Limberg, Heiko

Article

Potenziale von Clustering-Algorithmen für die Plausibilisierung im Außenhandel

WISTA – Wirtschaft und Statistik

Provided in Cooperation with:
Statistisches Bundesamt (Destatis), Wiesbaden

Suggested Citation: Limberg, Heiko (2024) : Potenziale von Clustering-Algorithmen für die Plausibilisierung im Außenhandel, WISTA – Wirtschaft und Statistik, ISSN 1619-2907, Statistisches Bundesamt (Destatis), Wiesbaden, Vol. 76, Iss. 1, pp. 54-65

Heiko Limberg

ist Data Scientist und Referent im Referat „Grundsatzfragen, Qualitätssicherung, Verbreitung“ der Gruppe Außenhandel des Statistischen Bundesamtes. Er beschäftigt sich mit Modellen zur Automatisierung der Aufbereitung von Außenhandelsdaten.

POTENZIALE VON CLUSTERING-ALGORITHMEN FÜR DIE PLAUSIBILISIERUNG IM AUSSENHANDEL

Heiko Limberg

↘ **Schlüsselwörter:** Ausreißeranalyse – Clustering – Isolation Forest – DBSCAN – Local Outlier Factor – Kerneldichteschätzung

ZUSAMMENFASSUNG

Der Warenverkehr Deutschlands mit dem Ausland ist Gegenstand der vom Statistischen Bundesamt durchgeführten Außenhandelsstatistik, für die die auskunftspflichtigen Unternehmen ihre Exporte und Importe melden. Mithilfe aufwendiger Prüfschritte werden fehlerhafte Angaben weitestgehend identifiziert und bereinigt. Ein wichtiger Teil dieses Plausibilisierungsprozesses besteht darin, ungewöhnlich hohe oder niedrige Werte zu kontrollieren. Der Beitrag beschreibt die Erprobung verschiedener nichtparametrischer Clustering-Modelle, um auffällige Werte aufzudecken und nach verschiedenen Gütekriterien deren Wirkung zu bewerten. Als geeignet erweisen sich die Modelle Isolation Forest und Kerneldichteschätzung.

↘ ***Keywords:*** *outlier detection – clustering – Isolation Forest – DBSCAN – local outlier factor – kernel density estimation*

ABSTRACT

In foreign trade statistics, the Federal Statistical Office compiles data on Germany's trade in goods with other countries from the export and import declarations submitted by businesses that are required to report data. Complex edit checks are employed to detect and correct erroneous data to the greatest possible extent. An important part of this data editing process involves checking for unusually high or low values, known as outliers. This article describes the evaluation of various non-parametric clustering models to detect outliers, and assesses their effectiveness using different quality criteria. Isolation Forest and kernel density estimation are found to be suitable models.

1

Einleitung

Das Statistische Bundesamt veröffentlicht monatlich detaillierte Informationen zum Warenverkehr Deutschlands mit dem Ausland. Die Außenhandelsstatistik weist das jeweilige Partnerland des Warenverkehrs sowie die Art der gehandelten Ware nach sowie für alle möglichen Kombinationen dieser Merkmale den jeweiligen statistischen Wert und die Menge auf Monatsbasis.[1] Die Art der gehandelten Ware wird auf der detailliertesten Ebene durch eine der fast 10 000 achtstelligen Warennummern des Warenverzeichnisses für die Außenhandelsstatistik ausgedrückt. Diese Daten ermöglichen unter anderem Untersuchungen zur Entwicklung des deutschen Außenhandels und zu dessen gesamtwirtschaftlicher Bedeutung über die Zeit.

Die Außenhandelsstatistik beruht im Warenverkehr mit anderen Mitgliedstaaten der Europäischen Union (EU) (sogenannter Intrahandel) auf einer Vollerhebung mit Anmeldeschwelle. Die meldepflichtigen Unternehmen melden ihre Exporte und Importe direkt an das Statistische Bundesamt. Im Warenverkehr mit Ländern außerhalb der EU (sogenannter Extrahandel) erfolgt die Datenerhebung durch die Bundesfinanzverwaltung, die dem Statistischen Bundesamt die statistisch relevanten Daten aus Zollanmeldungen übermittelt (Kruse und andere, 2021). Um das hohe Qualitätsniveau der Außenhandelsstatistik laufend zu gewährleisten, sind monatlich etwa 60 Millionen zugrunde liegende Datensätze aufwendig zu plausibilisieren und fehlerhafte Angaben bestmöglich zu korrigieren. Diese Aufgabe ist nur durch eine weitgehend automatisierte Identifikation unplausibler Angaben zu bewältigen.

Der manuellen[2] und maschinellen[3] Korrektur der Meldungen vorangestellt sind maschinelle Prüfverfahren, um Auffälligkeiten in den Daten zu erkennen. Bei dieser maschinellen Prüfung werden die gemeldete Eigenmasse, der gemeldete Statistische Wert und die gemeldete Besondere Maßeinheit der Waren mit einer definierten Toleranzmarge verglichen. Wenn die Angaben in einer Meldung aus diesem Toleranzbereich fallen, wird der Datensatz als auffällig gekennzeichnet, auf seine Richtigkeit überprüft und gegebenenfalls korrigiert. Die Toleranzmargen nennen sich „Stammbänder“ und definieren plausible Wertebereiche. Die Plausibilisierung mithilfe der Stammbänder entspricht einer Ausreißeranalyse.

Derzeit werden diese Stammbänder größtenteils manuell und anlassbezogen angepasst, wenn Auffälligkeiten in den Daten entdeckt werden. Dieses Vorgehen ist mit einem hohen Aufwand verbunden. Dieser Artikel stellt verschiedene Clustering-Modelle zur Unterstützung der Stammbandpflege vor und untersucht, ob sie im Plausibilisierungsprozess der Außenhandelsdaten nutzbar sind. Ziel ist, die Anpassung der Stammbänder durch Clustering-Verfahren effizienter zu gestalten, indem die plausiblen Wertebereiche automatisch gesetzt werden.

Kapitel 2 erläutert, welche Herausforderungen sich bei der Ausreißeranalyse zur Plausibilisierung der Außenhandelsdaten ergeben. Des Weiteren beschreibt das dritte Kapitel die für diesen Beitrag benutzte Datenbasis und Details ihrer Aufbereitung. Kapitel 4 stellt die gewählten Parameter der angewendeten Clustering-Verfahren kurz vor, ebenso Evaluierungsmöglichkeiten dieser Modelle. Das fünfte Kapitel enthält die ausgewerteten Resultate und erörtert die Nutzbarkeit der Modelle für die Plausibilisierung der Außenhandelsdaten. Der Beitrag endet mit einem Fazit.

