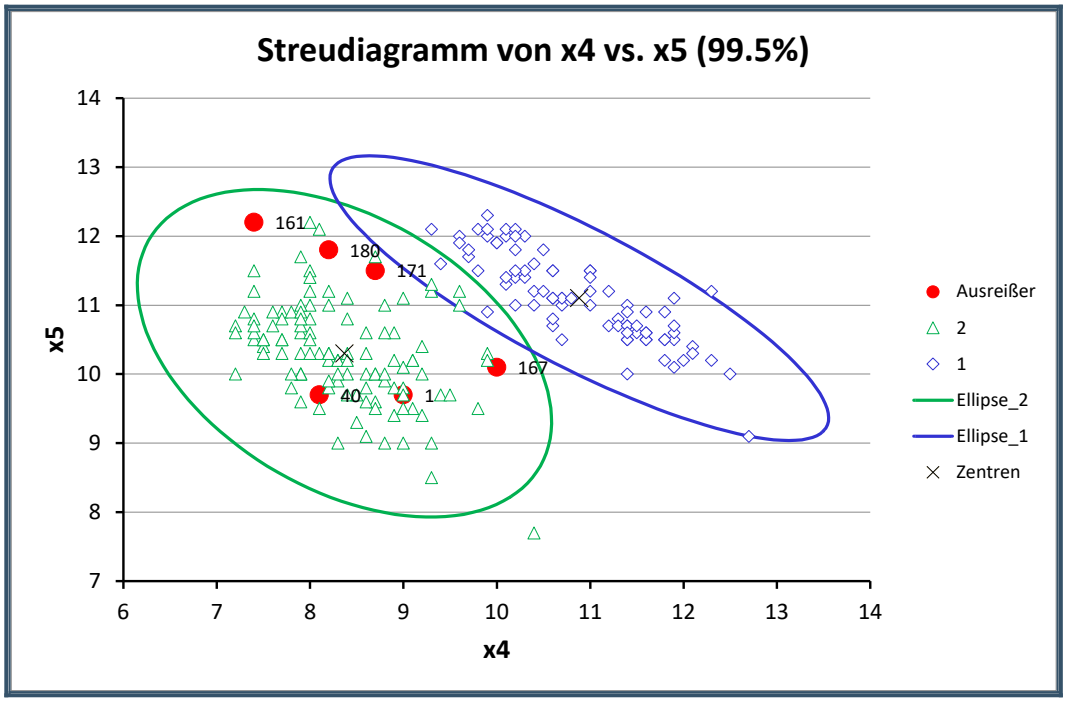
Clusteranalyse

mit intervallskalierten Variablen

mittels Ward-Algorithmus und Mahalanobis-Distanz



Eckehardt Spenhoff

Inhaltsverzeichnis

1 Clusteranalyse

Man unterscheidet in der Statistik zwischen strukturprüfenden und strukturbildenden Verfahren. Während strukturprüfende Verfahren bestehende Hypothesen überprüfen, dienen strukturbildende Verfahren der explorativen Analyse unbekannter Datenstrukturen. Die Clusteranalyse gehört zu den strukturbildenden Verfahren.

Weiterhin wird zwischen Clusterformation und der eigentlichen Clusteranalyse unterschieden.

- **Clusterformation:**
  Will man beispielsweise anhand von Körpermaßen Konfektionsgrößen festlegen, die für möglichst viele Kunden eine passende Auswahl ermöglichen, so ist die Clusterformation die geeignete Methode. In diesem Fall wird die Anzahl der Konfektionsgrößen (Cluster) vorgegeben und die Objekte werden diesen Gruppen möglichst optimal zugeordnet.
- **Clusteranalyse:**
  In der Clusteranalyse werden hingegen keine Gruppen vorgegeben. Ziel ist es, durch Bündelung der Objekte natürliche Gruppenstrukturen zu identifizieren. Dabei gilt, dass die Objekte innerhalb eines Clusters möglichst ähnlich sein sollen, während sich die Cluster untereinander möglichst deutlich unterscheiden. Typische Anwendungsfelder sind unter anderem die Segmentierung von Internetnutzern, die Kundensegmentierung oder die Bildung von Innovationstypen.

In vielen industriellen und sicherheitsrelevanten Anwendungen besteht die Aufgabe darin, anhand mehrerer kontinuierlicher Messgrößen unterschiedliche Objektklassen zu identifizieren. Ein klassisches Beispiel ist die Gruppierung von Banknoten auf Basis geometrischer Merkmale. Führt die Clusterbildung in diesem Fall zu einer klaren Trennung der Objekte, so nährt dies den Verdacht, dass die Unterschiede der Cluster auf echte und gefälschte Banknoten zurückgeführt werden können.

Die vorliegende Aufgabe ist typisch für eine Clusteranalyse mit intervallskalierten Variablen:

- mehrere metrische Merkmale
- unbekannte Klassenstruktur
- mögliche Ausreißer
- korrelierte Variablen

Ziel ist es, aus einer multivariaten Datenmatrix automatisch natürliche Gruppen zu identifizieren und diese statistisch korrekt zu validieren.

In diesem Kapitel wird eine vollständige Clusteranalyse mit dem Ward-Algorithmus und einer Verbesserung mittels k-Means-Methodik, der euklidischen Distanz, der Mahalanobis-Distanz, einer automatischen Bestimmung der optimalen Clusterzahl sowie einer statistischen Ausreißererkennung durchgeführt.

1.1 Voraussetzungen der Clusteranalyse

Wichtige Voraussetzungen, die bei der Durchführung einer Clusteranalyse beachtet werden sollten, sind:

- Die Analyse kann für unterschiedliche Datentypen (kategoriale und metrische Daten) genutzt werden. Hierzu wurden zahlreiche Ähnlichkeits- und Distanzmaße (Proximitätsmaße) definiert. OQM-Stat beschränkt sich bewusst auf metrische Daten mit den Distanzmaßen „euklidische Distanz“ und „Mahalanobis-Distanz“.
- Fehlende Werte und Ausreißer sollten vorab beseitigt werden, da sie die Analyseergebnisse deutlich verzerren können. OQM-Stat nutzt deshalb einen statistischen Ausreißertest, bei dem Ausreißer bei der Clusterbildung nicht berücksichtigt werden. Die Ausreißer können jedoch weiterhin dargestellt und ihr Ursprung beurteilt werden.
- Weisen die verwendeten Variablen große Unterschiede bezüglich ihres Wertebereichs auf, so sollten diese auf ein einheitliches Niveau transformiert werden. Die Mahalanobis-Distanz berücksichtigt dies automatisch durch die sogenannte Whitening-Transformation.

Bei der Berechnung der Cluster wird nach bestimmten Regeln entschieden, wie die Objekte zu Gruppen zusammengefasst werden. Das Ergebnis dieses Prozesses hängt nicht nur von der Wahl des Clustering-Algorithmus ab, sondern auch davon, wie die Distanzen zwischen den Objekten bestimmt werden.

Deshalb werden in OQM-Stat ausschließlich der Ward-Algorithmus und ein partitionierendes k-Means-Verfahren zur Verfeinerung der Gruppenzugehörigkeit eingesetzt.

Die hier eingesetzte Methodik setzt voraus:

- metrisch skalierte Variablen
- sinnvolle Abstandsdefinition
- Mittelwerte und Varianzen sind interpretierbar
- Korrelationen zwischen Variablen möglich

Die Daten liegen auf Intervallskalenniveau vor und erfüllen damit die Voraussetzungen für eine multivariate Clusteranalyse.

1.2 Distanzmaße in der Clusteranalyse

Ein zentrales Element jeder Clusteranalyse ist die Definition eines geeigneten Distanzmaßes. Das Distanzmaß bestimmt, wie ähnlich oder unähnlich sich zwei Objekte im Merkmalsraum sind. Die Qualität der resultierenden Cluster hängt wesentlich von der Wahl dieses Maßes ab.

In der Literatur wurden zahlreiche Proximitätsmaße für unterschiedliche Skalenniveaus entwickelt. Für metrische Daten haben sich insbesondere die euklidische Distanz und die Mahalanobis-Distanz etabliert.

OQM-Stat beschränkt sich bewusst auf diese beiden Distanzmaße, da sie für intervallskalierte Daten mathematisch fundiert, numerisch stabil und in der industriellen Praxis bewährt sind.

1.2.1 Euklidische Distanz

Die euklidische Distanz ist das klassische geometrische Distanzmaß im p-dimensionalen Raum. Für zwei Beobachtungsvektoren

$$x_i = \left(x_{i1}, x_{i2}, \ldots, x_{ip}\right), \quad x_j = \left(x_{j1}, x_{j2}, \ldots, x_{jp}\right)$$

lautet sie:

$$d_E\left(x_i, x_j\right) = \sqrt{\sum_{k=1}^{p}\left(x_{ik} - x_{jk}\right)^2}$$

Geometrisch entspricht diese Distanz der Länge der Verbindungsstrecke zwischen zwei Punkten im p-dimensionalen Merkmalsraum. Die euklidische Distanz besitzt folgende Eigenschaften:

- sie ist nicht negativ:

$$d_E\left(x_i, x_j\right) \geq 0$$

- sie ist symmetrisch:

$$d_E\left(x_i, x_j\right) \geq 0$$

- sie erfüllt die Dreiecksungleichung
- sie ist invariant gegenüber Translationen

Trotz ihrer einfachen geometrischen Interpretation weist die euklidische Distanz einige Schwächen auf, die in der Praxis beachtet werden müssen:

- **Unterschiedliche Skalen der Variablen**
  Variablen mit großem Wertebereich dominieren die Distanzberechnung.
- **Korrelationen zwischen Variablen**
  Stark korrelierte Variablen gehen mehrfach in die Distanz ein und verfälschen damit die tatsächliche Struktur.
- **Unterschiedliche Varianzen**
  Variablen mit hoher Streuung erhalten ein größeres Gewicht als solche mit geringer Streuung.

Diese Effekte können zu verzerrten Clusterstrukturen führen, insbesondere bei technisch oder physikalisch gemessenen Größen.

1.2.2 Standardisierung und Whitening

Um die genannten Nachteile zu kompensieren, werden in der Praxis häufig Transformationsverfahren eingesetzt. Eine einfache Möglichkeit ist die Standardisierung jeder Variablen:

$$z_{ik} = \frac{x_{ik} - \mu_k}{\sigma_k}$$

mit μ_k = Mittelwert der k-ten Variablen
σ_k = Standardabweichung der k-ten Variablen

Diese Transformation führt jedoch nur zu einer Skalierung, berücksichtigt aber keine Korrelationen zwischen den Variablen. Eine vollständige Lösung bietet das sogenannte **Whitening**, bei dem zusätzlich die Kovarianzstruktur eliminiert wird.

1.2.3 Mahalanobis-Distanz

Die Mahalanobis-Distanz wurde von P. C. Mahalanobis eingeführt und stellt eine kovarianzgewichtete Distanz dar. Sie berücksichtigt sowohl die Skalierung als auch die Korrelationen der Variablen.

Für zwei Beobachtungen xi und xj lautet sie:

$$d_M\left(x_i, x_j\right) = \sqrt{\left(x_i - x_j\right)^T \Sigma^{-1} \left(x_i - x_j\right)}$$

mit Σ = Kovarianzmatrix der Daten.

Die Mahalanobis-Distanz misst die Distanz zweier Punkte relativ zur Streuung der Datenwolke. Punkte entlang einer stark gestreckten Hauptachse gelten als näher beieinander als Punkte mit gleicher euklidischer Distanz in Richtung geringer Varianz. Geometrisch beschreibt die Mahalanobis-Distanz Ellipsen (bzw. Hyperellipsoide) konstanter Dichte im Merkmalsraum.

1.2.4 Whitening-Transformation

Die Mahalanobis-Distanz kann durch eine lineare Transformation auf eine euklidische Distanz im transformierten Raum zurückgeführt werden. Sei die Kovarianzmatrix gegeben durch die Cholesky-Zerlegung:

$$\Sigma = LL^T$$

Dann gilt für den transformierten Vektor:

$$z = L^{-1}\left(x - \mu\right)$$

mit μ = Mittelwertvektor.

