IT Center

RWTH Aachen

Anomalieerkennung in heterogenen Metadaten am Beispiel des Datensicherungssystems Commvault

Autor:
Alina Engers

Betreuung:
Dr. rer. nat. Thomas Eifert
Prof. Dr. Florian Heinrichs

Aachen, 4. November 2026

Zusammenfassung

Die Früherkennung von Cyberangriffen wie Ransomware stellt in der IT-Sicherheit eine zunehmende Herausforderung dar. Die Analyse von Datensicherungsmetadaten zur Anomalieerkennung bietet hierfür einen vielversprechenden Ansatz. Im Rahmen des hochschulübergreifenden Projekts Datensicherung.NRW, das auf der Software Commvault basiert, zeigt sich jedoch, dass die anfallenden Metadaten sehr heterogen, stark schiefverteilt und zeitlich abhängig sind. Da klassische Distanzverfahren bei diesen Eigenschaften und den ausgeprägten Long-Tails oft an ihre Grenzen stoßen, wird in dieser Arbeit ein gestaffeltes Frühwarnsystem entwickelt und evaluiert.

Um die statistischen Voraussetzungen für die genutzten Methoden zu erfüllen, erfolgt zunächst eine robuste Datenvorverarbeitung. Diese umfasst ein gezieltes Feature-Engineering zur Verhältnisbildung sowie eine Trendbereinigung (mittels Log-1-p-Transformation und EWMA). Auf dieser Basis wird ein Ensemble-Ansatz des unüberwachten Lernens implementiert, der vier komplementäre Methoden kombiniert: den Isolation Forest, die multivariate t-Verteilung, empirische Copulas (COPOD) und die Extremwerttheorie. Diese Kombination ermöglicht es, sowohl Struktur- als auch Korrelationsbrüche in den Daten zu identifizieren.

Die Evaluation anhand künstlich induzierter Anomalien, wie beispielsweise simulierten Verschlüsselungsprozessen, bestätigt die Wirksamkeit der Methodik durch hohe Werte für Precision und Recall sowie eine hohe Rank Power. Bei der anschließenden Anwendung auf den ungelabelten produktiven Gesamtdatensatz der RWTH Aachen mit über drei Millionen Jobs konnte die Menge der kritischen Alarme auf 7.514 hochgradig auffällige Vorgänge reduziert werden. Ein Alarm wird hierbei nur ausgelöst, wenn mindestens drei der vier Methoden eine Anomalie melden.

Dadurch bietet die vorliegende Arbeit eine praxistaugliche und skalierbare Grundlage für ein System, das Administratoren nicht nur frühzeitig vor kompromittierten Clients warnt, sondern durch die drastische Reduktion von Fehlalarmen auch effektiv vor Alert Fatigue schützt.

Inhaltsverzeichnis

Abkürzungsverzeichnis

1 Einleitung **2**
- 1.1 Motivation . . . 2
- 1.2 Forschungsfrage . . . 2
- 1.3 Methodik . . . 3
- 1.4 Struktur . . . 3

2 Hintergrund **4**
- 2.1 Überblick der Softwarelösung Commvault . . . 4
  - 2.1.1 Metadaten . . . 4
- 2.2 Anomalieerkennung . . . 5
  - 2.2.1 Anomalie . . . 5
  - 2.2.2 Eingabedaten . . . 5
  - 2.2.3 Ausreißertypen . . . 6
  - 2.2.4 Besonderheiten von Zeitreihen . . . 6
  - 2.2.5 Ansätze der Anomalieerkennung . . . 7
  - 2.2.6 Probleme der vorliegenden Daten . . . 7
- 2.3 Bisherige Lösungsansätze . . . 8
- 2.4 Verwandte Arbeiten . . . 8

3 Methodik **10**
- 3.1 Abgrenzung zur Vorarbeit . . . 10
- 3.2 Verwendete Daten . . . 10
  - 3.2.1 Beschaffung der Daten . . . 10
  - 3.2.2 Untersuchte Daten . . . 10
  - 3.2.3 Filtern der Daten . . . 11
  - 3.2.4 Datenverarbeitung . . . 11
  - 3.2.5 Handhabung der Zeitabhängigkeit durch Datentransformation . . . 14
- 3.3 Verfahren zur Anomalieerkennung . . . 17
  - 3.3.1 Isolation Forest . . . 17
  - 3.3.2 Multivariate t-Verteilung . . . 18
  - 3.3.3 Copula-Modelle . . . 19
  - 3.3.4 Extreme Value Theory über Schwellenwerte . . . 21
- 3.4 Vergleich der Methoden . . . 22
- 3.5 Evaluationsmetriken . . . 23
- 3.6 Datenpipeline . . . 24

4 Analyse **25**
- 4.1 Isolation Forest . . . 25
- 4.2 Multivariate t-Verteilung . . . 27
- 4.3 Copulas . . . 29
- 4.4 Extremwerttheorie . . . 31
- 4.5 Erkannte Anomalien . . . 35
- 4.6 Validierung anhand von Test-Clients . . . 36

5 Diskussion **39**

5.1 Einordnung der Datenvorverarbeitung . . . 39
5.2 Verhalten der Algorithmen . . . 40
5.2.1 Isolation Forest . . . 40
5.2.2 Multivariate t-Verteilung . . . 41
5.2.3 COPOD . . . 42
5.2.4 Extremwerttheorie . . . 43
5.2.5 Gemeinsame Nutzung . . . 43
5.3 Entwicklung eines Frühwarnsystems . . . 44
5.4 Validierung mithilfe der Test-Clients . . . 44
5.5 Qualitative Experten-Evaluation . . . 45
5.6 Limitierungen der Arbeit . . . 46
5.6.1 Granulare Filterung . . . 46
5.6.2 Fehlende Label . . . 46
5.6.3 Echtzeit-Verarbeitung . . . 46
5.6.4 Handhabung der Zeitabhängigkeit . . . 47
5.6.5 Heuristische Schwellenwerte . . . 47
5.6.6 Mindestanforderungen an die Datenmenge . . . 47
5.6.7 Merkmalskorrelation und theoretische Annahmen . . . 47
5.6.8 Unterscheidung der Anomalietypen . . . 48
5.7 Synthese . . . 48

6 Fazit und Ausblick **49**
6.1 Fazit . . . 49
6.2 Ausblick . . . 50

Anhang **57**

A Code **58**

B Test-Client Visualisierung **58**

Abkürzungsverzeichnis

AKF Autokorrelationsfunktion 11, 13, 15, 39

API Application Programming Interface 2–4, 10, 46, 50

COPOD Copula-based Outlier Detector 1, 8, 20–22, 29–31, 35–38, 40–45, 47, 49, 50

DaSi.NRW Datensicherung NRW 2–4, 10, 49, 51

ECDF empirische kumulative Verteilungsfunktionen 20, 42

EM-Algorithmus Expectation-Maximization-Algorithmus 19

EWMA Exponentially Weighted Moving Average 15, 22, 39

GPD Generalized Pareto Distibution 14, 21, 22, 32, 43, 47, 49

i.i.d. independent identically distributed 14, 15, 21, 22, 39, 47, 49, 51

IQR Interquartilsabstand 8

JET Job Elapsed Time 4, 6, 10, 11, 15, 17, 25–27, 29, 31, 34, 35, 40, 41, 43, 44, 46, 48–50

MLE Maximum-Likelihood-Schätzer 14, 19, 22, 32, 43

POT Peaks-over-Threshold 14, 32, 35, 37, 43, 49

RWTH Aachen Rheinisch-Westfälische Technische Hochschule Aachen 2, 25, 39, 45, 49

SOA Size of Application 4, 6, 10, 11, 15, 17, 25–27, 29, 31, 33–35, 39–43, 47, 49

SOMOD Size of Media on Disk 4, 6, 10, 11, 15, 17, 25–27, 29, 31, 33–35, 40–44, 49

TSLJ Times since last Job 17, 25, 26, 38, 40, 45, 48, 49

Abbildungsverzeichnis

1 Anzahl der Clients mit zeitlicher Abhängigkeit pro Gruppierung 13
2 Anzahl der Jobs mit zeitlicher Abhängigkeit pro Gruppierung 13
3 Anzahl der Clients mit zeitlicher Abhängigkeit pro Gruppierung nach Regression 16
4 Anzahl der Jobs mit zeitlicher Abhängigkeit pro Gruppierung nach Regression . 16
5 Datenpipeline Flussdiagramm 24
6 Anomalie-Score des Isolation Forest in Prozent 25
7 Gewichtung der Distanz der t-Verteilung in Prozent 27
8 Gewichtung der Distanz der t-Verteilung in Prozent für einen kleinen Ausschnitt 28
9 Anomalie-Score von COPOD in Prozent 30
10 Verteilung der p-Werte der Extremwertanalyse in Prozent 32
11 Upset-Diagramm der Ergebnisse der Extremwertanalyse 33
12 Anzahl erkannter Anomalien nach Attribut 33
13 Anzahl erkannter Anomalien pro Methode 36
14 Referenzclient 58
15 Löschung der Daten 59
16 Verschlüsselung 60
17 Schleichende Verschlüsselung 61

Tabellenverzeichnis

1 Theoretischer Vergleich der verwendeten Algorithmen hinsichtlich ihrer Stärken und Schwächen für die Anomalieerkennung in Datensicherungsmetadaten. . . . 23
2 Vergleich der Attribute des analysierten Jobs 36143207 mit dem Median und dem Durchschnitt des Clients 20049. 26
3 Vergleich der Attribute des analysierten Jobs 26790845 mit dem Median und dem Durchschnitt der inkrementellen Backups des Clients 16269. 26
4 Vergleich der Attribute des analysierten Jobs 30917297 mit dem Median und Durchschnitt der inkrementellen Backups des Clients 7205. 29
5 Vergleich der Attribute des analysierten Jobs 35576387 mit dem Median und dem Durchschnitt des Clients 12479. 29
6 Vergleich der Attribute des analysierten Jobs 23176354 mit dem Median und dem Durchschnitt der Log-Only-Backups des Clients 2028. 31
7 Vergleich der Attribute des analysierten Jobs 42445386 mit dem Median und dem Durchschnitt der inkrementellen Backups des Clients 14435. 31
8 Vergleich der Attribute des analysierten Jobs 37860534 mit dem Median und dem Durchschnitt der inkrementellen Backups des Clients 6727. 34
9 p-Wert der EVT 34
10 Vergleich der Attribute des analysierten Jobs 23538796 mit dem Median und dem Durchschnitt der inkrementellen Backups des Clients 11205. 35
11 Evaluationsmetriken des Test-Clients 20975. Ein Strich (-) signalisiert mathematisch nicht definierte Werte, da keine echten Anomalien existieren ($d_t = 0$). . 37
12 Evaluationsmetriken des Test-Clients 20989 ($d_t = 2$) 37
13 Evaluationsmetriken des Test-Clients 20995 ($d_t = 2$) 38
14 Evaluationsmetriken des Test-Clients 20932 ($d_t = 12$) 38

1 Einleitung

1.1 Motivation

Die vorliegende Arbeit befasst sich mit der statistischen Analyse von Metadaten am Beispiel des Datensicherungssystems Commvault zur Anomalieerkennung. Während in einer vorherigen Seminararbeit zunächst die grundsätzliche Machbarkeit einer solchen Analyse untersucht wurde [1], ist das Ziel dieser Arbeit, verschiedene datengetriebene Verfahren zur Anomalieerkennung systematisch zu analysieren, zu vergleichen und hinsichtlich ihrer Eignung für die Anomalieerkennung der gegebenen Daten zu bewerten. Darauf aufbauend soll der Grundstein für die Entwicklung eines Frühwarnsystems gelegt werden, das potenziell kompromittierte Systeme frühzeitig identifizieren kann. Hierbei beschränkt sich die vorliegende Arbeit auf die Analyse historischer Daten. Der Fokus liegt auf der zuverlässigen Anomalieerkennung, nicht auf der Prävention oder der Vorhersage zukünftiger Ereignisse. Die vorliegenden Daten stellen zudem eine Herausforderung dar. Sie sind stark heterogen, besitzen teilweise Heavy Tails (also ausgeprägte Verteilungsenden) und sind schief verteilt [1]. Es gibt eine Vielzahl möglicher Anomalieerkennungsmethoden [2], allerdings ist die Erforschung möglicher Kombinationen dieser Methoden, die sich für heterogene Backup-Daten eines hochschulübergreifenden Datensicherungsdienstes eignen, in der Literatur bislang nicht oder nur unzureichend adressiert. Ziel dieser Arbeit ist die Evaluierung einer möglichen Kombination.

Als praxisnaher Anwendungsfall dienen die Commvault-Metadaten des hochschulübergreifenden Datensicherungsdienstes des Projekts DaSi.NRW [1]. Hier bieten nur wenige Dienstleister den Datensicherungsdienst für viele Hochschulen an, weshalb diese zentrale Struktur im Falle sicherheitsrelevanter Auffälligkeiten besondere Anforderungen an ein Frühwarnsystem stellt [3].

Da die RWTH Aachen als Dienstleister fungiert, besteht Zugriff auf die API von Commvault, worüber die Metadaten der Sicherungsvorgänge extrahiert werden können [4]. Die Metadaten stammen nicht nur von verschiedenen Hochschulen, sondern umfassen auch unterschiedliche Betriebssysteme, Anwendungsfälle und Backup-Strategien. Auf Basis dieser vielschichtigen Datenstrukturen werden in dieser Arbeit datengetriebene Verfahren zur Anomalieerkennung angewendet und hinsichtlich ihrer Robustheit, Sensitivität und ihres praktischen Nutzens für ein Frühwarnsystem untersucht. Dieses Frühwarnsystem soll in der Lage sein, auf Basis abnormalen Metadatenverhaltens auf potenzielle Angriffe wie Ransomware hinzuweisen. Ein solcher Angriff spiegelt sich in den Metadaten beispielsweise durch den plötzlichen Einbruch der Deduplikationsrate oder signifikante Sprünge im gesicherten Datenvolumen wider. Damit verbindet die Arbeit eine methodische Analyse mit einem konkreten Anwendungsfall im Hochschulumfeld.

1.2 Forschungsfrage

Die zentrale Forschungsfrage dieser Arbeit lautet: „Welche statistischen Methoden und ihre Kombinationen eignen sich, um in stark heterogenen Metadaten, beispielsweise in denen des Datensicherungssystems Commvault, Anomalien mit einer optimalen Balance zwischen hoher Sensitivität und hoher Präzision zu identifizieren?“ Zur Beantwortung dieser Forschungsfrage werden folgende Teilfragen formuliert:

[1] `https://datensicherung.dh.nrw/de/`

1. Wie müssen Rohdaten aufbereitet werden, um die Methoden sinnvoll anwenden zu können?
2. Welche Auswirkungen haben die spezifischen Dateneigenschaften auf die Leistungsfähigkeit und Robustheit der ausgewählten Methoden?
3. Wie lässt sich aus der Kombination der Methoden ein gestaffeltes Warnsystem erstellen, das Administratoren zuverlässig alarmiert, ohne Alert Fatigue [5] auszulösen?

Zur Beantwortung dieser Fragen werden verschiedene statistische Modelle und Algorithmen des unüberwachten Lernens auf die gegebenen Daten angewendet.

1.3 Methodik

Zur Beantwortung der Forschungsfrage werden Metadaten (3.2) aus Datensicherungen des Projekts DaSi.NRW verwendet, die über die Commvault-API extrahiert und in einer lokalen Datenbank gespeichert werden. Die Daten werden anhand ausgewählter Merkmale strukturiert und in vergleichbare logische Gruppierungen (basierend auf Client-Id und Art des Backups) unterteilt, um die Heterogenität der Daten zu berücksichtigen.

Innerhalb dieser Gruppierungen werden die Sicherungsvorgänge betrachtet und auf Anomalien hin untersucht. Hierzu kommt ein Ensemble-Ansatz [2, Kapitel 2] zum Einsatz, der vier komplementäre Verfahren kombiniert: den Isolation Forest, die multivariate t-Verteilung, Copulas sowie die Extremwerttheorie. Ziel ist es, auffällige Abweichungen vom normalen Verhalten zu identifizieren und die Eignung der jeweiligen Verfahren hinsichtlich Robustheit, Datenempfindlichkeit sowie praktischer Anwendbarkeit zu bewerten. Die Bewertung erfolgt sowohl anhand der Auswertung künstlich erstellter Anomalien als auch durch Befragungen eines Expertenteams aus Administratoren des DaSi.NRW-Projekts.

Die Ergebnisse der verschiedenen Verfahren werden analysiert und miteinander verglichen, um deren Eignung für den Einsatz in einem potenziellen Frühwarnsystem zu beurteilen. Eine detaillierte Beschreibung der verwendeten Daten sowie der Methoden erfolgt in Kapitel 3.

1.4 Struktur

Im folgenden Kapitel Hintergrund wird zunächst das Speichersystem Commvault erläutert. Anschließend folgen eine Erklärung der Anomalieerkennung sowie eine Vorstellung der in der Vorarbeit bereits verwendeten Methoden. Abschließend wird auf verwandte Arbeiten eingegangen. Im Kapitel Methodik werden die verwendeten datengetriebenen Verfahren zur Anomalieerkennung vorgestellt. Die daraus resultierenden Ergebnisse werden im Kapitel Analyse betrachtet und im Kapitel Diskussion eingeordnet und bewertet. Im letzten Kapitel Fazit und Ausblick werden die Ergebnisse zusammengefasst und ein Ausblick auf zukünftige Arbeiten gegeben.

2 Hintergrund

2.1 Überblick der Softwarelösung Commvault

Commvault ist ein hierarchisch aufgebautes Speichersystem. Die einzelnen Komponenten der Software sind jedoch wabenförmig strukturiert, deren Gesamtheit als CommCell bezeichnet wird. Jede CommCell verfügt über einen CommServe, der den Mittelpunkt des Systems bildet. Mit diesem CommServe kommunizieren verschiedene Media Agents, welche wiederum mit weiteren Agents kommunizieren, die auf den Zielsystemen laufen. [6]

Für den Kontext dieser Arbeit ist primär relevant, dass jede Hochschule in Form einer Organisation dargestellt wird und Systeme sichert. Die gesicherten Systeme werden Clients genannt und können unter anderem Server, Laptops oder virtuelle Maschinen sein. Die Sicherung erfolgt über sogenannte Agents in Form von Jobs. Diese Jobs haben verschiedene Eigenschaften, die in der Arbeit untersucht werden. Die Daten werden mithilfe des CommServe auf dem zugrunde liegenden Objektspeicher gespeichert. Zur Visualisierung und Überprüfung der gesicherten Systeme dient das CommandCenter, eine Weboberfläche, die basierend auf den jeweiligen Benutzerberechtigungen die entsprechenden Systeme anzeigt. Zusätzlich stehen weitere Möglichkeiten zur Verfügung, wie beispielsweise das Anpassen der Pläne/Jobs sowie Monitoring und Reporting. [6]

Zudem stellt Commvault eine API zur Verfügung, über die Metadaten abgerufen werden können. Diese API enthält mehrere hundert mögliche Operationen, mit denen man sowohl Metadaten abfragen als auch Vorgänge automatisieren kann. [4]

Für diese Arbeit sind vor allem die Metadaten der Jobs sowie der Clients von Interesse. Diese Daten wurden bereits im Rahmen eines vorherigen Projekts[2] über die API extrahiert und bereinigt, sodass die vorliegende Arbeit auf einem strukturierten Datensatz aufbauen kann.

2.1.1 Metadaten

Die hier betrachteten Metadaten enthalten sowohl Informationen über Jobs als auch über Clients. Clients werden beispielsweise in verschiedene Klassen eingeordnet, etwa Server, VM, Hypervisor oder persönliche Arbeitsstation (z. B. Laptop oder Computer). Die Daten der Jobs hingegen sind deutlich komplexer. Es gibt Daten wie den Beginn, das Ende und die Durchlaufzeit eines Jobs, aber auch Daten über die Art des Jobs, beispielsweise Full Backup, Synthetic Backup, Incremental Backup oder Restore. Zusätzlich gibt es Informationen über die Anzahl der gespeicherten Bytes auf dem Client sowie auf dem Speichersystem. Diese Werte unterscheiden sich oftmals stark aufgrund der Deduplikation von Commvault. Deduplikation bedeutet, dass dieselbe Information nicht mehrfach gespeichert wird, sondern nur einmal, und alles Weitere aus Verweisen auf die gespeicherte Information besteht. Dadurch haben Jobs Abhängigkeiten zu vergangenen Jobs sowie zu anderen Clients. [4, 6]

Im Gegensatz zur Seminararbeit werden nun mehr Metadaten verwendet, die es ermöglichen, Jobs zu gruppieren. Dies führt zu einer Analyse vergleichbarer Ergebnisse, welche die Anomalieerkennung erleichtert. Zudem werden hauptsächlich die Anzahl gespeicherter Bytes auf dem Start- (SOA) sowie auf dem Zielsystem (SOMOD) und die Durchlaufzeit der Jobs (JET) betrachtet, da anhand dieser Daten Anomalien zu erkennen sind.

[2] Reporting-Website DaSi.NRW: `https://dasireport.itc.rwth-aachen.de/ui/de`

2.2 Anomalieerkennung

Im Rahmen dieser Arbeit werden mithilfe statistischer Methoden sowie Algorithmen Anomalien innerhalb der beschriebenen Metadaten identifiziert. Der Fokus liegt dabei auf der Erkennung ungewöhnlicher Muster in den Sicherungsvorgängen, die auf technische Fehler, atypisches Systemverhalten oder sicherheitsrelevante Ereignisse hinweisen können.

2.2.1 Anomalie

Unter einer Anomalie versteht man eine „substantielle Abweichung von der Norm“ (eigene Übersetzung) [2, Kapitel 1]. Um sie zu identifizieren, werden neue Daten mit bereits bekannten Daten verglichen oder auf diese Prognosen angewendet. Was genau als Anomalie gilt, hängt stark vom verwendeten Datensatz ab. Anomalieerkennung ist ein wichtiger Forschungsbereich und wird in vielen Disziplinen eingesetzt, etwa in der Cybersecurity, der Home Security sowie im Gesundheitswesen. [2, Kapitel 1]

Die Definition der Norm spielt eine entscheidende Rolle. Je nach Norm und Abweichungsmaß kann ein Datenpunkt als normal bzw. abnormal gelten. Als Norm kann beispielsweise der Mittelwert oder der Median verwendet werden, und als Abweichungsmaß die euklidische Distanz. Aufgrund dessen wird die Anomalieerkennung eher als Indikator als als Maß gewertet. Allerdings wird aufgrund der Eigenschaften der vorliegenden Daten die Definition der Norm über diese Standardmaße unmöglich, weshalb andere Maße wie beispielsweise die Mahalanobis-Distanz (3.3.2) genutzt werden. [2, Kapitel 2]

Es gibt einen Zusammenhang zwischen den Daten und dem Prozess, der die Daten erstellt hat. Deshalb müssen vier Möglichkeiten bei Bewertung der Anomalieerkennungsmethoden betrachtet werden, welche oftmals in Form einer Konfusionsmatrix dargestellt sind:

1. Echt-positive Klassifizierung
2. Echt-negative Klassifizierung
3. Falsch-positive Klassifizierung
4. Falsch-negative Klassifizierung

Eine echt-positive Klassifizierung identifiziert in diesem Kontext eine tatsächliche Anomalie als Anomalie. Eine echt-negative Klassifizierung bedeutet, dass ein Datenpunkt nicht als Anomalie klassifiziert wird, wenn er auch keine ist. Bei einer falsch-positiven Klassifizierung wurde eine Anomalie in den Daten erkannt, obwohl im Prozess alles normal ablief. Das Gegenteil ist die falsch-negative Klassifizierung. Dort wurde keine Anomalie in den Daten erkannt, im Prozess ist jedoch eine Anomalie aufgetreten. [2, Kapitel 2] [7]

Insbesondere im Kontext der gegebenen Daten ist die Definition einer Anomalie eng mit den operativen Abläufen des Systems verknüpft, wodurch sowohl technische Fehler als auch ungewöhnliches Nutzerverhalten zu berücksichtigen sind.

2.2.2 Eingabedaten

Die Eingabedaten sind eine Sammlung von Datenpunkten. Diese Daten können diskret oder kontinuierlich sein. Zudem kann ein einzelner Datenpunkt aus einem oder mehreren Attributen bestehen, die im Kontext des maschinellen Lernens als Dimensionen bezeichnet werden. Besteht

der Datenpunkt aus einem Attribut, ist er univariat. Umfasst er mehrere Attribute, ist er multivariat. [8]

Dies gilt auch, wenn die vorliegenden Daten Zeitreihen sind. Jeder Datenpunkt wird durch einen Beobachtungsvektor beschrieben [9]. Die vorliegenden Daten sind eine multivariate Zeitreihe, denn zu einem Zeitpunkt werden sowohl die Daten für SOA als auch für SOMOD und JET erhoben, die jedoch keine äquidistanten Abstände zwischen den Zeitpunkten besitzen.

2.2.3 Ausreißertypen

Es wird zwischen drei Ausreißertypen unterschieden: Punkt-, Kontext- und Kollektivausreißer. Punktausreißer bezeichnen Daten, die im Vergleich zum Rest der Daten auffällig sind. Kontext-Ausreißer sind auffällig, wenn man das Verhalten eines Datenpunkts in einem zuvor definierten Kontext betrachtet. Kollektiv-Ausreißer treten nur in Datensätzen auf, in denen die Attribute zueinander in Beziehung stehen. Wenn die Kombination der Attributwerte ein anormales Verhalten aufweist, liegt ein Kollektiv-Ausreißer vor. [8]

Wird der Datensatz spezifisch als Zeitreihe betrachtet, lässt sich diese Kategorisierung übertragen. Dabei werden die Ausreißer häufig als Punkt-, Segment- oder Zeitreihenausreißer bezeichnet. Zudem wird zwischen globalen und lokalen Ausreißern unterschieden. Globale Ausreißer weisen im Vergleich zur restlichen Zeitreihe ein anderes Verhalten auf, während lokale Ausreißer nur im Verhältnis zu ihren benachbarten Daten auffallen. [9]

Punktausreißer sind einzelne Datenpunkte, die ein anderes Verhalten aufweisen, während Segmentausreißer Segmente einer Zeitreihe sind, die sich nicht wie erwartet verhalten. Beide Arten können sowohl in uni- als auch in multivariaten Zeitreihen auftreten. Zeitreihenausreißer hingegen treten nur in multivariaten Zeitreihen auf. Hier verhält sich ein Teil der multivariaten Zeitreihe anders als die anderen betrachteten Teile. [9]

Mithilfe der verwendeten Methoden lassen sich sowohl Punkt-, Kontext- als auch Kollektivausreißer identifizieren.