2

Herausforderungen

Das Statistische Bundesamt veröffentlicht monatlich Statistiken zum deutschen Außenhandel im Wesentlichen nach Warennummer, Richtung (Import und Export) und Bestimmungs- (im Fall von Exporten) beziehungsweise Ursprungsland (im Fall von Importen). Das Warenverzeichnis mit fast 10 000 Warennummern für das Jahr 2023 erschwert eine automatische Plausibilisierung. Dieses Kapitel beschreibt einige Aspekte, weshalb sich die Anwendung von Verfahren des Maschinellen Lernens

1 Als Mengenangaben stehen die Eigenmasse in Kilogramm – definiert als Gewicht der Ware ohne alle Umschließungen – und gegebenenfalls die sogenannte Besondere Maßeinheit, beispielsweise die Stückzahl der Ware, zur Verfügung.

2 Einen Teil der Meldungen korrigieren die zuständigen Bearbeiterinnen und Bearbeiter durch Expertenwissen und durch Rücksprachen mit den meldenden Unternehmen.

3 Ein Teil der Meldungen wird maschinell mittels eines Algorithmus behandelt (Blang/Helmert, 2008).

für die Plausibilisierung der gemeldeten Daten herausfordernd gestaltet.

2.1 Anzahl der verfügbaren Meldungen

Der Einsatz von automatisierten Verfahren, beispielsweise des Maschinellen Lernens, bei der Plausibilitätsprüfung erfordert eine ausreichend große Anzahl von Lerndatensätzen je zu prüfender Einheit. Die hohe Granularität der Außenhandelsstatistik führt dazu, dass sich der veröffentlichte Wert auf der detailliertesten Ebene, also der Kombination aus achtstelliger Warennummer, Richtung, Statistik und Partnerland, mitunter aus einer nur sehr geringen Zahl von Einzeldatensätzen bildet.

Generell werden Ausreißeranalysen durch Stammbänder in der Außenhandelsstatistik je nach Bedarf ausgeführt. Die hohe und über die Jahre wachsende Anzahl an Stammbändern zeigt ↘ Tabelle 1.

Beispielsweise gab es für das Berichtsjahr 2019 für die Kombination der Merkmale Richtung, Statistik[4] und Warennummer (WA) 38 233 Stammbänder. Für das Berichtsjahr 2023 hat sich deren Zahl auf 39 136 erhöht. Dagegen gab es für das Berichtsjahr 2019 für die Kombination der Merkmale Richtung, Statistik, Warennummer und Bestimmungsland (Export) beziehungsweise Ursprungsland (Import) 670 Stammbänder, deren Zahl sich bis 2023 auf 847 erhöht hat. Es zeigt sich, dass Stammbänder immer für die Kombination aus Richtung, Statistik und Warennummer definiert sind. Entsprechend wird für diese Merkmalskombinationen generell eine Ausreißeranalyse durchgeführt. In Kombination mit diesen drei Merkmalen prüfen Stammbänder zusätzlich die Merkmale Unternehmen, Bestimmungsland beziehungsweise Ursprungsland (BLD/ULD) sowie Ursprungs- beziehungsweise Bestimmungsbundesland (UBLD/BBLD).

Ein erheblicher Anteil aller Stammbänder nimmt zusätzlich Bezug auf Unternehmen, um Ausreißerprüfungen spezifisch auf Unternehmensebene durchführen zu können. Ein kleiner Anteil der Stammbänder garantiert ebenfalls eine Ausreißeranalyse für das Merkmal Bestimmungs- beziehungsweise Ursprungsland.[5]

Der erhöhte Bedarf an Stammbändern ist vor allem auf die steigende Anzahl von Warennummern zurückzuführen. Das Warenverzeichnis für die Außenhandelsstatistik ändert sich jährlich, da Warennummern neu hinzukommen, wegfallen oder in einer oder mehreren anderen Warennummern aufgehen können.

↘ Grafik 1 gibt einen Gesamtüberblick über die dynamische Entwicklung des Warenverzeichnisses seit 2009. Das linke Teildiagramm bildet die Anzahl der Warennummern ab, für die die Laufzeit im rechten Teildiagramm zutrifft. So gibt es beispielsweise ausweislich der obersten schwarzen Linie eine Zahl größer 10^3 (= nämlich genau 7 375) an Warennummern, die über den gesamten betrachteten Zeitraum gültig sind. Zugleich zeigt die nächstniedrigere schwarze Linie, dass 779 Warennummern 2011 weggefallen sind. Die blaue Linie zeigt

4 Das Merkmal Statistik hat lediglich zwei Ausprägungen, die zwischen Datensätzen unterscheiden, die dem Extrahandel oder dem Intrahandel zuzuordnen sind.

5 Da es sich bei den beiden Erhebungsmerkmalen Bestimmungsland im Export und Ursprungsland im Import um die wesentlichen Partnerlandkategorien für die Veröffentlichung handelt, werden diese in einem einzigen internen Merkmal Bestimmungs- beziehungsweise Ursprungsland (BLD/ULD) abgespeichert und plausibilisiert.