Im transformierten Raum gilt dann:

$$d_M\left(x_i, x_j\right) = \left\| z_i - z_j \right\|$$

und für einen Punkt:

$$MD^2 = (x-\mu)^T \Sigma^{-1} (x-\mu) = \|z\|^2$$

Die Mahalanobis-Distanz entspricht somit der euklidischen Norm im whitened space. Die Mahalanobis-Distanz bietet entscheidende Vorteile gegenüber der euklidischen Distanz:

- automatische Skalierung aller Variablen
- vollständige Berücksichtigung von Korrelationen
- physikalisch sinnvolle Gewichtung
- objektive Ausreißerdefinition
- direkte Verbindung zur multivariaten Normalverteilung

Sie ist daher das bevorzugte Distanzmaß für multivariate Qualitäts-, Mess- und Prozessdaten.

1.2.5 Distanzmaß in OQM-Stat

OQM-Stat unterstützt zwei Betriebsarten:

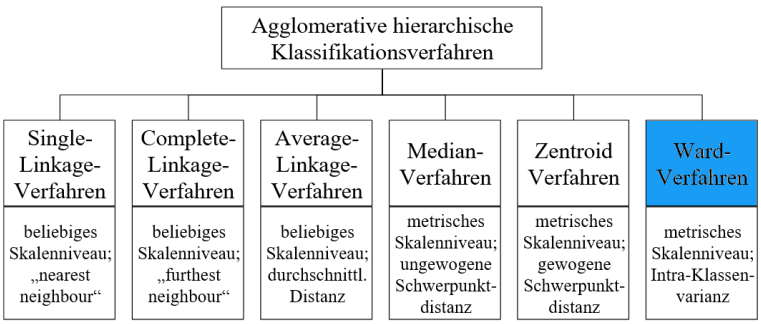
- **Euklidischer Modus:**
  klassische Ward-Clusterung im Originalraum
- **Mahalanobis-Modus:**
- Ward-Clusterung im whitened space (entspricht Mahalanobis-Distanzen im Originalraum)

Die Whitening-Transformation erfolgt automatisch auf Basis der Stichprobenkovarianzmatrix. Optional kann eine robuste Kovarianzschätzung verwendet werden. Damit wird sichergestellt, dass sowohl skalierungs- als auch korrelationsbedingte Verzerrungen ausgeschlossen werden.

1.3 Der Ward-Algorithmus (hierarchische Clusteranalyse)

Der Ward-Algorithmus ist ein **hierarchisches, agglomeratives Clustering-Verfahren** und gehört zu den am häufigsten eingesetzten Methoden der Clusteranalyse für metrische Daten. Er zeichnet sich durch eine besonders stabile und kompakte Clusterbildung aus und wird daher häufig in industriellen und technischen Anwendungen eingesetzt.

Im Gegensatz zu partitionierenden Verfahren, bei denen die Anzahl der Cluster vorab festgelegt werden muss, erzeugt der Ward-Algorithmus eine vollständige Hierarchie der Datenstruktur. Dadurch steht dem Anwender die gesamte Fusionshistorie zur Verfügung, und die optimale Clusterzahl kann im Nachhinein bestimmt werden.



1.3.1 Grundprinzip des Ward-Verfahrens

Der Ward-Algorithmus ist ein agglomeratives Verfahren, das schrittweise ausgehend von Einzelelementen größere Cluster bildet:

- Zu Beginn bildet jede Beobachtung ein eigenes Cluster.
- In jedem Schritt werden genau die beiden Cluster fusioniert, deren Vereinigung den geringsten Zuwachs der Gesamtstreuung verursacht.
- Dieser Prozess wird fortgesetzt, bis alle Objekte in einem einzigen Cluster vereinigt sind.

Das Verfahren minimiert in jedem Fusionsschritt den Anstieg der Within-Cluster-Streuquadratsumme (SSE). Die Gesamtstreuung der Daten lässt sich zerlegen in:

$$T = W + B$$

mit T = Gesamtstreuung (Total Sum of Squares)
W = Within-Cluster-Streuung
B = Between-Cluster-Streuung

Die Gesamtstreuung lautet:

$$T = \sum_{i=1}^{n} \left\| xi - \mu \right\|^2$$

Die Within-Cluster-Streuung ist:

$$W = \sum_{c=1}^{k} \sum_{x_i \in C_c} \left\| x_i - \mu_c \right\|^2$$

und die Between-Cluster-Streuung:

$$B = \sum_{c=1}^{k} nc \left\| \mu_c - \mu \right\|^2$$

mit k = Anzahl der Cluster
n_c = Größe des Clusters
μ_c = Schwerpunkt des Clusters
μ = Gesamtschwerpunkt

1.3.2 Fusionskriterium nach Ward

Beim Ward-Verfahren wird in jedem Schritt jene Fusion gewählt, die den geringsten Anstieg der Within-Cluster-Streuung verursacht. Für zwei Cluster C_a und C_b ergibt sich der Streuungszuwachs:

$$\Delta W_{ab} = \frac{n_a n_b}{n_a + n_b} \cdot \left\| \mu_a - \mu_b \right\|^2$$

mit na, nb = Größen der beiden Cluster
μa, μb = deren Zentren

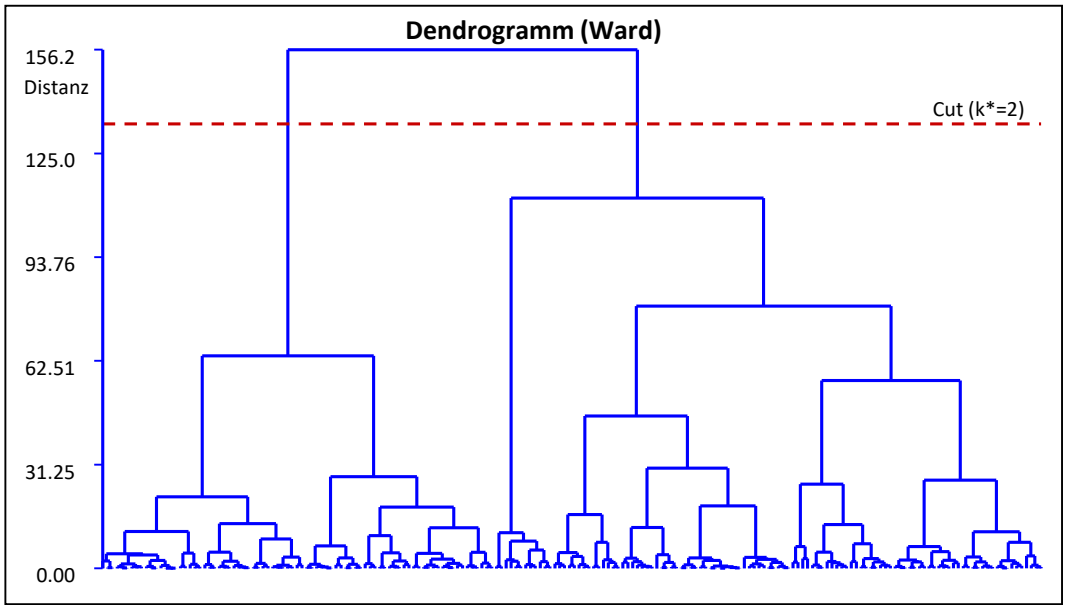
Diese Formel zeigt, dass sowohl der Abstand der Clusterzentren als auch deren Größen in die Fusionsentscheidung eingehen. Geometrisch bedeutet das Ward-Kriterium:

- Es werden bevorzugt Cluster mit naheliegenden Schwerpunkten fusioniert.
- Große Cluster werden stärker gewichtet als kleine.
- Die entstehenden Cluster sind möglichst kompakt und kugelförmig.

Im Mahalanobis-Modus erfolgt die Berechnung im whitened space, sodass die Clusterbildung kovarianzbereinigt erfolgt.

1.3.3 Dendrogramm

Das Ergebnis des Ward-Verfahrens ist ein Dendrogramm, das die gesamte Fusionshierarchie grafisch darstellt.



- Auf der horizontalen Achse stehen die Objekte oder Cluster.
- Auf der vertikalen Achse ist der Fusionsabstand (Streuungszuwachs) dargestellt.
- Jeder Knoten repräsentiert eine Fusion zweier Cluster.

Durch das Setzen einer horizontalen Schnittlinie im Dendrogramm kann eine gewünschte Clusteranzahl bestimmt werden.

1.3.4 Vorteile des Ward-Algorithmus

Der Ward-Algorithmus besitzt mehrere Vorteile, die ihn für industrielle Anwendungen besonders geeignet machen:

- Er liefert sehr stabile und reproduzierbare Cluster.
- Die Cluster sind kompakt und gut interpretierbar.
- Die Methode ist robust gegenüber Rauschen.
- Die gesamte Hierarchie bleibt verfügbar.
- Keine Vorab-Festlegung der Clusterzahl erforderlich.

Dadurch eignet sich der Ward-Algorithmus hervorragend für explorative Analysen komplexer Messdaten.

Bezüglich des Ward-Verfahrens sei zudem darauf hingewiesen, dass eine Untersuchung von S. Berg (1981) gezeigt hat, dass das Ward-Verfahren im Vergleich zu anderen Fusionsalgorithmen in den meisten Fällen sehr gute Partitionen liefert und die Objekte mit hoher Zuverlässigkeit den „richtigen“ Gruppen zuordnet. Das Ward-Verfahren kann somit als ein sehr leistungsfähiger und robuster Fusionsalgorithmus angesehen werden.

1.3.5 Ward-Algorithmus in OQM-Stat

OQM-Stat implementiert den Ward-Algorithmus vollständig numerisch stabil und effizient. Je nach gewähltem Distanzmaß arbeitet das Verfahren:

- im Originalraum mit euklidischer Distanz
- im whitened space mit Mahalanobis-Distanz

Dabei werden folgende Schritte automatisch durchgeführt:

- Berechnung der Distanzmatrix
- Hierarchische Fusion nach Ward
- Aufbau der vollständigen Hierarchie
- Erzeugung des Dendrogramms
- Speicherung aller Fusionsschritte

Optional kann auf Basis der Ward-Zentren ein k-Means-Verfahren zur Verfeinerung der Clusterzuordnung gestartet werden. Das Ward-Verfahren verbindet die Vorteile hierarchischer Verfahren mit einer klaren statistischen Interpretation. Es stellt damit eine ideale Grundlage für eine objektive Clusterbildung dar. In Verbindung mit der Mahalanobis-Distanz entsteht ein Verfahren, das sowohl geometrisch als auch statistisch optimal an die Struktur der Daten angepasst ist.

1.4 Bestimmung der optimalen Clusterzahl

Ein zentrales Problem der Clusteranalyse besteht in der Bestimmung der „richtigen“ Anzahl von Clustern. Während hierarchische Verfahren wie der Ward-Algorithmus eine vollständige Hierarchie erzeugen, liefern sie zunächst keine eindeutige Entscheidung über die optimale Clusterzahl.

Um aus der Hierarchie eine sinnvolle Partition abzuleiten, müssen zusätzliche Kriterien herangezogen werden, welche die Trennschärfe der Cluster bewerten. In OQM-Stat wird hierzu das Calinski-Harabasz-Kriterium (CH-Kriterium) eingesetzt.

1.4.1 Grundidee der Clusterzahlbestimmung

Die Clusteranalyse ist ein exploratives Verfahren. Im Gegensatz zu klassischen Hypothesentests existiert keine a-priori bekannte Gruppenstruktur, die überprüft werden könnte. Stattdessen wird versucht, aus den Daten selbst eine natürliche Gruppierung zu extrahieren.

Dabei steht man vor der grundlegenden Frage:

- Wie viele Cluster sind in den Daten tatsächlich enthalten?

