

Bachelorarbeit  
Einsatz von Prozessanalysen für maschinelle Lernverfahren  
zur Anomalieerkennung in Multibeam Daten

-

Sommersemester 2022

Erstbetreuer: Tobias Ziolkowski

Manuel Krebs

22. August 2022

# Inhaltsverzeichnis

<b>1</b>	<b>Einführung</b>	<b>1</b>
<b>2</b>	<b>Literaturübersicht</b>	<b>2</b>
<b>3</b>	<b>Methodik</b>	<b>3</b>
3.1	Interquartilsabstand . . . . .	3
3.1.1	Funktionsweise . . . . .	3
3.1.2	Anwendung . . . . .	4
3.1.3	Auswertung . . . . .	4
3.2	Isolation Forest . . . . .	5
3.2.1	Funktionsweise . . . . .	5
3.2.2	Anwendung . . . . .	5
3.2.3	Vorteile . . . . .	5
3.2.4	Auswertung . . . . .	5
3.2.5	Limits/Challenges? . . . . .	5
3.3	Local Outlier Factor . . . . .	6
3.3.1	Funktionsweise . . . . .	6
3.3.2	Anwendung . . . . .	6
3.3.3	Auswertung . . . . .	6
3.3.4	Limits/Challenges? . . . . .	6
3.4	One-Class Support Vector Machine . . . . .	7
3.4.1	Funktionsweise . . . . .	7
3.4.2	Anwendung . . . . .	7
3.4.3	Auswertung . . . . .	7
<b>4</b>	<b>Ergebnisse</b>	<b>8</b>
<b>5</b>	<b>Diskussion</b>	<b>9</b>
<b>6</b>	<b>Fazit</b>	<b>10</b>

---

# Abbildungsverzeichnis

3.1	Beispiel eines Boxplots . . . . .	4
-----	-----------------------------------	---

# Kapitel 1

## Einführung

- Multibeam Daten -> Definition
- Prozessanalysen für maschinelle Lernverfahren -> Definition und warum wichtig
- Anomalien -> Definition, Beispiele und Erklärung, dass Anomalie = Outlier
- Sagen, was ich mit dem Paper erreichen möchte -> Forschungsfrage
- <https://sci-hub.hkvisa.net/10.3182/20130902-3-cn-3020.00044> -> Sliding Window

Outliers are the patterns which are not in the range of normal behavior [5]

Moving window pattern!  
Chandola, V., Banerjee, A., Kumar, V.: Outlier detection: a survey

”what techniques were used in the past and how machine learning could help us now and in the future for a better bathymetric data processing”

Unterschied lokale und globale Outlier!

Outlier detection (or data cleaning) is an important step in data processing because, if an outlier data point is used during data mining, it is likely to lead to inaccurate outputs -> Ramírez-Gallego, S.; Krawczyk, B.; García, S.; Woźniak, M.; Herrera, F. A survey on data preprocessing for data stream mining: Current status and future directions. Neurocomputing 2017, 239, 39–57. [CrossRef]

## Kapitel 2

# Literaturübersicht

Andere Anwendungsgebiete!

Erklären, wie Multibeamdaten erfasst werden Erklären, wie Outlier meistens erkannt werden.

## Kapitel 3

# Methodik

### 3.1 Interquartilsabstand

Der Interquartilsabstand (im folgenden mit IQR angekürzt) ist ein Streuungsmaß, welcher in statistischen Analysen dabei hilft, durch Verteilungen Rückschlüsse über einen Datensatz zu ziehen. Visuell werden diese meist durch einen Boxplot dargestellt, welche zu den am weitesten verbreiteten Werkzeugen in der statistischen Praxis, insbesondere in der Phase der explorativen Datenanalyse [3].