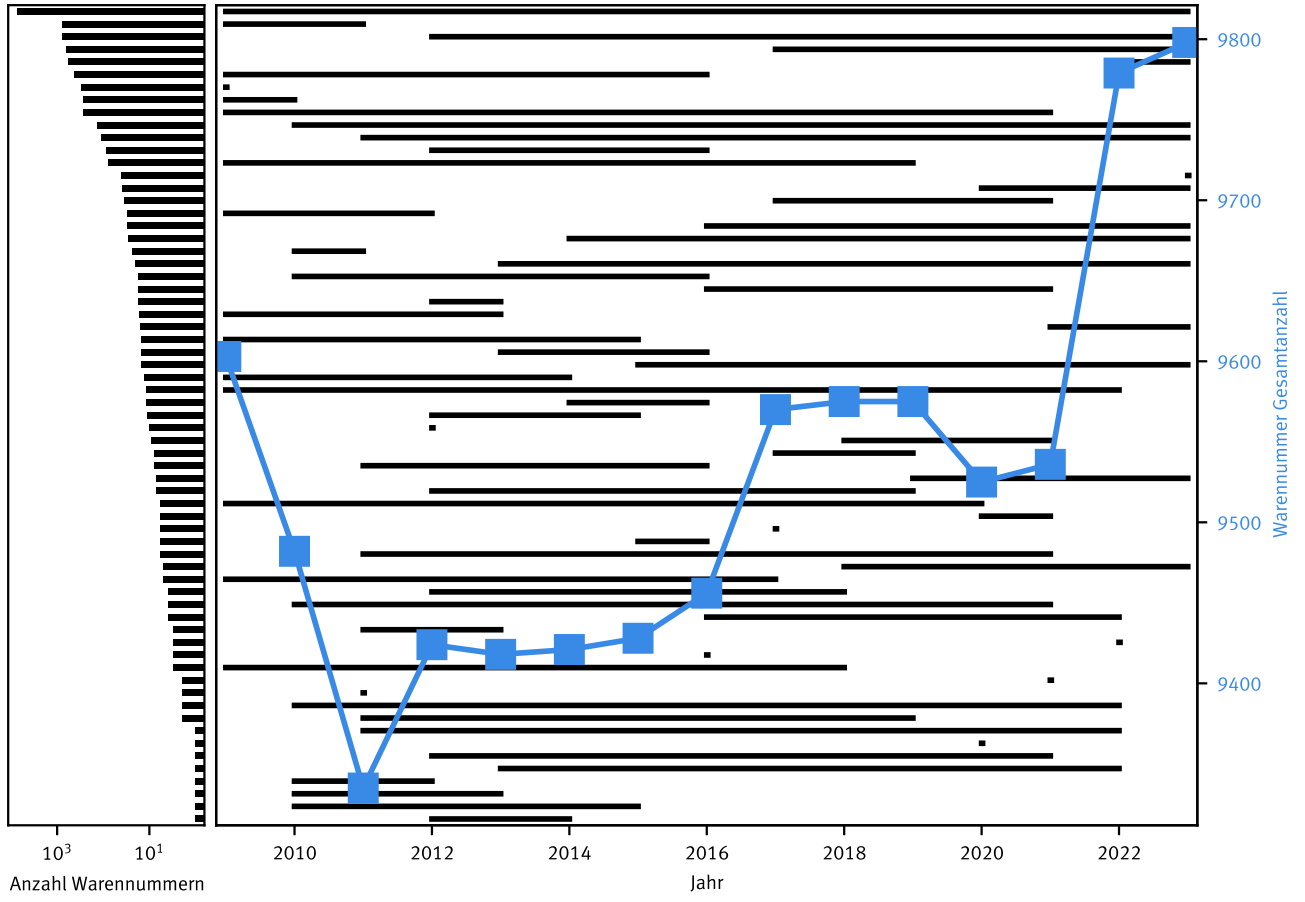
Tabelle 1
Verwendung von Stammbändern in der Außenhandelsstatistik für die Plausibilisierung verschiedener Merkmale

Richtung	Statistik	WA	Unternehmen	BLD/ULD	UBLD/BBLD	2019	2020	2021	2022	2023
						38 233	38 043	38 095	39 056	39 136
						10 713	13 095	14 970	16 478	17 028
						670	781	851	875	847
						346	523	548	577	563
						276	442	474	485	465
						26	32	36	30	31
					Insgesamt	50 264	52 916	54 974	57 501	58 070

WA: Warennummer; BLD/ULD: Bestimmungsland/Ursprungsland; UBLD/BBLD: Ursprungs-/Bestimmungsbundesland.
Rot eingefärbte Zellen geben an, dass Stammbänder sich auf Ausprägungen der angegebenen Merkmale beziehen. Grau eingefärbte Zellen geben an, dass sich Stammbänder, die in der Zeile gezählt werden, nicht auf ein Merkmal beziehen.

Grafik 1

Anzahl von Warennummern im Warenverzeichnis für die Außenhandelsstatistik je Gültigkeitszeitraum



Anzahl von Gruppen von Warennummern mit gleichen Gültigkeitszeiträumen (links dargestellt) und entsprechendem Gültigkeitszeitraum als horizontale Linien im rechten Teildiagramm. Hierbei markiert der Anfang bzw. das Ende einer horizontalen schwarzen Linie den Anfang bzw. das Ende der Gültigkeit einer Warennummer. In blau ist die Gesamtanzahl von Warennummern im angegebenen Jahr mit entsprechender blauer Skala dargestellt.

wiederum an, wie viele Warennummern die jeweiligen Jahresausgaben des Warenverzeichnisses für die Außenhandelsstatistik insgesamt enthielten. So verfügt die Ausgabe 2023 über 9798 einzelne Warennummern, denen die meldepflichtigen Unternehmen ihre exportierten beziehungsweise importierten Waren zuordnen müssen.

Grafik 1 zeigt weiter, dass die Komplexität des Warenverzeichnisses über die Zeit tendenziell zugenommen hat. Größere Revisionen des Warenverzeichnisses erfolgen im Abstand von fünf Jahren (Statistisches Bundesamt, 2023). Zu sehen ist das beispielsweise an der größeren Anzahl von Warennummern, die zu den Jahreswechseln 2016/2017 und 2021/2022 hinzugekommen sind, während gleichzeitig wenige Warennummern ihre Gültigkeit verloren haben. Gleichzeitig ist zu erkennen, dass die weitaus größte Gruppe von Warennummern aus solchen besteht, die im gesamten betrachteten Zeitraum und somit langfristig gültig sind.

Dieser Beitrag untersucht Ausreißeranalysen mittels Clustering-Verfahren auf der Ebene von Richtung, Statistik und WA, da diese Merkmalskombination den größten Anteil der Stammbänder ausmacht. Neu hinzugekommene Warennummern, die in einem anfänglichen Zeitraum nicht genügend Meldungen aufweisen, um Clustering-Algorithmen testen zu können, wurden ausgeschlossen.

2.2 Heterogenität der Daten

Außenhandelsdaten zeigen in der Differenzierung nach Warennummer, Richtung und Statistik unterschiedliche Verteilungen von Kenngrößen wie Statistischer Wert je Eigenmasse, Statistischer Wert je Besondere Maßeinheit und Eigenmasse je Besondere Maßeinheit. Hierbei weisen die Verteilungen verschiedene Eigenheiten auf.

Multimodalität: Es gibt Warennummern, die eine weite Bandbreite von Eigenmasse und Statistischem Wert aufweisen. Eine Ausreißeranalyse sollte daher nicht nur auf unimodale Verteilungen abzielen, sondern mit Multimodalitäten umgehen können.

Schiefe: Die Verteilung der Durchschnittswerte einiger Warennummern ist oft asymmetrisch. Es kommt häufig vor, dass diese eine rechtsschiefe Gestalt annimmt. Die Annahme einer symmetrischen Verteilung, etwa der Normalverteilung, als Grundlage für ein Modell wäre hier inadäquat.