Eine triviale Lösung wäre, jedes Objekt als eigenes Cluster zu betrachten. Ebenso trivial wäre ein einziges Cluster für alle Objekte. Beide Lösungen sind jedoch inhaltlich bedeutungslos. Gesucht ist daher eine Clusterzahl, bei der:

- die Objekte innerhalb der Cluster möglichst ähnlich sind.
- sich die Cluster untereinander möglichst deutlich unterscheiden.

Wie bereits gezeigt, lässt sich die Gesamtstreuung der Daten zerlegen in:

$$T = W_k + B_k$$

mit T = Gesamtstreuung
W_k = Streuung innerhalb der Cluster für k Cluster
B_k = Streuung zwischen den Clustern für k Cluster

Mit wachsender Clusterzahl nimmt die Streuung innerhalb der Cluster ab, während die Streuung zwischen den Clustern zunimmt. Bei sehr vielen Clustern wird W_k sehr klein, jedoch verliert die Clusterlösung dann ihre interpretierbare Bedeutung. Ziel ist es daher, einen Kompromiss zwischen guter Trennung und sinnvoller Gruppierung zu finden.

1.4.2 Calinski-Harabasz-Kriterium

Das Calinski-Harabasz-Kriterium (auch Varianzquotient genannt) wurde 1974 von Calinski und Harabasz vorgeschlagen und gehört zu den am häufigsten eingesetzten Kriterien zur Bestimmung der optimalen Clusterzahl.

Es ist definiert als:

$$CH(k) = \frac{B_k \,/\, (k-1)}{W_k \,/\, (n-k)}$$

mit k = Anzahl der Cluster
n = Anzahl der Objekte
B_k = Streuung zwischen den Clustern
W_k = Streuung innerhalb der Cluster

Das Kriterium ist somit ein normierter Quotient aus:

- mittlerer Streuung zwischen den Clustern pro Freiheitsgrad
- mittlerer Streuung innerhalb der Cluster pro Freiheitsgrad

Ein hoher CH-Wert bedeutet:

- große Trennung zwischen den Clustern
- geringe Streuung innerhalb der Cluster

Das Maximum von $CH(k)$ wird als optimale Clusterzahl k^* interpretiert:

$$k^* = arg \max_k CH(k)$$

Die benötigten Größen lauten:

Gesamtmittelwert:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Clusterzentren:

$$\mu_c = \frac{1}{n_c}\sum_{x_i \in C_c} x_i$$

Within-Cluster-Streuung:

$$W_k = \sum_{c=1}^{k}\sum_{x_i \in C_c} \left\| x_i - \mu_c \right\|^2$$

Between-Cluster-Streuung:

$$B_k = \sum_{c=1}^{k} n_c \left\| \mu_c - \mu \right\|^2$$

Diese Zerlegung ist identisch zur Varianzanalyse und besitzt eine klare statistische Interpretation.

1.4.3 CH-Kriterium in OQM-Stat

OQM-Stat berechnet das CH-Kriterium für alle Clusterzahlen im Bereich:

$$k_{\min} \le k \le k_{\max}$$

Der Anwender kann diesen Bereich explizit vorgeben. Standardmäßig wird ein sinnvoller Bereich gewählt, der von der Stichprobengröße abhängt. Die CH-Werte werden tabellarisch ausgegeben und zusätzlich grafisch dargestellt. Das Maximum markiert die empfohlene Clusterzahl. Optional kann der Anwender das CH-Kriterium deaktivieren und die Clusterzahl manuell vorgeben. Wie jedes streuungsbasierte Kriterium besitzt auch das CH-Kriterium Einschränkungen, insbesondere bei kleinen Stichprobenumfängen.

Problematisch wird das CH-Kriterium insbesondere bei:

- sehr kleinen Stichproben (z. B. n < 30)
- geringer Dimensionalität
- schwacher oder kontinuierlicher Struktur
- stark schiefen Verteilungen
- Vorliegen eines kontinuierlichen Gradienten anstelle echter Gruppen

In diesen Fällen kann es zu folgenden Effekten kommen:

- flache CH-Kurven ohne ausgeprägtes Maximum
- instabile Maxima bei zufälligen Schwankungen

- überoptimistische Bewertung kleiner Clusterzahlen
- numerische Instabilitäten bei sehr kleinen W_k

Beispiel: Berufsdatensatz

Beruf	Einkommen	Markenbewusstsein
Arzt	6861	21765
Ingenieur	5150	28245
Chemiker	5474	25179
Manager	7389	19048
Professor	5152	24608
CEO	12810	27611
Anwalt	7203	21536
Koch	4162	24823
Architekt	6779	22499
Forstwart	3204	7465
Physiker ETH	5335	17471
Lehrer	4311	14735
Bauarbeiter	3949	17921
Fischer	2132	8822
Servicemitarbeiter	3018	12201

Der Datensatz mit 15 Berufen, Einkommen und Markenbewusstsein stellt ein typisches Grenzbeispiel dar:

- sehr kleine Stichprobe
- nur zwei Dimensionen
- eher kontinuierlicher sozioökonomischer Gradient
- keine natürliche Gruppierung

In solchen Fällen existieren keine echten Cluster im statistischen Sinne. Das CH-Kriterium liefert zwar mathematische Werte, diese besitzen jedoch keine inhaltlich belastbare Interpretation. Die Clusteranalyse erzeugt hier lediglich eine künstliche Segmentierung eines kontinuierlichen Merkmalsraums.

Bei kleinen Datensätzen sollte das CH-Kriterium daher nur unterstützend verwendet werden. In diesen Fällen ist eine inhaltliche Interpretation zwingend erforderlich. Empfohlen wird:

- Begrenzung des Suchbereichs auf wenige Cluster (z. B. k = 2 oder k = 3)
- grafische Analyse mittels Scatterplots
- Beurteilung der Trennschärfe
- Plausibilitätsprüfung anhand fachlicher Kriterien

Die Clusteranalyse wird hier zu einem explorativen Hilfsmittel und nicht zu einem automatischen Klassifikationswerkzeug.

1.4.4 Einordnung des CH-Kriteriums

In der Literatur wurden eine Vielzahl statistischer Kriterien entwickelt, die unter dem Begriff der sogenannten Stopping Rules zusammengefasst werden. Diese liefern statistische und damit weitgehend objektive Anhaltspunkte zur Bestimmung der optimalen Clusterzahl bei Anwendung hierarchi-

scher Clusterverfahren. Ziel dieser Stopping Rules ist es, aus der Fusionshierarchie jene Clusterzahl zu bestimmen, die der „wahren“ Gruppenstruktur der Daten möglichst nahekommt.

Im Rahmen einer umfangreichen Simulationsstudie untersuchten Milligan und Cooper (1985) insgesamt 30 solcher Stopping Rules. Die Autoren generierten Datensätze mit unterschiedlich trennscharfen Clusterstrukturen (mit 2 bis 5 Clustern) und testeten anschließend, inwieweit verschiedene hierarchische Verfahren – darunter Single-Linkage, Complete-Linkage, Average-Linkage und Ward – in der Lage waren, die vorgegebene wahre Gruppenzahl korrekt zu identifizieren.

Die Auswertung zeigte, dass das Kriterium von Calinski und Harabasz unter allen untersuchten Stopping Rules die beste Leistungsfähigkeit aufwies. In über 90 % der untersuchten Fälle konnte mit dem Calinski-Harabasz-Kriterium die wahre Clusterstruktur korrekt erkannt werden.

Damit besitzt das CH-Kriterium eine hervorragende empirische Absicherung und kann als eines der leistungsfähigsten und zuverlässigsten Verfahren zur Bestimmung der optimalen Clusterzahl bei hierarchischen Clusteranalysen angesehen werden.Das Calinski-Harabasz-Kriterium ist ein leistungsfähiges Werkzeug zur automatischen Bestimmung der Clusterzahl bei ausreichend großen und strukturierten Datensätzen. In industriellen Anwendungen mit mehreren hundert oder tausend Beobachtungen liefert es in der Regel sehr stabile und reproduzierbare Ergebnisse.

Bei kleinen Stichproben ersetzt es jedoch nicht die fachliche Interpretation und sollte stets im Zusammenhang mit grafischen Darstellungen und inhaltlicher Expertise betrachtet werden.

1.5 Statistische Ausreißererkennung mit Mahalanobis-Distanz

In multivariaten Datensätzen treten häufig Beobachtungen auf, die nicht zur eigentlichen Datenstruktur gehören. Solche Ausreißer können unterschiedliche Ursachen haben:

- Messfehler
- Eingabefehler
- fehlerhafte Sensorik
- besondere Prozesszustände
- echte Sonderfälle

Ausreißer können die Clusterbildung erheblich verzerren und führen häufig zu instabilen oder verfälschten Clusterlösungen. Eine statistisch fundierte Ausreißererkennung ist daher ein zentraler Bestandteil jeder robusten Clusteranalyse.

In univariaten Analysen werden Ausreißer meist über z-Werte oder Boxplot-Kriterien identifiziert. Diese Methoden berücksichtigen jedoch nur eine einzelne Variable.

In multivariaten Datensätzen können jedoch Beobachtungen auftreten, die in keiner einzelnen Variable auffällig sind, aber im Merkmalsraum dennoch weit von der Datenwolke entfernt liegen. Solche Punkte werden als multivariate Ausreißer bezeichnet.

Zur Identifikation multivariater Ausreißer ist daher ein Distanzmaß erforderlich, das die gesamte Kovarianzstruktur der Daten berücksichtigt. Das geeignete Distanzmaß zur Identifikation multivariater Ausreißer ist die Mahalanobis-Distanz:

$$MD^2(x) = (x - \mu)^T \Sigma^{-1} (x - \mu)$$

mit x = Beobachtungsvektor
μ = Mittelwertvektor
Σ = Kovarianzmatrix

Die Mahalanobis-Distanz misst die normierte Entfernung eines Punktes vom Zentrum der Datenwolke. Unter der Annahme einer multivariaten Normalverteilung gilt:

$$MD^2 = \chi_p^2$$

mit p = Anzahl der Dimensionen.

Damit lässt sich ein objektiver statistischer Ausreißertest formulieren:

Ein Punkt gilt als Ausreißer, wenn

$$MD^2 > \chi_{\alpha,p}^2$$

wobei α = Konfidenzniveau (z. B. 0.99)

Wie bereits gezeigt, kann die Mahalanobis-Distanz über eine Whitening-Transformation auf eine euklidische Norm im transformierten Raum zurückgeführt werden:

$$z = L^{-1}(x - \mu) \text{ mit } \sum = LL^T$$

Dann gilt:

$$MD^2 = \|z\|^2$$

Im whitened space liegen die Daten kugelförmig verteilt um den Ursprung. Ausreißer erscheinen als Punkte mit besonders großem Abstand zum Zentrum.

1.5.1 Ausreißerbehandlung in OQM-Stat

OQM-Stat verwendet standardmäßig eine Mahalanobis-basierte Ausreißererkennung mit Chi²-Schwellenwert. Das Verfahren umfasst:

- Berechnung der Kovarianzmatrix
- Whitening-Transformation
- Berechnung der Mahalanobis-Distanzen
- Vergleich mit Chi²-Quantil
- Markierung statistischer Ausreißer

Ausreißer werden bei der Clusterbildung ausgeschlossen, können jedoch weiterhin grafisch dargestellt und in der Analyse ausgewertet werden.

Die Trennung von regulären Beobachtungen und Ausreißern ist entscheidend für die Qualität der Clusterlösung:

- Ausreißer verzerren Zentren
- sie beeinflussen die Distanzmatrix
- sie führen zu instabilen Fusionen
- sie können künstliche Cluster erzeugen

Durch die statistische Ausreißererkennung wird sichergestellt, dass die Clusteranalyse auf einer stabilen, repräsentativen Datenbasis erfolgt. Die Chi²-basierte Ausreißererkennung setzt voraus:

- annähernde Multinormalität
- ausreichende Stichprobengröße
- stabile Kovarianzmatrix

Bei sehr kleinen Stichproben kann die Kovarianzschätzung instabil werden. In diesen Fällen sollte der Ausreißertest nur als Orientierungshilfe verwendet und stets grafisch überprüft werden.