2.2.4 Besonderheiten von Zeitreihen

Stationarität Bei Stationarität wird zwischen strikter und schwacher Stationarität unterschieden. Eine Zeitreihe X_t wird als schwach stationär bezeichnet, wenn ihr Erwartungswert über die Zeit konstant bleibt, also

$$E[X_t] = \mu, \tag{1}$$

und ihre Autokovarianz nur von der zeitlichen Differenz abhängt:

$$E[X_{t_1} X_{t_2}] - E[X_{t_1}]E[X_{t_2}] = r(|t_1 - t_2|). \tag{2}$$

[10]

Strikte Stationarität bedeutet, dass eine Zeitreihe die Kriterien der schwachen Stationarität erfüllt und die Wahrscheinlichkeitsverteilung gegenüber zeitlicher Veränderung invariant ist. Entsprechend ist die Stationarität der Zeitreihe unabhängig vom Beobachtungszeitraum und weist über die Zeit hinweg identische Verteilungseigenschaften auf [10]. Im Folgenden wird unter dem Begriff Stationarität schwache Stationarität verstanden, da die Prüfung schwacher Stationarität empirisch einfacher ist.

Trend, Saisonalität und Residuen Zeitreihen bestehen häufig aus Trends, Saisonalität, Zyklen und Residuen [11]. Trends wirken über einen langen Zeitraum auf Zeitreihen ein, während Saisonalitäten periodisch und Zyklen zyklisch auf sie einwirken [11]. Residuen hingegen folgen keinem Schema, sondern sind sogenannte „irreguläre Fluktuationen“ [11]. Sie repräsentieren die Daten, die nach der Bereinigung einer Zeitreihe von Trends, Saisonalitäten und Zyklen verbleiben.

Bei der Zerlegung von Zeitreihen wird zwischen additiven und multiplikativen Modellen unterschieden [12, Kapitel 9]. Da die Backup-Daten im Laufe der Zeit oft exponentiell ansteigen, wird im weiteren Verlauf von einem multiplikativen Zusammenhang zwischen den Komponenten ausgegangen. Dieser wird später mittels einer Log-1-p-Transformation in ein additives Modell überführt [13].

2.2.5 Ansätze der Anomalieerkennung

Es gibt verschiedene Arten der Anomalieerkennung. Darunter fallen abstandsbasierte, Clustering- und modellbasierte Verfahren. Klassifikationsalgorithmen erweisen sich oft als ungeeignet für die Anomalieerkennung, da ein sehr kleiner Anteil der Daten (falls überhaupt vorhanden) Anomalien sind und Klassifikationsalgorithmen zu viele falsch-positive Klassifizierungen liefern. [2, Kapitel 2]

Zusätzlich spielen die Daten bei der Wahl der passenden Methode oder des passenden Algorithmus eine entscheidende Rolle. Dichte- und distanzbasierte Verfahren eignen sich beispielsweise nicht, wenn die Daten nur spärlich vorliegen. Dafür sind sie beispielsweise unabhängig von der Datenverteilung. Eine weitere Möglichkeit sind rangbasierte Verfahren, die die Daten transformieren und entsprechend einteilen. Um die Vorteile der genannten Methoden auszuschöpfen, werden häufig sogenannte Ensemble-Methoden eingesetzt, bei denen mehrere Verfahren kombiniert werden, um bessere Ergebnisse zu erzielen. [2, Kapitel 2]

Falls die Daten in Form von Zeitreihen vorliegen, gibt es spezielle Methoden und Algorithmen, um die Informationen zur Zeitabhängigkeit zu bewahren. Anstatt die Daten als Menge zu betrachten, werden sie nun als Sequenz verwendet. [2, Kapitel 9]

Die Wahl des passenden Verfahrens hängt somit stark von der Datenstruktur und dem Anwendungsfall ab, sodass häufig mehrere Ansätze geprüft oder kombiniert werden, um die Zuverlässigkeit der Anomalieerkennung zu erhöhen. In dieser Arbeit wird ein solcher Ensemble-Ansatz gewählt, um sowohl Algorithmen des unüberwachten Lernens als auch distanz- und rangbasierte Verfahren miteinander zu kombinieren und ein breites Spektrum möglicher Anomalietypen zu identifizieren.

2.2.6 Probleme der vorliegenden Daten

Die vorliegenden Daten weisen eine starke Heterogenität auf. Sie sind schiefverteilt und besitzen ausgeprägte Long-Tails oder Heavy-Tails. Zudem gibt es sowohl stationäre als auch nicht-stationäre Zeitreihen. [1]

Dass sich die Daten in ihrer Art unterscheiden, ist dem System geschuldet. Es gibt beispielsweise Clients, die schon seit mehreren Jahren existieren, während andere erst vor kurzem erstellt wurden. Außerdem gibt es Clients, die sich ändernde Daten sichern; andere sind nahezu invariant.

Es wurde erkannt, dass klassische abstandsbasierte Verfahren bei stark schiefverteilten Daten mit Long-Tails oft an ihre Grenzen stoßen. Dagegen wird vermutet, dass die Betrachtung der Daten als Zeitreihen gute Ergebnisse liefern könnte. [1]

Diese Herausforderungen bilden die Grundlage für die Wahl und Anpassung geeigneter Verfahren, die im folgenden Kapitel näher erläutert werden.

2.3 Bisherige Lösungsansätze

In der vorangegangenen Seminararbeit wurden zumeist distanzbasierte Verfahren angewendet. Dabei wurden beispielsweise Verfahren wie der robuste Z-Score sowie der IQR mithilfe eines vordefinierten Schwellenwertes angewendet. Diese Verfahren waren aufgrund der Heterogenität der Daten sowie der fehlenden Gruppierung sehr empfindlich. [1]

Vielversprechender waren Methoden, die auf Zeitreihen basieren, wie die Residuenanalyse sowie die CUSUM- und PELT-Algorithmen. Diese Methoden müssen jedoch weiter analysiert und ihre Daten verarbeitet werden, bevor sie im Rahmen eines Frühwarnungssystems sinnvoll eingesetzt werden können. [1]

2.4 Verwandte Arbeiten

Die Anomalieerkennung ist ein zentraler Forschungsschwerpunkt. Zu dem Thema existieren Standardwerke wie beispielsweise von Mehrotra, Mohan und Huang [2]. Die Anwendung von Anomalieerkennungsmethoden auf IT-Systeme wird zudem zu einem stetig wachsenden Forschungsschwerpunkt. Während sich einige Forschungsarbeiten auf die Echtzeitanalyse mittels Online-Algorithmen fokussieren (vgl. Ahmad et al. [14]), konzentriert sich die vorliegende Arbeit auf die Auswertung historischer Daten.

In der Praxis spiegelt sich diese Relevanz in der Integration dieser Methoden in Datensicherungssysteme wider. Führende Anbieter setzen dabei auf unterschiedliche Ansätze. Das Datensicherungssystem NetBackup des Anbieters Veritas verfügt über eine Anomalieerkennung, die beispielsweise auf Basis von Clustering und dem Z-Score agiert oder künstliche Intelligenz zum Scannen der Daten verwendet [15, 16]. Eine weitere Datensicherungssoftware, Veeam, bietet für ihr Produkt Ransomware Detection an [17]. Der genutzte Algorithmus ist Random Cut Forest, der von Amazon entwickelt wurde [17]. Commvault selbst nutzt auch Anomalieerkennungsmethoden, listet diese jedoch nicht auf [6]. Zudem wird die Anomalieerkennung im Rahmen ihres Cybersecurity-Pakets angeboten[3]. Um die oftmals einseitigen Ansätze der Praxis zu erweitern, kommen in der vorliegenden Arbeit spezifische statistische Methoden und Algorithmen des maschinellen Lernens zum Einsatz, die auf unterschiedliche Anomaliearten prüfen. Der Isolation Forest sowie COPOD wurden explizit zur Anomalieerkennung entwickelt [18, 19], während die Extremwerttheorie zumeist zur Vorhersage von Naturkatastrophen genutzt wird, da sie darauf ausgelegt ist, die Extremwerte einer Verteilung zu modellieren [20]. Sie wird jedoch zunehmend auch zur Anomalieerkennung genutzt [21]. Auch die multivariate t-Verteilung wird bereits in der Anomalieerkennung genutzt [22], da sie eine robustere Modellierung komplexer Daten ermöglicht.

Obwohl die Literatur zur Anwendung der Methoden auf Datensicherungssysteme begrenzt ist, wurden einige davon auf strukturell ähnlichen Daten angewendet. Der Isolation Forest wurde beispielsweise auf Network-Traffic-Daten getestet [23], während die Extremwerttheorie auf

[3] `https://www.commvault.com/explore/anomaly-detection`

Log-Daten eines Cloud-Anbieters angewendet wurde [24]. Die spezielle Anwendung der hier verwendeten Methoden wurde in dieser Kombination im Rahmen der Literaturrecherche nicht identifiziert.

Während die generelle Anomalieerkennung in IT-Systemen gut erforscht ist, bedarf die Analyse von Datensicherungsmetadaten weiterer Forschung. Obwohl die Relevanz der Analyse von Backup-Daten unumstritten ist, setzen bestehende Ansätze häufig nur eine Methode ein, wodurch manche Anomalienformen nicht identifiziert werden können. Die Literatur (z.B. Aggarwal) zeigt jedoch, dass stark heterogene Daten von der Anwendung mehrerer Methoden auf den Datensatz profitieren [25]. Die Literatur zur Kombination von Isolation Forest, der multivariaten t-Verteilung, Copulas und der Extremwerttheorie ist jedoch begrenzt. In der vorliegenden Arbeit werden daher diese vier komplementären Methoden angewendet und deren Relevanz wird überprüft. Dadurch werden die Daten aus verschiedenen Perspektiven betrachtet und mehrere mögliche Anomalien identifiziert.

3 Methodik

3.1 Abgrenzung zur Vorarbeit

In der zugrunde liegenden Seminararbeit wurden bereits verschiedene statistische Methoden auf eine Teilmenge der Daten angewendet. Darunter explorative Methoden wie der robuste Z-Score oder der Interquartilsabstand, Zeitreihenanalysen wie Change-Point-Detection oder die STL-Dekomposition sowie Relationsanalysen wie das Verhältnis zwischen SOA und SOMOD. [1]

Diese Methoden konnten, aufgrund der Verteilungseigenschaften der Daten, größtenteils keine aussagekräftigen Ergebnisse erzielen [1]. Aufgrund dessen werden für die folgende Analyse weitere Datenmerkmale herangezogen und logisches Gruppieren anhand dieser Merkmale angewendet, um möglichst vergleichbare Daten zu erhalten.

3.2 Verwendete Daten

Die in dieser Arbeit verwendeten Daten stammen aus den aufbereiteten Daten des Projekts DaSi.NRW. Seit 2023 ist es in Betrieb und seitdem werden durchgehend neue Clients hinzugefügt, die ihre Daten über Commvault sichern. Entsprechend verändern sich die Daten fortlaufend, und es gibt Unterschiede innerhalb des Datensatzes, beispielsweise durch die Beschaffung des Clients, den Bestand des Clients, die Ausführung der Jobs sowie die Art der Jobs.

3.2.1 Beschaffung der Daten

Die Metadaten aus Commvault wurden bereits im Rahmen eines anderen Projekts aufbereitet und gespeichert. Auf diese Daten wird mithilfe einer GraphQL-API zugegriffen. Dadurch ist es möglich, nur die zur Analyse relevanten Daten zu erhalten. Aufgrund der großen Datenmenge von über 15.000 Clients und 4 Millionen Jobs des Projekts DaSi.NRW wird eine containerisierte lokale Postgres-Datenbank verwendet, um die abgefragten Daten zunächst zu speichern, bevor sie weiterverwendet werden. Anschließend werden die Daten nach ausgewählten Merkmalen wie Client-Type, Job-Type oder Job-Status gruppiert, um eine vergleichbare Datengrundlage zu erhalten. Innerhalb dieser Gruppierungen werden statistische Methoden und Verfahren des maschinellen Lernens zur Anomalieerkennung eingesetzt. Die Ergebnisse werden entsprechend auch in der Datenbank gespeichert und visualisiert. All dies geschieht mithilfe eines Python-Skripts sowie der dafür verwendeten Bibliotheken:

- API und Datenbanken: gql, sqlalchemy [26, 27]
- Berechnungen: pandas, numpy, scikit-learn, statsmodels, PyOD, scipy [28, 29, 30, 31, 32, 33]
- Visualisierungen: kaleido, matplotlib, seaborn, plotly [34, 35, 36, 37]

3.2.2 Untersuchte Daten

Erneut werden hauptsächlich die Merkmale SOA, SOMOD und JET betrachtet. SOA gibt an, wie viele Bytes der Client sichert, während SOMOD angibt, wie viele Bytes tatsächlich auf der Festplatte gespeichert werden. Diese Zahlen unterscheiden sich durch Commvaults Deduplikation deutlich. JET hingegen gibt an, wie lange ein Job zur Ausführung benötigt hat. Diese Zeit ist in Sekunden angegeben.

Mithilfe dieser Werte lassen sich Anomalien erkennen. Bleibt beispielsweise SOA konstant, springen SOMOD und JET jedoch sprunghaft an, was auf die Verschlüsselung von Daten und damit auf einen potenziellen Ransomware-Angriff hindeuten kann. Wenn sowohl SOA als auch SOMOD und JET stark steigen, deutet dies darauf hin, dass neue große Datenmengen gespeichert wurden.

Um die Limitierungen bisheriger Analysen zu überwinden, werden in dieser Arbeit mehr Merkmale berücksichtigt. Diese dienen hauptsächlich der Gruppierung. Clients werden daraufhin geprüft, ob sie noch aktiv sind. Auch bei den Jobs werden mehr Metadaten herangezogen. Um die Zeiten zu modellieren, werden sowohl die Job-Startzeit als auch die Endzeit mitgegeben. Zudem wird zwischen verschiedenen Job-Typen, Status und Backuplevel-Namen unterschieden. Mithilfe des Gruppierens nach diesen Kriterien lassen sich die gegebenen Daten so vergleichbar wie möglich halten. Die Granularität der Gruppierungen kann individuell angepasst werden. Dadurch ist es möglich, ausschließlich Jobs von Clients desselben Typs zu analysieren. Es ist jedoch auch möglich, Jobs zu analysieren, die Clients derselben Art angehören, den Status Completed aufweisen, als Backup klassifiziert sind und zudem ein Synthetic Full Backup sind.

3.2.3 Filtern der Daten

Zur Analyse wurden nur Daten verwendet, die eine eindeutige Job-ID sowie eine Zuordnung zu einem Client enthalten. Des Weiteren wurden nur Jobs mit dem Status „Completed“ betrachtet. Für die Analyse ist es nicht sinnvoll, Jobs zu analysieren, die fehlgeschlagen sind und daher abnorme Werte in den Merkmalen aufweisen, wodurch die Normdefinition verzerrt werden könnte. Zudem werden nur Backup-Jobs berücksichtigt; mögliche Restores bleiben außen vor. Falls eine Normalisierung oder Skalierung erforderlich ist, erfolgt sie als Bestandteil der jeweiligen Methoden.

3.2.4 Datenverarbeitung

Um zu beurteilen, welche Methoden auf die vorliegenden Daten angewendet werden können, müssen vorab verschiedene Eigenschaften untersucht werden. Eine wichtige Eigenschaft, die für einige der zu verwendenden Methoden vorausgesetzt wird, ist die zeitliche Unabhängigkeit der einzelnen Datenpunkte.

Deshalb wird vor der Durchführung der Methoden auf den Daten geprüft, ob zeitliche Abhängigkeiten vorliegen. Dies ist bei realen Daten oftmals der Fall [38, Kapitel 2]. Dazu wurde die AKF auf die Daten angewendet. Die AKF misst die Kovarianz zwischen zwei Zeitpunkten einer Zeitreihe, die durch eine vordefinierte Verschiebung, auch als Lag bezeichnet, getrennt sind, und normiert sie anschließend durch die Varianz [12, Kapitel 2].

Für die Beurteilung der zeitlichen Abhängigkeit wurden mehrere Lags betrachtet. Liegen die entsprechenden Autokorrelationswerte nahe Null, deutet dies darauf hin, dass keine signifikante lineare zeitliche Abhängigkeit zwischen den Beobachtungen besteht. Dieses Verhalten ist konsistent mit einem White-Noise-Prozess, bei dem die Autokorrelation für alle Lags ungleich Null verschwindet. Ein solcher Prozess beschreibt eine zeitliche Abfolge unabhängiger und identisch verteilter Zufallsvariablen, deren Autokovarianz für alle Lags $k \neq 0$ exakt Null ist [12, Kapitel 2].

Um die AKF auf die vorliegenden Daten anwenden zu können, werden die Daten der jeweiligen Clients zunächst chronologisch nach der Jobstartzeit sortiert. Für die Beurteilung der zeitlichen

Abhängigkeit werden anschließend mehrere Lags auf einmal berechnet. Die maximale Anzahl der zu betrachtenden Lags wird dabei dynamisch an die Stichprobengröße angepasst. Die maximale Anzahl beträgt 100 Lags; bei weniger Datenpunkten wird ein Viertel der verfügbaren Datenpunkte verwendet. Dadurch ist es möglich, sowohl die Aussagekraft zu sichern als auch die rechnerische Effizienz zu gewährleisten.

In der Praxis wird für die Annahme von White Noise häufig ein Konfidenzintervall von $\pm 2/\sqrt{n}$ verwendet, wobei n die Stichprobengröße ist. Dies basiert auf dem Standardfehler $se \simeq \frac{1}{\sqrt{n}}$, welcher bei großen Werten von n asymptotisch normalverteilt ist und entsprechend für das 95%-Konfidenzintervall mithilfe des gerundeten Z-Wertes von 2 approximiert werden kann. [12, Kapitel 2]

Die vorliegenden Metadaten weisen jedoch eine hohe Varianz bei der Anzahl der Jobs pro Client auf, die von einigen wenigen bis zu mehreren tausenden reicht. Bei klassischen Signifikanztests führt dies zum sogenannten Large-N-Problem. Das Konfidenzintervall konvergiert mit wachsendem N gegen Null, weshalb bereits sehr geringe Autokorrelationen als signifikant eingestuft werden. [39]

Um dem entgegenzuwirken, wurde sich für einen hybriden Schwellenwert entschieden, um Signifikanz mit Effektstärke zu kombinieren. Dieser lautet:

$$|\rho| > \max\left(\frac{2}{\sqrt{n}}, 0{,}29\right) \tag{3}$$

In diesem Fall entspricht n der Anzahl der analysierten Jobs pro Client. Der Schwellenwert von 0,29 wurde als heuristischer Schwellenwert gewählt, weil Werte unter 0,1 einen schwachen Effekt aufweisen und häufig durch Rauschen entstehen. Werte unter 0,3 haben höchstens einen mittleren Effekt, weshalb der Schwellenwert knapp unterhalb dieser Grenze gewählt wurde. Da die Annahme besteht, dass die vorliegenden Daten von Rauschen behaftet sind, wurde die Grenze auf 0,29 festgelegt, um dieses Rauschen herauszufiltern, sodass die Daten einen nutzbaren Effekt haben, ohne möglicherweise relevante Korrelationen zu filtern. [40, Kapitel 3]

Für die finale Klassifikation, ob ein Client zeitabhängig ist oder nicht, wird die Anzahl der signifikanten Lags berücksichtigt. Da ein 95%-Konfidenzintervall verwendet wird, ist zu erwarten, dass etwa 5% der Lags als signifikant eingestuft werden, jedoch statistischem Rauschen entsprechen. Entsprechend wird die Annahme der zeitlichen Unabhängigkeit nur dann für einen Client verworfen, wenn mehr als 5% der berechneten Lags den definierten hybriden Schwellenwert überschreiten.

Die durchgeführten Tests zeigen jedoch, dass trotz des toleranteren hybriden Schwellenwertes, insbesondere bei inkrementellen Backups, eine signifikante zeitliche Abhängigkeit besteht. Insgesamt gibt es 13.709 verschiedene Gruppierungen, die für jedes Attribut auf zeitliche Abhängigkeiten hin geprüft wurden. Von diesen 41.127 entstandenen Überprüfungen sind 9.046 zeitabhängig:

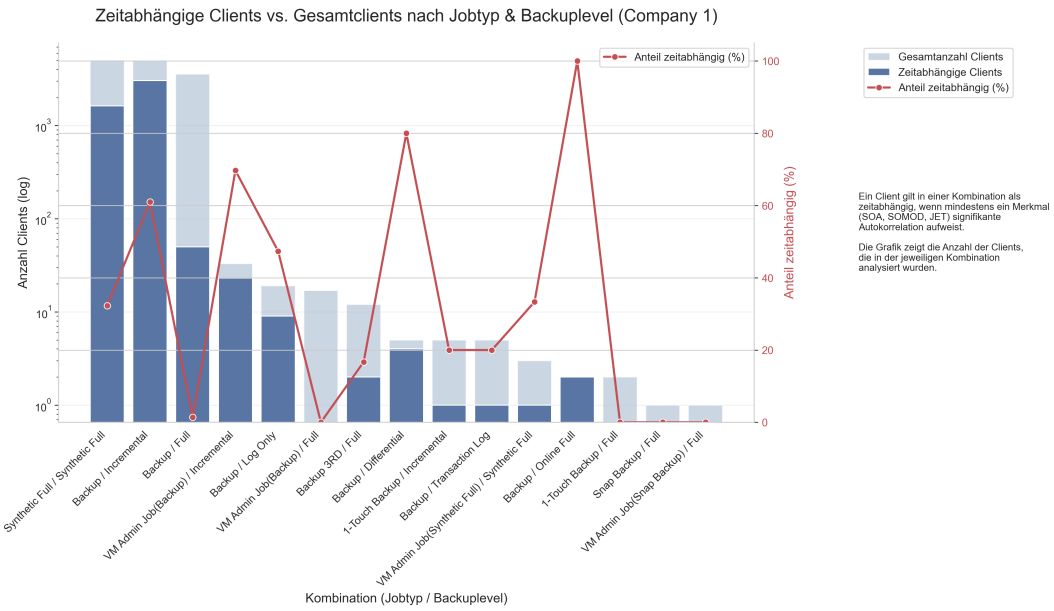


Abbildung 1: Anzahl der Clients mit zeitlicher Abhängigkeit pro Gruppierung

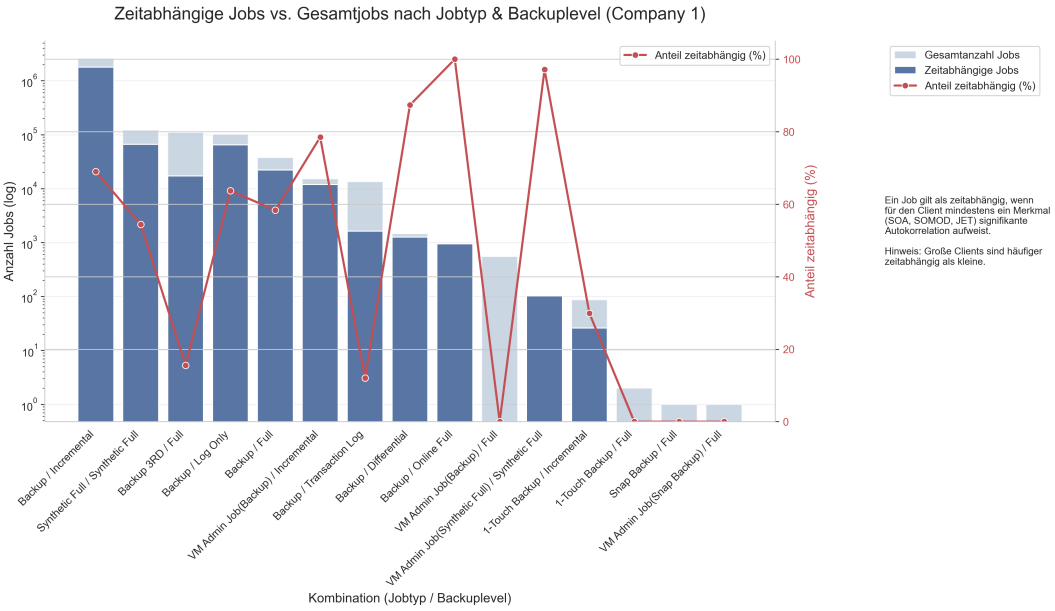


Abbildung 2: Anzahl der Jobs mit zeitlicher Abhängigkeit pro Gruppierung

Auf den ersten Blick gibt es wesentlich mehr zeitunabhängige Clients als Jobs, aber dies liegt unter anderem daran, dass je mehr Jobs ein Client besitzt, desto größer ist die Anzahl der zeitabhängigen Jobs. Die Annahme, dass keine zeitlichen Abhängigkeiten zwischen den einzelnen Jobs bestehen, muss entsprechend verworfen werden. Dies ist unter anderem darauf zurückzuführen, dass die Jobs nicht zwingend in äquidistanten Abständen ausgeführt wurden (auch als ungleichmäßige Zeitreihe bezeichnet [41]), was jedoch für die AKF gewünscht ist [12, Kapitel 2]. Daher können durch die unterschiedlichen Abstände Trends in den Daten entstehen,

die sich mithilfe einer einfachen oder wiederholten Differenzierung nicht beheben lassen, ohne die zugrunde liegende Verteilung der Daten zu zerstören.

Um diese Limitierungen zu überwinden und die i.i.d.-Voraussetzungen der statistischen Modelle zu erfüllen, wird in dieser Arbeit anstelle einer klassischen Differenzierung ein spezifisches Feature-Engineering in Kombination mit einer robusten Trendbereinigung angewendet.

3.2.5 Handhabung der Zeitabhängigkeit durch Datentransformation

Obwohl der zuvor durchgeführte Test zeitliche Abhängigkeiten bestätigt hat, bleibt das primäre Ziel dieser Arbeit die Identifikation von strukturellen und globalen Ausreißern. Während komplexe Zeitreihenmodelle darauf ausgelegt sind, zeitliche Dynamiken abzubilden und vorherzusagen, liegt der Fokus hier auf hoher Skalierbarkeit für große Client-Mengen sowie auf der zuverlässigen Erkennung von Punktausreißern.

Die hier verwendeten statistischen Methoden basieren im Kern auf der Annahme, dass die Daten i.i.d. sind. Die multivariate t-Verteilung nutzt die MLE, die die i.i.d.-Eigenschaften der Daten voraussetzt [38, Kapitel 2]. Copulas-Modelle erfordern für die Schätzung der empirischen Randverteilung ebenfalls i.i.d. Zufallsvariablen [42]. Es gibt Copulas, die mit zeitabhängigen Daten arbeiten können (sogenannte Conditional Copulas), doch diese benötigen zum einen einen hohen Rechenaufwand, da die Schätzung der Randverteilung beispielsweise über GARCH-Modelle erfolgt, und zum anderen eignen sie sich besser für Vorhersagen als für die Anomalieerkennung [43]. Auch die Extremwerttheorie nutzt insbesondere für POT das Pickands-Balkema-de-Haan-Theorem, das i.i.d.-Bedingungen fordert [38, Kapitel 3 und 4].

