X. (2013): *Applied Logstic Regression*, 3. Auflage, John Wiley & Sons Inc., Hoboken, New Jersey, ISBN 978-0-470-58247-3.

[10] Jackson, B., Scargle, J. D., Barnes, D., Arabhi, S., Alt, A., Gioumousis, P., Gwin, E., Sangtrakulcharoen, P., Tan, L. und Tsai, T. T. (2005): An Algorithm for Optimal Partitioning of Data on an Interval, *IEEE Signal Processing Letters* **12**(2), 105-108

[11] Killick, R., Fearnhead, P. und Eckley, I. A. (2012): Optimal Detection of Changepoints With a Linear Computational Cost, *Journal of the American Statistical Association* **107** (500), 1590 - 1598

[12] Killick, R., Heynes, K. und Eckley, I. A. (2016): changepoint: An R Package for Changepoint Analysis. R package Version 2.2.2, `https://CRAN.R-project.org/package=changepoint`

[13] Kohlenbach, J. (2017): Verkehrsschätzung mittels Rissweitenänderung bei einer Brücke in Bochum, *Projektbericht im Rahmen der Veranstaltung Fallstudien II*. (unveröffentlicht)

[14] Radio Bochum (2018): Lokalnachrichten: *Die Brücke Wittener Straße über den Sheffieldring ist ab heute auch stadteinwärts gesperrt.* `https://www.radiobochum.de/bochum/lokalnachrichten/lokalnachrichten/article/-2c87eee164.html` (Abfrage am 08.01.2019)

[15] R Core Team (2017): R: A Language and Environment for Statistical Computing, *R Foundation for Statistical Computing*, Vienna, Austria. `https://www.R-project.org/`

[16] Tukey, J. W. (1977): *Exploratory Data Analysis*, Addison-Wesley Publishing Company, ISBN: 0-201-07616-0.

[17] Zhang, N. R. und Siegmund, D. O. (2007): A Modified Bayes Information Criterion with Applications to the Analysis of Comparative Genomic Hybridization Data, *Biometrics* **63**, 22 -32