1.6 Refinement der Clusterzuordnung mittels k-Means-Verfahren

Der Ward-Algorithmus liefert eine hierarchische Struktur der Daten und erzeugt kompakte, gut interpretierbare Cluster. Das Ergebnis basiert jedoch auf einer schrittweisen Fusion und ist damit nicht direkt auf eine globale Optimierung der Clusterzuordnung ausgerichtet.

Zur weiteren Verfeinerung der Gruppenzugehörigkeit kann daher ein partitionierendes Verfahren eingesetzt werden, das die Clusterzentren iterativ optimiert. Hierzu hat sich insbesondere das k-Means-Verfahren bewährt.

In OQM-Stat wird das k-Means-Verfahren optional als Refinement-Stufe eingesetzt, wobei die vom Ward-Algorithmus ermittelten Clusterzentren als Startwerte verwendet werden.

1.6.1 Grundprinzip des k-Means-Verfahrens

Das k-Means-Verfahren ist ein partitionierendes Clustering-Verfahren, bei dem die Anzahl der Cluster k vorab festgelegt wird. Ziel ist es, die Objekte so auf die Cluster zu verteilen, dass die Summe der quadratischen Abstände zu den jeweiligen Clusterzentren minimal wird. Gegeben sei eine Menge von Beobachtungen

$$x_1, x_2, \ldots, x_n \in \mathbb{R}^p$$

Gesucht ist eine Partition in k Cluster $C_1, \ldots, C_k$ mit Zentren $\mu_1, \ldots, \mu_k$, sodass folgende Zielfunktion minimiert wird:

$$SSE = \sum_{c=1}^{k} \sum_{x_i \in C_c} \left\| x_i - \mu_c \right\|^2$$

Diese Größe wird als Within-Cluster-Streuquadratsumme bezeichnet. Das klassische k-Means-Verfahren ein **Iteratives Optimierungsverfahren** besteht aus den folgenden Schritten:

- Initialisierung der Clusterzentren $\mu_1, \ldots, \mu_k$
- Zuordnung jedes Objekts zum nächstgelegenen Zentrum
- Neuberechnung der Clusterzentren als Mittelwerte der zugeordneten Punkte
- Wiederholung der Schritte 2 und 3 bis zur Konvergenz

Der Algorithmus konvergiert zu einem lokalen Minimum der Zielfunktion.

1.6.2 Initialisierung durch Ward-Zentren

Ein bekanntes Problem des k-Means-Verfahrens besteht in seiner Abhängigkeit von der Initialisierung. Ungünstige Startwerte können zu schlechten lokalen Minima führen. Um dieses Problem zu vermeiden, verwendet OQM-Stat die vom Ward-Algorithmus ermittelten Clusterzentren als Initialisierung. Diese Vorgehensweise besitzt mehrere Vorteile:

- sehr stabile Startwerte
- keine zufällige Initialisierung
- reproduzierbare Ergebnisse
- schnelle Konvergenz
- globale Struktur bereits berücksichtigt

Damit verbindet sich die hierarchische Struktur des Ward-Verfahrens mit der globalen Optimierung des k-Means-Verfahrens. Im euklidischen Modus basiert die Zuordnung auf der euklidischen Distanz:

$$d_E\left(x, \mu_c\right) = \left\| x - \mu_c \right\|$$

Im Mahalanobis-Modus erfolgt die Zuordnung im whitened space:

$$z = L^{-1} \left\| x - \mu \right\|$$

$$d_M\left(x, \mu_c\right) = \left\| z - z_c \right\|$$

Damit bleibt die kovarianzgewichtete Geometrie auch im k-Means-Refinement vollständig erhalten. Der k-Means-Algorithmus wird iteriert, bis sich die Zuordnung der Objekte nicht mehr ändert oder der Rückgang der Zielfunktion unter eine vorgegebene Schwelle fällt.

In OQM-Stat wird zusätzlich eine maximale Iterationszahl vorgegeben, um eine garantierte Terminierung sicherzustellen.

1.6.3 Ergebnis des Refinements

Das Ergebnis des k-Means-Refinements ist eine stabile Partition der Daten mit:

- finalen Clusterzentren
- finaler Gruppenzuordnung (NeuCluster)
- minimaler Within-Cluster-Streuung

Diese Partition bildet die Grundlage für:

- grafische Darstellung
- Konfidenzellipsen
- statistische Tests
- weitere multivariate Analysen

Das k-Means-Refinement ist kein Ersatz für den Ward-Algorithmus, sondern eine sinnvolle Ergänzung. Während Ward die globale Struktur der Daten erfasst, sorgt k-Means für eine lokale Optimierung der Gruppenzuordnung. In Kombination entsteht ein sehr leistungsfähiges Verfahren, das sowohl explorativ als auch statistisch fundiert ist. Diese Kombination aus hierarchischer und partitionierender Clusteranalyse wird in der Literatur häufig als Best Practice empfohlen.

Auch das k-Means-Verfahren besitzt Einschränkungen:

- es setzt konvexe Clusterformen voraus
- es ist empfindlich gegenüber Ausreißern
- es minimiert nur eine quadratische Zielfunktion
- es liefert nur lokale Optima

Durch die vorgeschaltete Ward-Analyse und die statistische Ausreißererkennung werden diese Einschränkungen in OQM-Stat jedoch weitgehend kompensiert.

1.7 Grafische Analyse der Clusterstruktur

Die numerische Bestimmung von Clusterstrukturen liefert eine objektive und reproduzierbare Gruppenzuordnung. Für die Interpretation der Ergebnisse ist jedoch eine grafische Darstellung unverzichtbar. Erst durch geeignete Visualisierungen wird die geometrische Struktur der Daten, die Lage der Clusterzentren, die Streuung innerhalb der Cluster sowie mögliche Überlappungen sichtbar. Die grafische Analyse erfüllt dabei mehrere Funktionen:

- Plausibilitätskontrolle der numerischen Clusterlösung
- visuelle Beurteilung der Trennschärfe
- Identifikation möglicher Grenzfälle
- Erkennung verbliebener Ausreißer
- Interpretation der Clustergeometrie

In OQM-Stat werden hierfür interaktive 2D-Streudiagramme mit Zentren, Ausreißern und Konfidenzellipsen bereitgestellt.

1.7.1 Streudiagramme im Merkmalsraum

Für jede Variablenkombination xi, xj können zweidimensionale Streudiagramme erzeugt werden. Jeder Punkt repräsentiert eine Beobachtung, farblich kodiert nach Clusterzugehörigkeit. Die Streudiagramme ermöglichen eine direkte visuelle Beurteilung:

- der Clusterlage im Merkmalsraum
- der internen Streuung
- der Überlappung zwischen Clustern
- der relativen Clustergröße

Damit wird sichtbar, ob die numerische Clusterlösung geometrisch sinnvoll ist. Zusätzlich zu den Einzelpunkten werden die Clusterzentren in den Streudiagrammen dargestellt. Die Zentren entsprechen den Mittelwertvektoren der Cluster:

$$\mu_c = \frac{1}{n_c} \sum_{x_i \in C_c} x_i$$

Die Darstellung der Zentren erlaubt:

- eine schnelle Orientierung im Merkmalsraum
- den Vergleich der relativen Lage der Cluster
- die Beurteilung der Trennung entlang einzelner Dimensionen

Die Zentren dienen zudem als Referenzpunkte für die Konfidenzellipsen.

1.7.2 Konfidenzellipsen der Cluster

Zur Visualisierung der Streuung innerhalb eines Clusters werden Konfidenzellipsen dargestellt. Diese basieren auf der zweidimensionalen Kovarianzmatrix des jeweiligen Clusters. Für ein Cluster mit Mittelwert μ und Kovarianzmatrix Σ gilt für die Ellipse:

$$(x-\mu)^T \Sigma^{-1} (x-\mu) = \chi^2_{\alpha,2}$$

Dabei ist α das gewählte Konfidenzniveau (z. B. 95 %, 99 % oder 99.5 %). Alle Punkte innerhalb der Ellipse gehören mit Wahrscheinlichkeit α zur multivariaten Normalverteilung des Clusters. Die Ellipsen liefern wichtige geometrische Informationen:

- Größe der Ellipse → Streuung des Clusters
- Orientierung → Korrelation der Variablen
- Lage → Clusterzentrum
- Überlappung → Trennschärfe der Cluster

Stark überlappende Ellipsen deuten auf schlecht trennbare Cluster hin, während klar getrennte Ellipsen eine stabile Clusterstruktur anzeigen.

Wird die Clusteranalyse mit Mahalanobis-Distanzen durchgeführt, so erfolgt die eigentliche Clusterbildung im sogenannten whitened space. In diesem Raum sind alle Variablen standardisiert und entkorreliert, sodass die euklidische Distanz der Mahalanobis-Distanz im Originalraum entspricht. Die Whitening-Transformation lautet:

$$z = L^{-1}(x-\mu)$$

mit

$$\Sigma = LL^T$$

wobei μ der Mittelwertvektor,
Σ die Kovarianzmatrix und
L die untere Cholesky-Zerlegung von Σ ist.

Die Clusterbildung, die Distanzberechnung sowie die Ausreißererkennung erfolgen vollständig im z-Raum. Für die grafische Darstellung der Ergebnisse ist jedoch der Originalraum entscheidend. Nur dort besitzen die Achsen eine inhaltliche Bedeutung (z. B. Länge, Breite, Randabstand etc.). Daher müssen für die Visualisierung:

- die Clusterzentren
- die Konfidenzellipsen
- die Distanzgeometrie

aus dem Whitening-Raum wieder in den Originalraum zurücktransformiert werden. Die Rücktransformation erfolgt über:

$$x = \mu + Lz$$

Damit lassen sich:

- Zentren aus dem z-Raum korrekt in den Originalraum projizieren
- Ellipsenachsen geometrisch korrekt darstellen
- Clustergeometrien realitätsgetreu abbilden

Ohne diese Rücktransformation wären die dargestellten Ellipsen und Zentren zwar mathematisch korrekt, aber inhaltlich nicht interpretierbar, da sie sich in einem künstlichen, dimensionslosen Koordinatensystem befinden würden. Erst durch die Rücktransformation erhält man:

- Ellipsen in den physikalischen Einheiten der Messgrößen
- korrekte Streuungsgeometrie im Merkmalsraum
- interpretierbare Achsen
- fachlich nachvollziehbare Visualisierung

1.7.3 Darstellung der Ausreißer

Ausreißer werden in den Streudiagrammen gesondert dargestellt. Sie erscheinen typischerweise außerhalb der Konfidenzellipsen und besitzen große Mahalanobis-Distanzen. Die grafische Darstellung erlaubt:

- visuelle Bestätigung der statistischen Ausreißererkennung
- Beurteilung möglicher Messfehler
- Identifikation besonderer Prozesszustände

Ausreißer werden bei der Clusterbildung ausgeschlossen, bleiben jedoch für die Analyse sichtbar. Nicht jeder Datensatz besitzt eine natürliche Clusterstruktur. Insbesondere bei kleinen Stichproben oder kontinuierlichen Gradienten können Streudiagramme dennoch wertvolle Informationen liefern.

In solchen Fällen zeigen die Streudiagramme häufig:

- kontinuierliche Trends
- schiefe Verteilungen
- Streuungsstrukturen
- Rangordnungen

Die grafische Analyse dient dann primär der explorativen Datenanalyse und nicht der Bestätigung einer Clusterlösung.

1.7.4 Grafische Analyse in OQM-Stat

OQM-Stat stellt ein interaktives Plot-Modul zur Verfügung, das folgende Elemente kombiniert:

- Streupunkte nach Clusterzugehörigkeit
- Clusterzentren
- statistische Ausreißer
- Konfidenzellipsen mit frei wählbarem Konfidenzniveau
- Achsenbeschriftungen und Titel
- Legende

OQM-Stat führt die Rücktransformation automatisch durch:

- Clusterzentren werden aus dem Whitening-Raum zurücktransformiert
- Konfidenzellipsen werden geometrisch korrekt im Originalraum konstruiert
- Streudiagramme werden stets im Originalraum dargestellt

Der Anwender arbeitet somit immer mit interpretierbaren Grafiken, unabhängig davon, ob Euklidische oder Mahalanobis-Distanzen verwendet werden. Die Grafiken werden direkt im Excel-Arbeitsblatt erzeugt und können für Berichte, Präsentationen und Dokumentationen verwendet werden.