Es gibt andere Ansätze, um die Bedingungen der Extremwerttheorie zu erfüllen. Dazu ist gefordert, dass zwei Datenpunkte, die größer als der Schwellenwert sind, einen großen zeitlichen Abstand aufweisen, sodass keine direkte zeitliche Abhängigkeit besteht [38, Kapitel 5]. Hierfür wird häufig das sogenannte Declustering als Ansatz gewählt [38, Kapitel 5]. Dabei wird aus einem zeitlich definierten Cluster lediglich das Maximum extrahiert, das anschließend zur Modellierung der GPD verwendet wird [38, Kapitel 5].

Dieses Verfahren ist jedoch für das Ziel dieser Arbeit ungeeignet. Falls beispielsweise über mehrere Tage hinweg anomale Werte erzeugt werden, würde in diesem Ansatz nur das Maximum berücksichtigt werden, wodurch mögliche weitere Anomalien mit anderem Hintergrund verschleiert würden. Ein weiteres Problem ist, dass die Sicherungen nicht in gleichmäßigen Abständen angelegt werden.

Um Verfahren wie die multivariate t-Verteilung, Copula-Modelle oder die Extremwerttheorie effektiv auf die Daten anwenden zu können, ohne die Annahme unabhängiger Beobachtungen zu verletzen, ist eine Datenvorverarbeitung erforderlich. In der Zeitreihenanalyse werden nichtstationäre Prozesse, die durch Differenzierung in einen stationären Zustand überführt werden können, als integrierte Prozesse bezeichnet [12, Kapitel 4]. Der einfachste Ansatz wäre, lediglich die Differenz zum vorherigen Job zu bilden. Dies würde jedoch nur die Abhängigkeiten vom Lag 1 entfernen, während längere Trends oder Saisonalitäten im Datensatz weiterhin bestehen würden. Von einer erneuten Differenzierung der verbleibenden abhängigen Daten wird jedoch abgesehen, da die Risiken einer Überdifferenzierung den potenziellen Nutzen übersteigen. Durch Überdifferenzierung wird ungewollt ein künstlicher Moving-Average-Prozess in die Zeitreihe induziert, der die zugrunde liegende Verteilung der Daten grundlegend verändert [44]. Ein Moving-Average-Prozess der Ordnung q modelliert eine Zeitreihe nicht auf Basis ihrer vergangenen Werte, sondern als Linearkombination von White Noise und den q vergangenen

Fehlern [12, Kapitel 3]. Falls eine bereits unabhängige Zeitreihe erneut differenziert wird, verliert sie diese Unabhängigkeit.

Da eine einfache Differenzierung folglich nicht ausreicht und eine mehrfache Differenzierung die Datenstruktur zerstört, wird in dieser Arbeit stattdessen eine Kombination aus Feature-Engineering und Trendbereinigung angewendet, um Stationarität zu erreichen.

Feature Engineering Da die einzelnen Jobs teilweise zeitabhängig sind, werden sie in relative Attribute überführt:

- Deduplikationsverhältnis: Berechnet den Quotienten von SOA und SOMOD ($\frac{SOA}{\max(1,SOMOD)}$). Bei einem starken Abfall dieses Verhältnisses lässt sich darauf schließen, dass Daten verschlüsselt wurden.
- Durchsatz: Berechnet den Quotienten von SOA und JET ($\frac{SOA}{\max(1,JET)}$). Ein drastischer Einbruch deutet auf Systemausfälle hin. Es wurde sich explizit für SOA im Zähler entschieden, denn dadurch ist es möglich, die Verarbeitungsgeschwindigkeit der Deduplikation mitzuverfolgen.

Die somit erzeugten Verhältnisse sind zwar nicht statistisch unabhängig, jedoch größtenteils von Zeitabhängigkeit befreit. Eine vollständige Erfüllung der i.i.d.-Bedingungen ist mit den vorliegenden Daten schwer zu erreichen.

Zudem wird der zeitliche Abstand zum vorherigen Job in Minuten als weiteres Feature berechnet. Dazu wird die Endzeit des älteren Jobs von der Startzeit des neueren subtrahiert. Ein großer Sprung der Datenmenge nach mehreren Tagen ist weniger auffällig als derselbe Sprung nach wenigen Stunden.

Residuenberechnung durch Log-Transformation und EWMA Die Attribute SOA, SOMOD und JET müssen von ihrem natürlichen Trend bereinigt werden. Da die Daten Long Tails aufweisen, wird vor der Residuenberechnung eine Log-1-p-Transformation durchgeführt ($f(x) = \log(1 + x)$), damit die Trends für die folgende EWMA erkennbar werden [13].

Um die Residuen zu berechnen, wird der mithilfe der EWMA gewichtete log-transformierte Wert vom ursprünglichen log-transformierten Wert subtrahiert ($res = \log(x) - EWMA(\log(x))$). Der EWMA wird meist zur Glättung von Zeitreihen verwendet. Je länger ein Datenpunkt zurückliegt, desto weniger wird er gewichtet. Aufgrund dieser Eigenschaften eignet er sich zur Trendbereinigung der vorliegenden Daten. [45]

Validierung der Zeitunabhängigkeit Nach der Residuenberechnung werden die Daten mithilfe der AKF überprüft, und Datensätze, die den Schwellenwert von 0,3 weiterhin überschreiten, werden für die folgende Analyse ausgeschlossen. Dadurch wird sichergestellt, dass die Methoden mit adäquaten Daten arbeiten.

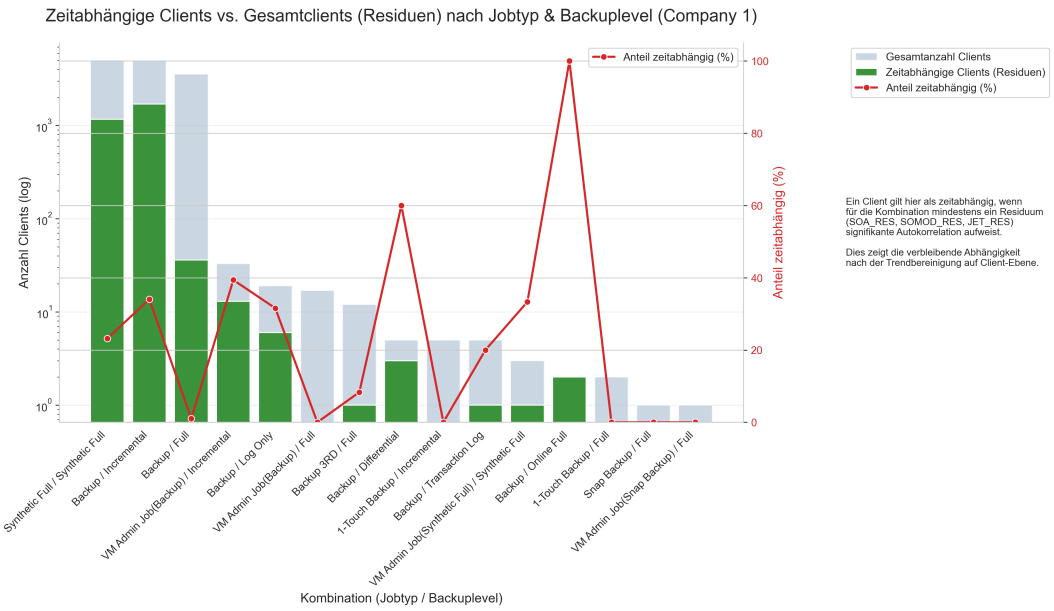


Abbildung 3: Anzahl der Clients mit zeitlicher Abhängigkeit pro Gruppierung nach Regression

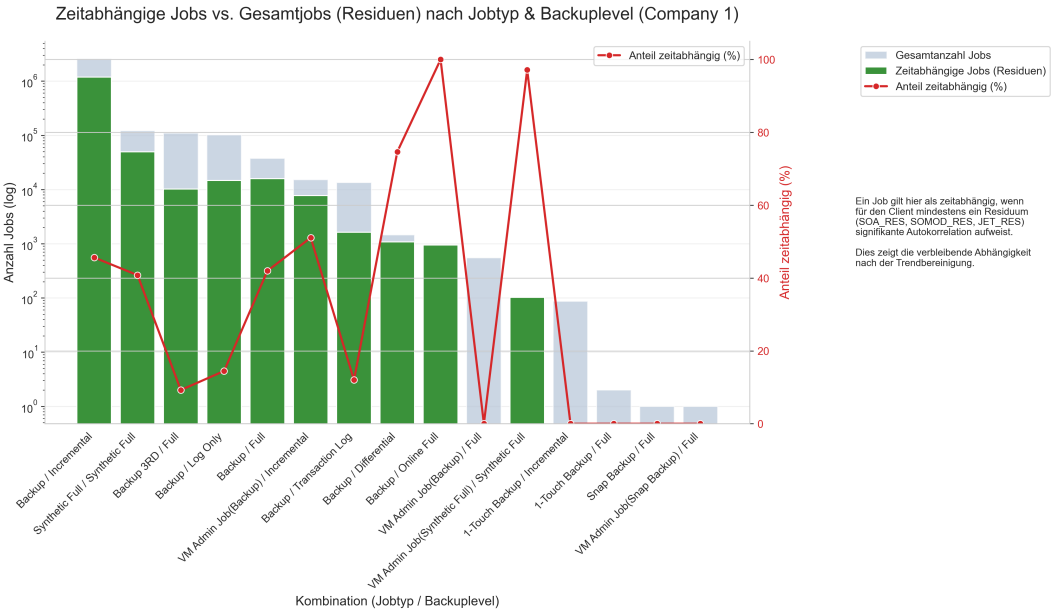


Abbildung 4: Anzahl der Jobs mit zeitlicher Abhängigkeit pro Gruppierung nach Regression

Wie in den Abbildungen 3 und 4 zu sehen ist, sind weiterhin einige Daten zeitabhängig. Diese werden in der Extremwertanalyse entsprechend nicht betrachtet.

Modellspezifische Datenübergabe Da die verwendeten Methoden unterschiedliche Stärken besitzen, werden ihnen unterschiedliche Attribute zugewiesen:

- Isolation Forest: Diesem Algorithmus werden die meisten Attribute zugewiesen. Die Zeitabhängigkeit wird durch die Angabe der Zeit des letzten Jobs umgangen. Somit erhält er SOA, SOMOD, JET, TSLJ, die Stunde, zu der der Job gelaufen ist, den Wochentag, an dem der Job gelaufen ist, das Deduplikationsverhältnis und den Durchsatz.
- Multivariate t-Verteilung und Copulas: Diese Modelle erhalten die bereinigten Verhältnisse.
- Extremwerttheorie: Dieses Verfahren wird univariat auf die zeitunabhängigen Residuen von SOA, SOMOD und JET sowie auf das Deduplikationsverhältnis und den Durchsatz angewendet, um tatsächliche Abweichungen vom Normalverhalten strukturiert zu modellieren.

Mithilfe dieser Verfahren ist es möglich, die Daten skalierbar und ressourceneffizient für die anschließenden Methoden aufzubereiten.

3.3 Verfahren zur Anomalieerkennung

Um eine zuverlässige Anomalieerkennung anhand der vorliegenden Daten zu gewährleisten, werden mehrere komplementäre Methoden miteinander verknüpft. Es kommen sowohl ein baumbasierter maschineller Lernalgorithmus als auch mehrere robuste, multivariate statistische Modelle zum Einsatz. Durch diese Kombination lassen sich unterschiedliche Anomalietypen identifizieren und evaluieren.

Beim maschinellen Lernen fiel die Wahl auf den Isolation Forest. Dieses Verfahren bietet eine robuste Grundlage für die Identifikation von Punktausreißern, da es Anomalien gezielt isoliert und somit weniger anfällig für die Schiefe der Daten ist als klassische distanzbasierte Verfahren (beispielsweise kNN-Algorithmen). [18]

Ergänzend wird eine mathematische Modellierung der Daten durchgeführt. Aufgrund der Schiefe der vorliegenden Daten wurde sich für eine Modellierung mithilfe der multivariaten t-Verteilung entschieden. Diese Verteilung eignet sich zur Modellierung von Daten mit Long Tails. Die Extremwerte können somit besser abgefangen werden und die Verteilung passt sich den Gegebenheiten besser an. [46]

Um die Abhängigkeiten der betrachteten Merkmale und die daraus resultierenden Anomalien besser einzuordnen, werden Copula-Modelle verwendet. Durch Copula-Modelle lassen sich multivariate Anomalien identifizieren, bei denen die einzelnen Merkmale für sich genommen unauffällig sein mögen, deren Kombination jedoch stark vom erwarteten Muster abweicht. [47, Kapitel 1], [19]

Zuletzt werden die Methoden um die statistische Extremwertanalyse erweitert. Da im Kontext möglicher Cyber-Attacken besonders Extremwerte von Bedeutung sind, ist es sinnvoll, diese zu modellieren und zu differenzieren, mit welchen Wahrscheinlichkeiten sie auftreten. Dies ist mit der Extremwertanalyse möglich, wodurch zwischen normalen Backup-Verhalten und Anomalien besser unterschieden werden kann. [38, Kapitel 1 und 4]

3.3.1 Isolation Forest

Die meisten modellbasierten Verfahren zur Anomalieerkennung definieren zunächst Normalverhalten, um anschließend anormales Verhalten davon abzugrenzen. Dies führt jedoch häufig sowohl zu falsch-positiven als auch zu falsch-negativen Ergebnissen. Der Isolation Forest hingegen isoliert gezielt Anomalien. [18]

Dazu werden zwei Annahmen über Anomalien getroffen. Zum einen machen Anomalien nur einen kleinen Teil der gegebenen Daten aus. Und zum anderen zeigen ihre Werte eine starke Abweichung von der Norm. Aufgrund dieser Eigenschaften werden Anomalien bereits relativ nah an der Baumwurzel isoliert, während die Isolation der normalen Datenpunkte erst später erfolgt. Dadurch ist das Verfahren sehr recheneffizient, und anhand der Pfadlänge lässt sich erkennen, ob eine Anomalie vorliegt. [18]

Genutzt werden sogenannte iTrees. Dies sind Binärbäume, die durch die rekursive Aufteilung von Daten entstehen. Terminiert wird dieser Algorithmus, wenn der Baum eine maximale Höhe erreicht, die Knoten nur noch einen Datenpunkt enthalten oder alle Daten denselben Wert besitzen. Die Daten werden anhand eines zufällig ausgewählten Attributs und eines zufällig ausgewählten Trennwertes aufgeteilt. Bei einer zufälligen Aufteilung anhand eines zufällig ausgewählten Attributs haben Anomalien im Durchschnitt eine kürzere Pfadlänge als normale Werte. [18]

Um das Verfahren robust zu gestalten, werden jedoch nicht nur ein Baum, sondern mehrere erstellt, sodass Anomalien anhand des kürzesten durchschnittlichen Pfades erkannt werden. Zudem benötigt der Algorithmus nur die Anzahl der zu erstellenden Bäume und die Teilstichprobengröße. Der Algorithmus ermöglicht es, mit einer kleinen Anzahl an Bäumen und Teilstichprobengrößen gute Ergebnisse zu erzielen, und das mit hoher Recheneffizienz, da die Bäume nicht vollständig ausgebildet werden müssen, um Anomalien zu identifizieren. Zudem eignet er sich für große Datenmengen. [18]

Der in diesem Algorithmus verwendete Anomalie-Score berechnet sich wie folgt:

$$s(x, n) = 2^{-\frac{E(h(x))}{c(n)}} \tag{4}$$

mit $h(x)$ als Pfadlänge, $E(h(x))$ als der Durchschnitt der Pfadlängen mehrerer Isolation Trees und $c(n)$ als Normierungsfaktor, der die durchschnittliche Pfadlänge eines binären Suchbaums mit n Knoten darstellt. Wenn $s(x, n)$ nahe 1 liegt, ist der Wert sehr wahrscheinlich eine Anomalie; falls $s(x) < 0,5$ ist, ist der Wert wahrscheinlich keine Anomalie. Falls für alle Werte $s(x, n) \approx 0,5$ gilt, liegen in den Daten keine auffälligen Anomalien vor. [18]

Aufgrund dieser Eigenschaften eignet sich der Isolation Forest für die vorliegenden Daten. Es ist ein Algorithmus des unüberwachten Lernens, der recheneffizient ist und mit großen Datenmengen funktioniert. Durch das zuvor durchgeführte Gruppieren sollte es möglich sein, abnormale Datenpunkte schnell zu isolieren und mithilfe eines Schwellenwertes zu klassifizieren.

3.3.2 Multivariate t-Verteilung

Zur Modellierung von Daten werden häufig Normalverteilungen verwendet, doch sie sind gegenüber Ausreißern nicht sehr robust [46]. Dass sie sich nicht zur Modellierung der vorliegenden Daten eignet, wurde bereits gezeigt [1]. Aufgrund der unterschiedlichen Größe der Backups werden Mittelwert und Standardabweichung stark verzerrt, und mögliche Anomalien werden nicht erkannt [1]. Unter Robustheit versteht man in diesem Kontext, dass kleine Abweichungen von den Vermutungen keinen großen Einfluss auf die verwendeten Methoden haben [48, Kapitel 1]. Die Vermutungen können beispielsweise über die Verteilung, Zufälligkeit oder Unabhängigkeit erfolgen [48, Kapitel 1].

Anders als bei der Normalverteilung werden für die t-Verteilung drei Parameter benötigt:

1. μ, der den typischen Zustand des Systems, in diesem Fall der Metadaten von Commvault, beschreibt,
2. ψ, die Skalierungsmatrix, welche die Streuung der Daten und die Korrelation unter den Merkmalen beschreibt, sowie
3. ν, die Freiheitsgrade, welche die Schwere der Tails steuert [46]

ν kann a priori eingestellt werden und dient als Parameter, mit dem man die Robustness feinjustieren kann. Allerdings ist es mit genügend Daten auch möglich, ν mithilfe eines MLE zu schätzen, wodurch man eine „adaptive robuste Prozedur“ [46] erhält. Solange mehr als 15 Daten vorhanden sind, wird die adaptive Schätzung von ν verwendet. [46]

Zur Schätzung der Parameter wird der EM-Algorithmus verwendet. Dies ist ein iterativer Algorithmus, der das Problem unvollständiger oder verborgener Daten löst, indem er abwechselnd einen Schätzschritt (Estimation) und einen Maximierungsschritt ausführt [49]. Bei der Schätzung wird für jeden Datenpunkt ein individuelles Gewicht w_i geschätzt, das angibt, wie stark dieser Punkt der angenommenen Verteilung entspricht [46]. Die Gewichtung erfolgt mithilfe der Freiheitsgrade ν, der Anzahl der Komponenten/Dimensionen des Vektors y_i (v_i), sowie der Mahalanobis-Distanz δ_i^2 [46]:

$$w_i = \frac{\nu + v_i}{\nu + \delta_i^2} \tag{5}$$

Unter der Mahalanobis-Distanz versteht man eine robuste Distanz, die die Korrelationsstruktur der Daten berücksichtigt, indem sie den Abstand eines Datenpunktes zum Mittelpunkt normiert durch die Kovarianzmatrix misst [50].

In diesem Fall erfolgt die Anomalieerkennung durch Gewichtung. Die Gewichtung gibt an, wie relevant die t-Verteilung für einen einzelnen Datenpunkt bei der Bestimmung der Wahrscheinlichkeitsverteilung ist. Falls das Gewicht eines Datenpunktes sehr klein ist, ist die Wahrscheinlichkeit, dass es sich um eine Anomalie handelt, relativ hoch. Zur Klassifizierung der Anomalien ergeben sich nun zwei Möglichkeiten. Entweder die Nutzung eines festen Gewichts-Grenzwertes oder ein Vergleich mit anderen Punkten mit geringer Gewichtung. Dies lässt sich mithilfe von Visualisierungen darstellen; für das spätere Frühwarnsystem werden diejenigen Datenpunkte als Anomalien klassifiziert, deren Mahalanobis-Distanz zu anderen Punkten größer ist.

Mithilfe der multivariaten t-Verteilung ist es zum einen möglich, Anomalien in den Verhältnissen der Attribute untereinander zu identifizieren, zum anderen globale Punktausreißer zu identifizieren und zu klassifizieren.

Die Modellierung von Daten mit Tails, die kleiner sind als bei der Normalverteilung oder deren Länge sich in den verschiedenen Attributen stark unterscheidet, ist nicht vorteilhaft [46]. Der erste Fall kommt in den vorliegenden Daten selten vor [1] und um den zweiten Fall zu minimieren, werden im Folgenden Copula-Modelle auf die Daten angewendet.

3.3.3 Copula-Modelle

Copulas sind Funktionen, die die Abhängigkeiten zwischen den einzelnen Dimensionen in einem multivariaten Datensatz darstellen. Dabei ist zu beachten, dass jedes Attribut eine eigene Verteilung aufweisen kann. Diese wird als Randverteilung bezeichnet. Die Randverteilungen sowie die Copula besitzen den Wertebereich $[0; 1]$. [47, Kapitel 2]

Der Name Copula verdeutlicht die Kopplung einer gemeinsamen Verteilungsfunktion zu ihren univariaten Randverteilungen. Gemäß dem Theorem von Sklar existiert für jede gemeinsame Verteilungsfunktion ($H(x, y)$) eine Copula (C), die die Randverteilungen (F, G) für alle x, y beschreibt:

$$H(x, y) = C(F(x), G(y)) \tag{6}$$

[47, Kapitel 2]

Das Theorem besagt, dass es für jede gemeinsame Verteilung (mit Eingabewerten $\in \overline{R}$ (erweiterte reelle Zahlen)) eine solche Copula gibt. Falls die Randverteilungen zudem stetig sind, ist diese Copula eindeutig. Falls nicht, dann ist die Copula für das kartesische Produkt der Wertebereiche ($Ran(F) \times Ran(G)$) eindeutig. Dies gilt auch im Umkehrschluss. Wenn eine Copula der Randverteilungen gegeben ist, dann beschreibt sie deren gemeinsame Verteilungsfunktion. [47, Kapitel 2]

Dieses Theorem gilt zudem für die Zufallsvariablen (X, Y). Mithilfe von Copulas lässt sich auch bestimmen, ob diese Zufallsvariablen abhängig oder unabhängig sind. Falls die Copula dem Produkt ihrer Randverteilungen entspricht, dann sind die Variablen unabhängig: $C_{XY} = \Pi$. Diese Definitionen gelten ebenso für multivariate Daten mit mehr als zwei Zufallsvariablen. [47, Kapitel 2]

Copulas werden bereits zur Anomalieerkennung eingesetzt. Ein etabliertes Verfahren ist der COPOD-Algorithmus, der eine empirische Copula erstellt und mit dieser die Wahrscheinlichkeiten der Extremität der Datenpunkte darstellt. Dieser Algorithmus ist parameterfrei und zudem recheneffizient. [19]

Anders als andere Algorithmen zur Anomalieerkennung benötigt COPOD weder Lernen noch Trainieren, was zu sichereren Ergebnissen führen kann. Andere Algorithmen, die verschiedene Parameter benötigen, liefern oftmals unterschiedliche Ergebnisse je nach Wahl der Hyperparameter. Zudem besitzt COPOD den Vorteil, dass die Anzahl der Attribute die Recheneffizienz des Algorithmus nicht sonderlich beeinträchtigt. Dies ist für den vorliegenden Datensatz mit drei betrachteten Attributen nicht von großer Bedeutung, erleichtert jedoch eine mögliche Skalierung. [19]

Der Algorithmus berechnet zunächst für die Randverteilungen die ECDF. Dies ist notwendig, da die tatsächlichen Verteilungen der einzelnen Attribute meist unbekannt sind. Mithilfe der ECDF werden sogenannte empirische Copulas erstellt, um die Wahrscheinlichkeiten an den Rändern der Verteilungen, auch Tails genannt, zu approximieren. [19]

Für jeden Datenpunkt werden sowohl die Wahrscheinlichkeit für den linken als auch für den rechten Tail berechnet. Es wird entsprechend ermittelt, welcher Anteil der Datenpunkte kleiner oder gleich (linker Tail) bzw. größer oder gleich (rechter Tail) dem betrachteten Wert x_i ist. Ein Datenpunkt gilt als Anomalie, wenn er in einer oder mehreren Dimensionen in einem der beiden Tails extrem kleine Wahrscheinlichkeiten aufweist. [19]

Um dem sogenannten „Curse of Dimensionality“ entgegenzuwirken, macht sich COPOD die Eigenschaften der Unabhängigkeit zunutze. Anstatt die Distanz in einem hochdimensionalen Raum zu berechnen, evaluiert der Algorithmus die Wahrscheinlichkeiten für jedes Attribut eindimensional und unabhängig voneinander. Um den finalen Anomalie-Score eines Datenpunktes zu berechnen, werden die negativen Logarithmen dieser univariaten Tail-Wahrscheinlichkeiten über alle Dimensionen hinweg aufsummiert. Je größer dieser Score, desto unwahrscheinlicher ist

das Auftreten des Datenpunktes und desto sicherer handelt es sich bei ihm um eine Anomalie. [19]

Aufgrund dieser Eigenschaften wurde entschieden, den COPOD-Algorithmus auf die vorliegenden Daten anzuwenden. Zwar stellt die Anzahl der Dimensionen für den untersuchten Datensatz aktuell kein Problem dar, jedoch ist die genaue Verteilung der einzelnen Attribute unbekannt. Zudem liegt die Vermutung nahe, dass die Attribute im Normalfall gleichbleibende Beziehungen zueinander aufweisen, die bei einem abnormalen Verhalten des Systems (wie beispielsweise einem Ransomware-Angriff) gebrochen werden. Da COPOD parameterlos und recheneffizient arbeitet, lassen sich die vorliegenden großen Datenmengen schnell verarbeiten, wodurch sich der Algorithmus bei entsprechenden Ergebnissen ideal für den Einsatz in einem zukünftigen Frühwarnsystem eignet.

3.3.4 Extreme Value Theory über Schwellenwerte

Die Extremwertanalyse basiert auf der Erkenntnis, dass sogenannte extreme Daten (also Daten aus nichtstationären Zeitreihen oder Datensätzen, die einen Schwellenwert überschreiten) als Stichproben aus einer Gruppe von Extremwertverteilungen betrachtet werden können, die sich aus den beobachteten Daten statistisch schätzen lassen [20, Kapitel 2]. Ziel ist es, das stochastische Verhalten dieser Extreme zu quantifizieren. Dazu müssen oft Wahrscheinlichkeiten für Ereignisse geschätzt werden, die extremer sind als die bisher gemessenen Werte.