Trends: Die Annahme einer stationären Verteilung über die Zeit scheitert bei saisonalen Waren und bei Waren, bei denen die entsprechenden Kenngrößen längerfristige Trends aufweisen. So kann der durchschnittliche Wert der Waren, die einer bestimmten Warennummer zuzuordnen sind, über die Zeit zurückgehen. Dadurch wäre es problematisch anzunehmen, dass die Verteilungen konstant bleiben.

Aufgrund der genannten Eigenschaften werden nichtparametrische Clustering-Verfahren getestet, denen keine Verteilung zugrunde gelegt wird. Des Weiteren wird das Clustering Monat für Monat vorgenommen. Ebenfalls werden die monatlichen Verteilungen so transformiert, dass 90 % der Werte zwischen 0 und 1 liegen. Dies soll den Effekt von Trends vermindern, da die betrachteten Werte auf eine vergleichbare Skala je Berichtsmonat zurückgeführt werden.

3

Datenbasis

Die hier vorgestellten Algorithmen werden an einer Auswahl von Warennummern getestet, die aufgrund der Struktur der darunter gemeldeten Daten als relevante Use Cases für die Erprobung der Clustering-Verfahren erachtet werden können. Insgesamt werden 83 Kombinationen von Statistik, Richtung und Warennummer über die Berichtszeiträume Dezember 2022 bis einschließlich August 2023 monatlich auf Ausreißer untersucht.

Stammbänder überprüfen im Plausibilisierungsprozess für die Außenhandelsdaten, ob relative Größen einer Meldung in einen plausiblen Bereich fallen. Die Ausreißeranalysen werden in diesem Artikel anhand des Statistischen Wertes je Eigenmasse getestet. Hierbei werden die Merkmale Statistik, Richtung, Warennummer, Statistischer Wert und Eigenmasse verwendet. Bereits korrigierte Werte ersetzen fehlende Angaben der Warennummern in den Meldungen, da die Imputation fehlender Angaben nicht im Fokus dieses Beitrags steht.

Die Modelle sollen für jeden Berichtsmonat und auf verschiedene Warennummern, die verschiedene Skalen der numerischen Größen enthalten, angewendet werden. Daher wird der Statistische Wert je Eigenmasse in jedem Monat skaliert, um die gemeldeten Werte auf vergleichbaren Skalen darstellen zu können. Es wird eine lineare Skalierung verwendet, sodass Werte auf dem 5-%-Perzentil 0 und Werte auf dem 95-%-Perzentil 1 sind. Dadurch ist es möglich, die Parameter der Modelle ebenfalls auf vergleichbaren Skalen für verschiedene Warennummern zu wählen.

4

Methodik

Wie im Kapitel 2 bereits beschrieben, können die verschiedenen Verteilungen der Daten sehr heterogen sein. Daher werden in diesem Artikel ausschließlich nichtparametrische Verfahren zur Ausreißerdetektion verwendet.

Aufgabe der Stammbänder ist es, eine Meldung anhand relativer numerischer Größen als auffällig oder nicht auffällig zu markieren. Daher wird das Resultat der Ausreißeranalyse eine binäre Zahl sein. Ein Resultat von 0 beziehungsweise 1 für einen Datenpunkt heißt, dass es sich um einen unplausiblen beziehungsweise plausiblen Datenpunkt handelt. Es ergeben sich daraufhin Bereiche, in denen Datenpunkte plausibel sind und folglich mit 1 klassifiziert werden. Diese werden „Inlier-

Bereiche“ genannt. Im Gegensatz dazu werden Bereiche, in denen Datenpunkte mit 0 klassifiziert werden, „Outlier-Bereiche“ genannt.

Das bisherige manuelle Einstellen der Stammbänder lässt Raum für subjektive Einflüsse auf diese Einteilung. So könnte die Qualität der Adjustierung der Outlier-Bereiche von der bearbeitenden Person abhängig sein. Aus diesem Grund ist es nicht ratsam, einen Algorithmus so zu trainieren, dass dieser Ergebnisse bisheriger Stammbänder reproduziert.

Stattdessen werden Algorithmen genutzt, die diese Aufgabe ausschließlich aufgrund der Daten bewältigen. Mit anderen Worten: unüberwachtes Lernen wird zur Detektion von auffälligen Meldungen verwendet. Aus diesem Grund werden hier Clustering-Verfahren genutzt[6] und im Folgenden kurz dargestellt.

4.1 Verwendete Modelle

Kerneldichteschätzung

Mit Kernel Density Estimation (KDE) (Węglarczyk, 2018) lässt sich ein Schätzer für eine Wahrscheinlichkeitsdichtefunktion (probability density function – PDF) erhalten. Die PDF ergibt sich als Summe von Kernels $k_h(x, x_i)$, die an der Stelle x ausgewertet werden. Entscheidend ist die Bandbreite h des Kernels. Bei einer gegebenen Bandbreite h nimmt die Schätzung der PDF $\hat{f}(x)$ folgende Form an:

$$\hat{f}(x) = \frac{1}{N}\sum_{i=1}^{N} k_h(x, x_i) \tag{1}$$

x_i sind die Positionen der N Datenpunkte. Je größer die Bandbreite, umso glatter erscheint die Schätzung der PDF.

Ausreißer können mit der KDE-Methode identifiziert werden, indem ein Schwellenwert eingeführt wird. Liegt der Wert der geschätzten PDF an einer Stelle x unter diesem Wert, wird diese Stelle als Ausreißer klassifiziert. Dieser Schwellenwert wird im Folgenden „Cut“ genannt.

Da sich eine PDF zu Eins integriert, würde dieser Grenzwert jedoch stark von der Verteilung der Daten abhängen. Zum Beispiel würde eine Gauss-Verteilung an ihrem Maximum einen höheren Dichtewert erreichen als zwei Gauss-Verteilungen mit derselben Standardabweichung. Aus diesem Grund wird die Dichteschätzung für das Ausreißer-Modell nicht auf 1 normiert, sondern auf ihre maximale Dichte.