Die grafische Analyse ist ein unverzichtbarer Bestandteil jeder Clusteranalyse. Sie ergänzt die numerischen Kriterien und ermöglicht eine inhaltliche Interpretation der Ergebnisse.

Nur durch die Kombination aus:

- numerischer Optimierung
- statistischer Validierung
- grafischer Plausibilitätsprüfung

entsteht eine belastbare und fachlich fundierte Clusterlösung.

1.8 Statistische Validierung der Clusterlösung

Die eigentliche Clusterbildung stellt nur den ersten Schritt einer multivariaten Analyse dar. Eine Clusterlösung ist zunächst lediglich eine hypothesengenerierende Struktur. Erst durch eine anschließende statistische Validierung lässt sich beurteilen, ob die gefundenen Gruppen tatsächlich signifikant voneinander verschieden sind oder ob sie lediglich zufällige Artefakte der Datenstruktur darstellen. Eine valide Clusteranalyse muss daher folgende Fragen beantworten:

- Sind die Cluster statistisch signifikant verschieden?
- Lassen sich die Gruppen multivariat voneinander trennen?
- Ist die gefundene Clusterstruktur stabil?
- Sind die Gruppen geometrisch kompakt und klar separiert?

OQM-Stat integriert hierzu eine vollständige statistische Validierung auf Basis multivariater Testverfahren.

Hierarchische Clusterverfahren wie der Ward-Algorithmus liefern stets eine Partition – unabhängig davon, ob tatsächlich eine natürliche Gruppenstruktur existiert oder nicht. Insbesondere bei kleinen Stichproben, schwach separierten Gruppen oder hoch korrelierten Variablen besteht die Gefahr, dass scheinbar plausible Cluster rein zufällig entstehen. Daher ist eine statistische Überprüfung der Clusterlösung zwingend erforderlich. Eine valide Clusterlösung sollte folgende Eigenschaften besitzen:

- hohe Intra-Cluster-Homogenität
- hohe Inter-Cluster-Heterogenität
- signifikante multivariate Trennung
- stabile Zentren

Nur wenn diese Kriterien erfüllt sind, kann von einer belastbaren Klassifikation gesprochen werden.

1.8.1 Multivariate Trennung: Hotelling-T²-Test (k = 2)

Liegt eine Zweiklassenlösung vor, so bietet sich der Hotelling-T²-Test als multivariates Analogon zum t-Test an. Der Test überprüft die Hypothese:

$$H_0 : \mu_1 = \mu_2$$

gegen

$$H_1 : \mu_1 \neq \mu_2$$

wobei μ_1 und μ_2 die multivariaten Mittelwertvektoren der beiden Cluster darstellen. Die Teststatistik lautet:

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} \left(\bar{x}_1 - \bar{x}_2 \right)^T S_p^{-1} \left(\bar{x}_1 - \bar{x}_2 \right)$$

mit der gepoolten Kovarianzmatrix

$$S_p = \frac{\left(n_1 - 1 \right) S_1 + \left(n_2 - 1 \right) S_2}{n_1 + n_2 - 2}$$

Die T²-Statistik lässt sich in eine F-Verteilung überführen:

$$F = \frac{n_1 + n_2 - p - 1}{p \left(n_1 + n_2 - 2 \right)} T^2$$

mit Freiheitsgraden:

$$df_1 = p, \quad df_2 = n_1 + n_2 - p - 1$$

Ein signifikanter p-Wert zeigt, dass sich die beiden Cluster multivariat signifikant unterscheiden. Die Cluster sind dann statistisch eindeutig trennbar. Gerade bei industriellen Klassifikationsproblemen (z. B. Gut-/Schlecht-Teile, Original/Fälschung, IO/NIO) ist dieser Test von zentraler Bedeutung.

1.8.2 Mehrklassenfall: Wilks-Lambda-Test (MANOVA)

Liegt eine Clusterlösung mit mehr als zwei Gruppen vor, so wird eine multivariate Varianzanalyse (MANOVA) durchgeführt. Als Teststatistik dient das **Wilks-Lambda-Kriterium**:

$$\Lambda = \frac{|E|}{|E+H|}$$

wobei E die Fehlervarianzmatrix (within-clusters) und
H die Hypothesenvarianzmatrix (between-clusters) bezeichnet.

Kleine Werte von Λ sprechen für eine starke Trennung der Gruppen. Die Teststatistik wird über eine Chi-Quadrat-Approximation ausgewertet:

$$\chi^2 = -\left(N - 1 - \frac{p+g}{2}\right) ln(\Lambda)$$

mit Freiheitsgraden:

$$df = p(g-1)$$

Ein signifikanter Wilks-Test zeigt, dass mindestens zwei Cluster multivariat signifikant voneinander verschieden sind. Die Clusterlösung ist damit statistisch abgesichert. Neben der rein statistischen Trennung ist auch die geometrische Struktur der Cluster von Bedeutung. OQM-Stat bewertet hierzu:

- Intra-Cluster-Streuung (Within-SSE)
- Inter-Cluster-Streuung (Between-SSE)
- Distanz der Objekte zu ihren Zentren
- Überlappung der Konfidenzellipsen

Eine gute Clusterlösung zeichnet sich aus durch:

- kleine Within-SSE
- große Between-SSE
- kompakte Ellipsen
- geringe Überlappung

Diese Größen sind direkt interpretierbar und lassen sich grafisch nachvollziehen.

1.8.3 Stabilität der Clusterlösung durch k-Means-Refinement

Der Ward-Algorithmus liefert eine hierarchische Startpartition. Diese wird in OQM-Stat optional durch ein partitionierendes k-Means-Verfahren verfeinert. Dabei werden:

- die Ward-Zentren als Startwerte verwendet
- iterative Reallokationen durchgeführt
- die Summe der quadratischen Abweichungen minimiert

Eine stabile Clusterlösung zeigt dabei:

- nur wenige Reallokationen
- schnelle Konvergenz
- geringe Reduktion der SSE

Dies spricht für eine robuste natürliche Gruppenstruktur. Bei sehr kleinen Datensätzen (z. B. $n < 20$) ist die Aussagekraft multivariater Testverfahren eingeschränkt. Gründe hierfür sind:

- instabile Kovarianzschätzungen
- geringe Freiheitsgrade
- reduzierte Teststärke
- hohe Sensitivität gegenüber Ausreißern

Dennoch können selbst bei kleinen Datensätzen hochsignifikante Trennungen auftreten, sofern die Gruppen stark separiert sind. Ein signifikanter Hotelling-T^2-Test bei kleinen Stichproben ist daher durchaus möglich, wenn:

- die Mittelwerte stark differieren
- die Streuungen klein sind
- die Gruppen geometrisch klar getrennt liegen

In solchen Fällen liefert der Test trotz kleiner Fallzahlen valide Evidenz für eine echte Gruppenstruktur. Die statistische Validierung ist ein unverzichtbarer Bestandteil einer seriösen Clusteranalyse. Erst durch multivariate Tests lässt sich beurteilen, ob die gefundenen Cluster tatsächlich reale Objektklassen repräsentieren. OQM-Stat stellt hierzu ein vollständiges Validierungssystem bereit:

- Hotelling-T^2-Test für Zweiklassenlösungen
- Wilks-Lambda-Test für Mehrklassenlösungen
- geometrische Clusteranalyse
- k-Means-Stabilitätsprüfung

Damit wird aus einer explorativen Clusteranalyse ein statistisch abgesichertes Klassifikationsmodell.

1.9 Industrielle Interpretation der Clusterlösung

Die Clusteranalyse ist ein exploratives, strukturbildendes Verfahren. Ihr primäres Ziel besteht darin, aus einer multivariaten Datenmatrix natürliche Gruppen zu identifizieren, ohne dass eine Klassenzugehörigkeit vorab bekannt ist. In industriellen Anwendungen stellt die Clusteranalyse damit häufig den ersten Schritt einer systematischen Klassifikationskette dar.

Erst durch die anschließende Interpretation und Modellbildung wird aus einer Clusterlösung ein praxistaugliches Identifikationssystem. In technischen Anwendungen lautet die eigentliche Fragestellung nicht:

„Welche Gruppen existieren in den Daten?“

sondern:

„Zu welcher bekannten Klasse gehört ein neues Objekt?“

Typische industrielle Klassifikationsprobleme sind:

- Gutteil / Schlechtteil
- Original / Fälschung
- Konform / Nicht-konform
- Freigabe / Sperrung
- Prozess stabil / Prozess instabil

Die Clusteranalyse liefert hierzu die notwendige Strukturinformation:

- Anzahl der Objektklassen
- geometrische Trennung
- relevante Merkmalskombinationen
- Streuungsstruktur
- Ausreißercharakteristik

Damit bildet sie die Grundlage für den Aufbau eines deterministischen Klassifikationsmodells. Die Clusteranalyse bildet den Einstieg in die multivariate Klassifikation. Erst durch die anschließende Modellbildung entsteht ein industriell nutzbares Identifikationssystem. Die Kombination aus:

- Ward-Clusteranalyse
- Calinski-Harabasz-Stopping-Rule
- Hotelling-T^2-Validierung
- Wilks-Lambda-Test
- Mahalanobis-Distanzen

stellt ein leistungsfähiges, statistisch abgesichertes Klassifikationsframework dar. In den folgenden Kapiteln wird gezeigt, wie diese Methodik auf reale Produktionsprozesse übertragen werden kann — von der Datenerfassung über die Modellbildung bis zur Online-Entscheidung.

1.10 Beispiel: Markenbewusstsein

Dies ist ein Lehrbeispiel; daher wurde der Datensatz bewusst klein gehalten. Dennoch lässt sich anschaulich demonstrieren, wie die Clusteranalyse in OQM-Stat arbeitet. Es werden zwei Cluster erwartet, die sich inhaltlich wie folgt charakterisieren lassen:

Ein Cluster ist durch hohes Einkommen und hohes Markenbewusstsein gekennzeichnet, während der andere Cluster niedriges Einkommen mit geringem Markenbewusstsein kombiniert. Aufgrund der kleinen Stichprobe sind keine stabilen und belastbaren Ergebnisse zu erwarten, was hier ausdrücklich demonstriert werden soll. Gleichzeitig sollen alternative Lösungsansätze aufgezeigt werden, mit denen sich trotz dieser Einschränkungen einfachere und besser interpretierbare Ergebnisse erzielen lassen.

1.10.1 Der Datensatz zum Markenbewusstsein

Der Datensatz stammt aus einem im Internet veröffentlichten Artikel:

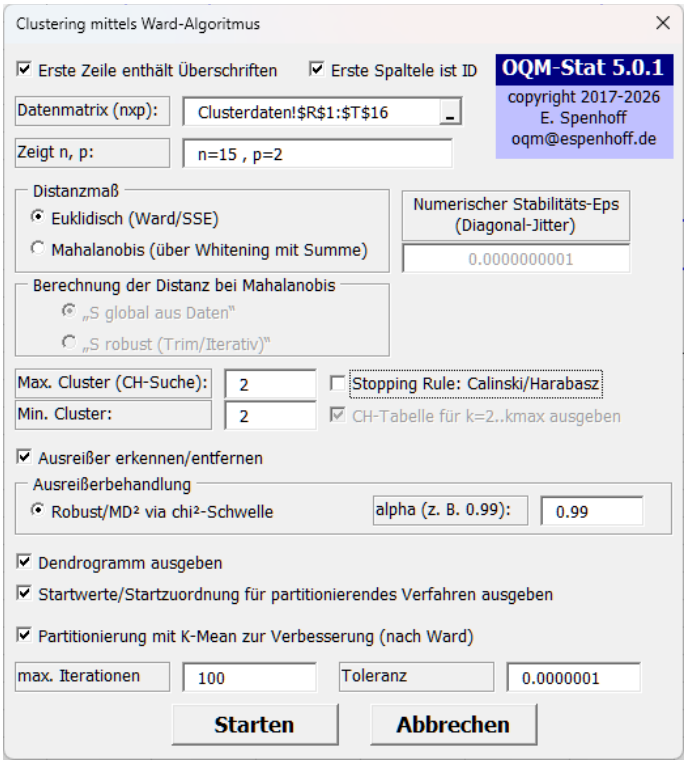
Clusteranalyse – Methodenberatung UZH – Universität Zürich
(https://www.methodenberatung.uzh.ch › cluster).