Klassischerweise nutzt man das Block-Maxima-Verfahren. Es werden nur die absoluten Maxima aus einem Beobachtungszeitraum betrachtet:

$$M_n = \max\{X_1, ..., X_n\} \tag{7}$$

Hierbei sind $X_1, ..., X_n$ eine Folge von i.i.d.-Variablen, die gewöhnlich in diskreten, gleichbleibenden Zeitabständen gemessen werden. Aufgrund der Beschaffenheit der Metadaten eignet sich dieser Ansatz jedoch nicht. Da pro Zeitblock nur das absolute Maximum betrachtet wird, fallen zu viele potenziell relevante Extremwerte aus der Untersuchung heraus, wodurch Informationen über das Datenverhalten verloren gehen. [38, Kapitel 1 und 3]

Um diese Dateneffizienz zu überwinden, wird in dieser Arbeit ein Schwellenwertansatz gewählt. Die Werte, die einen ausreichend hohen Schwellenwert u überschreiten, können asymptotisch mit der GPD modelliert werden. [38, Kapitel 4]

Die Wahl dieses Schwellenwertes ist entscheidend, denn wählt man ihn zu klein, besteht das Risiko, dass die Annahme des asymptotischen Verhaltens verletzt wird. Wählt man ihn zu groß, bleiben jedoch zu wenige Datenpunkte übrig, um die GPD verlässlich zu schätzen. [38, Kapitel 4]

Die zugrunde liegenden Theoreme dieses Ansatzes setzen voraus, dass die Daten einer stetigen Verteilung folgen [51]. Die vorliegenden Daten sind diskret, doch aufgrund des großen Wertebereichs kann in der Modellierung eine Quasi-Stetigkeit angenommen werden. Eine weitere Annahme ist, dass die betrachteten Variablen strikt positiv sind [52]. Da beim Schwellenwert-Ansatz jedoch nicht die absoluten Datenpunkte modelliert werden, sondern ausschließlich die Differenz zwischen den Datenpunkten, die größer als der (positiv definierte) Schwellenwert sind, und dem Schwellenwert, ist diese Bedingung garantiert. Da das Ziel dieser Arbeit die Erkennung anomal großer Metadaten ist, wird ausschließlich der rechte Tail der Verteilung betrachtet.

Das genaue Ausmaß des Tails wird durch den sogenannten Shape-Parameter (häufig mit c oder ξ bezeichnet) bestimmt. Wenn $c > 0$, dann entspricht die Tail-Verteilung einer Pareto-Verteilung. Ist $c = 0$, dann folgt der Tail einer exponentiellen Abnahme. Bei $c < 0$ besitzt die Verteilung eine absolute Obergrenze. [38, Kapitel 4], [51]

Um den Shape-Parameter sowie die Skalierung an die vorliegenden Daten anzupassen, wird zur Parameterschätzung der MLE genutzt, da sich dieser effizient und robust an die Verteilung anpasst [38, Kapitel 1].

Dieses Verfahren eignet sich für die vorliegenden Daten, da ein Extremwert, etwa ein sehr großes Backup, nicht zwangsläufig eine Anomalie ist. Die finale Klassifikation, ob ein Extremwert tatsächlich anomal ist, erfolgt anhand eines Signifikanzniveaus. Über die kumulative Verteilungsfunktion der GPD wird berechnet, mit welcher Wahrscheinlichkeit ein bestimmter Wert auftritt. Dafür wird die unbedingte Überschreitungswahrscheinlichkeit, im Folgenden als p-Wert bezeichnet, genutzt. Diese setzt sich wie folgt zusammen:

$$P(X > x) = P(X > x|X > u) \cdot P(X > u) \tag{8}$$

Dabei ist $P(X > u)$ der Anteil der Daten, die den Schwellenwert überschritten haben (oftmals als ζ_u bezeichnet), und $P(X > x|X > u)$ die von der GPD modellierte Wahrscheinlichkeit [38, Kapitel 4].

Durch den Abgleich mit dem definierten Signifikanzniveau wird eine mögliche Anomalie klassifiziert. Dies verhindert die Klassifikation zu vieler falsch-positiver Anomalien.

3.4 Vergleich der Methoden

In dieser Arbeit werden verschiedene Methoden verwendet, um Anomalien in den Metadaten der Datensicherung am Beispiel von Commvault zu erkennen. Diese Methoden decken verschiedene Bereiche ab und betrachten die Daten aus unterschiedlichen Perspektiven. Während der Isolation Forest, die multivariate t-Verteilung sowie Copula-Modelle den Datensatz multivariat analysieren, verfolgt die Extremwerttheorie einen univariaten Ansatz.

Zudem ergänzen sich die Verfahren hinsichtlich der erkannten Ausreißer. Copulas identifizieren Kollektiv-Ausreißer, es sei denn, die Abweichung in einer einzelnen Dimension ist gravierend genug, um den Datenpunkt als Anomalie zu klassifizieren. Der Isolation Forest und die multivariate t-Verteilung sind hingegen in der Lage, sowohl Kontext- als auch Punktausreißer zu identifizieren. Die Extremwerttheorie erkennt typischerweise univariate Punktausreißer. Falls diese jedoch in Folge auftreten, können im Kontext einer Zeitreihe auch Segmentausreißer identifiziert werden.

Die Eingabedaten für die jeweiligen Algorithmen werden gezielt angepasst, um die Stärken der Verfahren optimal auszuschöpfen. Dem Isolation Forest werden als Basisalgorithmus neben den Residuen nahezu alle generierten Merkmale übergeben, um komplexe Zusammenhänge zu isolieren. Die multivariate t-Verteilung sowie COPOD verwenden hingegen die Verhältnisse. Dies ermöglicht es, Strukturbrüche zu identifizieren, ohne die Daten zuvor skalieren zu müssen. Im Gegensatz dazu arbeitet die Extremwertanalyse univariat an den Residuen des EWMA. Dies stellt zum einen sicher, dass die für die EVT notwendige i.i.d.-Annahme erfüllt ist. Zum anderen werden hier bewusst keine Verhältnisse genutzt, da diese Anomalien maskieren könnten, die in den reinen Attributen (oder in diesem Fall deren Residuen) auffällig wären.

Alle verwendeten Methoden benötigen jedoch eine Mindestanzahl an Datenpunkten. Zudem werden heuristische Schwellenwerte zur Klassifikation von Anomalien verwendet, da die Methoden grundsätzlich auf Scores basieren.

Um die Eignung dieser Verfahren abzuwägen, fasst die folgende Tabelle 1 deren Stärken und Schwächen zusammen:

	Isolation Forest	**Multivariate t-Verteilung**	**Copulas (COPOD)**	**Extreme Value Theory (POT)**
Stärken	• Sehr recheneffizient • Identifiziert Punktausreißer • Keine Verteilungsannahmen • Wenig Parameter	• Berücksichtigt Korrelationen • Fängt Schiefe der Daten besser ab	• Parameterfrei • Sehr recheneffizient • Keine Verteilungsannahme	• Fundierte Wahrscheinlichkeiten • Fokus auf relevante Extremwerte
Schwächen	• Erkennt selten korrelierte Anomalien • Score wird über Mittelwert der Pfadlängen berechnet	• Feste Verteilungsannahme • Anfällig bei extremen Long-Tails	• Empirische Schätzung nötig	• Univariater Fokus • Stark abhängig von Schwellenwert u

Tabelle 1: Theoretischer Vergleich der verwendeten Algorithmen hinsichtlich ihrer Stärken und Schwächen für die Anomalieerkennung in Datensicherungsmetadaten.

Da die Methoden unterschiedliche Stärken und Schwächen aufweisen, wird in dieser Arbeit ein gestaffeltes Warnsystem vorgeschlagen. Ziel ist es, für das spätere Frühwarnsystem eine Stufenlogik zu etablieren. Wenn nur eine Methode eine Anomalie identifiziert, wird eine einfache Warnung angezeigt. Wenn zwei oder mehr Methoden einen Job als Anomalie klassifizieren, wird ein Alarm ausgelöst.

3.5 Evaluationsmetriken

Um die verwendeten Methoden evaluieren zu können, wurden mithilfe von Dummy-Clients künstliche Anomalien erzeugt. Dafür wurden vier Test-Clients verwendet:

- Client 1: Erschaffen einer Baseline, plötzliches Erstellen vieler neuer Daten, neue Baseline, Verschlüsselung
- Client 2: Erschaffen einer Baseline, plötzliches Erstellen vieler neuer Daten, neue Baseline, Daten plötzlich löschen
- Client 3: Erschaffen einer Baseline, plötzliches Erstellen vieler neuer Daten, neue Baseline, schleichende Verschlüsselung von 10 % pro Job

- Client 4: Referenzclient, läuft nur mit Baseline

Dadurch wurden gelabelte Daten erzeugt, weshalb die Konfusionsmatrix zur Evaluierung der Algorithmen genutzt werden kann [7]. Mithilfe dieser werden Precision-Recall und der daraus resultierende RankPower-Score berechnet. Precision berechnet den Anteil der tatsächlichen Anomalien an allen vom Algorithmus als anomal klassifizierten Datenpunkte ($Pr = \frac{m_t}{m}$, wobei m_t die tatsächlichen Anomalien in den erkannten Anomalien und m die erkannten Anomalien sind). Recall hingegen prüft, wie viele der tatsächlichen Anomalien identifiziert werden ($Re = \frac{m_t}{d_t}$ wobei d_t die tatsächlichen Anomalien im Datensatz sind). Diese beiden Scores werden im RankPower vereint ($RP = \frac{m_t(m_t+1)}{2 \cdot \sum_{i=1}^{m_t} R_i}$, wobei R_i der Rang des Datenpunktes ist). Die Rankpower wird verwendet, um die Gewichtung der Scores mit einzubeziehen. Je höher der Score, desto höher ist die Wahrscheinlichkeit, dass es sich um eine Anomalie handelt. Diese Erkenntnis wird in der Bewertung berücksichtigt. [2, Kapitel 2]

3.6 Datenpipeline

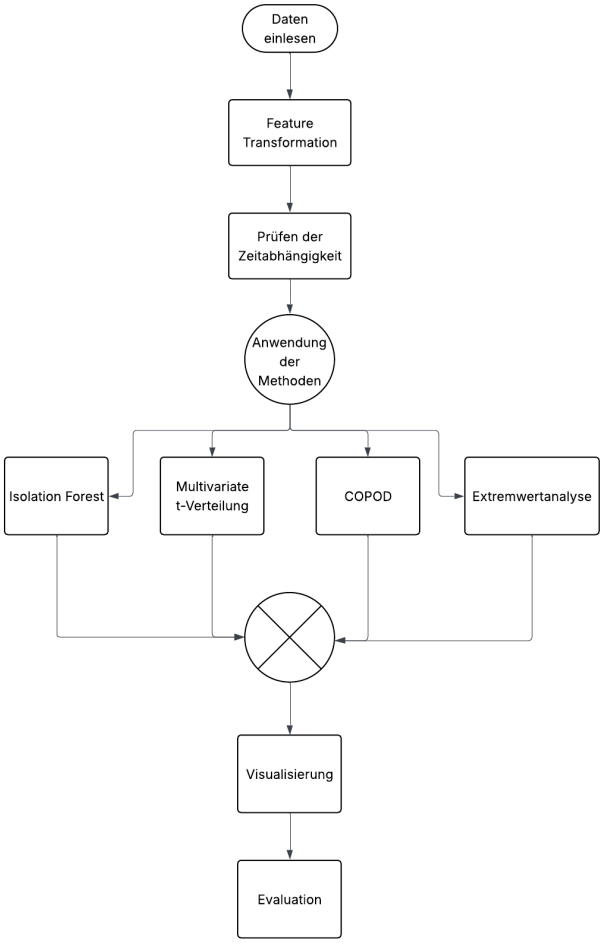


Abbildung 5: Datenpipeline Flussdiagramm

4 Analyse

Zur Analyse wurden ausschließlich Daten der RWTH Aachen genutzt, wobei alle aufgezeichneten Sicherungsvorgänge bis einschließlich den 14.06.2026 berücksichtigt wurden. Es wurden ausschließlich Jobs berücksichtigt, die den Status „Completed“ aufweisen und als Backup klassifiziert sind.

4.1 Isolation Forest

Die Anwendung des Isolation Forests basiert auf den Features SOA, SOMOD, JET, TSLJ, der Stunde, dem Wochentag sowie dem Deduplikationsverhältnis und dem Durchsatz. Dadurch deckt der Isolation Forest den Großteil der Daten ab.

Um eine valide Grundlage für den Isolation Forest zu gewährleisten, ist zu beachten, dass für die Analyse einer Gruppe mindestens 15 Jobs erforderlich sind, da zur Berechnung des Scores eine gewisse Pfadlänge benötigt wird. Aufgrund dessen wurden insgesamt 21.556 Jobs nicht berücksichtigt, da in einer Gruppe nicht genügend Jobs zur Analyse vorlagen.

Insgesamt wurden 2.970.020 Jobs betrachtet, von denen 14.755 einen Anomalie-Score über 0,7 erreichten. Dies sind 0,5 % der Gesamtdatenmenge. Der Schwellenwert von 0,7 wurde gewählt, da Werte nahe 1 mit hoher Wahrscheinlichkeit eine Anomalie identifizieren [18].

Wie in Abbildung 6 ersichtlich, ist die resultierende Verteilung relativ links zentriert und weist im rechten Bereich Long-Tails auf. Da die Verteilung schief ist und Long Tails in Richtung 1 aufweist (statt sich um den Wert 0,5 zu konzentrieren), lässt sich schlussfolgern, dass Anomalien in den Daten vorliegen [18]. Der Anomalie-Score wird durch die durchschnittliche Pfadlänge bestimmt; je kürzer diese ist, desto höher ist der Score [18]. Der hier genutzte Schwellenwert von 0,7 ist ein heuristischer Schwellenwert, der eine hohe Wahrscheinlichkeit dafür hat, dass der betrachtete Datenpunkt tatsächlich eine Anomalie ist.

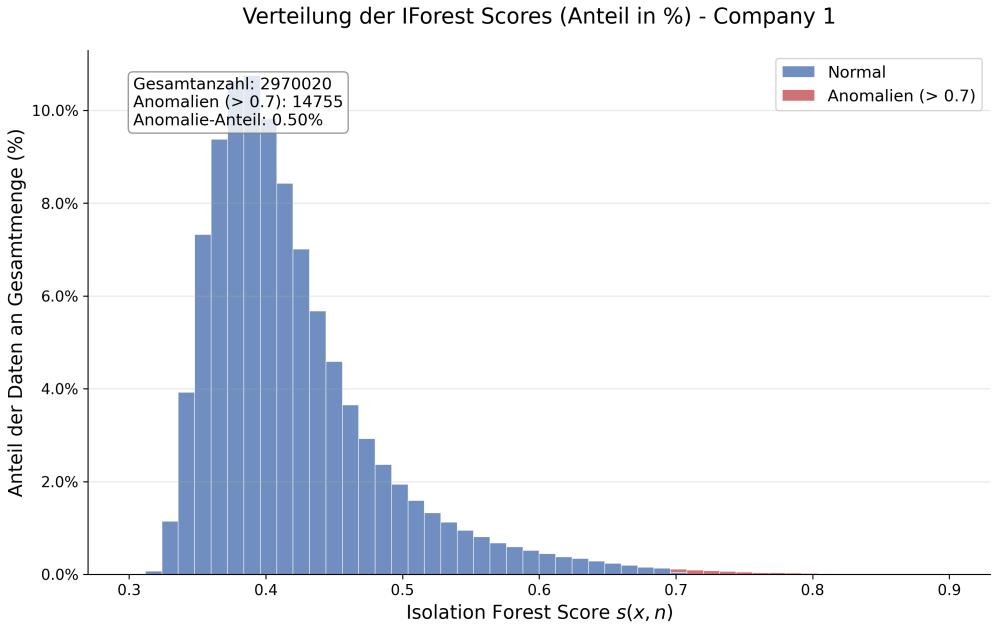


Abbildung 6: Anomalie-Score des Isolation Forest in Prozent

Zur qualitativen Validierung werden im Folgenden zwei konkrete, vom Algorithmus identifizierte Anomalien im Detail analysiert:

Fallbeispiel 1: Anomalie durch Konfiguration Der Job 36143207 des Clients 20049 ist einer von 181 abgeschlossenen Jobs und weist einen Anomalie-Score von 0,9 auf. Es handelt sich um ein inkrementelles Backup mit einer Laufzeit von 8 Stunden und 24 Minuten. Von diesen 181 Jobs wurden 2 vom Isolation Forest als Anomalie klassifiziert.

Attribut	Analysierter Job	Median	Durchschnitt
SOA	1.839,5 GB	4,9 MB	10,2 GB
SOMOD	814,8 GB	18,5 MB	4,5 GB
JET	8h 24m	4m	6m 42s
Deduplikationsverhältnis	2,6	0,3	0,4
Durchsatz	60,9 MB/s	21,9 KB/s	0,4 MB/s

Tabelle 2: Vergleich der Attribute des analysierten Jobs 36143207 mit dem Median und dem Durchschnitt des Clients 20049.

Die Ursache für die Klassifikation als Anomalie liegt, neben den hohen Werten in den anderen Attributen, daran, dass der Job im Attribut TSLJ den Wert 0 besitzt. Es ist der erste Backup-Job, den der Client durchführt. Standardmäßig würde als erster Job ein Full-Backup laufen, da dies jedoch nicht der Fall ist, klassifiziert der Algorithmus diesen Job korrekterweise als Anomalie. Auffällig ist zudem, dass SOMOD beim Median größer ist als SOA. Dies und der geringe Deduplikationswert weisen möglicherweise auf große Datenveränderungen sowie auf Probleme bei der Deduplikation hin. Von Interesse ist zudem, dass das Deduplikationsverhältnis für diesen ersten Job sehr hoch ist, während der Median sowie der Durchschnitt deutlich niedrigere Werte aufweisen.

Fallbeispiel 2: Anomalie durch abnormale Werte in Attributen Ein weiteres Beispiel ist der Job 26790845 des Clients 16269. Dieser hat einen Anomalie-Score von 0,9 und ist einer von 341 abgeschlossenen Jobs, von denen der Isolation Forest 2 als Anomalie identifiziert hat. Es ist ein inkrementelles Backup, das 168 Stunden und 18 Minuten lief. Es vergingen 96 Stunden und 58 Minuten seit dem letzten Backup, das ein Synthetic-Full war.

Tabelle 3 stellt die Charakteristika dieses Jobs dem Median und dem Durchschnitt der regulären inkrementellen Backups dieses Clients gegenüber. Alle Attribute sind bei diesem Ausreißer stark erhöht oder im Fall des Deduplikationsverhältnisses signifikant verringert.

Attribut	Analysierter Job	Median	Durchschnitt
SOA	9.023,4 GB	10,2 GB	44,5 GB
SOMOD	5.263,1 GB	0,2 GB	19 GB
JET	168h 18m	7m	43m 18s
Deduplikationsverhältnis	1,7	54	57
Durchsatz	14,9 MB/s	29,1 MB/s	31,3 MB/s

Tabelle 3: Vergleich der Attribute des analysierten Jobs 26790845 mit dem Median und dem Durchschnitt der inkrementellen Backups des Clients 16269.

Zusätzlich sind die Werte jedes untersuchten Attributs bei diesem inkrementellen Backup größer als die des vorangegangenen Synthetic-Fulls (das 778,3 GB bei SOA, 562,2 bei SOMOD sowie 33 Stunden und 55 Minuten bei JET benötigt hat). Auch das Deduplikationsverhältnis (1,4) und der Durchsatz (6,3 MB/s) des Synthetic-Fulls wurden übertroffen. Das niedrigere Deduplikationsverhältnis des Synthetic-Fulls weist jedoch darauf hin, dass dieses inkrementelle Backup erneut Daten speichert, die bereits im Commvault-Speichersystem vorhanden sind.

4.2 Multivariate t-Verteilung

Anders als beim Isolation Forest wird zur Modellierung der multivariaten t-Verteilung auf Basis der zuvor transformierten Daten gearbeitet; dabei werden nur das Deduplikationsverhältnis und der Durchsatz berücksichtigt. Zur Modellierung der Verteilung werden mindestens 15 Jobs benötigt. Da diese in einer Gruppe nicht immer gegeben waren, wurden insgesamt 21.556 der 2.991.576 transformierten Jobs nicht berücksichtigt. Wie in Abbildung 7 zu erkennen ist, wurden insgesamt 2.970.020 Jobs betrachtet, wobei 34.114 Anomalien identifiziert wurden. Dies entspricht 1,15 % der analysierten Jobs. Die Anzahl der erkannten Anomalien ist in der Definition eines mitgegebenen Parameters festgelegt. Als heuristischer Richtwert für Anomalien wurde 1% gewählt, um zu viele falsch-positive Ergebnisse auszuschließen, da Anomalien per Definition selten auftreten. Der Wert ist in diesem Fall minimal erhöht, da 1 % der Daten einer Gruppierung (Kombination aus Client, Jobtype und Backuplevelname) als Anomalie erkannt wurden, was eine geringfügige Verzerrung der Gesamtanzahl verursacht:

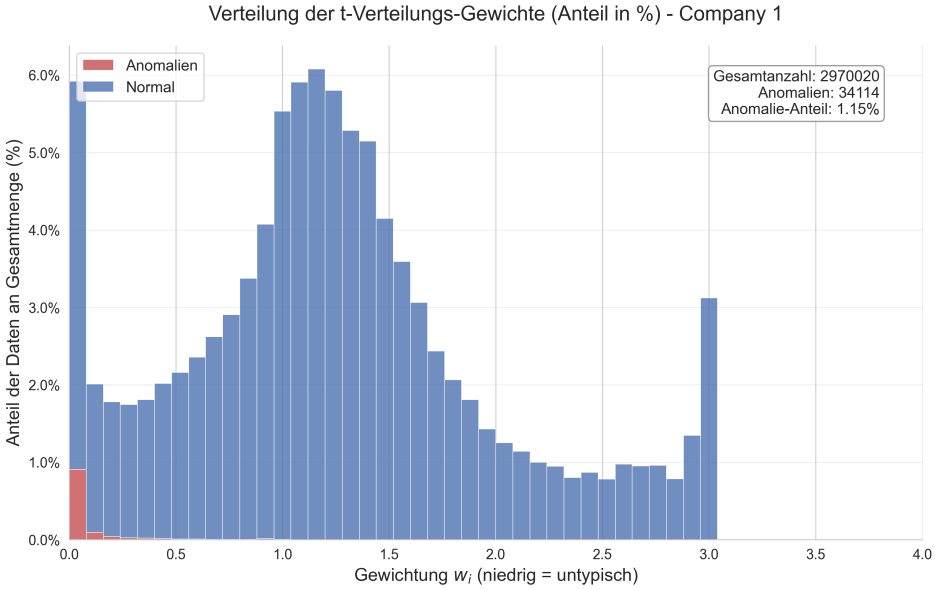


Abbildung 7: Gewichtung der Distanz der t-Verteilung in Prozent

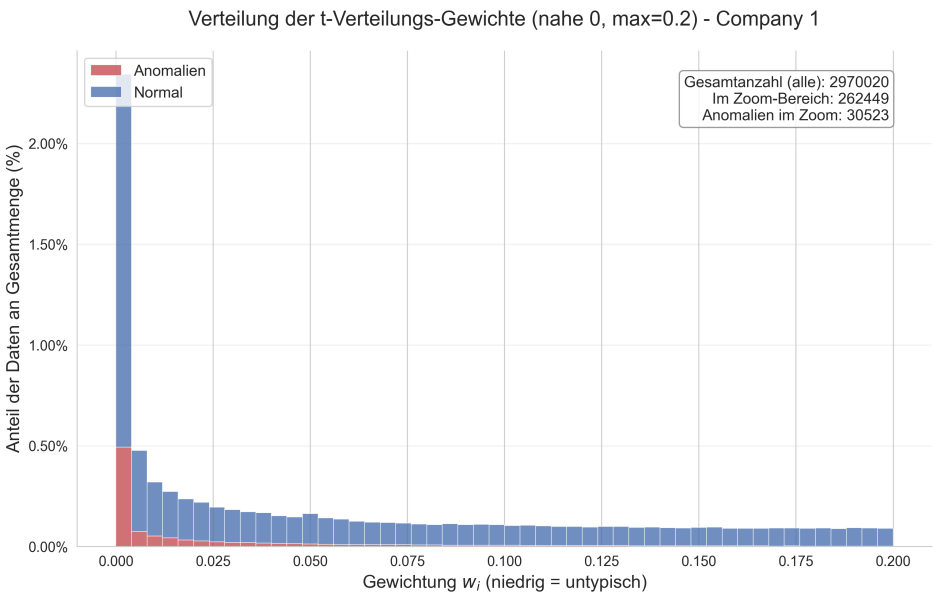


Abbildung 8: Gewichtung der Distanz der t-Verteilung in Prozent für einen kleinen Ausschnitt

Die daraus hervorgehende Verteilung der Gewichtung ist bimodal. Sie weist im Gegensatz zum Isolation Forest keine Long Tails auf. Die Gewichtung von 0 ist ein Peak (der ungefähr 10 % der Daten ausmacht), der zweite liegt bei ungefähr 1,3 (mit ca. 3,8 % der Daten) und ist rechtsschief. Zudem sind in diesem Fall die niedrigsten Gewichte die untypischsten. Werte wie 4 signalisieren beispielsweise, dass sie sehr typisch sind, während in 8 zu erkennen ist, dass viele Jobs das Gewicht 0 haben. Entsprechend werden viele Jobs als sehr unwahrscheinlich (Peak bei 0) und viele Jobs als sehr wahrscheinlich (Peak bei 1,3) eingestuft.

Die multivariate t-Verteilung benötigt, anders als die empirischen Copulas, die drei Parameter μ, ψ und ν. μ wird zuerst als das arithmetische Mittel geschätzt, jedoch wird im Laufe des EM-Algorithmus robust geschätzt. ψ wird zunächst als Standard-Kovarianz geschätzt, ebenso wie μ, jedoch während des EM-Algorithmus robust neu geschätzt. ν hingegen wird initial auf den Wert 5 gesetzt, anschließend jedoch mithilfe einer Maximum-Likelihood-Optimierung angepasst.

Auch hier werden wieder zwei verschiedene als Anomalien markierte Jobs betrachtet:

Fallbeispiel 1: Anomalie durch Fehlkonfiguration und abnormale Werte Betrachtet wird Job 30917297 vom Client 7205. Dieser Client besitzt 4971 Jobs, von denen insgesamt 52 als Anomalie durch die multivariaten t-Verteilung klassifiziert werden. Die Mahalanobis-Distanz beträgt $4 \cdot 10^{32}$, was eine Gewichtung von $10 \cdot 10^{-33}$ zur Folge hat. Es ist ein inkrementelles Backup, bei dem nahezu keine Deduplikation möglich war.