#### 3.1.1 Funktionsweise

Zunächst wird das erste und das dritte Quartil berechnet. Der IQR bildet sich nun aus der Differenz des der beiden Quartile, also  $\text{IQR} = \text{Quartil } 3 - \text{Quartil } 1$ . Um dies als graphische Repräsentation zu verdeutlichen zeigt Grafik 3.1 einen Boxplot, welcher die 'z'-Koordinate der ersten 1000 Messdaten veranschaulicht. Der IQR lässt sich nun durch die blau eingefärbte Fläche zeigen, welche 50 Prozent der Messdaten enthält. Alle Messwerte, welche außerhalb oberen und unteren Grenze liegen, lassen sich nun als potenzielle Outlier identifizieren. Diese Grenzen werden typischerweise auf das 1,5-fache des IQR gesetzt. Die untere Grenze lässt sich also durch  $\text{Quartil } 1 - (1.5 * \text{IQR})$  und die obere Grenze durch  $\text{Quartil } 3 + (1.5 * \text{IQR})$  berechnen [5].

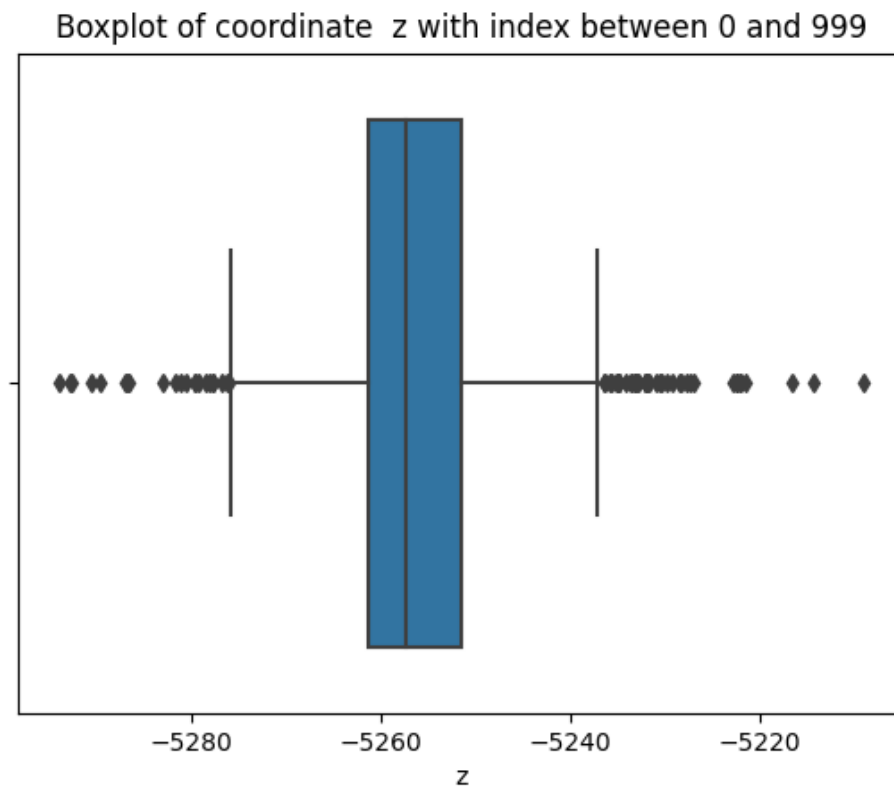


Abbildung 3.1: Beispiel eines Boxplots

Wie in der Grafik 3.1 nun zu erkennen ist, befinden sich einige der Messwerte nur knapp außerhalb der Grenzen. bei den Werten weit außerhalb der Grenzen ist die Wahrscheinlichkeit am höchsten, dass diese Ausreißer sind.

### 3.1.2 Anwendung

Bevor die Methodik angewendet werden kann muss der Datensatz angepasst werden. Dies geschieht mithilfe dem zuvor beschrieben Ansatz des Moving-Window-Patterns. Nachdem der Datensatz aufgeteilt wurde, wird die Methode mit den einzelnen Chunks aufgerufen. Hier hat man nun mehrere Möglichkeiten:

- Händische Identifizierung von Outliern mithilfe der Ausgabe mehreren Boxplots.
- Berechnung der oberen und unteren Grenze und automatische Filterung aller Werte, welche sich außerhalb befinden.