Local Outlier Factor

Für den Local Outlier Factor (LOF) (Breunig und andere, 2000) werden die sogenannten nächsten Nachbarn eines Datenpunktes herangezogen, um Ausreißer aufzudecken. Dabei handelt es sich um Datenpunkte, die die kürzeste Distanz zu einem betrachteten Datenpunkt haben. Auf Basis dieser Nachbarn wird die lokale Dichte quantifiziert. Die lokalen Dichten werden anschließend miteinander verglichen. Datenpunkte mit geringerer lokaler Dichte als ihre Nachbarn werden als Ausreißer klassifiziert. Der Kontaminations-Parameter des LOF beeinflusst hierbei, ab welchen Dichteunterschieden ein Datenpunkt als Ausreißer erachtet wird. Kontamination gibt an, wie hoch der zu erwartende Anteil an Ausreißern am vorliegenden Datensatz ist.

Durch die Betrachtung der Distanz zum n-ten Nachbarn werden lokale Fluktuationen geglättet. Demnach hat eine Erhöhung der betrachteten Nachbarn einen glättenden Effekt auf die Bereiche, die auffällige von nicht auffälligen Werten unterscheiden. Der kontrollierende Parameter wird im Folgenden n-Nachbarn genannt, wobei n die Anzahl der betrachteten nächsten Nachbarn ist.

Isolation Forest

Isolation Forests (Liu und andere, 2008) teilen den Wertebereich sukzessive in kleinere Teilmengen auf, bis alle Datenpunkte isoliert in verschiedenen Teilmengen vorliegen. Da Ausreißer in weniger dichten Bereichen liegen, ist die Anzahl der Partitionen, die der Isolation-Forest-Algorithmus benötigt, um einen anomalen Datenpunkt zu isolieren, geringer als bei den Datenpunkten, die keine Ausreißer sind.

Auf Basis der Anzahl der Partitionierungen wird jedem Datenpunkt ein Score zwischen 0 und 1 zugeordnet. Je näher ein Score der Zahl Eins ist, desto wahrscheinlicher ist es, dass dieser einen Ausreißer darstellt.

Der Isolation Forest besitzt einen ausschlaggebenden Parameter. Dieser nennt sich Kontamination und setzt

6 Per Definition sind Clustering-Verfahren Methoden zur Einteilung eines Datensatzes in distinkte Gruppen ohne Zutun eines bereits eingeteilten Datensatzes als Referenz (Tan und andere, 2005).

fest, mit welchem Anteil von Ausreißern im betrachteten Datensatz zu rechnen ist. Dieser Parameter hat einen direkten Einfluss auf die Abschneidegrenze für den angesprochenen Score.

DBSCAN

DBSCAN (Density-based spatial clustering of applications with noise) (Ester und andere, 1996) klassifiziert Datenpunkte in Kernobjekte, Dichte-erreichbare Objekte und Ausreißer. Wenn sich mindestens n Nachbarn in einer Umgebung ε um einen Datenpunkt befinden, so ist dieser ein Kernobjekt. Alle Punkte in dieser Umgebung ε, die nicht Kernobjekte sind, sind Dichte-erreichbare Objekte. Datenpunkte, die weder Kernobjekte noch Dichte-erreichbare Objekte sind, werden als Ausreißer deklariert.

Das DBSCAN-Verfahren besitzt zwei Parameter, die Einfluss auf das Clustering-Ergebnis haben. Zum einen ist die Mindestanzahl der Objekte in der ε-Umgebung (hier n-Nachbarn genannt) von Interesse, zum anderen der Radius ε selbst. Je höher der Parameter n-Nachbarn, desto schwieriger ist es für Objekte, Kernobjekte zu werden. Eine Erhöhung von ε hat den gegensätzlichen Effekt.

4.2 Parameterwahl

Zur optimalen Parameterwahl wird das Kartesische Produkt aus verschiedenen Parametern je Algorithmus getestet. Hierbei wird eine Iteration durch die verschiedenen Kombinationen von Parametern der verschiedenen Algorithmen durchgeführt und die entstehenden Modelle auf die Daten je Monat trainiert. Die ausgewählten Parameter sind ↘ Tabelle 2 zu entnehmen.

Tabelle 2
Angewendete Modelle und die Wahl ihrer Parameter in den angegebenen Intervallen Minimum bis Maximum

Modellname	Parameter	Minimum	Maximum
Kerneldichteschätzung (KDE)	Bandbreite	0.01	0.5
	Cut	0.01	0.5
Density-based spatial clustering of applications with noise (DBSCAN)	n-Nachbarn	4	14
	ε	0.01	0.5
Isolation Forest	Kontamination	0.025	0.2
Local Outlier Factor (LOF)	n-Nachbarn	4	14
	Kontamination	0.025	0.2

4.3 Bewertung der Modelle

Da es keine Referenzdaten als Vergleichswerte gibt und die Clustering-Algorithmen autonom auf Basis der Verteilungen des Statistischen Werts je Eigenmasse entscheiden, welche Meldungen überprüft werden sollen, müssen Gütekriterien gefunden werden, um die Ergebnisse zu validieren. Diese Kriterien sollten möglichst folgende Voraussetzungen erfüllen:

- Sie benötigen keine Referenzdaten: Dies ist wichtig, da keine Vergleichsdaten, die ein Optimum darstellen, vorhanden sind.
- Skalen-Invarianz: Da die ursprünglichen Werte unterschiedlich normiert werden, soll diese Normierung keine Auswirkung auf die Gütekriterien haben.
- Beschränkter Definitionsbereich: Die Gütekriterien sollten möglichst ein definiertes Maximum und Minimum besitzen. Ein Kriterium, welches gegen unendlich geht, ist hierbei nicht wünschenswert.
- Interpretierbarkeit: Die Gütekriterien sollten leicht zu interpretieren sein.

Im Folgenden werden verschiedene Aspekte zur Einschätzung der Qualität und deren Quantifizierungsmöglichkeiten betrachtet.

Entstehender Aufwand

Die durch Stammbänder als auffällig erachteten Meldungen müssen entweder manuell oder maschinell überprüft werden. Um den Arbeitsaufwand den Kapazitäten anzupassen und zu verhindern, dass Meldungen überkorrigiert[7] werden, sollte sich die Anzahl der auffälligen Datensätze in bestimmten Grenzen halten und quantifizieren lassen. Daher wird die Zahl der erkannten Ausreißer im Verhältnis zur Gesamtzahl der Meldungen als Arbeitsaufwand definiert. Ausreißer sind anormale Werte, die sich vom Rest der Werte maßgeblich unterscheiden. Ein zu großer Arbeitsaufwand von über 0.5 würde aussagen, dass die Hälfte der Werte bearbeitet werden. Ein so hoher Wert würde sowohl den Bearbeitungsaufwand unnötig erhöhen als auch infrage stellen, ob es sich um Ausreißer handelt.