Wie Tabelle 4 veranschaulicht, zeigt dieser Client ein extremes Verhalten hinsichtlich der Verteilung seiner historischen Daten. Der Vergleich zwischen dem Median und dem arithmetischen Mittelwert verdeutlicht eine Rechtsschiefe in den regulären Backup-Mustern dieses Clients, weshalb extreme Ausreißer in der Historie dennoch zur Norm gehören. Viele Jobs speichern beispielsweise überhaupt keine Daten.

Attribut	Analysierter Job	Median	Durchschnitt
SOA	29.009,9 GB	0,0 GB	25,1 GB
SOMOD	28.719,3 GB	0,0 GB	10,4 GB
JET	187h 20m	28s	10m 16s
Deduplikationsverhältnis	1,0	0	11,3
Durchsatz	43,0 MB/s	0 MB/s	0,9 MB/s

Tabelle 4: Vergleich der Attribute des analysierten Jobs 30917297 mit dem Median und Durchschnitt der inkrementellen Backups des Clients 7205.

Es ist demnach nicht der erste Job dieses Clients, der eine so große Menge an Daten speichert. Allerdings weichen insbesondere das Deduplikationsverhältnis und der Durchsatz bei diesem spezifischen Job deutlich von den durchschnittlichen Werten ab, wobei das Deduplikationsverhältnis nach unten und der Durchsatz nach oben weichen.

Fallbeispiel 2: Anomalie aufgrund hoher Datenmenge Der Job 35576387 des Clients 12479, der 679 Jobs besitzt, ist ein inkrementelles Backup, das eine Woche nach dem letzten Synthetic-Full-Backup durchgeführt wurde. Die Mahalanobis-Distanz beträgt $5,7 \cdot 10^{20}$, woraus eine Gewichtung von $5,3 \cdot 10^{-21}$ resultiert. Insgesamt hat die multivariate t-Verteilung 8 Anomalien bei diesem Client identifiziert.

In Tabelle 5 werden die Werte dieses Ausreißers der regulären Datenverteilung des Clients gegenübergestellt. Im Vergleich zu den bisher betrachteten Jobs liegen die Werte bei SOMOD und JET nahe am Median. SOA hingegen ist erhöht. Besonders auffällig sind jedoch das extrem hohe Deduplikationsverhältnis und der hohe Durchsatz. Beide liegen um ein Vielfaches über dem Median oder dem Durchschnitt.

Attribut	Analysierter Job	Median	Durchschnitt
SOA	16,5 GB	0,0 MB	25,3 MB
SOMOD	21,8 MB	0,0 MB	0,2 MB
JET	1m 19s	3s	54s
Deduplikationsverhältnis	754	0	1,3
Durchsatz	208 MB/s	0 MB/s	0,3 MB/s

Tabelle 5: Vergleich der Attribute des analysierten Jobs 35576387 mit dem Median und dem Durchschnitt des Clients 12479.

4.3 Copulas

Wie auch die multivariate t-Verteilung arbeitet COPOD mithilfe der Attribute des Deduplikationsverhältnisses und des Durchsatzes. Auch für die Copulas werden mindestens 10 Jobs benötigt, weshalb insgesamt 21.556 der ungefähr 3 Millionen Jobs nicht berücksichtigt wurden. Zudem wurden bei 2.791 Jobs immer dieselben Daten erhoben, weshalb keine Copulas gebildet werden konnten.

Wie in Abbildung 9 zu erkennen ist, wurden 2.967.229 Jobs betrachtet, von denen 34.845 Jobs als Anomalie klassifiziert wurden, was 1,15 % der Daten entspricht. Wie auch bei der multivariaten t-Verteilung wurde bei der Methode 1 % als heuristischer Richtwert gewählt, der durch die Betrachtung der Daten in Gruppierungen zu leichten Abweichungen geführt hat.

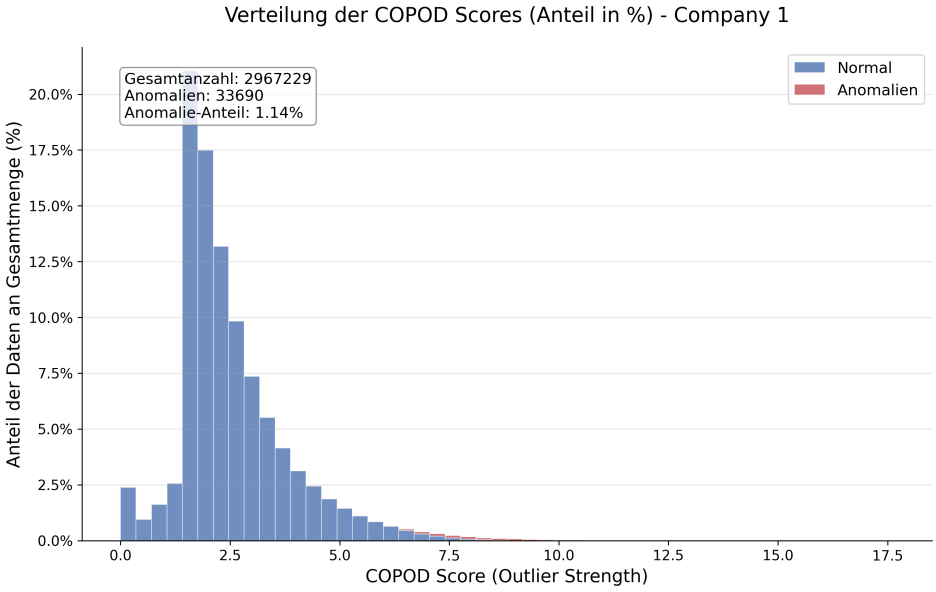


Abbildung 9: Anomalie-Score von COPOD in Prozent

Die Verteilung der Scores ähnelt der des Isolation Forests. Auch hier gibt es eine stark rechtsschiefe Verteilung mit Long Tails. Je höher der Score ist, desto wahrscheinlicher ist, dass es sich um eine Anomalie handelt. Auch wenn es in 9 nicht gut zu erkennen ist, gibt es Scores mit einer Höhe von 17,6, die jedoch so selten sind, dass sie in dieser Grafik nicht ersichtlich sind.

Es folgt die Betrachtung zweier Fallbeispiele:

Fallbeispiel 1: Anomalie aufgrund hohen Deduplikationsverhältnis und Durchsatz
Der Job 23176354 des Clients 2028 ist einer von 28.150 transformierten Jobs, von denen 284 von COPOD als Anomalie eingestuft wurden. Der COPOD-Score beträgt 18,9 und der Job ist ein Log-Only-Backup. Tabelle 6 veranschaulicht, inwiefern dieser spezifische Ausreißer vom regulären Verhalten der historischen Log-Only-Backups dieses Clients abweicht. Während die regulären Log-Backups im Median und im Durchschnitt lediglich Datenmengen im niedrigen Kilobyte-Bereich sichern, speichert dieser Job mit 11,5 MB ein Vielfaches der sonst üblichen Datenmenge.

Attribut	Analysierter Job	Median	Durchschnitt
SOA	11,5 MB	41,7 KB	116,3 KB
SOMOD	0,2 MB	3,7 KB	3,5 KB
JET	2m 30s	1m 4s	2m 55s
Deduplikationsverhältnis	64,4	29,8	28,4
Durchsatz	100 KB/s	1,6 KB/s	1,9 KB/s

Tabelle 6: Vergleich der Attribute des analysierten Jobs 23176354 mit dem Median und dem Durchschnitt der Log-Only-Backups des Clients 2028.

Das Deduplikationsverhältnis des betrachteten Jobs liegt weitaus über dem Median bzw. dem Mittelwert (mehr als doppelt so hoch). Dasselbe gilt für den Durchsatz, der ungefähr um den Faktor 100 höher ist als der Median oder der Mittelwert. Die Kombination dieser beiden Attribute ist demnach sehr ungewöhnlich und führt zu einem hohen Anomalie-Score.

Fallbeispiel 2: Anomalie aufgrund großen Datentransfers mit sehr hoher Deduplikation Ein weiteres Beispiel ist Job 42445386 des Clients 14435. Er ist einer von 850 transformierten Jobs, von denen 10 Jobs von COPOD als Anomalie gekennzeichnet wurden. Der COPOD-Anomaliescore beträgt 13,4, und es handelt sich um ein inkrementelles Backup.

Tabelle 7 stellt die betrachteten Jobs gegenüber dem Median und dem Durchschnitt des Clients 14435 dar. Wie zu sehen ist, laufen viele Jobs, die eine recht große Datenmenge speichern, jedoch für gewöhnlich nicht so groß wie der betrachtete Job. Außerdem ist zu erkennen, dass bei dem betrachteten Job wenig neue Daten gespeichert werden, da SOMOD im Vergleich zu SOA klein ist und zudem ein sehr hohes Deduplikationsverhältnis aufweist.

Attribut	Analysierter Job	Median	Durchschnitt
SOA	27,1 GB	3,8 GB	20,9 GB
SOMOD	18,4 MB	306,8 MB	997,4 MB
JET	2m 35s	5m 49s	11m 18s
Deduplikationsverhältnis	1475	17,4	94
Durchsatz	174,7 MB/s	19,8 MB/s	31,5 MB/s

Tabelle 7: Vergleich der Attribute des analysierten Jobs 42445386 mit dem Median und dem Durchschnitt der inkrementellen Backups des Clients 14435.

Bei den betrachteten Merkmalen wird ersichtlich, dass der untersuchte Job ein deutlich höheres Deduplikationsverhältnis und einen deutlich höheren Durchsatz aufweist als der Durchschnitt. Es werden entsprechend viele Daten gespeichert, die bereits im Speichersystem hinterlegt sind.

4.4 Extremwerttheorie

Anders als bei den zuvor betrachteten Methoden werden bei der Extremwertanalyse die Daten univariat analysiert. Zudem wird an den zeitunabhängigen Residuen von SOA, SOMOD und JET sowie an den Verhältnissen gearbeitet.

Wie in Kapitel 3 erläutert, betrachtet die Extremwertanalyse nur Werte, die einen definierten Schwellenwert überschreiten (POT-Ansatz). Diese Schwellenwerte werden dynamisch für jede Gruppe, also eine Kombination aus Client, Jobtyp, Backuplevelnamen sowie dem betrachteten Attribut, berechnet. Dafür wird das 95 %-Quantil genutzt. Die obersten 5 % der Daten werden somit betrachtet und mithilfe der GPD modelliert. Zur Modellierung dieser benötigt man den Shape-Parameter, der die Gewichtung der Tails definiert, sowie einen Skalierungsparameter, der die Skalierung der Verteilung beschreibt. Beide Parameter werden mithilfe des MLE geschätzt.

Um die GPD modellieren zu können, werden ausreichend viele Daten benötigt. Aufgrund dessen müssen in einer Gruppierung mindestens 50 Datenpunkte vorhanden sein, damit die oberen 5 % zur Modellierung der GPD verwendet werden können. Dies ist jedoch nicht immer der Fall, sodass 40.833 Gruppierungen übersprungen werden. Die übrigen Daten werden mithilfe des MLE geschätzt, bei 7548 Gruppierungen ist diese Schätzung jedoch fehlgeschlagen, weshalb sie nicht betrachtet werden.

Wie in Kapitel 3 definiert, wurden mithilfe der GPD die p-Werte berechnet. Bei einem Signifikanzniveau von 0,001 ergaben sich dabei 12.754 identifizierte Anomalien:

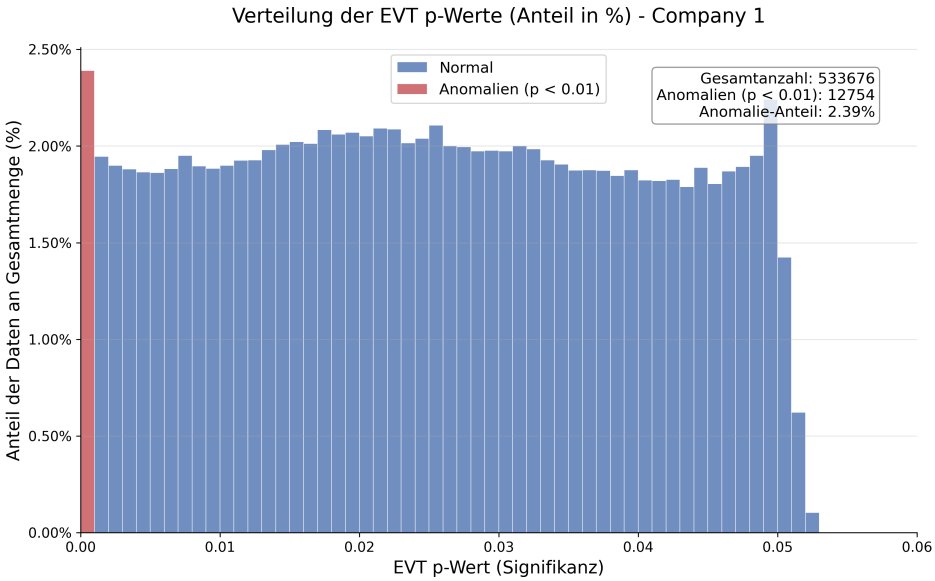


Abbildung 10: Verteilung der p-Werte der Extremwertanalyse in Prozent

Wie in Abbildung 10 zu erkennen ist, ist die Verteilung der p-Werte zwischen 0 und 0,5 annähernd gleichverteilt, weist jedoch leichte Peaks bei 0 und 0,05 auf. Anstelle der ansonsten 3 Millionen betrachteten Jobs, die bei der Extremwertanalyse durch die univariate Betrachtung der 3 Residuen (zumindest für zeitunabhängige Jobs) und 2 Verhältnisse auf ungefähr 10 Millionen betrachteten Datenpunkten ansteigen, übersteigen insgesamt nur 533.676 Werte den Schwellenwert, und davon werden 12.754 Werte als Anomalie klassifiziert. Dies entspricht 2,39 % der über dem Schwellenwert liegenden Daten.

Zusätzlich ist es interessant, ob die Extremwertanalyse auch mehrere Attribute in einem Job als auffällig markiert:

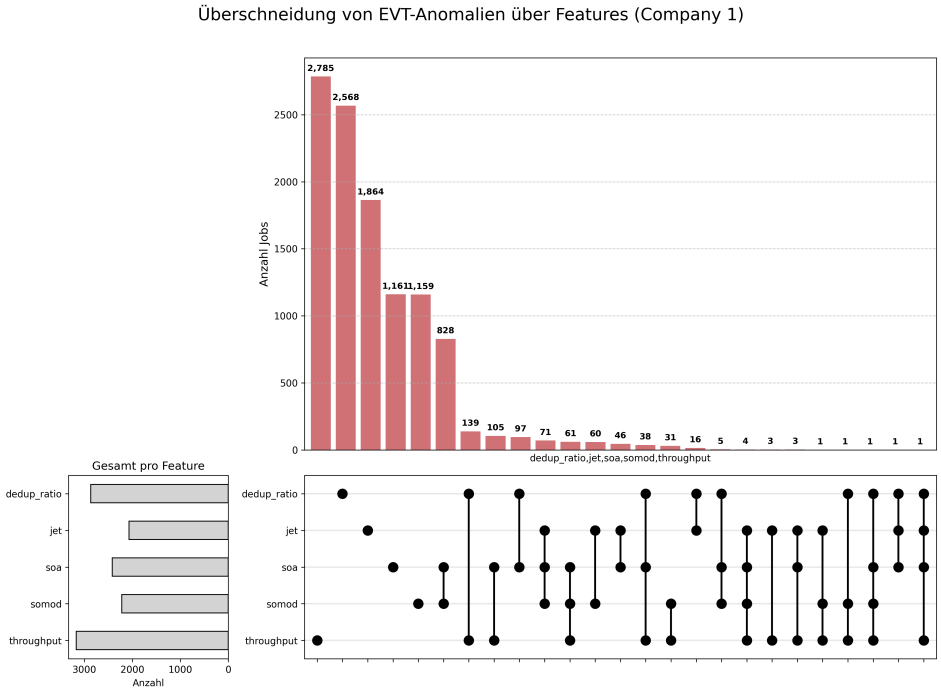


Abbildung 11: Upset-Diagramm der Ergebnisse der Extremwertanalyse

Wie in Abbildung 11 zu erkennen ist, ist die Anzahl der Jobs, die in mehr als zwei Attributen auffällig sind, selten. Die meisten Kombinationen sind in SOA und SOMOD zu finden. Generell werden die meisten Anomalien am Durchsatz und am Deduplikationsverhältnis erkannt. Dies war zu erwarten, denn hier wurden alle Jobs betrachtet und keine, wie die zeitabhängigen Residuen, ausgeschlossen. Anzumerken ist jedoch, dass kein Job bei allen betrachteten Merkmalen als Anomalie identifiziert wird. Insgesamt werden auch nur 6 Jobs bei vier Attributen gleichzeitig erkannt.

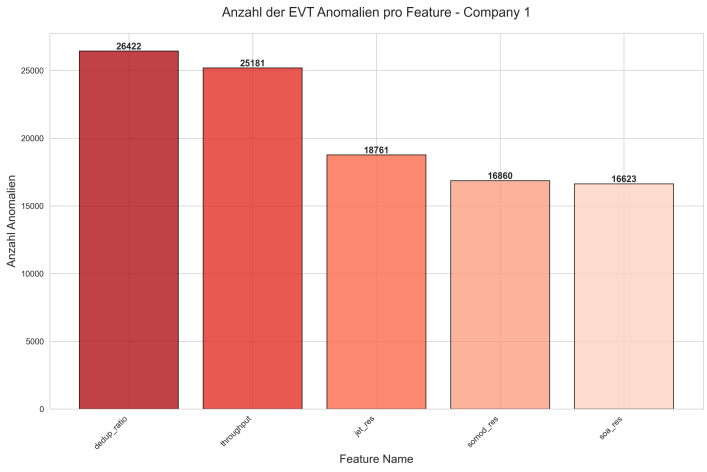


Abbildung 12: Anzahl erkannter Anomalien nach Attribut

Dies hat, wie man in Abbildung 12 erkennen kann, den Grund, dass bei den Residuen aufgrund der geringeren Datengrundlage auch grundsätzlich weniger Anomalien identifiziert wurden.

Zwei dieser Anomalien werden besonders betrachtet:

Fallbeispiel 1: Anomalie durch hohe Werte Betrachtet wird Job 37860534 des Clients 6727. Dieser Client besitzt 320 Jobs, bei denen 2 Anomalien gekennzeichnet wurden. Der Job 37860534 wurde sowohl von den SOA- als auch von den SOMOD-Residuen als Anomalie identifiziert. Der betrachtete Job ist ein inkrementelles Backup. Der p-Wert beträgt beim SOA-Residuum $4,5 \cdot 10^{-15}$, beim SOMOD-Residuum 0,0005, während der p-Wert bei den anderen Attributen mindestens 0,002 beträgt.

Attribut	Analysierter Job	Median	Durchschnitt
SOA	8.405,4 GB	0,03 MB	168,3 GB
SOMOD	6.864 GB	0,7 GB	48,9 GB
JET	47h 3m 2s	25m 49s	1h 59m 38s
Deduplikationsverhältnis	1,2	0,004	22,8
Durchsatz	49,6 MB/s	19 B/s	49 MB/s

Tabelle 8: Vergleich der Attribute des analysierten Jobs 37860534 mit dem Median und dem Durchschnitt der inkrementellen Backups des Clients 6727.

Tabelle 8 verdeutlicht, dass der betrachtete Job im Vergleich zu den anderen betrachteten Werten deutlich höhere Werte in allen Attributen aufweist. Insbesondere ist dies bei SOA und SOMOD auffällig, was dazu geführt hat, dass der Job in diesen beiden Werten als Anomalie identifiziert wurde. Zudem ist auffällig, dass der Mittelwert und der Median bei diesem Client relativ weit voneinander entfernt sind, weshalb darauf zu schließen ist, dass bei diesem Client häufig kleine Werte gespeichert werden, Ausreißer jedoch (wenn auch nicht in diesem Ausmaß) keine Seltenheit sind.

Fallbeispiel 2: Anomalie aufgrund von erhöhten Werten Ein weiteres Beispiel ist der Job 23538796. Er ist ein inkrementelles Backup des Clients 11205. Interessant ist dieser Job vor allem, weil vier der fünf möglichen Attribute in der Extremwertanalyse als Anomalien identifiziert wurden. Darunter zählen die Residuen von SOA und SOMOD, das Deduplikationsverhältnis sowie der Durchsatz. Die jeweiligen p-Werte dazu sind:

Attribut	p-Wert
SOA-Residuum	0,00011
SOMOD-Residuum	0,00014
Deduplikationsverhältnis	0,00024
Durchsatz	0,0001

Tabelle 9: p-Wert der EVT

Zum Vergleich: Das JET-Residuum weist nur einen p-Wert von 0,011 auf und liegt damit über dem Signifikanzniveau.

Attribut	Analysierter Job	Median	Durchschnitt
SOA	15,9 GB	0,4 GB	0,4 GB
SOMOD	300 MB	98,6 MB	98,6 MB
JET	11m 18s	2m 52s	4m 30s
Deduplikationsverhältnis	46,3	3,8	4
Durchsatz	23,7 MB/s	2,1 MB/s	2,2 MB/s

Tabelle 10: Vergleich der Attribute des analysierten Jobs 23538796 mit dem Median und dem Durchschnitt der inkrementellen Backups des Clients 11205.

Wie in Tabelle 10 ersichtlich, sind die Daten in allen Attributen, mit Ausnahme von JET, stark erhöht. Zudem ist ersichtlich, dass dieser Client grundsätzlich nur geringen Schwankungen in den Attributen unterliegt, was an dem geringen Unterschied zwischen Median und Mittelwert erkennbar ist.

4.5 Erkannte Anomalien

Die vier verwendeten Methoden prüfen die vorliegenden Daten jeweils auf unterschiedliche Weise und klassifizieren Datenpunkte entsprechend aufgrund verschiedener Eigenschaften als Anomalien. Entsprechend wird nicht jede Anomalie von allen verwendeten Methoden als solche klassifiziert.

Wie in Abbildung 13 zu erkennen ist, wurden unterschiedlich viele Anomalien pro Methode erkannt. Dabei wurden bei der multivariaten t-Verteilung ähnlich viele Anomalien wie bei COPOD erkannt, was auf den heuristischen Schwellenwert von 1 % zurückzuführen ist. Bei der Extremwertanalyse gilt ein Schwellenwert von 0,001, doch da nur wenige Jobs die Bedingungen von POT erfüllen, werden auch nur wenige dieser Jobs als Anomalie markiert.

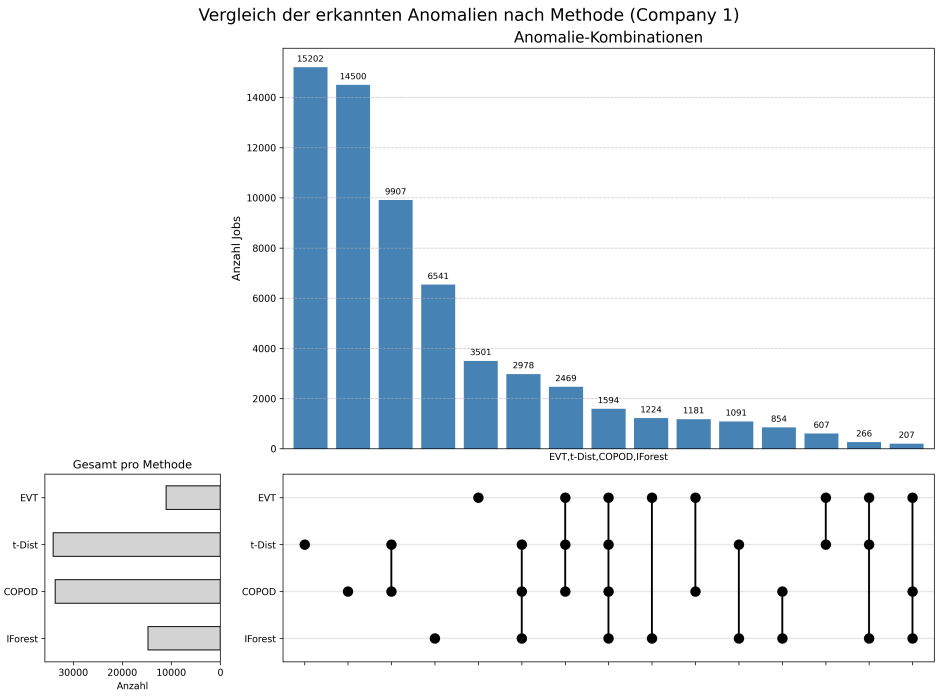


Abbildung 13: Anzahl erkannter Anomalien pro Methode

Auch wenn eine ähnliche Anzahl an Anomalien erkannt wurde, bedeutet dies nicht automatisch, dass dieselben Datenpunkte markiert wurden. Es werden jedoch mehrere Anomalien als solche von mehr als einer Methode klassifiziert. Erkennbar ist, dass die multivariate t-Verteilung und COPOD die größten Überschneidungen aufweisen. Dies ist unter anderem darauf zurückzuführen, dass sie dieselben transformierten Attribute betrachten.

Für das in Kapitel 3 vorgeschlagene gestaffelte Warnsystem sind jedoch insbesondere die Schnittmengen aller vier Modelle von Relevanz. Abbildung 13 verdeutlicht, dass 7514 Jobs von mindestens drei oder sogar von allen vier Methoden simultan als Anomalie klassifiziert wurden. Diese Schnittmengen minimieren das Risiko falschpositiver Alarme und identifizieren die Anomalien, die im folgenden Frühwarnsystem einen Alarm auslösen müssen. Die verbleibenden einfach oder doppelt erkannten Anomalien hingegen werden primär als Warnung genutzt.

4.6 Validierung anhand von Test-Clients

Wie in Kapitel 3 beschrieben, wurden vier Test-Clients genutzt, die jeweils unterschiedliche Arten von künstlichen Anomalien erzeugen:

- Client 20975 (siehe Abb. 14): Referenzclient
- Client 20989 (siehe Abb. 15): Erstellen einer sehr großen Datenmenge, gefolgt von einer verzögerten Löschung dieser Daten.
- Client 20995 (siehe Abb. 16): Erstellen einer großen Datenmenge, gefolgt von der plötzlichen Verschlüsselung dieser Daten.

- Client 20932 (siehe Abb. 17): Erstellen einer großen Datenmenge mit anschließender schleichender Verschlüsselung (kontinuierlich ansteigender Verschlüsselungsgrad über mehrere Backups hinweg).

Die Extremwerttheorie kann in dieses Evaluationsszenario nicht einbezogen werden. Der POT-Ansatz setzt voraus, dass mindestens 5 % der Daten den dynamischen Schwellenwert überschreiten, wofür mindestens 50 historische Datenpunkte pro Gruppierung verlangt werden. Da die Test-Clients zum Zeitpunkt der Auswertung wenige große Datensätze erzeugen,überschreiten nicht genug Jobs den Schwellenwert.