Über die Entstehung der Variablen Einkommen und Markenbewusstsein werden keine näheren Angaben gemacht. Es ist nicht bekannt, ob es sich um eine Zufallsstichprobe oder um einen konstruierten Beispieldatensatz handelt. Ebenso bleibt unklar, wie das Markenbewusstsein konkret erhoben bzw. gemessen wurde.

Für die nachfolgende Analyse und Beurteilung wird daher angenommen, dass es sich um einen realen Datensatz handelt.

Beruf	Einkommen	Markenbewusstsein
Arzt	6861	21765
Ingenieur	5150	28245
Chemiker	5474	25179
Manager	7389	19048
Professor	5152	24608
CEO	12810	27611
Anwalt	7203	21536
Koch	4162	24823
Architekt	6779	22499
Forstwart	3204	7465
Physiker ETH	5335	17471
Lehrer	4311	14735
Bauarbeiter	3949	17921
Fischer	2132	8822
Servicemitarbeiter	3018	12201

1.10.2 Eingabemenü von OQM-Stat



Die Daten werden standardmäßig mit Kopfzeile (Namen der Variablen) und ID-Spalte (Bezeichnung der Objekte bzw. Beobachtungen) eingelesen. Dies ist sehr sinnvoll, da sowohl die Variablennamen als auch die Objektbezeichnungen in der Ergebnisausgabe weiterverwendet werden. Nachdem der Eingabebereich definiert wurde, werden zur Kontrolle

- n = Anzahl der Objekte und
- p = Anzahl der Variablen

ausgegeben.

Im nächsten Schritt erfolgt die Auswahl des Distanzmaßes. Wir wählen zunächst die e**uklidische Distanz**, werden jedoch später auch zeigen, welchen Einfluss die **Mahalanobis-Distanz** auf die Ergebnisse hat.

Eine besonders wichtige Option ist die Suche nach der optimalen Clusterzahl k^*. Hierzu wird die **Calinski-Harabasz-Stopping-Rule** verwendet, die als eines der besten Kriterien zur Bestimmung der Clusterzahl gilt.

Für diesen Datensatz wird jedoch kein eindeutiges k^* gefunden, da für keine Clusteranzahl ein Maximum des CH-Kriteriums auftritt. Dies zeigt deutlich, dass die geringe Größe des Datensatzes keine statistisch signifikante Bestimmung der optimalen Clusterzahl erlaubt. Aus diesem Grund muss der Suchbereich auf

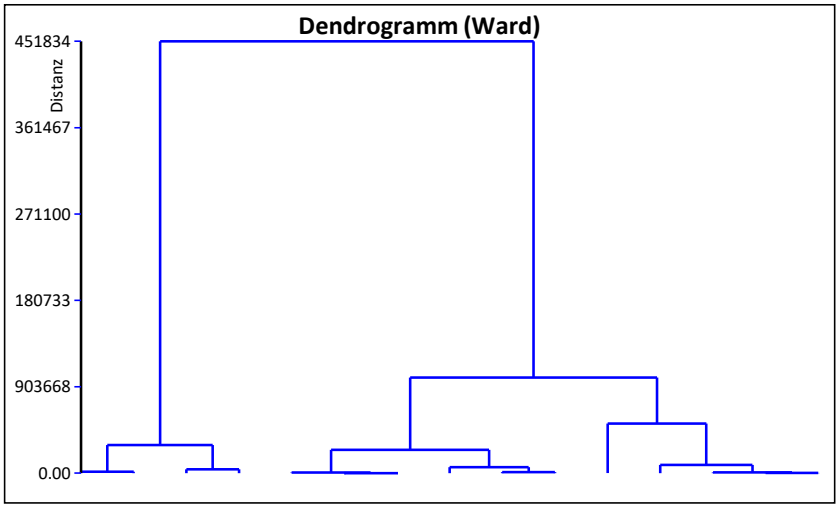
min. Cluster = 2 und
max. Cluster = 2

eingeschränkt werden. Zusätzlich wird untersucht, was geschieht, wenn max. Cluster = 3 gesetzt wird. Da die automatische Bestimmung der Clusterzahl nicht erfolgreich ist, muss die „Stopping Rule“ deaktiviert werden, da andernfalls unsinnige oder instabile Ergebnisse entstehen können.

Der nächste Schritt betrifft die Definition eines **Ausreißertests**, mit dem mögliche Ausreißer identifiziert werden. Es handelt sich hierbei um einen einfachen Test, dessen Sensitivität über die Wahrscheinlichkeit α eingestellt wird. Kleinere Werte von α erhöhen die Sensitivität, größere Werte verringern sie. Geeignete Werte für α sind beispielsweise: 0,9; 0,95; 0,975; 0,99; 0,995 und 0,999. Die voreingestellten Standardwerte können für eine erste Analyse problemlos übernommen werden.

Der letzte Schritt betrifft die Ausgabe der Ergebnisse sowie die Verbesserung der Zuordnung der Objekte mithilfe des **k-Means-Algorithmus**. Alle entsprechenden Optionen sollten aktiviert werden, da andernfalls das wichtige Scatterplot nur eingeschränkt dargestellt werden kann. Nach diesen Einstellungen wird die Clusteranalyse gestartet.

1.10.3 Das Dendrogramm und andere Analyseergebnisse



Die Analyse erzeugt ein Dendrogramm, das auf das Vorhandensein von zwei Clustern schließen lässt. Dabei ist jedoch zu berücksichtigen, dass dieses Ergebnis durch die Einstellung max. Cluster = 2 erzwungen wurde. Die eingelesene Datenmatrix wird erneut ausgegeben und um mehrere Ergebnisspalten ergänzt:

- die Spalte **Cluster**, welche die ursprüngliche Zuordnung durch den Ward-Algorithmus enthält,
- die Spalte **Ausreißer**, in der erkannte Ausreißer dokumentiert werden,
- die Spalte **NeuCluster**, welche die veränderte Zuordnung der Objekte nach der Partitionierung mit dem k-Means-Algorithmus anzeigt,
- sowie die Spalte **Distanzmaße**, in der für jedes Objekt die euklidische Distanz gemäß der gewählten Definition ausgegeben wird.

	Urdaten		Zuordnung			
Objekt	Einkommen	Markenbewusstsein	Cluster	Ausreißer	NeuCluster	Distanzmaße
Arzt	6861	21765	2	0	2	1130.345994
Ingenieur	5150	28245	2	0	2	5592.217967
Chemiker	5474	25179	2	0	2	2556.385085
Manager	7389	19048	2	0	2	3875.07175
Professor	5152	24608	2	0	2	2196.960312
CEO	12810	27611	2	0	2	8029.618488
Anwalt	7203	21536	2	0	2	1496.991642
Koch	4162	24823	2	0	2	3013.409059
Architekt	6779	22499	2	0	2	488.5643837
Forstwart	3204	7465	1	0	1	3340.963278
Physiker ETH	5335	17471	2	0	2	5423.585519
Lehrer	4311	14735	1	0	1	4092.610185
Bauarbeiter	3949	17921	2	0	2	5446.859168
Fischer	2132	8822	1	0	1	2237.17168
Servicemitarbeiter	3018	12201	1	0	1	1403.103925

Wir sehen Cluster 1 mit den Berufen Forstwart, Lehrer, Fischer und Servicemitarbeiter und im Cluster 2 befinden sich alle elf anderen Berufe. Auch der CEO mit seinem sehr hohen Gehalt gehört

zu dieser Gruppe. Keines der 15 Objekte wurde als Ausreißer erkannt. Die Partitionierung ergab keine Veränderung.

Clustering mittels Ward- (und k-Mean-Algorithmus)			
Bedingungen für Ward-Algorithmus			
Distanzmaß:	Euklidisch		
Ausreißertest:	MD²/Chi²		
alpha:	0.99		
Cluster-Suche (CH):	2 .. 2		
k*(optimal):	2		
Anzahl Variabler p:	2		
n (original):	15		
n (ohne Outlier):	15		
Rechenzeit (s):	0.016		
Partitionierung mit k-Means nach Ward-Resultaten			
max. Iterationen:	100		
Toleranz:	0.0000001		
Iterationen:	1		
Umklassifikationen:	0		
SSE (Start):	228932276.8		
SSE (Ende):	228932276.8		

Rechts neben der Matrix werden weitere Ergebnisse ausgegeben: Im oberen Block der Ausgabe befinden sich die Bedingungen und Ergebnisse des Ward-Algorithmus, im zweiten Block die Bedingungen und Ergebnisse der k-Means-Partitionierung. Die **SSE-Werte (Sum of Squared Errors)** stellen dabei das zentrale Ergebnis der Clusteranalyse dar, da sie den zu minimierenden Zielwert sowohl für das Ward-Verfahren als auch für den k-Means-Algorithmus repräsentieren.

Im unteren Block werden – bei Verwendung der Mahalanobis-Distanzen – zusätzlich die Ergebnisse statistischer Tests ausgegeben:

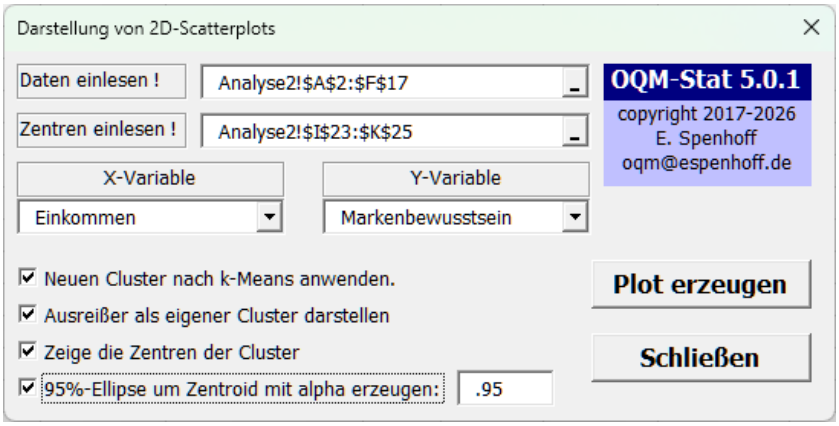
- der **Hotelling-T^2-Test** für zwei Cluster (vergleichbar mit dem univariaten t-Test),
- bzw. der **Wilks-Λ-Test** für mehr als zwei Cluster (vergleichbar mit der Streuungszerlegung im univariaten Fall).

Ein signifikantes Testergebnis bedeutet hierbei nicht, dass die Cluster „richtig“ sind, sondern lediglich, dass sich die Mittelwertvektoren der Cluster statistisch signifikant voneinander unterscheiden. Die anschließende Ausgabe betrifft im ersten Block die **Mittelwertvektoren der Cluster**, die im Scatterplot als Zentren dargestellt werden. Gleichzeitig dienen diese Zentren als Startpunkte für den partitionierenden k-Means-Algorithmus. Dadurch wird die Anzahl der erforderlichen Iterationen deutlich reduziert, da die vom Ward-Algorithmus gelieferten Startwerte bereits nahezu optimal sind.

In den folgenden Blöcken werden die **Kovarianzmatrizen der einzelnen Cluster** ausgegeben. Diese werden benötigt, um in Scatterplots mit zwei Variablen **Konfidenz-Ellipsen** (Ellipsoide) darstellen zu können.