7 Dies tritt auf, wenn zu viele Datenpunkte als auffällig erachtet werden und dadurch eine hohe Anzahl an Korrekturen der Meldungen hervorgerufen wird. In diesem Fall wird von einer „Überkorrektur“ gesprochen.

Lokalisierung der Outlier-Bereiche

Ausreißer sind Datenpunkte, die sich in Bereichen niedriger Datendichte befinden, jedoch ist dies der Kern von vielen Clustering Algorithmen und Verfahren für Ausreißerdetektion. Eine einfache Methode, um zu quantifizieren, ob Daten relativ dicht in einem Inlier-Bereich allokiert sind, stützt sich auf ein Maß, das dem errechneten Arbeitsaufwand aus Abschnitt „Entstehender Aufwand“ (siehe unten) ähnelt. Statt jedoch die Anzahl der als auffällig erachteten Werte durch die Gesamtzahl der Werte in einem Bezugszeitraum zu teilen, wird die Länge der Inlier-Bereiche verwendet. Die gemittelte Dichte im Inlier-Bereich D_I ist die Anzahl der Datenpunkte in den Inlier-Bereichen im Verhältnis zur Länge der Inlier-Bereiche. Die gemittelte Gesamtdichte D_Ω ist die Gesamtzahl der Daten geteilt durch Maximum Minus Minimum der betrachteten Werte, definiert wird der gemittelte Dichte-Index als

$$(2) \qquad Gemittelte\ Dichte = \frac{D_I - D_\Omega}{\max\{D_I, D_\Omega\}}$$

Dieser Index ist zwischen −1 und 1 normiert. Je höher sein Wert, desto dichter sind die Datenpunkte in den Inlier-Bereichen. Bei einem Wert von 0 führt das Clustering zu keiner Erhöhung der Dichte bei den Datenpunkten, die nicht als Ausreißer klassifiziert werden. Eine hohe Dichteänderung spricht daher dafür, dass ein Clustering-Verfahren dichte Bereiche als Cluster erachtet. Aus diesem Grund wird angestrebt, dass dieses Gütemaß möglichst hoch ist.

Robustheit der Inlier-Bereiche

Die bisherigen Teilaspekte der Güte fokussierten lediglich individuell auf jeden einzelnen Monat. Allerdings suggeriert der Begriff „Stammband“, dass ebenfalls ein Zusammenhang zwischen den Resultaten angrenzender Monate bestehen sollte. Mit anderen Worten: Inlier- und Outlier-Bereiche sollten von einem auf den anderen Monat nicht völlig unterschiedlich sein. Dies soll garantieren, dass die Plausibilisierung nachvollziehbar und einfach zu handhaben bleibt. Ebenfalls soll verhindert werden, dass kleine Änderungen an den Verteilungen andere Resultate in Bezug auf Inlier- und Outlier-Bereiche generieren.

Eine Möglichkeit, dies zu quantifizieren, ist, die Spannen der Inlier-Bereiche von Monat zu Monat zu vergleichen. Da diese Bereiche Intervalle auf der gesamten Domäne der möglichen vorkommenden Werte sind, wird ein Ähnlichkeitsmaß für Mengen benötigt. Hierbei wird der Jaccard-Index verwendet, der den Anteil der sich überschneidenden Inlier-Bereiche von zwei Monaten in Hinsicht auf den gesamten Inlier-Bereich der zwei Monate zusammengenommen berechnet. Der Jaccard-Index (Winter/Lewandowsky, 1971) zweier Mengen $\mathbb{B}_1$ und $\mathbb{B}_2$ wird wie folgt berechnet.

$$(3) \qquad J(\mathbb{B}_1, \mathbb{B}_2) = \frac{|\mathbb{B}_1 \cap \mathbb{B}_2|}{|\mathbb{B}_1 \cup \mathbb{B}_2|}$$

Ein Wert von 0 gibt an, dass die Inlier-Bereiche sich nicht überschneiden. Ein Wert von 1 zeigt an, dass die Inlier-Bereiche in zwei aufeinanderfolgenden Monaten identisch sind. Da eine gewisse Konsistenz der Inlier-Bereiche Robustheit garantiert, sollte der Jaccard-Index möglichst hoch sein.

5

Ergebnisse

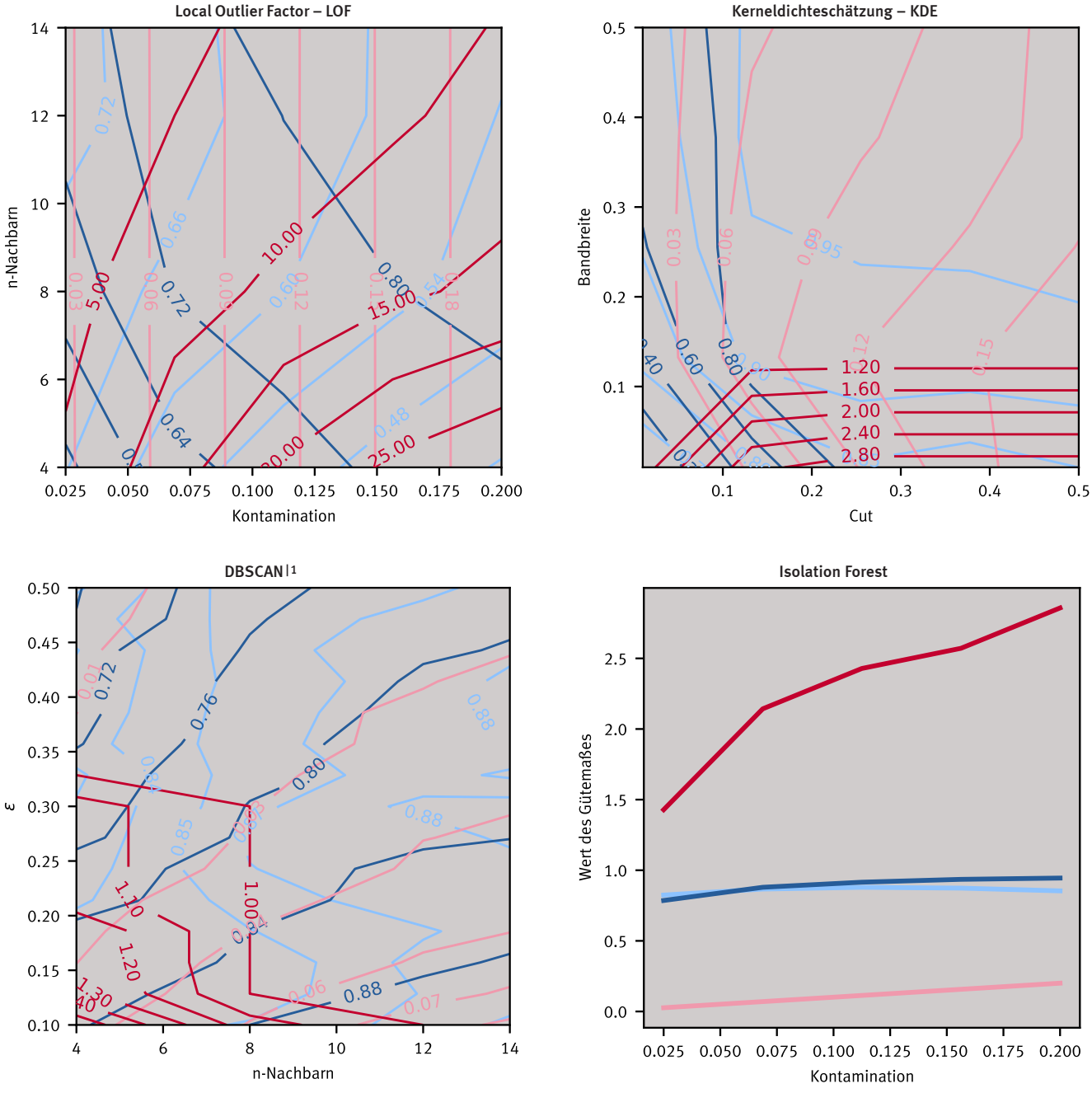
Sowohl durch die Wahl des Algorithmus als auch die Kombination verschiedener Parameter entsteht eine Bandbreite von Modellen, die dazu dienen, Ausreißer zu entdecken. Im Folgenden werden die Ergebnisse dieser Modelle über eine Vielzahl von Merkmalskombinationen aus Statistik, Richtung und Warennummer durch Anwendung der bereits beschriebenen Gütemaße ausgewertet.