Referenzclient Der Referenzclient 20975 besitzt 74 reguläre Jobs ohne induzierte Anomalien. Die multivariate t-Verteilung sowie COPOD klassifizieren korrekterweise keinen der Jobs als Anomalie. Der Isolation Forest hingegen deklariert den ersten Job nach der initialen Baseline-Erstellung fälschlicherweise als Anomalie. Bei genauerer Betrachtung lässt sich dies auf eine minimal erhöhte Laufzeit in Kombination mit einem verhältnismäßig großen zeitlichen Abstand zum vorherigen Job zurückführen.

Methode	Precision	Recall	Rank Power
Isolation Forest	0	-	0
Multivariate t-Verteilung	-	-	-
COPOD	-	-	-

Tabelle 11: Evaluationsmetriken des Test-Clients 20975. Ein Strich (-) signalisiert mathematisch nicht definierte Werte, da keine echten Anomalien existieren ($d_t = 0$).

Löschen der Daten Bei Client 20989 wurden zwei künstliche Anomalien in einer Reihe von 74 Jobs erzeugt: Zunächst die Erstellung eines großen Datensatzes und zeitversetzt dessen Löschung. Der Isolation Forest identifiziert insgesamt drei Anomalien, wobei zwei davon die korrekten künstlichen Anomalien sind. Die multivariate t-Verteilung sowie COPOD erkennen jeweils nur die erste Anomalie (die Erstellung der Daten). Die Nichterkennung der zweiten Anomalie (Löschung) ist hierbei jedoch dem Schwellenwert geschuldet: Bei 74 Jobs erlaubt das 1 %-Quantil lediglich die Klassifikation eines einzigen Jobs als Anomalie. Die internen Scores der Modelle weisen jedoch eindeutig auf die zweite Anomalie hin: So fällt die Gewichtung der t-Verteilung für die Löschung auf $3 \cdot 10^{-6}$ (Normalwert ≈ 3) und der COPOD-Score steigt auf 7,2 (Normalwert $\approx 0,03$).

Methode	Precision	Recall	Rank Power
Isolation Forest	0,67	1	0,6
Multivariate t-Verteilung	1	0,5	1
COPOD	1	0,5	1

Tabelle 12: Evaluationsmetriken des Test-Clients 20989 ($d_t = 2$)

Verschlüsseln der Daten Für den Client 20995 wurden 66 Jobs analysiert. Darunter befinden sich zwei künstliche Anomalien: Das Hinzufügen von 40 GB an neuen Daten zu

einer etablierten Baseline und eine darauffolgende abrupte Verschlüsselung. Der Isolation Forest identifiziert exakt diese beiden Jobs fehlerfrei als Anomalien. Erneut deckt der feste Schwellenwert der multivariaten t-Verteilung und von COPOD lediglich die erste Anomalie ab. Doch auch hier verzeichnen die internen Metriken für den Verschlüsselungs-Job signifikante Ausschläge: Die Gewichtung der t-Verteilung stürzt auf $2 \cdot 10^{-17}$ ab und der COPOD-Score erreicht einen Wert von 6.

Methode	Precision	Recall	Rank Power
Isolation Forest	1	1	1
Multivariate t-Verteilung	1	0,5	1
COPOD	1	0,5	1

Tabelle 13: Evaluationsmetriken des Test-Clients 20995 ($d_t = 2$)

Schleichende Verschlüsselung Dieses Szenario auf Client 20932 (74 Jobs) stellt den anspruchsvollsten Test dar: Nach der Datenerschaffung wurde der Bestand über mehrere Sicherungen hinweg nur inkrementell verschlüsselt (insgesamt $d_t = 12$ betroffene Jobs). Wie erwartet, sinkt hier der Recall. Der Isolation Forest und die multivariate t-Verteilung erkennen das initiale Hinzufügen der Daten. Den Beginn der Verschlüsselung erkennen COPOD und der Isolation Forest. Letzterer markiert zudem fälschlicherweise den allerersten Job der Sicherungsreihe (aufgrund eines unverhältnismäßig hohen Wertes bei TSLJ). Die anschließenden, kleineren Verschlüsselungsschritte werden von den Algorithmen nicht als Anomalie klassifiziert. Eine Analyse der Gewichtungen der multivariaten t-Verteilung und der COPOD-Scores zeigt jedoch auch hier, dass sich jeder einzelne Verschlüsselungsschritt von der Baseline abhebt.

Methode	Precision	Recall	Rank Power
Isolation Forest	0,67	0,17	1,0
Multivariate t-Verteilung	1,0	0,08	1,0
COPOD	1,0	0,08	1,0

Tabelle 14: Evaluationsmetriken des Test-Clients 20932 ($d_t = 12$)

Der Wert bei der Rank Power entsteht, weil die Clients mit dem höchsten Score entsprechend als Anomalie identifiziert wurden. Der Rank Power ist entsprechend nur in Verbindung mit Precision und Recall aussagekräftig.

5 Diskussion

5.1 Einordnung der Datenvorverarbeitung

Um die verwendeten Methoden auf die vorliegenden Daten anwenden zu können, mussten diese zuvor verarbeitet werden. Die Daten eines produktiven Datensicherungssystems sind sehr heterogen, und zwar auf mehreren Ebenen. Zum einen unterscheiden sich die Clients generell meist stark voneinander. Hier werden nur die Daten der RWTH Aachen betrachtet, allerdings nutzen die verschiedenen Institute ihre Clients für unterschiedliche Zwecke. Selbst innerhalb eines Instituts kann es zu unterschiedlichen Anwendungsszenarien kommen. Auch die Beschaffenheit der Clients an sich unterscheidet sich. Es gibt beispielsweise Server, virtuelle Maschinen sowie Hypervisoren.

Zudem unterscheiden sich die Jobs, die ein Client ausführt. Die Hauptunterscheidung liegt zwischen Backup und Restore, aber auch hier gibt es Unterkategorien. Ein Backup kann beispielsweise ein Full-Backup oder ein inkrementelles Backup sein. Außerdem ändern sich die Jobs auch innerhalb einer Kategorie je nachdem, wie die Nutzung ausfällt.

Aufgrund dessen ist die einfache Verwendung statistischer Modelle oder Algorithmen des maschinellen Lernens unmöglich oder führt vielmehr zu fehlerhaften und wenig aussagekräftigen Ergebnissen. Die Datenvorverarbeitung musste zudem an die verwendeten Methoden angepasst werden.

Dazu wurden die Daten zuvor mithilfe von Feature-Engineering sowie Trendbereinigung an die benötigten i.i.d.-Bedingungen angenähert. Mithilfe des Feature-Engineerings werden nun zusätzlich Verhältnisse berücksichtigt. Die Kombination der Log-1-p-Transformation und des EWMA dient der Trendbereinigung, sodass im besten Fall zeitlich unabhängige Residuen übrig bleiben. Wie in Kapitel 4 jedoch deutlich geworden ist, sind nicht alle transformierten Datenpunkte nach der Trendbereinigung zeitunabhängig. Dies hat zwei Ursachen. Zum einen wurde zur Prüfung der Zeitunabhängigkeit die AKF verwendet. Diese geht jedoch grundlegend davon aus, dass die Daten einen äquidistanten Abstand zueinander haben, was hier jedoch nicht zwingend der Fall ist. Zudem können die Daten nicht nur trend-, sondern auch saisonalitätenbehaftet sein. Dies ist beispielsweise der Fall, wenn Backup-Jobs jeweils von Montag bis Freitag laufen, am Wochenende jedoch pausieren. Diese Saisonalitäten lassen sich mithilfe der verwendeten Trendbereinigung nicht abdecken. Da verschiedene Methoden zur Bereinigung der Daten von der Zeitabhängigkeit getestet wurden und die vielversprechendste verwendet wurde, wurden die noch zeitabhängigen Residuen aus der Extremwertanalyse ausgeschlossen. Dies dient dazu, die strengeren i.i.d.-Bedingungen der Methode zu gewährleisten, auch wenn dadurch ein Teil der Datenbasis verloren geht.

Eine auch nach der Trendbereinigung zeitabhängige Gruppierung ist beispielsweise das Residuum von SOA des Clients 2574. Es handelt sich um inkrementelle Backups, und der AKF-Score beträgt 0,98, was weitaus höher als der vorgegebene Schwellenwert von 0,29 ist. Zu dieser Gruppierung gehören 971 Jobs, die zumeist sehr regelmäßig täglich zwischen 16:00 Uhr und 17:00 Uhr laufen. Ab einem bestimmten Zeitpunkt erfolgt jedoch um 15:00 Uhr ein zusätzliches inkrementelles Backup, das nahezu jedes Mal 17,8 MB speichert. Anschließend speichert das typische inkrementelle Backup jedoch ungefähr 25,7 MB, läuft jedoch eine Stunde nach dem neu hinzugefügten Backup. Aufgrund der Regelmäßigkeit lässt sich darauf schließen, dass dies eine vorgenommene Konfiguration ist. Aber infolge dieses zusätzlichen Jobs kann die Trendbereinigung keinen klaren Trend mehr berechnen, da dieser zusätzliche Job eine

Saisonalität aufweist, die nicht herausgefiltert werden kann. Daraus resultiert der hohe Score der Zeitabhängigkeit, nicht zuletzt aufgrund des fehlenden äquidistanten Abstands zwischen den Jobs; dies ist keinesfalls verwunderlich.

5.2 Verhalten der Algorithmen

Wie in Kapitel 3 beschrieben, untersuchen die verwendeten Methoden komplementäre Eigenschaften der Daten und erkennen daraufhin unterschiedliche Anomalietypen.

5.2.1 Isolation Forest

Der Isolation Forest zeichnet sich dadurch aus, dass er eine große Anzahl von Attributen simultan verarbeitet. Neben den Kernattributen SOA, SOMOD und JET fließen das Deduplikationsverhältnis, der Durchsatz sowie die zeitlichen Kontextvariablen TSLJ, die Stunde und der Wochentag der Ausführung mit ein. Dank dieser Vielzahl an Attributen ist es dem Isolation Forest möglich, insbesondere Konfigurationsfehler sowie Prozessbrüche zu identifizieren.

Ein prägnantes Beispiel für diese Funktionalität liefert das Fallbeispiel 2, in dem eine rein strukturelle Anomalie identifiziert wird. Da ein inkrementelles Backup als erster Job eines Clients ausgeführt wird und zusätzlich deutlich größere Werte in den drei ursprünglichen Attributen (sowohl im Vergleich zum Median als auch zum Durchschnitt) aufweist, kann dieser Punkt vom Algorithmus entsprechend schnell isoliert werden. Besonders auffällig ist zudem das generell geringe Deduplikationsverhältnis, obwohl es in diesem Job deutlich höher ist als der Durchschnitt und der Median. Dies deutet darauf hin, dass bei diesem Client ständig wechselnde Daten gespeichert werden. Eine andere Möglichkeit wäre, dass die Deduplikation nicht korrekt funktioniert. Das könnte dann der Fall sein, wenn beispielsweise Metadaten einer Datenbank gesichert werden.

Ein solcher Job sollte dem zuständigen Administrator gemeldet werden, damit die Konfiguration des Clients erneut geprüft werden kann. Mithilfe des vorgeschlagenen Alarmsystems würde dieser Job definitiv als Anomalie gemeldet werden, da er nicht nur vom Isolation Forest, sondern auch von der multivariaten t-Verteilung sowie COPOD als Anomalie klassifiziert wurde. Dadurch wird jedoch ein Problem bezüglich des möglichen Frühwarnsystems ersichtlich. Anomalien können zwar klassifiziert werden; anhand der Identifikation allein ist es jedoch äußerst fehleranfällig zu unterscheiden, ob es sich um einen möglichen Angriff oder um einen Konfigurationsfehler handelt. Aufgrund dessen ist die Prüfung durch einen Menschen für das vorliegende System unerlässlich.

Ein potenzieller Angriff wurde jedoch in 3 identifiziert. Zumindest wäre dies eine logische Erklärung für die vorliegenden Daten sowie die Unterschiede des betrachteten Jobs gegenüber dem Median und dem Durchschnitt. In diesem Job werden in den drei ursprünglichen Attributen deutlich höhere Werte erfasst, teilweise über 900-mal größer als die des Medians. Allerdings sinken sowohl der Durchsatz als auch das Deduplikationsverhältnis. Letzteres entspricht nur noch knapp 3 % des Medians. Eine Möglichkeit wäre, dass viele der alten Daten plötzlich gelöscht und mehrere Terabyte neuer Daten hinzugefügt wurden. Dies ist jedoch relativ unwahrscheinlich, da erst wenige Tage zuvor ein Synthetic Full Backup lief und die Erstellung mehrerer Terabyte an Forschungsdaten meist mehr Zeit in Anspruch nimmt. Eine weitere Möglichkeit wäre, dass die Daten verschlüsselt wurden, weshalb das Deduplikationsverhältnis eingebrochen ist und wesentlich mehr Daten bei SOA und SOMOD gespeichert wurden als zuvor. Dass bei diesen Datenmengen zusätzlich JET gestiegen ist, ist die direkte Konsequenz.

Auch dieser Job würde im vorgeschlagenen gestaffelten Warnsystem direkt als Alarm an den zuständigen Administrator weitergeleitet werden, denn nicht nur der Isolation Forest, sondern auch die Extremwertanalyse haben den vorliegenden Job als Anomalie klassifiziert. Im Fall der Extremwertanalyse sowohl bei den Residuen von SOA als auch bei SOMOD, was verhältnismäßig selten ist, wie in Abbildung 11 zu sehen ist, besitzt sie daher eine hohe Aussagekraft.

5.2.2 Multivariate t-Verteilung

Im Gegensatz zum baumbasierten Isolation Forest basiert die multivariate t-Verteilung auf der Annahme einer festen mathematischen Verteilungsstruktur. Durch die Einbeziehung der Freiheitsgrade (ν) ist dieses Modell jedoch deutlich robuster gegenüber Extremwerten als eine klassische Normalverteilung. In dieser Arbeit wird die t-Verteilung auf das Deduplikationsverhältnis und den Durchsatz angewendet. Dadurch ist das Verfahren in der Lage, Brüche in der multidimensionalen Korrelationsstruktur beider Kennzahlen zu identifizieren.

Wie in Abbildung 7 ersichtlich ist, erhalten viele Jobs eine relativ geringe Gewichtung, was auf die starken Schwankungen der produktiven Backup-Daten zurückzuführen ist. Da der Algorithmus die Freiheitsgrade adaptiv anpasst, fängt er diese extremen Long-Tails als reguläres, wenn auch seltenes Verhalten ein. Dass durch den gesetzten Schwellenwert letztlich nur etwa 1 % der Datenpunkte als Anomalie klassifiziert werden, unterstreicht die mathematische Robustheit des Modells: Es reagiert nicht auf gewöhnliches Rauschen oder bekannte langanhaltende Ränder, sondern isoliert ausschließlich die auffälligsten Strukturbrüche. Die Extreme der Gewichte lassen jedoch darauf schließen, dass viele der Datensätze meist gleichbleibende Datenmengen speichern, was die Varianz gegen Null verschiebt, wodurch die Gewichte verschwindend klein werden.

Im Normalfall gilt jedoch Folgendes: Je geringer das Deduplikationsverhältnis ist, desto mehr Daten müssen im System gespeichert werden, was JET erhöht und den Durchsatz senkt. Die multivariate t-Verteilung schlägt genau dann an, wenn diese typische Korrelation verletzt wird.

Ein extremes Beispiel für einen solchen Bruch liefert das Fallbeispiel 4. Hier speichert ein inkrementeller Job über 29 Terabyte bei SOA, obwohl der Median des Clients bei Null liegt. Die eigentliche Klassifikation als Anomalie erfolgt jedoch anhand der transformierten Merkmale. Das Deduplikationsverhältnis bricht auf 1,0 ein, während der Durchsatz auf 43 MB/s ansteigt. Diese Kombination ist sehr selten, da der Einbruch des Deduplikationsverhältnisses den Durchsatz aufgrund der langen Laufzeit des Jobs drosseln müsste. Die t-Verteilung führt für diesen Punkt zu einer Gewichtung von praktisch Null (10^{-33}), wodurch er vollständig aus der normalen Verteilungsdichte ausgeschlossen wird. Im vorgeschlagenen Warnsystem löst dieser Job einen Alarm aus, da er zusätzlich vom Isolation Forest und COPOD markiert wird. Zudem validiert die Extremwertanalyse diesen Vorfall, indem sie bei allen drei Residuen (SOA, SOMOD, JET) gleichzeitig anschlägt — ein Ereignis, das laut dem Upset-Diagramm (Abb. 11) statistisch eine absolute Seltenheit darstellt.

Ein strukturell ähnliches Verhalten zeigt die t-Verteilung im Fallbeispiel 5. Die Datenmengen weichen jedoch, außer bei SOA, nicht sonderlich von den Werten des Medians oder des Durchschnitts ab. Das Deduplikationsverhältnis hingegen steigt von einem Durchschnitt von 1,3 auf 754. Auch der Durchsatz steigt von 0,3 MB/s auf 208 MB/s. Dies zeigt, dass die Deduplikation sehr schnell möglich ist. Obwohl bei SOA eine große Datenmenge erfasst wird, werden im eigentlichen Speichersystem nahezu keine neuen Daten gesichert. Dieses Beispiel

zeigt genau die andere Verteilungsmöglichkeit. Während der Strukturbruch bei 4 durch einen Abfall des Deduplikationsverhältnisses und einen Anstieg des Durchsatzes offensichtlich wird, ist bei 5 das Zusammenspiel eines so hohen Deduplikationsverhältnisses mit einem hohen Durchsatz sehr unwahrscheinlich. Da sowohl die multivariate t-Verteilung als auch der Isolation Forest und COPOD den Job als Anomalie klassifizieren, wird eine Alarmierung des Administrators vorgenommen, ohne dass mögliche Long Tails der Verteilung zu falsch-positiven Klassifizierungen führen.

5.2.3 COPOD

Wie auch die multivariate t-Verteilung wird COPOD in dieser Arbeit auf das Deduplikationsverhältnis und den Durchsatz angewendet. Anstatt jedoch eine feste Verteilungsform vorzuschreiben, schätzt COPOD die gemeinsamen Randverteilungen mithilfe der ECDF. Dadurch arbeitet der Algorithmus vollkommen parameterfrei und benötigt lediglich die Angabe der erwarteten Kontaminationsrate (in diesem Fall 1 %) als Schwellenwert für die Klassifikation.

COPOD berechnet die Extremität für jedes Attribut separat, um dem Curse of Dimensionality zu entgehen, weshalb Unabhängigkeit zwischen den Dimensionen angenommen wird. Aufgrund dessen geht die direkte Korrelation zwischen den Attributen verloren. Dennoch ist das Modell in der Lage, Strukturbrüche und untypische Attributskombinationen zu identifizieren. Dies liegt an der Berechnung des Anomalie-Scores, bei der die negativen Logarithmen der univariaten Tail-Wahrscheinlichkeiten über alle Dimensionen (in diesem Fall 2) hinweg aufsummiert werden. Wenn ein Job folglich in beiden betrachteten Attributen gleichzeitig extreme Werte aufweist, addieren sich diese Abweichungen zu einem hohen Gesamtscore. Ein hoher COPOD-Score signalisiert damit eine Attributskombination, die im Normalbetrieb nahezu unmöglich ist.

Wie in Abbildung 9 zu sehen ist, sind die meisten Jobs bei einem Score von etwa 2,4 konzentriert, der verhältnismäßig niedrig ist. Ein Score über 6 wird je nach betrachteter Gruppierung bereits als Anomalie klassifiziert. Einige wenige Jobs haben zudem Scores von über 10. Diese extremen Werte spiegeln wider, dass der betroffene Job an den äußersten Rändern der gelernten Verteilung liegt.

Ein Beispiel einer solchen Kombination wird in Fallbeispiel 6 beschrieben. Hier liegen sowohl der Durchsatz als auch das Deduplikationsverhältnis weit über dem Normalzustand. Während das Deduplikationsverhältnis etwa das Doppelte des Medians und des Durchschnitts beträgt, übersteigt der Durchsatz diesen jedoch um das 100-Fache. Ähnliches gilt für das Fallbeispiel 7, bei dem die Werte noch erheblich weiter vom Normalzustand entfernt sind. Der Durchsatz beträgt hier das 8-Fache des Medians und das 5-Fache des Durchschnitts, während das Deduplikationsverhältnis sogar das 86-Fache des Medians und das 15-Fache des Durchschnitts erreicht. Da in diesem zweiten Fallbeispiel beide Attribute extrem auffällig sind, addieren sich die negativen Logarithmen zu einem hohen Gesamtscore von 13,4.

Die Zuverlässigkeit dieser Ergebnisse wird durch die starke Übereinstimmung mit den anderen Methoden untermauert. Das Fallbeispiel 6 wurde nicht nur von COPOD, sondern auch von Isolation Forest, der multivariaten t-Verteilung sowie der Extremwertanalyse (in den Residuen von SOA, SOMOD sowie im Durchsatz) als Anomalie identifiziert. Fallbeispiel 7 dagegen wurde von COPOD, dem Isolation Forest, der multivariaten t-Verteilung und der Extremwertanalyse hinsichtlich Durchsatz und Deduplikationsverhältnis als Anomalie identifiziert. Diese kollektive Klassifizierung einer Anomalie belegt, dass der parameterfreie Ansatz von COPOD im geplanten Staffelsystem verlässliche Schnittmengen liefert.

5.2.4 Extremwerttheorie

Bei der Extremwertanalyse werden, anders als bei der multivariaten t-Verteilung und COPOD, nicht nur das Deduplikationsverhältnis und der Durchsatz, sondern auch die zeitunabhängigen Residuen von SOA, SOMOD und JET betrachtet. Aufgrund der strengeren Mindestanzahl benötigter Jobs pro Gruppierung und einzelner Konvergenzprobleme bei der MLE-Schätzung der Parameter konnten hier jedoch einige Datenpunkte nicht berücksichtigt werden. Zudem ist die initiale Filterung mithilfe des 95 %-Quantils beim POT-Ansatz relativ konservativ, sodass ausschließlich die absoluten Spitzen der Verteilung modelliert werden. Da für die finale Klassifikation ein strenges Signifikanzniveau von $\alpha = 0,001$ gewählt wurde, isoliert die Extremwerttheorie nur die extremsten Ausreißer. Dies führt im univariaten Raum häufig dazu, dass ein Job nur in einem einzigen Attribut als Anomalie auffällt, da die Ränder der Merkmale unabhängig voneinander berechnet werden und andere Jobs in derselben Gruppierung in den verbleibenden Attributen noch extremere Abweichungen aufweisen. Dieses Verhalten wird im Upset-Diagramm (Abb. 11) veranschaulicht.

Aufgrund dieser mathematischen Struktur werden bei der Extremwertanalyse generell nur Werte betrachtet, die zu den obersten 5 % der Daten gehören, wodurch die GPD-Modellierung die unwahrscheinlichsten Spitzen als Anomalien filtert. Entsprechend weisen diese Jobs extreme Werte im jeweils identifizierten Merkmal auf. Dies wird im Fallbeispiel 8 ersichtlich. Dort übersteigt der Wert in SOA den Durchschnitt um das 50-Fache, der seinerseits bereits deutlich über dem Median liegt. Bei SOMOD lässt sich ein ähnliches Phänomen beobachten. Interessanterweise ist dieser Job auch bei der Laufzeit (JET) deutlich erhöht, jedoch nicht so stark, dass der berechnete p-Wert das Signifikanzniveau der Extremwerttheorie unterschreitet.

Dieser Effekt zeigt sich auch im Fallbeispiel 10, obwohl der betrachtete Client hier deutlich weniger Ausreißer aufweist. Infolgedessen wird ein Job, der im reinen Wertevergleich zu Fallbeispiel 8 weniger abnormal erscheint, vom Modell dennoch in vier von fünf Attributen gleichzeitig als Anomalie identifiziert. Daran wird deutlich, dass das vorherige Gruppieren eine entscheidende Rolle spielt: Es stellt sicher, dass Jobs nur in einer gleichbleibenden Systemumgebung verglichen werden. Dadurch können auch solche Abweichungen präzise als Anomalie erkannt werden, die im globalen Kontext aller Clients unauffällig gewirkt hätten, für diesen spezifischen Client jedoch einen massiven Strukturbruch darstellen.

5.2.5 Gemeinsame Nutzung

Jede der vier ausgewählten Methoden besitzt ihre eigenen Schwerpunkte, weshalb jeweils unterschiedliche Anomalien identifiziert werden. Während der Isolation Forest aufgrund der hohen Anzahl an einbezogenen Attributen als robuste Baseline fungiert, die komplexe strukturelle Prozessbrüche erkennt, identifiziert die multivariate t-Verteilung gezielt Verletzungen in der Korrelationsstruktur der Kennzahlen. COPOD wiederum ermittelt unabhängig voneinander extreme Abweichungen an den Rändern der empirischen Verteilungsfunktionen, was die Erkennung univariat unwahrscheinlicher Merkmalskombinationen ermöglicht. Ergänzend dazu filtert die Extremwerttheorie über den POT-Ansatz ausschließlich die absoluten Spitzenwerte der zeitunabhängigen Residuen, des Deduplikationsverhältnisses sowie des Durchsatzes heraus. Erst durch diese Kombination unterschiedlicher Ansätze entsteht eine verlässliche Grundlage für ein zukünftiges Frühwarnsystem.

5.3 Entwicklung eines Frühwarnsystems

Obwohl die Methoden gut zusammenspielen, zeigt Abbildung 13, dass jede Methode auch Anomalien identifiziert, die von keiner anderen Methode als solche eingestuft werden. Dies kann beispielsweise auf einen falsch gewählten Schwellenwert zurückzuführen sein, der zu falsch-positiven Meldungen führt. Es kann jedoch auch in der Komplementarität der Methoden begründet liegen. Bestimmte Anomalien lassen sich aufgrund der Eigenschaften der verschiedenen Methoden bei manchen Algorithmen nicht erfassen. So kann die Extremwerttheorie beispielsweise keine Anomalien, die COPOD aufgrund einer ungewöhnlichen Merkmalskombination, wie etwa eines starken Abfalls der Deduplikationsrate bei gleichbleibendem Durchsatz, identifizieren.

Ein praxistaugliches Frühwarnsystem im Hochschulumfeld muss sich jedoch auf dem schmalen Grat zwischen sensibler Anomalieerkennung und operativer Machbarkeit bewegen. Werden zu viele unschädliche Abweichungen, beispielsweise erkannte Konfigurationsprobleme, gemeldet, führt dies bei den zuständigen Administratoren unweigerlich zu einer Alert Fatigue. Dies birgt das Risiko, dass echte Warnungen im Ernstfall ignoriert oder abgeschaltet werden.