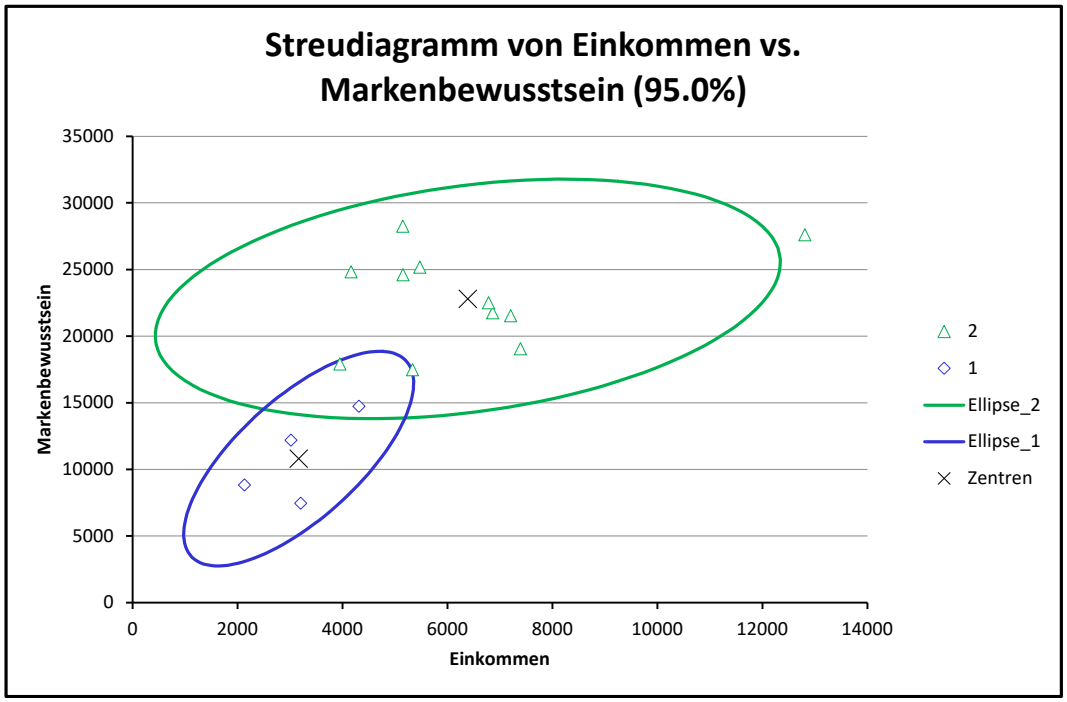
Bei Verwendung der **Mahalanobis-Distanzen** verdoppelt sich die Ausgabe, da die Ergebnisse sowohl im Whitening-Raum als auch – nach Rücktransformation – im Originalraum dargestellt werden.

1.10.4 Das 2D-Scatterplot



Um das Scatterplot darzustellen, muss zunächst ein separates Menü geöffnet und einige Eingaben vorgenommen werden. Zuerst wird die **Datenmatrix** definiert. Diese wird von der ersten Spalte (Objektname) bis zur vorletzten Spalte (**NeuCluster**) eingelesen, einschließlich der Kopfzeile mit den Variablennamen und bis zur letzten Datenzeile.

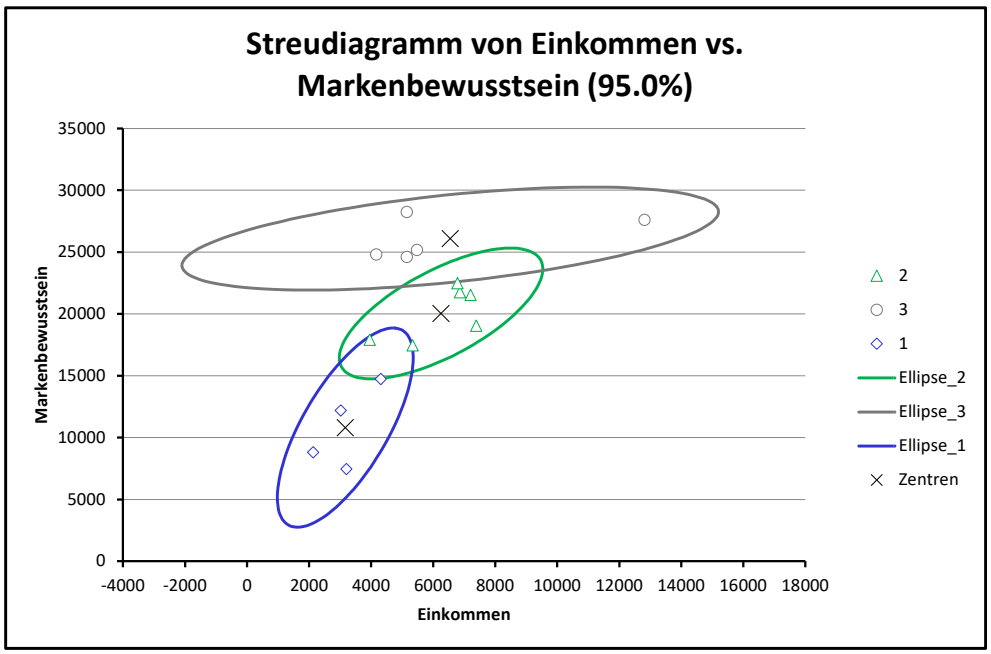
Alle anderen Eingaben führen entweder zu einer Fehlermeldung, zum Abbruch des Programms oder zu einer fehlerhaften grafischen Darstellung. Im nächsten Schritt werden die Startwerte im Originalraum erneut eingelesen, ebenfalls mit Kopfzeile und Clusterspalte. Diese entsprechen den Zentren der Cluster.



Anschließend kann ausgewählt werden, welche Variablen dargestellt werden sollen. Bei nur zwei Variablen – wie in diesem Beispiel – ist die mögliche Auswahl bereits vorgegeben. Zum Schluss werden alle Kontrollkästchen aktiviert, um sämtliche verfügbaren Informationen in der Grafik darzustellen. Danach wird die Grafikerzeugung gestartet. Das Ergebnis ist in der folgenden Abbildung dargestellt.In der Grafik fallen mehrere Punkte unmittelbar auf:

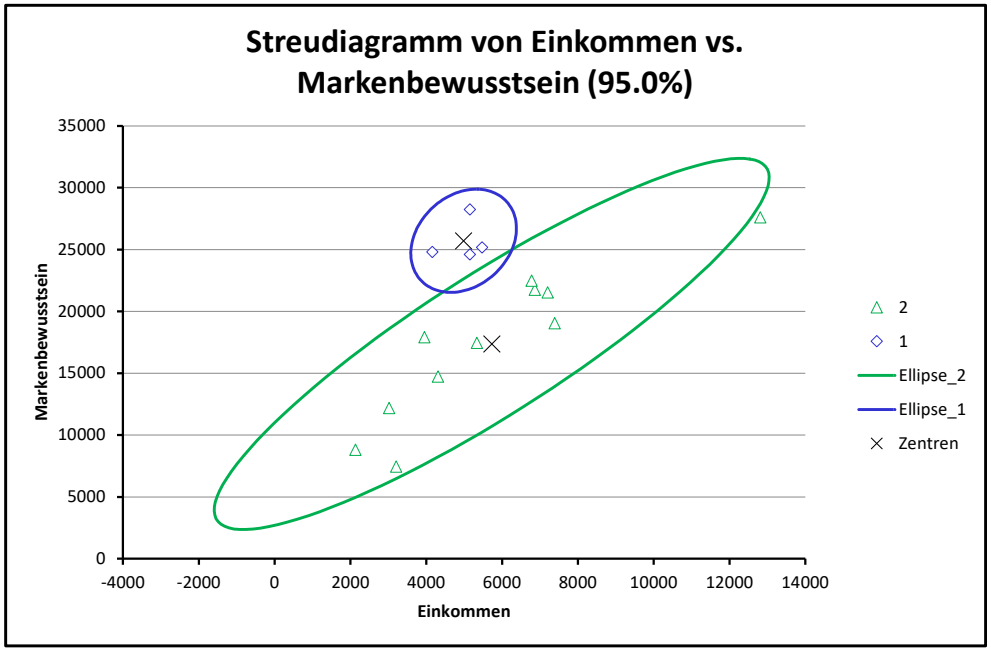
- Die Streuung von **Cluster 2** ist deutlich größer als die von **Cluster 1**.
- Die Cluster 1 und 2 besitzen eine ausgeprägte Schnittmenge; die Trennung der Cluster ist somit nicht perfekt.
- Der **CEO** liegt außerhalb der 95%-Ellipse, jedoch noch innerhalb der 99%-Ellipse und ist damit kein Ausreißer.
- Der CEO trägt in besonderem Maße zur Streuung von Cluster 2 bei.
- Zwei Objekte aus Cluster 2 liegen im Bereich von Cluster 1.
- Ein Objekt aus Cluster 1 liegt im Bereich von Cluster 2.

Um eine Verbesserung zu erreichen, wird eine 3-Cluster-Lösung untersucht. Dazu wurde lediglich die maximale Clusterzahl auf 3 gesetzt; alle übrigen Einstellungen blieben unverändert. Auch hier treten Auffälligkeiten auf, die eine sinnvolle Interpretation erschweren:



- In Cluster 3 wird ein negatives Einkommen mit hohem Markenbewusstsein kombiniert, was inhaltlich keinen Sinn ergibt.
- Bei Cluster 1 und 2 zeigt sich eine nahezu lineare Aneinanderreihung der Objekte.
- Die Streuungen von Cluster 1 und 2 sind wieder ähnlich, während Cluster 3 eine deutlich größere Streuung aufweist, erneut maßgeblich verursacht durch den CEO.

Wir gehen zurück zu den ursprünglichen Definitionen und nutzen die Malahanonis-Distanz.

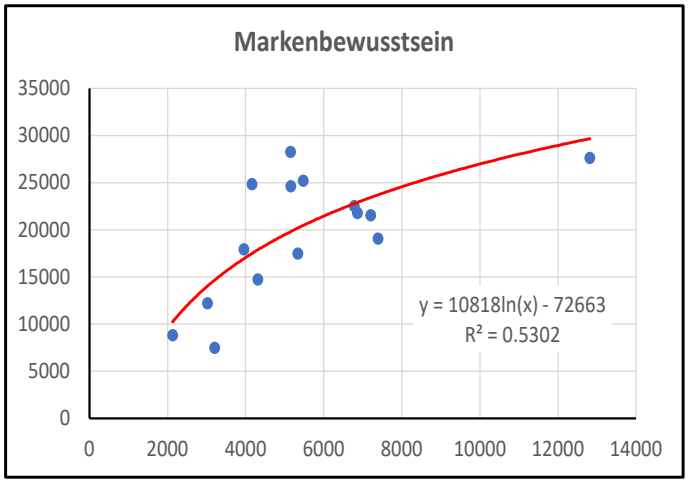


Auch diese Grafik liefert keine sinnvolle Interpretation:

- So ist der CEO mit hohem Markenbewusstsein in der gleichen Gruppe (Cluster 2) wie Berufe mit niedrigem Markenbewusstsein.
- Auch hier reicht die 95% Ellipse von Cluster 2 bis in den negativen Einkommensbereich.
- Die Streuungen der Cluster 1 und 2 sind sehr unterschiedlich.

Der Datensatz ist aus mehreren Gründen für eine Clusteranalyse ungeeignet. Beide Variablen decken einen sehr großen Wertebereich ab und sind zudem stark korreliert. Hinzu kommt, dass lediglich zwei Klassifikationsmerkmale (Variablen) zur Verfügung stehen und insgesamt viel zu wenige Objekte vorliegen. Dadurch kann der CEO weder als Ausreißer eindeutig erkannt noch als eigene Klasse sinnvoll abgegrenzt werden.

Die resultierende Lösung ist daher trivial, wie die nachfolgende Grafik einer nichtlinearen Regression zeigt.