Durch die Anwendung verschiedener Clustering-Methoden entstehen Bereiche, die eine höhere Datendichte aufweisen als andere (Inlier-Bereiche). Datenpunkte in diesen Bereichen werden als plausibel bezeichnet, Werte außerhalb des Inlier-Bereichs sind unplausibel und sind zu plausibilisieren.

↘ Grafik 2 ist zu entnehmen, dass es je Modell maximal zwei wählbare Parameter gibt. Die Grafik stellt dar, wie sich ihre Wahl auf die Detektion von Ausreißern auswirkt.

Der Effekt der verschiedenen Modelle auf die Glättung der Ausreißer-Ergebnisse ist an der Anzahl der Inlier-Bereiche als auch am Jaccard-Index zu erkennen. Für Algorithmen wie LOF und DBSCAN, die sich auf n-te Nachbarn beziehen, steigt die Anzahl der Inlier-Bereiche

Grafik 2
Gütemaße für die einzelnen Modelle, abhängig von den gewählten Parametern



Auf den Achsen sind die Parameter der Modelle verzeichnet. Die verschiedenfarbigen Konturlinien stellen die Gütemaße dar (rot: Anzahl Inlier-Bereiche, rosa: Arbeitsaufwand, hellblau: Jaccard-Index, dunkelblau: Gemittelte Dichte). Da für den Isolation Forest nur ein Parameter verändert wird, sind die Werte der Gütemaße auf der vertikalen Achse zu finden.

1 DBSCAN: Density-based spatial clustering of applications with noise.

mit sinkendem Parameter n-Nachbarn. Die Einschätzung der lokalen Dichte durch diese Algorithmen wird demnach lokaler und ist anfälliger für Fluktuationen.

Der Jaccard-Index wird ebenfalls reduziert bei Verringerung der n-Nachbarn. Allerdings ist dieser Effekt nicht stark ausgeprägt im Fall von DBSCAN und der Algorithmus zeigt nur geringe Schwankungen im Jaccard-Index

für die gewählten Kombinationen von Parametern. Dahingegen reagiert der LOF-Algorithmus äußerst sensibel auf die Wahl seiner beiden Parameter. Eine Verringerung des Kontaminations-Parameters resultiert in einem höheren Jaccard-Index und damit einem niedrigeren zeitlichen Zusammenhang zwischen Clustering-Ergebnissen von Monat zu Monat. Der Arbeitsaufwand müsste allerdings auf unter 0.03 fallen (das heißt, dass im Schnitt weniger als 3 % der Meldungen als auffällig erachtet werden), um eine Größe des Jaccard-Index zu erzielen, die mit den anderen Algorithmen vergleichbar ist. Die Inlier-Bereiche würden infolgedessen allerdings stets noch fünf Subintervalle umfassen.

Die gemittelte Dichte in den Inlier-Bereichen steigt bei Erhöhung des n-Nachbarn-Parameters im LOF- und DBSCAN-Verfahren an. Dieser Effekt scheint allerdings schwach, sodass die Wahl der übrigen Parameter (Kontamination und ε) größere Veränderungen der Dichte in den Inlier-Bereichen hervorruft. Genau wie beim Isolation Forest sorgt eine höhere Kontamination sowohl für mehr Inlier-Bereiche als auch einen Anstieg der gemittelten Dichte. Eine Erhöhung des ε im DBSCAN-Modell verringert diesen Dichte-Index, da hier tendenziell Bereiche mit weniger Datendichte zu plausiblen Wertebereichen zusammengefasst werden.

Der Parameter ε des DBSCAN-Modells reguliert, bis zu welcher Distanz Nachbarn eines Objekts gezählt werden. Damit kann ε als eine Art Regulator der Interaktion zwischen Datenpunkten gesehen werden. Ebenfalls verfügt der KDE-Schätzer über einen Parameter für den Einfluss der einzelnen Kernel-Instanzen. Die Bandbreite der KDE vergrößert den Einfluss jedes Kernels. Beide Parameter haben eine glättende Wirkung auf die Inlier-Bereiche, wenn sie erhöht werden. Somit verringert sich infolge einer Erhöhung die Anzahl der Inlier-Bereiche.

Diese beiden Parameter bestimmen ebenfalls vornehmlich den Arbeitsaufwand. Dieser ist im DBSCAN-Modell relativ gering und überschreitet die 0.1 für den gewählten Bereich der getesteten Parameter nicht. Per Definition skaliert der entstehende Arbeitsaufwand mit dem Kontaminations-Parameter der Isolation-Forest- und LOF-Methoden. An den tendenziell vertikal ausgerichteten Konturlinien desselben Scores im Fall des KDE-Modells ist zu erkennen, dass der Cut-Wert eine ähnliche Wirkung hat. Je höher dieser Parameter des KDE gewählt wird, desto mehr Werte werden als Ausreißer klassifiziert.