Um diesem Problem entgegenzuwirken, wird in dieser Arbeit ein gestaffeltes Warnsystem vorgeschlagen, welches die Anzahl der gleichzeitig anschlagenden Methoden als Filter nutzt:

- Stufe 1 (Einfache Warnung): Klassifizieren lediglich ein oder zwei Methoden einen Job als Anomalie, liegt mit hoher Wahrscheinlichkeit eine kurzzeitige Systemlast (beispielsweise durch die Speicherung großer neuer Datenmengen) oder eine Konfigurationsänderung vor. Das Frühwarnsystem generiert eine einfache Warnung, beispielsweise in Form eines täglichen Berichts oder einer E-Mail an den Administrator, um den Vorfall für eine spätere manuelle Überprüfung zu protokollieren.
- Stufe 2 (Kritischer Alarm): Schlagen hingegen drei oder alle vier Methoden gleichzeitig bei demselben Sicherungsvorgang an, liegt sehr wahrscheinlich ein Strukturbruch vor. Die statistische Wahrscheinlichkeit eines falsch-positiven Fehlers wird in dieser Schnittmenge deutlich reduziert. In diesem Fall muss das System eine sofortige automatisierte Alarmierung an die zuständigen Administratoren auslösen.

Ein konkretes Szenario für einen solchen Stufe-2-Alarm lässt sich direkt aus den mathematischen Zusammenhängen ableiten: Wenn die multivariate t-Verteilung und COPOD übereinstimmend anschlagen, weil das Deduplikationsverhältnis und parallel der Durchsatz einbrechen, und zeitgleich die Extremwerttheorie bei den Residuen von SOMOD sowie JET eine Überschreitung des Schwellenwerts meldet, liefert das System ein Indiz für einen aktiven Ransomware-Angriff. Durch die Verschlüsselung der Client-Dateien lassen sich die Daten nicht deduplizieren, was dazu führt, dass viele Daten erneut auf dem Speichersystem abgelegt werden, wodurch sich die Laufzeit des Jobs stark erhöht und der Durchsatz im Vergleich einbricht. Eine solche mehrdimensionale Evidenz ermöglicht es dem Frühwarnsystem, potenziell kompromittierte Systeme frühzeitig im Datenbestand zu identifizieren, bevor sich ein möglicher Schaden weiter ausbreiten kann.

5.4 Validierung mithilfe der Test-Clients

Wie in Kapitel 4 dargelegt, variieren die Ergebnisse der Methoden auf die simulierten Anomalien unterschiedlich. Die Extremwerttheorie konnte aufgrund der zu geringen Überschreitungen des Schwellenwerts nicht in diese Evaluierung einbezogen werden.

Es zeigt sich deutlich, dass die multivariate t-Verteilung und COPOD eine hohe Übereinstimmung aufweisen und oftmals dieselben multivariaten Strukturbrüche identifizieren. Dass bei den Test-Clients mit multiplen Anomalien (wie dem Löschen oder Verschlüsseln) jeweils nur das erste anomale Ereignis klassifiziert wird, ist auf den statischen Schwellenwert (Kontaminationsrate von 1 %) zurückzuführen. Zwar heben sich die Gewichtungen, bzw. Scores der weiteren Anomalien weichen signifikant vom Normalzustand ab; die starre Filterung limitiert die Erkennung bei $N < 100$ jedoch auf einen einzigen Datenpunkt. Ein pauschales Erhöhen des Schwellenwerts würde dieses Problem nicht lösen, sondern bei Clients ohne tatsächliche Anomalien unweigerlich zu Falsch-Positiven führen. Dies bestätigt die Notwendigkeit dynamischer Schwellenwerte für den produktiven Einsatz.

Der Isolation Forest erweist sich bei markanten Brüchen als das sensitivste Modell, scheitert jedoch weitestgehend an der Identifikation der schleichenden Verschlüsselung. Zudem ist er die einzige Methode, die Falsch-Positive erzeugt. Dies lässt sich auf die Einbeziehung der zeitlichen Kontextvariablen, insbesondere TSLJ, zurückführen. Während diese Attribute essenziell sind, um Abweichungen im Backup-Plan zu modellieren, führen sie bei reinen Terminverschiebungen ohne zugrunde liegenden Datenangriff zu hohen Anomalie-Scores. In einem produktiven Umfeld könnten solche Ausschläge Konfigurationsfehler aufdecken; für die reine Identifikation von Ransomware stellen sie im Rahmen dieser Validierung jedoch Falsch-Positive dar.

5.5 Qualitative Experten-Evaluation

Um die Ergebnisse einzuordnen, wurden Anomalie-Fallbeispiele einem Commvault-Administrator der RWTH Aachen zur qualitativen Evaluation vorgelegt. Dabei zeigt sich, dass die Validierung von Anomalien erheblich erschwert wird, da das Command-Center Jobs nur eine kurze Zeit anzeigt. Bei inkrementellen Backups besteht beispielsweise eine Vorhaltezeit von 30 Tagen, Synthetic Fulls werden jedoch drei Monate vorgehalten.

Zudem offenbarte das Gespräch Einschränkungen der betrachteten Attribute. Die Speicherunterschiede können beispielsweise durch die Verwendung von Subclients erklärt werden, die verschiedene Speicherorte sichern, aber meist demselben Speicherplan unterliegen. Dies erklärt auch, weshalb TSLJ teilweise negative Zahlen annehmen kann, obwohl im Grunde keine zwei Jobs desselben Clients gleichzeitig ausgeführt werden können.

Ein weiterer Punkt ist die Nutzung fester Zeitpläne der Erstellung der Jobs. Die meisten Clients nutzen einen von sechs definierten Sicherungsplänen. Entstehen Sicherungen außerhalb dieser Zeit, gibt es zumeist nur drei Möglichkeiten:

- Der Stream war belegt und der Client musste mit der Sicherung warten, bis dieser frei war
- Der Client oder der Server hatten Netzwerkprobleme
- Ein User hat ein Backup angestoßen

Zu beachten ist jedoch, dass keine dieser Möglichkeiten auf einen potenziellen Angriff hinweisen kann, da nur berechtigte Personen Datensicherungen erstellen können. Durch die Mitgabe der zeitabhängigen Attribute im Isolation Forest kann dieser solche Unregelmäßigkeiten oftmals als Anomalie deuten, was im Grunde der Fall ist, im Kontext eines Frühwarnsystems jedoch Mehraufwand für die zuständigen Administratoren bedeutet.

Da die Datensicherung mit Commvault ein produktives System ist, sind zudem die dadurch erzeugten Metadaten nicht immer zuverlässig. JET und damit einhergehend der Durchsatz sind zum Beispiel von vielen Faktoren stark abhängig. Es kommt beispielsweise darauf an, ob der Stream, der die Daten sichert, frei ist. Ob eine stabile Netzwerkverbindung besteht oder wie die Proximität zwischen dem zu sichernden Client und dem CommServe ist: ob VPN benötigt wird oder eine direkte Verbindung möglich ist.

Die Gespräche haben gezeigt, dass die Commvault-Administratoren Zugriff auf dieselben Daten wie die Anomalieerkennung haben, mit der Einschränkung, dass die Jobs eine kürzere Vorhaltezeit aufweisen. Das vorliegende System bietet also die Möglichkeit, die Menge der zu betrachtenden Jobs einzuschränken; die endgültige Klassifizierung einer Anomalie liegt jedoch weiterhin bei den für den Client zuständigen Administratoren.

5.6 Limitierungen der Arbeit

Obwohl die Kombination der gewählten Methoden vielversprechende Ergebnisse für ein potenzielles Frühwarnsystem liefert, unterliegt die vorliegende Arbeit methodischen und praktischen Limitierungen, die bei einer Überführung in die praktische Anwendung wichtig zu beachten sind.

5.6.1 Granulare Filterung

Eine Einschränkung, insbesondere für den Isolation Forest, ist die Nichtbeachtung von Subclients. Ein Client wird derzeit immer als eine Einheit gesehen, auf der Jobs laufen können. Ein Client kann jedoch unterschiedliche Subclients besitzen, die unterschiedliche Speicherpfade sichern, weshalb große Sprünge in den Metadaten nach der derzeitigen Gruppierung teilweise nicht auf Anomalien, sondern rein auf eine andere Konfiguration hinweisen.

5.6.2 Fehlende Label

Die wichtigste Einschränkung ist das Fehlen gelabelter Daten. Da es sich bei den vorliegenden Daten um Metadaten eines produktiven Datensicherungssystems handelt, bei denen keine Anomalien explizit markiert wurden, ist diese Arbeit im unüberwachten Lernen angesiedelt. Aufgrund dessen ist keine klassische Methodenevaluation möglich, bei der die detektierten Anomalien mit den gelabelten abgeglichen werden. Die erkannten Anomalien können zwar als signifikante Abweichung vom Normalverhalten bewiesen werden, allerdings ist keine konkrete Klassifizierung, ob es sich um einen Cyberangriff oder eine Fehlkonfiguration handelt, möglich. Diese Klassifikation ist nur mithilfe der zuständigen Administratoren möglich, die die Anomalie bewerten.

5.6.3 Echtzeit-Verarbeitung

Die vorliegende Arbeit nutzt ausschließlich historische Daten zur Analyse. Das System lässt sich mit relativ geringem Aufwand modifizieren, sodass täglich neue Daten hinzugefügt und analysiert werden können; ein Echtzeit-Frühwarnsystem wäre im jetzigen Stand jedoch nicht möglich. Für dieses müsste das System kontinuierlich an die Commvault-API gebunden sein. Dies erfordert eine Umstellung der Methoden auf Streaming-Algorithmen. Diese gibt es bereits für den Isolation Forest und die Extremwerttheorie [53, 21], allerdings würden sie die Komplexität der Berechnungen deutlich erhöhen, da zudem die Transformationen angepasst werden müssten.

5.6.4 Handhabung der Zeitabhängigkeit

Um die i.i.d.-Voraussetzungen der Methoden zu erfüllen und gleichzeitig ressourceneffizient zu arbeiten, wurden die Zeitabhängigkeiten der Jobs über Feature-Engineering und Residuenbildung gelöst. Dadurch gehen jedoch mögliche Abhängigkeiten verloren, die sich durch die Modellierung als Zeitreihe erkennen ließen.

5.6.5 Heuristische Schwellenwerte

Ein weiteres Problem ist die Sensibilität der Methoden. Die Klassifizierung als Anomalie basiert auf heuristischen Annahmen, wie dem Schwellenwert von 0,7 beim Isolation Forest, der angenommenen Anomalierate von 1 % bei der multivariaten t-Verteilung und COPOD sowie der Quantilswahl bei der Extremwertanalyse. Diese Werte wurden bewusst konservativ gewählt. Für ein funktionsfähiges Frühwarnsystem müssten diese Schwellenwerte über einen längeren Zeitraum hinweg im Live-Betrieb kalibriert werden. Ein zu niedrig angesetzter Schwellenwert führt zu einer hohen Anzahl an falsch-positiven Alarmen, was, wie bereits in 5.3 erläutert, zu Alert Fatigue führen kann.

5.6.6 Mindestanforderungen an die Datenmenge

Zudem haben die eingesetzten Methoden Limitationen hinsichtlich der benötigten Anzahl an Datenpunkten. Um valide Werte zu berechnen, wurden für den Isolation Forest, die multivariate t-Verteilung und COPOD mindestens 15 Datenpunkte pro Gruppierung vorausgesetzt. Da dies insbesondere bei neuen oder selten genutzten Clients nicht immer der Fall war, wurden Zehntausende Jobs von der Analyse ausgeschlossen. Bei der Extremwertanalyse werden sogar 50 Datenpunkte für die Analyse benötigt, weshalb hier weitaus mehr Daten nicht betrachtet werden. Zudem besteht das Problem, dass genügend Werte oberhalb des Schwellenwertes liegen müssen, da die Modellierung anhand der GPD ansonsten fehlschlägt. Für diese Datenpunkte bietet das aktuelle System keine Anomalieerkennung, sodass mögliche Anomalien nicht erkannt werden.

5.6.7 Merkmalskorrelation und theoretische Annahmen

Wie in Kapitel 3 dargelegt, setzen die meisten der verwendeten unüberwachten Methoden strikte i.i.d.-Bedingungen voraus. Obwohl durch das Feature-Engineering und die Trendbereinigung eine bestmögliche Annäherung an diese theoretischen Bedingungen erreicht wurde, lassen sie sich in einem produktiven Datensatz nicht vollständig erfüllen.

Insbesondere sind die durch das Feature-Engineering erzeugten Verhältnismerkmale (Deduplikationsverhältnis und Durchsatz) nicht stochastisch unabhängig, da beide auf dem gemeinsamen Attribut SOA basieren. Im univariaten Kontext der Extremwerttheorie ist diese multivariate Abhängigkeit jedoch unkritisch, da die Attribute separat betrachtet werden. Die multivariate t-Verteilung berücksichtigt diese Abhängigkeit wiederum explizit über die Skalierungsmatrix (Kovarianzmatrix) bei der Gewichtungsschätzung. Bei COPOD stellt die Annahme der Merkmalsunabhängigkeit hingegen einen bewussten Trade-off dar, um eine hohe Recheneffizienz zu gewährleisten.

Die für die Extremwerttheorie kritische Bedingung der zeitlichen Unabhängigkeit konnte für die analysierten Daten durch die EWMA-Trendbereinigung und das anschließende Filtern der verbleibenden autokorrelierten Gruppierungen erfolgreich erfüllt werden.

Es ist jedoch zu beachten, dass JET und damit einhergehend der Durchsatz sowie TSLJ generell sehr vielen Faktoren unterliegen, weshalb die damit identifizierten Anomalien weniger aussagekräftig als Anomalien, die auf anderen Attributen basieren, sind.

5.6.8 Unterscheidung der Anomalietypen

Zuletzt besteht das Problem, dass mithilfe der verwendeten Methoden verschiedene Arten von Anomalien identifiziert werden, jedoch nicht mithilfe der Methoden in potenzielle Angriffe oder reine Konfigurationsprobleme separiert werden können. Somit besteht erneut die Gefahr der Alert-Fatigue, denn die zuständigen Administratoren werden durch Konfigurationsprobleme, die von mehreren Methoden identifiziert wurden, ebenso alarmiert wie bei tatsächlichen Angriffen.

5.7 Synthese

Die Ergebnisse zeigen, dass keine der untersuchten Methoden allein sämtliche relevanten Anomalietypen zuverlässig identifizieren kann. Während der Isolation Forest insbesondere komplexe multivariate Strukturbrüche erkennt, erfassen die multivariate t-Verteilung und COPOD vor allem ungewöhnliche Merkmalskombinationen. Die Extremwerttheorie ergänzt diese Ansätze um die Identifikation statistisch außergewöhnlicher Einzelereignisse. Erst das Zusammenspiel dieser komplementären Verfahren ermöglicht eine robuste Anomalieerkennung in heterogenen Backup-Metadaten und bildet damit die Grundlage für ein praktikables gestaffeltes Frühwarnsystem.

6 Fazit und Ausblick

6.1 Fazit

Die vorliegende Arbeit demonstriert die grundsätzliche Machbarkeit eines Frühwarnsystems für das Projekt DaSi.NRW auf Basis statistischer Modelle und Algorithmen des unüberwachten Lernens. Die hier verwendeten Methoden wurden aufgrund ihrer Eigenschaften und Übereinstimmungen bezüglich der Beschaffenheit der Daten gewählt. Im Verlauf der Arbeit wurden die Forschungsfragen beantwortet:

Wie müssen die Rohdaten aufbereitet werden, um die Methoden sinnvoll anwenden zu können? Um die verwendeten Methoden sinnvoll auf die vorliegenden Daten anzuwenden, erfordert die Entwicklung eine Vorverarbeitung bestehend aus Feature-Engineering und Trendbereinigung, um die Bedingungen der Methoden, wie beispielsweise die i.i.d.-Annahme, zu erfüllen. Aus den drei ursprünglichen Attributen SOA, SOMOD und JET wurden zusätzlich die Attribute für das Deduplikationsverhältnis und den Durchsatz definiert. Die multivariate t-Verteilung sowie COPOD nutzen nur die neu erstellten Attribute. Der Isolation Forest arbeitet mit allen Attributen und zusätzlich mit zeitabhängigen Attributen wie TSLJ, der Stunde sowie dem Wochentag, an dem der Job lief. Die Extremwertanalyse arbeitet mit den neu erstellten Attributen sowie den zeitunabhängigen Residuen der ursprünglichen Attribute.

Zudem erfolgte eine logische Gruppierung nach der Art der Datensicherung, was die Vergleichbarkeit der betrachteten Jobs sicherstellt. Es wäre wenig zielführend gewesen, ein Full-Backup beispielsweise mit einem inkrementellen Backup zu vergleichen.

Welche Auswirkungen haben die spezifischen Dateneigenschaften auf die Leistungsfähigkeit und Robustheit der ausgewählten Methoden? Die Daten besitzen nicht nur zeitliche Abhängigkeiten, sondern zudem Long-Tails und weisen eine sehr schiefe Verteilung auf, welche die Anwendung klassischer Distanzverfahren ineffektiv macht. Außerdem müssen große Datenmengen berücksichtigt werden. Für die RWTH Aachen allein existieren bisher mehr als drei Millionen Jobs bei über 5000 Clients. Dabei repräsentiert die RWTH Aachen nur eine von 28 beteiligten Hochschulen im Projekt DaSi.NRW, was die Notwendigkeit skalierbarer Algorithmen unterstreicht.

Folglich wurden Methoden gewählt, die mit einem Großteil der Voraussetzungen funktionieren können. Alle gewählten Methoden sind sehr performant, sodass die Anomalieerkennung mit über drei Millionen Datenpunkten problemlos auf einem handelsüblichen Laptop durchgeführt werden konnte. Des Weiteren können die Methoden mit der stark schiefen Verteilung sowie den ausgeprägten Long Tails sehr gut umgehen. Der Isolation Forest identifiziert Strukturbrüche, während die multivariate t-Verteilung und COPOD die Daten modellieren können. Die Extremwerttheorie filtert mithilfe des POT-Ansatzes nur die extremsten Datenpunkte heraus und modelliert sie mittels der GPD. Die Zeitabhängigkeit war in diesem Fall ein Problem, das jedoch durch Feature-Engineering sowie Trendbereinigung weitestgehend gelöst werden konnte.

Wie lässt sich aus der Kombination der Methoden ein gestaffeltes Warnsystem erstellen, das Administratoren zuverlässig alarmiert, ohne eine Alert Fatigue auszulösen? Der Ansatz eines gestaffelten Warnsystems ist sinnvoll, da eine einzige Methode

zum einen nicht alle Anomalietypen abdeckt und zum anderen aufgrund der besonderen Beschaffenheit der Daten mit hoher Wahrscheinlichkeit ab und zu Anomalien als solche identifiziert, die keine sind. Durch die Staffelung ist es entsprechend möglich, die Wahrscheinlichkeit falschpositiver Anomalieerkennungen möglichst gering zu halten und somit die Administratoren vor einer Alert Fatigue zu schützen.

Mithilfe der Daten der Test-Clients ist ersichtlich, dass zumindest der Isolation Forest, die multivariate t-Verteilung und COPOD sensitiv auf Anomalien reagieren. Allerdings ist auch ersichtlich, dass die Methoden, in den Beispielen vor allem der Isolation Forest, auch stark auf Änderungen der zeitlichen Attribute reagieren, welche oftmals eher auf Konfigurationsfehler oder -änderungen hinweisen, die nicht direkt einen kritischen Alarm auslösen sollten. Daher ist die Verwendung einer gestaffelten Alarmschwelle durchaus sinnvoll. Zudem müssen die Schwellenwerte der Methoden deutlich angepasst werden, da gerade die Bedingung, dass 1% der Daten Anomalien sind, zu Problemen führen kann. Diese Anpassungen müssen jedoch im produktiven Betrieb feinjustiert werden.

Insgesamt wird deutlich, dass ein solches gestaffeltes Frühwarnsystem mit dem aktuellen Forschungsstand zwar funktionsfähig ist, jedoch noch Verbesserungspotenzial bietet. Aus den über drei Millionen zu betrachtenden Jobs wurden nur 7.514 Jobs als hochgradig auffällig markiert (drei oder mehr Methoden identifizieren eine Anomalie), was die Menge der Daten, die die Administratoren betrachten müssen, deutlich reduziert.

6.2 Ausblick

Diese Arbeit stellt lediglich den Grundstein für ein Frühwarnsystem dar; für die Entwicklung eines funktionsfähigen Prototyps bedarf es jedoch einer erweiterten Analyse.

Erweiterung der Gruppierungen und Einführung eines Rankings Wie sich im Expertengespräch herauskristallisiert hat, ist die hier verwendete Gruppierung nicht granular genug. Es muss zudem nach Subclients gruppiert werden. Außerdem wäre die Einführung eines Rankings der Anomalien sinnvoll, in dem beispielsweise die Anomalien, die bei JET oder dem Durchsatz identifiziert wurden, geringer gewichtet werden als die des Deduplikationsverhältnisses.

Anpassung der Attribute Durch die Evaluation der Experten ist zudem ersichtlich geworden, dass die Einbindung der zeitabhängigen Attribute im Isolation Forest verbesserungswürdig ist. Es wäre sinnvoll, die Uhrzeiten als fortlaufenden Zyklus zu betrachten. Dadurch wird der Tageswechsel vom Algorithmus nicht mehr als harter Bruch interpretiert, sodass ein manuelles Backup um 1 Uhr morgens im Vergleich zu den regulären Updates um 18 Uhr vom Modell realistischer eingeordnet wird.

Methodische Erweiterung Derzeit arbeiten die Methoden mit Daten, die aus der Commvault-API extrahiert und anschließend aufbereitet wurden. Somit sind die Daten immer um einen Tag im Verzug. Eine Möglichkeit, schneller auf Anomalien zu reagieren, wäre, entweder im Austausch oder zusätzlich zu den derzeit verwendeten Methoden solche zu nutzen, die auf Streams basieren. Wie in Kapitel 5 erwähnt, ist dies zumindest sowohl für den Isolation Forest als auch für die Extremwerttheorie möglich.

Eine weitere Verbesserungsmöglichkeit besteht in der Trendbereinigung, um anschließend zeitunabhängige Daten zu erhalten. Die Verwendung weiterer Dekompositionen wäre sinnvoll, um zusätzlich zur Trendbereinigung auch Saisonalitäten und zyklische Abhängigkeiten aus den Daten herauszufiltern, sodass die gewünschten i.i.d.-Bedingungen besser abgedeckt werden und mehr Daten analysiert werden können.

Automatisierung In Kapitel 5 wurde bereits erwähnt, dass die Wahl des heuristischen Schwellenwerts im Folgenden angepasst werden muss. Dies sollte, wenn möglich, im Pilotbetrieb des Frühwarnsystems erfolgen. Die testenden Administratoren müssen darauf eingestellt sein, dass zu Beginn wahrscheinlich zu viele Anomalien als solche identifiziert werden, ohne davon abgeschreckt zu werden und das Frühwarnsystem als untauglich einzuschätzen.

Dieser Pilotbetrieb hätte zusätzlich den Vorteil, dass die zuständigen Administratoren Anomalien kennzeichnen, die tatsächlich sicherheitsrelevant sind. Dadurch entsteht ein Set mit gelabelten Daten, auf denen weitere Algorithmen trainiert werden können. Dies kann dazu führen, dass das Frühwarnsystem von einem anfänglich unüberwachten in ein hybrides System überführt werden kann, was die Datenabsicherung weiter steigern wird.

Integration Dieses Frühwarnsystem ist vor allem deshalb interessant, weil für alle 28 Hochschulen, die am Projekt DaSi.NRW teilnehmen, nur 2 der Standorte einen CommServe betreiben, weshalb die Skalierung des Systems für alle Hochschulen ohne viel Mehraufwand möglich ist.

Zudem sollte eine Benutzeroberfläche für die zuständigen Administratoren bereitgestellt werden. Um möglichst wenig Mehraufwand zu verursachen und die Integration des Frühwarnsystems einfach zu halten, wäre die beste Möglichkeit, dieses Dashboard direkt im Commvault-CommandCenter anzuzeigen oder zumindest zu verlinken. Auf einem solchen Dashboard sollten unter anderem sowohl die Alarme als auch die Warnungen angezeigt werden und eine direkte Verlinkung zum anomalen Job bestehen.

Insgesamt ist ersichtlich, dass die vorliegende Arbeit die Anomalieerkennung ermöglicht; bis zu einem funktionierenden Frühwarnsystem sind jedoch noch einige Arbeitsschritte nötig.