### 3.1.3 Auswertung

## 3.2 Isolation Forest

Um Anomalien in Datensätzen zu erkennen werden meist erst normale Instanzen profiliert, anhand deren man dann nicht übereinstimmende Instanzen als Anomalien zu identifizieren [4]. Der Isolation Forest konzentriert sich hingegen auf das Erkennen der Anomalien selbst, indem es zunutze nimmt, dass Anomalien 1) die Minderheit des Datensatzes sind, und 2) sich die Attributwerte stark von denen normaler Instanzen unterscheidet. Sie sind somit "wenig und unterschiedlich" [4]. Isolation Forests basieren wie Random Forests auf Entscheidungsbäumen **QUELLE**. Und da es hier keine vordefinierten Labels gibt, handelt es sich um ein unüberwachtes Modell **QUELLE**.

### 3.2.1 Funktionsweise

Der Isolation Forest berechnet zu jeder Instanz einen Anomalie-Wert. Nach der zufälligen Auswahl eines Merkmals werden die Daten in einer Baumstruktur zufällig verarbeitet. Dies wird nun rekursiv fortgesetzt, bis alle Messwerte isoliert sind. Der Anomalie-Wert berechnet sich nun aus der Länge des Pfades zwischen der Wurzel und dem zu bewertenden Knoten. Anomalien werden daran erkannt, dass diese bereits nach wenigen Iterationen isoliert wurden und somit einen geringeren Anomalie-Wert besitzen als Knotenpunkte, die keine Outlier darstellen **QUELLE**..

### 3.2.2 Anwendung

### 3.2.3 Vorteile

Firstly, building iTrees only need to select subset of the training set randomly, the research result show that the reasonable number of sub-samplings is set to 256, which is a relative small number and reduce the swamping and masking effects effectively. Secondly, iForest utilizes no distance or density measures to detect anomaly, this eliminates computational cost significantly compared to the distance-based methods and density-based methods. Thirdly, iForest has a linear time complexity with low constant and a low memory requirement. Last but not the least, iForest algorithm is based on the ensemble idea, even if the efficiency of some iTrees are not very high, the ensemble algorithm always can turn the weak algorithm into strong algorithm.

### 3.2.4 Auswertung

### 3.2.5 Limits/Challenges?



### 3.3 Local Outlier Factor

Local Outlier Factor (fortgehend mit LOF abgekürzt) beschreibt die lokale Suche nach Anomalien. Die Methode beruht auf einer dichte-basierten Technik [1]. Die Idee ist, jedem Objekt ein Ausreißergrad zuzuordnen, welcher davon Abhängig ist, wie isoliert besagtes Objekt in Bezug auf die umgebende Nachbarschaft ist [2].

#### 3.3.1 Funktionsweise

#### 3.3.2 Anwendung

#### 3.3.3 Auswertung

#### 3.3.4 Limits/Challenges?

<https://sci-hub.hkvisa.net/10.3390/bdcc5010001>

## 3.4 One-Class Support Vector Machine

### 3.4.1 Funktionsweise

### 3.4.2 Anwendung

### 3.4.3 Auswertung

## Kapitel 4

# Ergebnisse

Erklärung der Ergebnisse und Analyse durch Disco im Anschluss.

## Kapitel 5

# Diskussion

Für Weiterführende Arbeit! <https://sci-hub.hkvisa.net/10.1145/3338840.3355641>

## Kapitel 6

## Fazit

# Literaturverzeichnis

- [1] Omar Alghushairy, Raed Alsini, Terence Soule, and Xiaogang Ma. A Review of Local Outlier Factor Algorithms for Outlier Detection in Big Data Streams. 5(1):1.
  - [2] Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. LOF: Identifying density-based local outliers. 29(2):93–104.
  - [3] Y. H. Dovoedo and S. Chakraborti. Boxplot-Based Outlier Detection for the Location-Scale Family. 44(6):1492–1513.
  - [4] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation Forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422. IEEE.
  - [5] H. P. Vinutha, B. Poornima, and B. M. Sagar. Detection of Outliers Using Interquartile Range Technique from Intrusion Dataset. In Suresh Chandra Satapathy, Joao Manuel R.S. Tavares, Vikrant Bhateja, and J. R. Mohanty, editors, *Information and Decision Sciences*, volume 701 of *Advances in Intelligent Systems and Computing*, pages 511–518. Springer Singapore.
-