Die gemittelte Dichte in den Inlier-Bereichen steigt mit Erhöhung der Kontamination (für LOF und Isolation Forest) beziehungsweise des Cut-Werts (für KDE), da die Modelle weniger tolerant gegenüber Werten in weniger dichten Bereichen des Wertebereichs werden. Somit bleiben nur noch die Bereiche mit der höchsten Datendichte als Inlier-Bereiche erhalten.

6

Fazit

In diesem Artikel wurden nichtparametrische Clustering-Methoden auf Meldungen im Außenhandel angewandt, um Ausreißer aufzudecken. Die Analyse der Clustering-Ergebnisse zeigt, dass die Parameter Parallelen aufweisen.

Zum einen besitzen die Modelle LOF und DBSCAN ordinale Parameter, die festlegen, aufgrund wie vieler Nachbarn eine Einschätzung der lokalen Dichte gemacht wird. Zum anderen wird den KDE- und DBSCAN-Modellen ebenfalls ein Parameter mitgegeben, der den Wirkungsbereich erhöht. Eine Zunahme dieser Parameter führt zur Glättung der gefundenen Inlier-Bereiche. Für einige Verfahren lässt sich die Anzahl der erwarteten Ausreißer einstellen. So haben sowohl das LOF-Verfahren als auch das Isolation-Forest-Verfahren einen Kontaminations-Parameter. Da hierdurch die Abschneidegrenze für den Score der beiden Modelle gesetzt wird, lässt sich dieser mit dem Cut-Wert des KDE verbinden.

Inwiefern die Modelle für die Plausibilisierung im Außenhandel geeignet sind, lässt sich aufgrund der Gütekriterien einschätzen.

Das LOF-Verfahren tendiert dazu, eine höhere Anzahl an Inlier-Bereichen zu generieren als die anderen Algorithmen. Damit geht eine größere Fluktuation von Monat zu Monat für die betrachteten Merkmalskombinationen aus Statistik, Richtung und Warennummer einher. Aus diesem Grund scheint das LOF-Verfahren für die Plausibilisierung im Außenhandel nicht verwendbar zu sein, da die analysierten Daten eine hohe Heterogenität aufzeigen. Ein Clustering-Algorithmus, der durch größere Schwankungen in seinen Resultaten hervorsticht, wäre nicht adäquat zu kontrollieren.

Das DBSCAN-Verfahren deckt eine äußerst geringe Zahl an Ausreißern auf. Für die Plausibilisierung von Meldewerten im Außenhandel sollten jedoch tendenziell mehr Warnhinweise für potenziell auffällige Werte generiert werden.

Daher kommen das KDE-Verfahren und das Isolation-Forest-Verfahren als geeignete Kandidaten für eine Automatisierung der Ausreißeranalyse infrage. Hierbei besitzt das KDE-Verfahren eher einen globalen Fokus, da sowohl die Bandbreite als auch der Cut-Wert nicht dazu führen, dass durch das Modell auf unterschiedlich dichte Datenansammlungen eingegangen wird. Das Isolation-Forest-Verfahren ist hingegen in der Lage, sowohl in Bereichen hoher Dichte als auch niedriger Dichte Ausreißer verschiedenartig zu erkennen. Hingegen kann das Resultat des KDE-Verfahrens einfacher in Bayessche Modelle eingebettet werden, da dieses Verfahren ebenfalls eine Schätzung der Wahrscheinlichkeitsdichte bietet.

Als nächster Schritt der Modellierung ist geplant, Inlier-Bereiche des Vormonats als Vorschlag für den aktuellen Monat zu verwenden. Diese zusätzliche Information könnte das in diesem Artikel beschriebene Problem der geringen Anzahl von Meldungen zumindest teilweise kompensieren. Hierbei wäre die probabilistische Interpretation des KDE-Verfahrens von Nutzen. Ebenfalls wird angestrebt, die Resultate der hier ausgewählten Modelle mit den bisherigen manuell adjustierten Stammbändern zu vergleichen.

LITERATURVERZEICHNIS

Blang, Dorothee/Helmert, Thomas. *Verwendung von Hot-Deck-Verfahren in der Außenhandelsstatistik.* In: Wirtschaft und Statistik. Ausgabe 11/2008, Seite 974 ff.

Breunig, Markus M./Kriegel, Hans-Peter/Ng, Raymond T./Sander, Jörg. *LOF: Identifying density-based local outliers*. In: ACM SIGMOD Record. Jahrgang 29. Ausgabe 2/2000, Seite 93 ff. DOI: 10.1145/335191.335388

Kruse, Hendrik W./Meyerhoff, Annette/Erbe, Anette. *Neue Methoden zur Mikrodatenverknüpfung von Außenhandels- und Unternehmensstatistiken*. In: WISTA Wirtschaft und Statistik. Ausgabe 5/2021, Seite 53 ff.

Liu, Fei Tony/Ting, Kai Ming/Zhou, Zhi-Hua. *Isolation Forest.* 2008 Eighth IEEE International Conference on Data Mining. DOI: 10.1109/ICDM.2008.17

Ester, Martin/Kriegel, Hans-Peter/Sander, Jörg/Xu, Xiaowei. *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise.* Association for the Advancement of Artificial Intelligence Press. 1996. KDD-96 Proceedings, Seite 226 ff.

Statistisches Bundesamt. *Warenverzeichnis für die Außenhandelsstatistik 2023*. 2022.

Tan, Pang-Ning/Steinbach, Michael/Kumar, Vipin. *Introduction to Data Mining*. First Edition. 2005.

Węglarczyk, Stanislaw. *Kernel density estimation and its application*. In: ITM Web of Conferences. Jahrgang 23. Artikel Nr. 37/2018. DOI: 10.1051/itmconf/20182300037

Winter, David K./Levandowsky, Michael. *Distance between Sets*. In: Nature. Ausgabe 234/1971, Seite 34 f. DOI: 10.1038/234034A0

Herausgeber
Statistisches Bundesamt (Destatis), Wiesbaden

Schriftleitung
Dr. Daniel Vorgrimler
Redaktion: Ellen Römer

Ihr Kontakt zu uns
www.destatis.de/kontakt

Erscheinungsfolge
zweimonatlich, erschienen im Februar 2024
Ältere Ausgaben finden Sie unter www.destatis.de sowie in der Statistischen Bibliothek.

Artikelnummer: 1010200-24001-4, ISSN 1619-2907