Literatur

1. Engers A. Statistische Analyse der Metadaten des Datensicherungssystem Commvault zur Anomalieerkennung. 2025
2. Mehrotra KG, Mohan CK und Huang H. Anomaly detection principles and algorithms. Terrorism, security, and computation. Cham: Springer, 2017
3. DH.NRW. datensicherung.dh.nrw: Datensicherung.NRW. o.J. Available from: https://datensicherung.dh.nrw/de/ [Accessed on: 2026 Jun 11]
4. Commvault. REST API - API Library. o.J. Available from: https://documentation.commvault.com/v11/software/rest_api_api_library.html [Accessed on: 2026 Jun 8]
5. Smith J und Doe A. Alert Fatigue Mitigation in Anomaly Detection Systems: A Comparative Study of Threshold Optimization and Alert Aggregation Strategies. Journal of Computing Innovations and Applications 2026. Available from: https://ciajournal.com/index.php/jcia/article/download/28/29 [Accessed on: 2026 Jul 7]
6. Commvault. Quick Start Guide. o.J. Available from: https://documentation.commvault.com/v11/essential/quick_start_guide.html [Accessed on: 2026 Jun 12]
7. Manai E, Mejri M und Fattahi J. Confusion Matrix Explainability to Improve Model Performance: Application to Network Intrusion Detection. *2024 10th International Conference on Control, Decision and Information Technologies (CoDIT)*. ISSN: 2576-3555. 2024 Jul :1–5. DOI: 10.1109/CoDIT62066.2024.10708595. Available from: https://ieeexplore.ieee.org/document/10708595/ [Accessed on: 2026 Jul 2]
8. Chandola V, Banerjee A und Kumar V. Anomaly detection: A survey. ACM Computing Surveys (CSUR) 2009 Jul; 41:15:1–15:58. DOI: 10.1145/1541880.1541882. Available from: https://dl.acm.org/doi/10.1145/1541880.1541882 [Accessed on: 2026 Jun 1]
9. Blázquez-García A, Conde A, Mori U und Lozano JA. A Review on Outlier/Anomaly Detection in Time Series Data. ACM Comput. Surv. 2021 Apr; 54:56:1–56:33. DOI: 10.1145/3444690. Available from: https://dl.acm.org/doi/10.1145/3444690 [Accessed on: 2026 Mar 3]
10. Taqqu MS. Weak Stationarity. *Encyclopedia of Statistical Sciences*. John Wiley & Sons, Ltd, 2006. DOI: 10.1002/0471667196.ess2895.pub2. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/0471667196.ess2895.pub2 [Accessed on: 2026 Feb 17]
11. Dudek G. STD: A Seasonal-Trend-Dispersion Decomposition of Time Series. 2022 Apr. Available from: https://arxiv.org/abs/2204.10398v1 [Accessed on: 2026 May 13]
12. Box GEP, Jenkins GM und Reinsel GC. Time series analysis: forecasting and control. Fourth edition. Wiley series in probability and statistics. Hoboken, NJ: Wiley, 2008. DOI: 10.1002/9781118619193
13. Limpert E, Stahel WA und Abbt M. Log-normal Distributions across the Sciences: Keys and Clues: On the charms of statistics, and how mechanical models resembling gambling machines offer a link to a handy way to characterize log-normal distributions, which can provide deeper insight into variability and probability—normal or log-normal: That is the question. BioScience 2001 May; 51:341–52. DOI: 10.1641/0006-3568(2001)051[0341:LNDATS]2.0.CO;2. Available from:

https://doi.org/10.1641/0006-3568(2001)051%5B0341:LNDATS%5D2.0.CO;2 [Accessed on: 2026 Jun 25]
14. Ahmad S, Lavin A, Purdy S und Agha Z. Unsupervised real-time anomaly detection for streaming data. Neurocomputing. Online Real-Time Learning Strategies for Data Streams 2017 Nov; 262:134–47. DOI: 10.1016/j.neucom.2017.04.070. Available from: https://www.sciencedirect.com/science/article/pii/S0925231217309864 [Accessed on: 2026 Jun 15]
15. VERITAS. NetBackup Anomaly Detection Configuration - Understanding data clustering and designing test patterns to validate anomaly detection configuration. 2022. Available from: https://www.veritas.com/content/dam/www/en_us/documents/white-papers/WP_netbackup_anomaly_detection_configuration_guide_V1572.pdf [Accessed on: 2026 Jun 9]
16. VERITAS. NetBackup 10 Anomaly Detection and Malware Scanning - Leverage the advanced capabilities of a unified resiliency platform to deliver ransomware anomaly detection and malware scanning. 2022. Available from: https://www.veritas.com/content/dam/www/en_us/documents/at-a-glance/SB_netbackup_anomaly_detection_V1351.pdf [Accessed on: 2026 Jun 9]
17. Kavousi P und Rossman Z. Veeam Ransomware Detection Service. 2026. Available from: https://www.veeam.com/whitepapers/ransomware-detection-service_wp.pdf [Accessed on: 2026 Jun 9]
18. Liu FT, Ting KM und Zhou ZH. Isolation Forest. 2008 Eighth IEEE International Conference on Data Mining 2008 Dec :413–22. DOI: 10.1109/ICDM.2008.17. Available from: https://ieeexplore.ieee.org/document/4781136 [Accessed on: 2026 Apr 22]
19. Li Z, Zhao Y, Botta N, Ionescu C und Hu X. COPOD: Copula-Based Outlier Detection. *2020 IEEE International Conference on Data Mining (ICDM)*. IEEE Computer Society, 2020 Nov :1118–23. DOI: 10.1109/ICDM50108.2020.00135. Available from: https://www.computer.org/csdl/proceedings-article/icdm/2020/831600b118/1r54CapFGjS [Accessed on: 2026 May 20]
20. Bousquet N und Bernardara P, Hrsg. Extreme Value Theory with Applications to Natural Hazards: From Statistical Theory to Industrial Practice. Cham: Springer International Publishing, 2021. DOI: 10.1007/978-3-030-74942-2. Available from: https://link.springer.com/10.1007/978-3-030-74942-2 [Accessed on: 2026 Apr 13]
21. Siffer A, Fouque PA, Termier A und Largouet C. Anomaly Detection in Streams with Extreme Value Theory. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Halifax NS Canada: ACM, 2017 Aug :1067–75. DOI: 10.1145/3097983.3098144. Available from: https://dl.acm.org/doi/10.1145/3097983.3098144 [Accessed on: 2026 Jun 15]
22. Lu YC, Chen F, Chen Y und Lu CT. A Generalized Student-t Based Approach to Mixed-Type Anomaly Detection. Proceedings of the AAAI Conference on Artificial Intelligence 2013 Jun; 27:633–9. DOI: 10.1609/aaai.v27i1.8581. Available from: https://ojs.aaai.org/index.php/AAAI/article/view/8581 [Accessed on: 2026 Jun 15]
23. Almansoori M und Telek M. Anomaly Detection using combination of Autoencoder and Isolation Forest. *1st Workshop on Intelligent Infocommunication Networks, Systems and Services (WI2NS2)*. 2023 :25–30. DOI: 10.3311/WINS2023-005. Available from: http://hdl.handle.net/10890/40706 [Accessed on: 2026 Jun 15]

24. Liu J, Huang J, Huo Y, Jiang Z, Gu J, Chen Z, Feng C, Yan M und Lyu MR. Log-based Anomaly Detection based on EVT Theory with feedback. 2023 Sep. DOI: 10.48550/arXiv.2306.05032. Available from: http://arxiv.org/abs/2306.05032 [Accessed on: 2026 Jun 15]
25. Aggarwal CC und Sathe S. Theoretical Foundations and Algorithms for Outlier Ensembles. ACM SIGKDD Explorations Newsletter 2015 Sep; 17:24–47. DOI: 10.1145/2830544.2830549. Available from: https://dl.acm.org/doi/10.1145/2830544.2830549 [Accessed on: 2026 Jun 15]
26. gql: GraphQL client for Python. Available from: https://github.com/graphql-python/gql [Accessed on: 2026 Jul 2]
27. SQLAlchemy. Available from: https://www.python4data.science/en/24.3.0/data-processing/postgresql/sqlalchemy.html [Accessed on: 2026 Jul 2]
28. McKinney W. Data Structures for Statistical Computing in Python. SciPy 2010 2010. DOI: 10.25080/Majora-92bf1922-00a. Available from: https://doi.org/10.25080/Majora-92bf1922-00a
29. Walt S van der, Colbert SC und Varoquaux G. The NumPy Array: A Structure for Efficient Numerical Computation. Computing in Science & Engineering 2011 Mar; 13:22–30. DOI: 10.1109/MCSE.2011.37. Available from: https://ieeexplore.ieee.org/document/5725236 [Accessed on: 2026 Jul 2]
30. Pedregosa F, Varoquaux G, Gramfort A, Michel V und Thirion B. Scikit-learn: Machine Learning in Python. 2018 Jun. DOI: 10.48550/arXiv.1201.0490. Available from: http://arxiv.org/abs/1201.0490 [Accessed on: 2026 Jun 22]
31. Seabold S und Perktold J. Statsmodels: Econometric and Statistical Modeling with Python. SciPy 2010 2010 May. DOI: 10.25080/Majora-92bf1922-011. Available from: https://proceedings.scipy.org/articles/Majora-92bf1922-011 [Accessed on: 2026 Jul 2]
32. Zhao Y, Nasrullah Z und Li Z. PyOD: A Python Toolbox for Scalable Outlier Detection. 2019 Jun. DOI: 10.48550/arXiv.1901.01588. Available from: http://arxiv.org/abs/1901.01588 [Accessed on: 2026 Jun 22]
33. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T und Cournapeau. SciPy 1.0–Fundamental Algorithms for Scientific Computing in Python. Nature Methods 2020 Mar; 17:261–72. DOI: 10.1038/s41592-019-0686-2. Available from: http://arxiv.org/abs/1907.10121 [Accessed on: 2026 Jul 2]
34. Zhang Z, Teng J, Yang Z, Cao T, Wang C, Gu X, Tang J, Guo D und Wang M. Kaleido: Open-Sourced Multi-Subject Reference Video Generation Model. arXiv:2510.18573 [cs.CV]. 2026 Mar. DOI: 10.48550/arXiv.2510.18573. Available from: http://arxiv.org/abs/2510.18573 [Accessed on: 2026 Jul 2]
35. Hunter JD. Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering 2007 May; 9:90–5. DOI: 10.1109/MCSE.2007.55. Available from: https://ieeexplore.ieee.org/document/4160265 [Accessed on: 2026 Jul 2]
36. Waskom ML. seaborn: statistical data visualization. Journal of Open Source Software 2021 Apr; 6:3021. DOI: 10.21105/joss.03021. Available from: https://joss.theoj.org/papers/10.21105/joss.03021 [Accessed on: 2026 Jul 2]
37. Kruchten N, Seier A und Parmer C. An interactive, open-source, and browser-based graphing library for Python. 2026 Jun. DOI: 10.5281/zenodo.14503524. Available from: https://github.com/plotly/plotly.py [Accessed on: 2026 Jul 2]
38. Coles S. An Introduction to Statistical Modeling of Extreme Values. Springer Series in Statistics. London: Springer, 2001. DOI: 10.1007/978-1-4471-3675-0. Available from:

http://link.springer.com/10.1007/978-1-4471-3675-0 [Accessed on: 2026 Apr 20]
39. Lin M, Lucas HC und Shmueli G. Research Commentary: Too Big to Fail: Large Samples and the p-Value Problem. Information Systems Research 2013; 24:906–17. Available from: https://www.jstor.org/stable/24700283 [Accessed on: 2026 May 11]
40. Cohen J. Statistical power analysis for the behavioral sciences. Second edition. Hillsdale, NJ: Lawrence Erlbaum Associates, publishers, 1988. DOI: 10.4324/9780203771587
41. Eckner A. Algorithms for Unevenly Spaced Time Series: Moving Averages and Other Rolling Operators. 2019. Available from: http://www.eckner.com/papers/Algorithms%20for%20Unevenly%20Spaced%20Time%20Series.pdf [Accessed on: 2026 Jun 13]
42. Patton A. A review of copula models for economic time series. Journal of Multivariate Analysis 2012 Sep; 110:4–18. DOI: 10.1016/j.jmva.2012.02.021. Available from: https://www.sciencedirect.com/science/article/pii/S0047259X12000826 [Accessed on: 2026 May 30]
43. Patton AJ. Modelling Asymmetric Exchange Rate Dependence. International Economic Review 2006; 47:527–56. Available from: https://www.jstor.org/stable/3663514 [Accessed on: 2026 Jun 3]
44. Plosser CI und Schwert GW. Estimation of a non-invertible moving average process: The case of overdifferencing. Journal of Econometrics 1977 Sep; 6:199–224. DOI: 10.1016/0304-4076(77)90015-X. Available from: https://www.sciencedirect.com/science/article/pii/030440767790015X [Accessed on: 2026 May 30]
45. Hunter JS. The Exponentially Weighted Moving Average. Journal of Quality Technology 1986 Oct; 18:203–10. DOI: 10.1080/00224065.1986.11979014. Available from: https://doi.org/10.1080/00224065.1986.11979014 [Accessed on: 2026 Jun 25]
46. Lange KL, Little RJA und Taylor JMG. Robust Statistical Modeling Using the t Distribution. Journal of the American Statistical Association 1989; 84:881–96. DOI: 10.2307/2290063. Available from: https://www.jstor.org/stable/2290063 [Accessed on: 2026 Apr 30]
47. Nelsen RB. An introduction to copulas. 2nd ed. Springer series in statistics. New York: Springer, 2006. 269 S.
48. Huber PJ und Ronchetti E. Robust statistics. Second edition. Hoboken, New Jersey: Wiley, 2009. DOI: 10.1002/9780470434697
49. Liu C und Rubin DB. ML Estimation of the t Distribution using EM and its extensions, ECM and ECME. Statistica Sinica 1995; 5:19–39. Available from: https://www.jstor.org/stable/24305551 [Accessed on: 2026 Jun 16]
50. Loureiro CP, Oliveira MR, Brito P und Oliveira L. Minimum Covariance Determinant Estimator and Outlier Detection for Interval-valued Data. 2026 Apr. DOI: 10.48550/arXiv.2604.26769. Available from: http://arxiv.org/abs/2604.26769 [Accessed on: 2026 May 1]
51. Pickands J. Statistical Inference Using Extreme Order Statistics. The Annals of Statistics 1975; 3:119–31. Available from: https://www.jstor.org/stable/2958083 [Accessed on: 2026 Jun 2]
52. Balkema AA und de Haan L. Residual Life Time at Great Age. The Annals of Probability 1974; 2:792–804. Available from: https://www.jstor.org/stable/2959306 [Accessed on: 2026 Jun 2]

53. Togbe MU, Barry M, Boly A, Chabchoub Y, Chiky R, Montiel J und Tran VT. Anomaly Detection for Data Streams Based on Isolation Forest using Scikit-multiflow. *The 20th International Conference on Computational Science and its Applications (ICCSA 2020)*. Springer, 2020. Available from: `https://hal.science/hal-02874869` [Accessed on: 2026 Jul 2]

Anhang

A Code

Der Code, auf dem die Ergebnisse der vorliegenden Arbeit basieren, ist in folgendem Repository zu finden:

`https://git.fh-aachen.de/ae2045s/anomaly-detection-in-backup-metadata`

B Test-Client Visualisierung

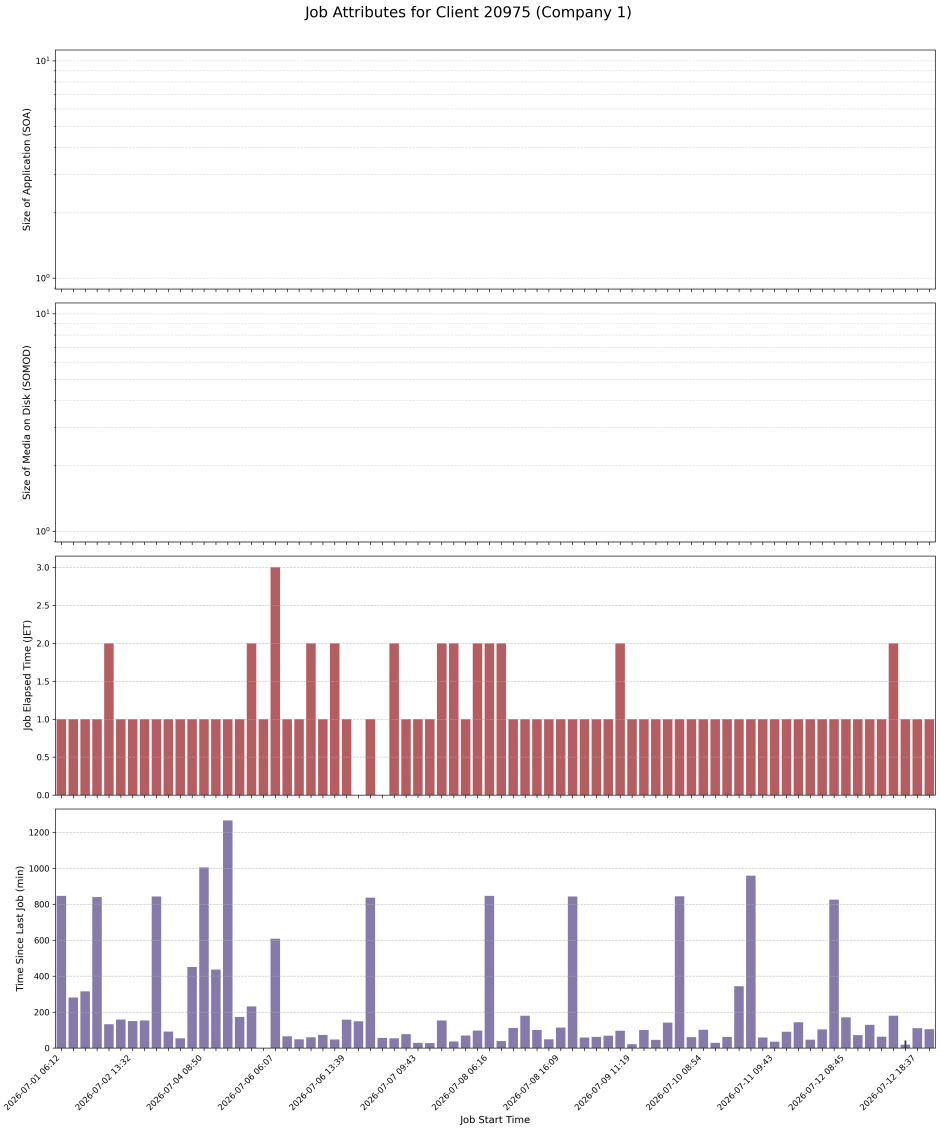


Abbildung 14: Referenzclient

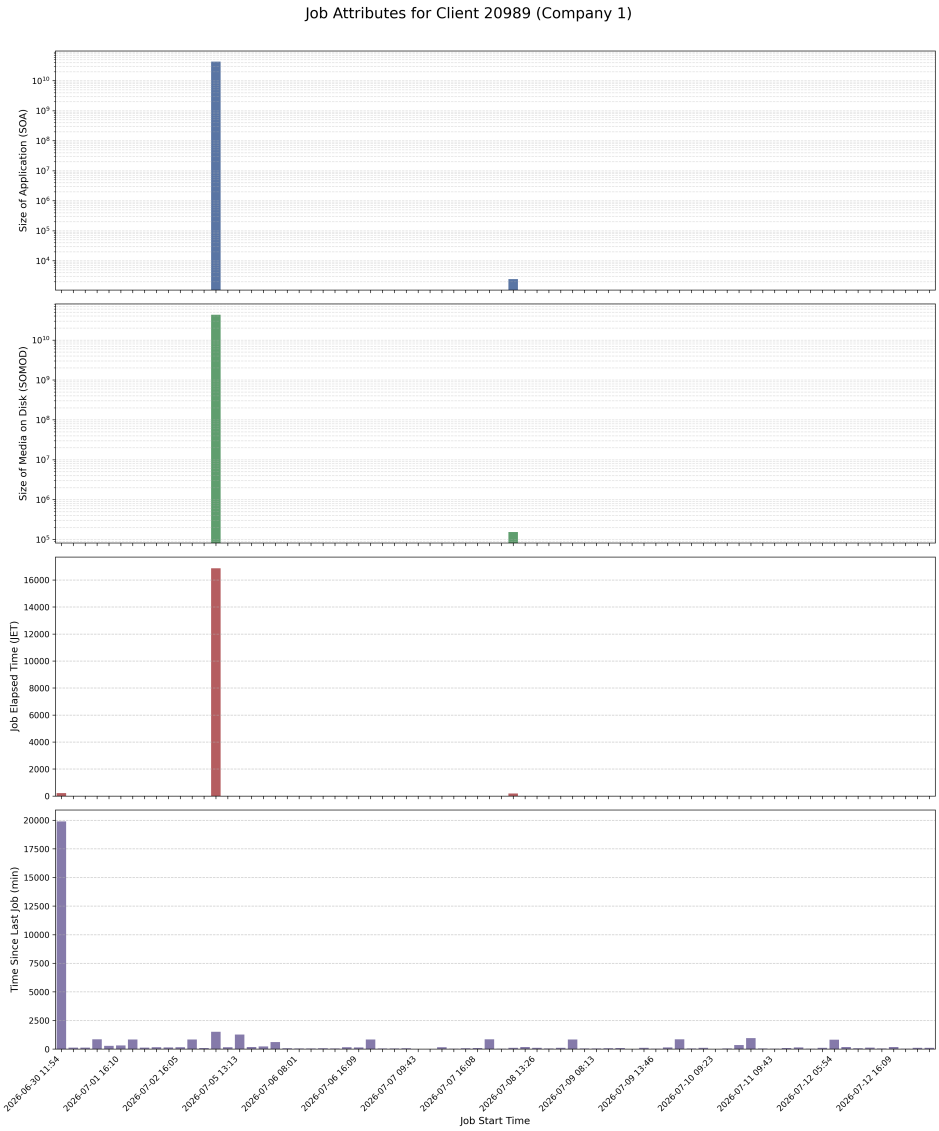


Abbildung 15: Löschung der Daten

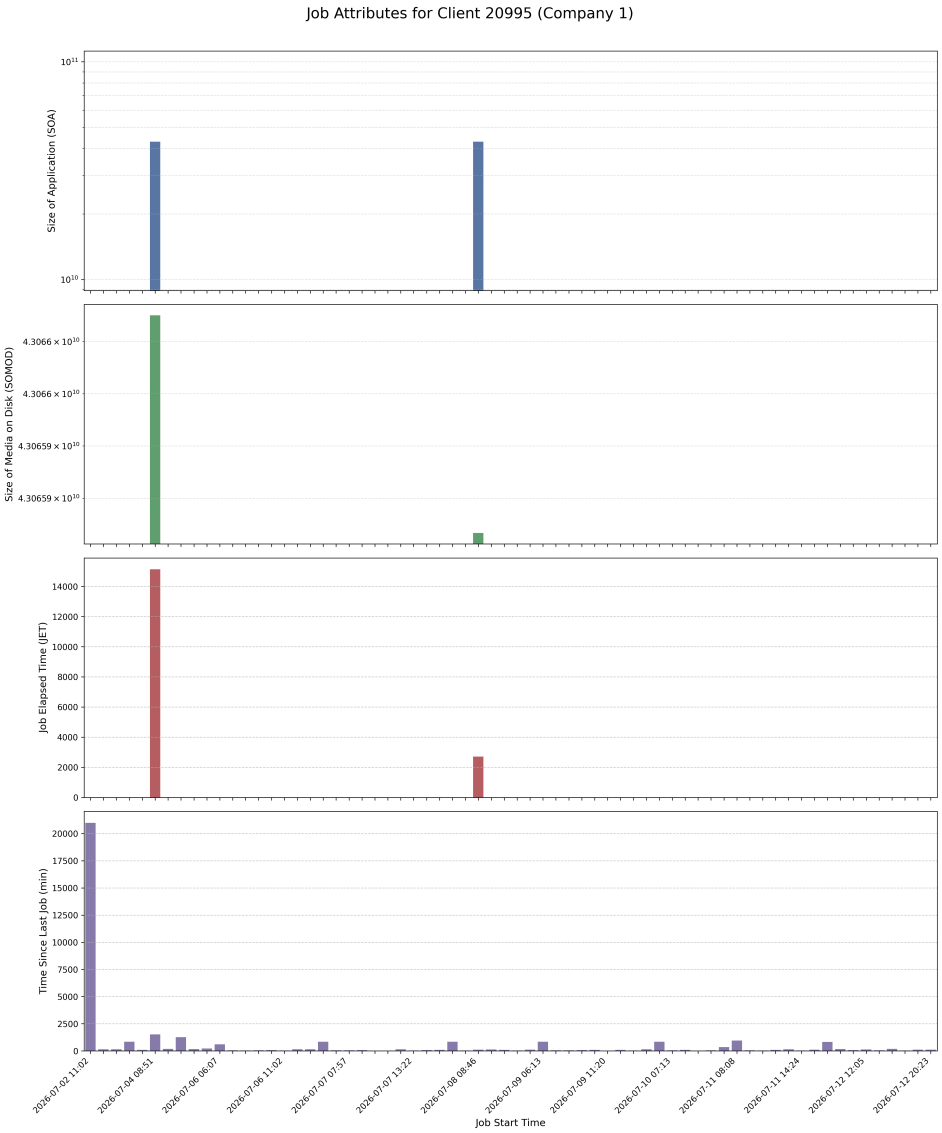


Abbildung 16: Verschlüsselung

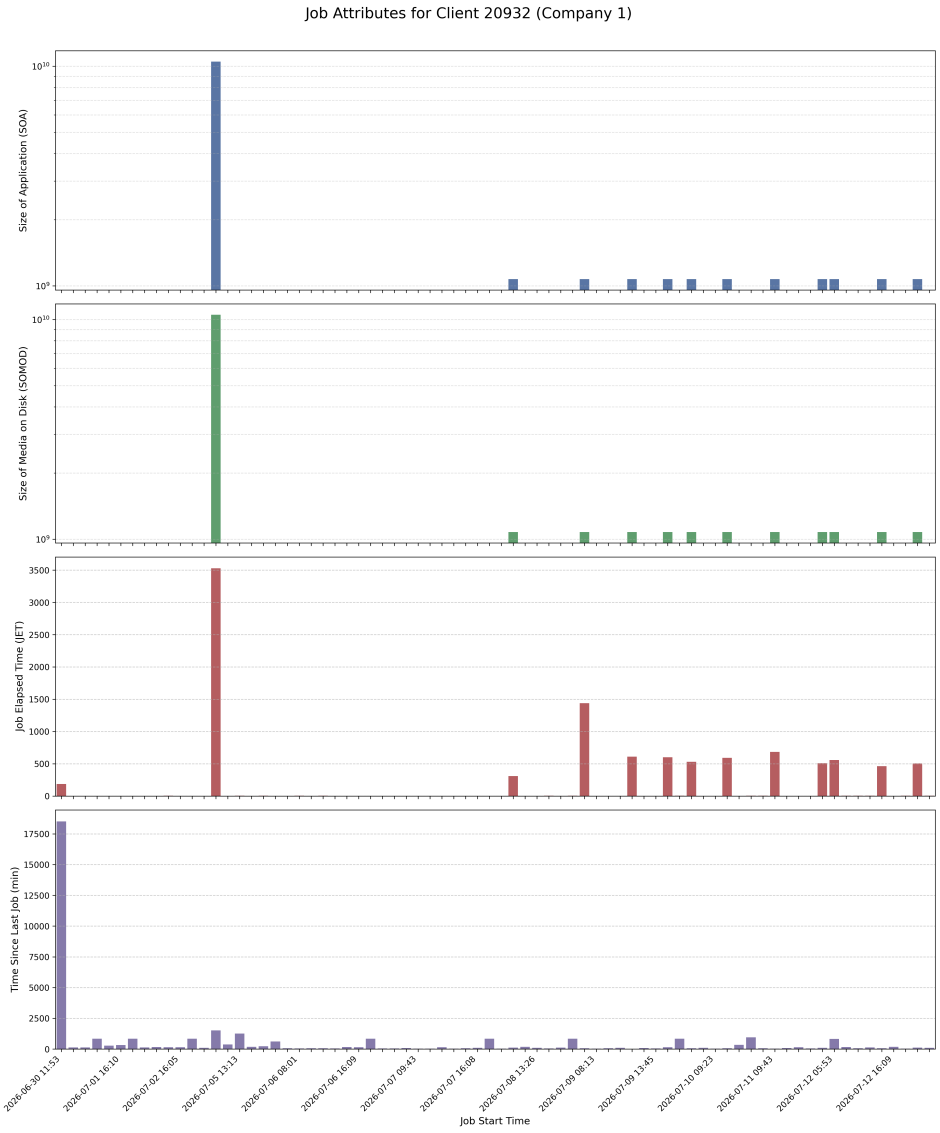


Abbildung 17: Schleichende Verschlüsselung