In Abhängigkeit von der Höhe des Einkommens kann nun das Markenbewusstsein geschätzt werden; es gilt demnach:

Je höher das Einkommen, desto höher ist in der Tendenz das Markenbewusstsein.

1.11 Beispiel: Banknoten

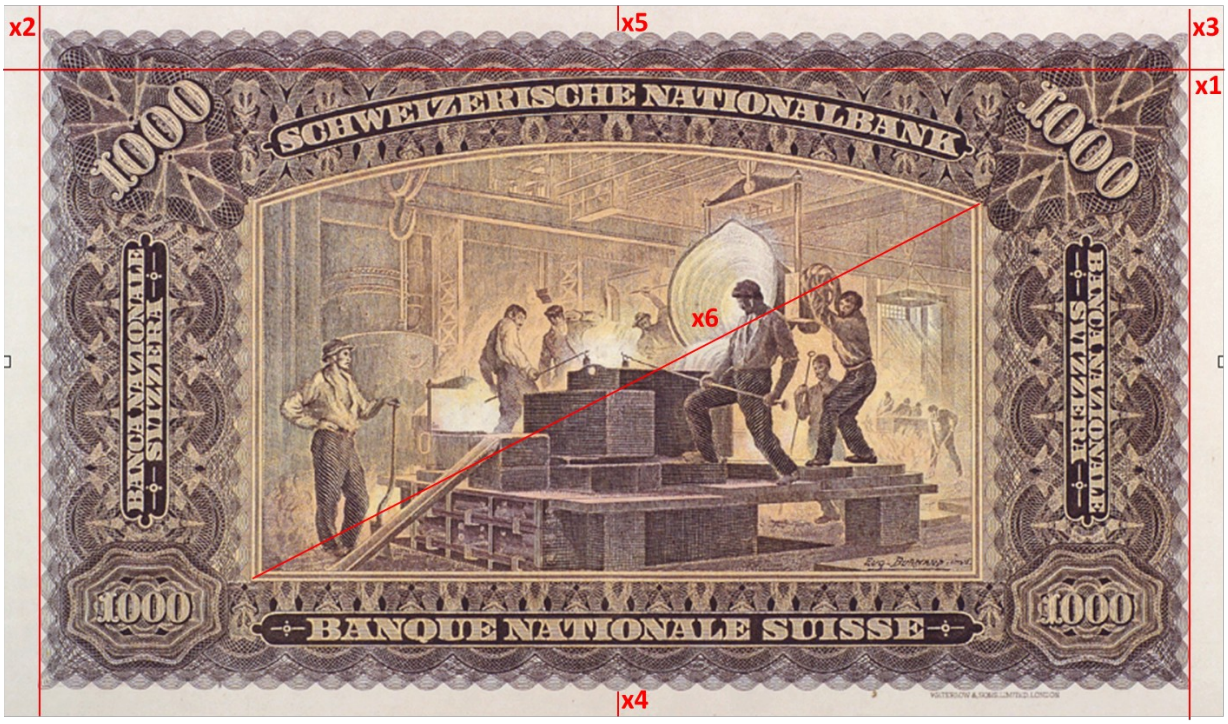
Wir wenden uns nun einen Datensatz von echten und gefälschten Banknoten zu, welcher in Kapitel 1 ausführlich beschrieben wurde. Eine Besonderheit ist, dass die zwei Gruppen (echte und gefälschte Banknoten) bekannt sind und deshalb ist dieser Datensatz geeignet, die Wirkungsweise und die Ergebnisse einer Clusteranalyse, wie in OQM-Stat etabliert ist, zu verifizieren.

Die Eingabemenüs sind im vorherigen Beispiel erklärt, so dass nur noch eine mögliche, zusätzliche Ausgabe der Analyse erklärt werden muss. Für das 2D-Scatterplot gilt das insgesamt 15 Grafiken erstellt werden müssen um alle Abhängigkeiten der Variablen beurteilen zu können. Dies ist ein durchaus lohnender Aufwand. Doch zuerst gilt es die Analyseergebnisse zu bewerten. Zur Analyse verwenden wir die robuste Mahalanobis-Distanz.

1.11.1 Datensatz: Echte und gefälschte Banknoten

Vermessen wurden 100 echte Schweizer Banknoten mit dem Nennwert von 1000 Franken, welche wegen Gebrauchsspuren aus dem Verkehr gezogen wurden. Zusätzlich wurden 100 gefälschte 1000-Franken-Banknoten vermessen.

Die Zielsetzung besteht darin, die Frage zu beantworten, ob es möglich ist, die Banknoten anhand einiger geometrischer Abmessungen automatisch in zwei Gruppen (echt und gefälscht) aufzuteilen, ohne dass diese Klassenzugehörigkeit im Vorfeld bekannt ist.



Der Datensatz besteht aus geometrischen Merkmalen von Banknoten. Für jede Banknote wurden mehrere kontinuierliche Merkmale gemessen, beispielsweise:

- Länge der Banknote; x1
- Breite der Banknote, links gemessen; x2
- Breite der Banknote, rechts gemessen; x3
- unterer Randabstand; x4
- oberer Randabstand; x5
- Länge der Bilddiagonalen; x6

Jede Banknote ist durch einen Merkmalsvektor beschrieben:

$$x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$$

mit p = Anzahl der Variablen
i = Index der Banknoten.

Die Messung gestaltete sich aufwendig. Um die erforderliche Messgenauigkeit zu erzielen, wurde ein Episkop mit 10-facher Vergrößerung auf einer Projektionsfläche verwendet. Die Rohdaten wurden aus dem Internet entnommen, stammen jedoch ursprünglich aus dem Buch

Bernhard Flury, Hans Riedwyl:
Angewandte multivariate Statistik, Gustav Fischer Verlag 1983

und liegen in tabellarischer Form vor:

Echte Banknoten													
id	x1	x2	x3	x4	x5	x6	id	x1	x2	x3	x4	x5	x6
1	214.8	131.0	131.1	9.0	9.7	141.0	51	214.6	129.8	129.4	7.2	10.0	141.3
2	214.6	129.7	129.7	8.1	9.5	141.7	52	215.3	130.6	130.0	9.5	9.7	141.1
3	214.8	129.7	129.7	8.7	9.6	142.2	53	214.5	130.1	130.0	7.8	10.9	140.9
4	214.8	129.7	129.6	7.5	10.4	142.0	54	215.4	130.2	130.2	7.6	10.9	141.6
5	215.0	129.6	129.7	10.4	7.7	141.8	55	214.5	129.4	129.5	7.9	10.0	141.4
6	215.7	130.8	130.5	9.0	10.1	141.4	56	215.2	129.7	129.4	9.2	9.4	142.0
7	215.5	129.5	129.7	7.9	9.6	141.6	57	215.7	130.0	129.4	9.2	10.4	141.2
8	214.5	129.6	129.2	7.2	10.7	141.7	58	215.0	129.6	129.4	8.8	9.0	141.1
9	214.9	129.4	129.7	8.2	11.0	141.9	59	215.1	130.1	129.9	7.9	11.0	141.3
10	215.2	130.4	130.3	9.2	10.0	140.7	60	215.1	130.0	129.8	8.2	10.3	141.4
11	215.3	130.4	130.3	7.9	11.7	141.8	61	215.1	129.6	129.3	8.3	9.9	141.6
12	215.1	129.5	129.6	7.7	10.5	142.2	62	215.3	129.7	129.4	7.5	10.5	141.5
13	215.2	130.8	129.6	7.9	10.6	141.4	63	215.4	129.8	129.4	8.0	10.6	141.5
14	214.7	129.7	129.7	7.7	10.9	141.7	64	214.5	130.0	129.5	8.0	10.8	141.4
15	215.1	129.9	129.7	7.7	10.8	141.8	65	215.0	130.0	129.8	8.6	10.6	141.5
16	214.5	129.8	129.8	9.3	8.5	141.6	66	215.2	130.6	130.0	8.8	10.6	140.8
17	214.6	129.9	130.1	8.2	9.8	141.7	67	214.6	129.5	129.2	7.7	10.3	141.3
18	215.0	129.9	129.7	9.0	9.0	141.9	68	214.8	129.7	129.3	9.1	9.5	141.5
19	215.2	129.6	129.6	7.4	11.5	141.5	69	215.1	129.6	129.8	8.6	9.8	141.8
20	214.7	130.2	129.9	8.6	10.0	141.9	70	214.9	130.2	130.2	8.0	11.2	139.6
21	215.0	129.9	129.3	8.4	10.0	141.4	71	213.8	129.8	129.5	8.4	11.1	140.9
22	215.6	130.5	130.0	8.1	10.3	141.6	72	215.2	129.9	129.5	8.2	10.3	141.4
23	215.3	130.6	130.0	8.4	10.8	141.5	73	215.0	129.6	130.2	8.7	10.0	141.2
24	215.7	130.2	130.0	8.7	10.0	141.6	74	214.4	129.9	129.6	7.5	10.5	141.8
25	215.1	129.7	129.9	7.4	10.8	141.1	75	215.2	129.9	129.7	7.2	10.6	142.1
26	215.3	130.4	130.4	8.0	11.0	142.3	76	214.1	129.6	129.3	7.6	10.7	141.7
27	215.5	130.2	130.1	8.9	9.8	142.4	77	214.9	129.9	130.1	8.8	10.0	141.2
28	215.1	130.3	130.3	9.8	9.5	141.9	78	214.6	129.8	129.4	7.4	10.6	141.0
29	215.1	130.0	130.0	7.4	10.5	141.8	79	215.2	130.5	129.8	7.9	10.9	140.9
30	214.8	129.7	129.3	8.3	9.0	142.0	80	214.6	129.9	129.4	7.9	10.0	141.8
31	215.2	130.1	129.8	7.9	10.7	141.8	81	215.1	129.7	129.7	8.6	10.3	140.6
32	214.8	129.7	129.7	8.6	9.1	142.3	82	214.9	129.8	129.6	7.5	10.3	141.0
33	215.0	130.0	129.6	7.7	10.5	140.7	83	215.2	129.7	129.1	9.0	9.7	141.9
34	215.6	130.4	130.1	8.4	10.3	141.0	84	215.2	130.1	129.9	7.9	10.8	141.3
35	215.9	130.4	130.0	8.9	10.6	141.4	85	215.4	130.7	130.2	9.0	11.1	141.2
36	214.6	130.2	130.2	9.4	9.7	141.8	86	215.1	129.9	129.6	8.9	10.2	141.5
37	215.5	130.3	130.0	8.4	9.7	141.8	87	215.2	129.9	129.7	8.7	9.5	141.6
38	215.3	129.9	129.4	7.9	10.0	142.0	88	215.0	129.6	129.2	8.4	10.2	142.1
39	215.3	130.3	130.1	8.5	9.3	142.1	89	214.9	130.3	129.9	7.4	11.2	141.5
40	213.9	130.3	129.0	8.1	9.7	141.3	90	215.0	129.9	129.7	8.0	10.5	142.0
41	214.4	129.8	129.2	8.9	9.4	142.3	91	214.7	129.7	129.3	8.6	9.6	141.6
42	214.8	130.1	129.6	8.8	9.9	140.9	92	215.4	130.0	129.9	8.5	9.7	141.4
43	214.9	129.6	129.4	9.3	9.0	141.7	93	214.9	129.4	129.5	8.2	9.9	141.5
44	214.9	130.4	129.7	9.0	9.8	140.9	94	214.5	129.5	129.3	7.4	10.7	141.5
45	214.8	129.4	129.1	8.2	10.2	141.0	95	214.7	129.6	129.5	8.3	10.0	142.0
46	214.3	129.5	129.4	8.3	10.2	141.8	96	215.6	129.9	129.9	9.0	9.5	141.7
47	214.8	129.9	129.7	8.3	10.2	141.5	97	215.0	130.4	130.3	9.1	10.2	141.1
48	214.8	129.9	129.7	7.3	10.9	142.0	98	214.4	129.7	129.5	8.0	10.3	141.2
49	214.6	129.7	129.8	7.9	10.3	141.1	99	215.1	130.0	129.8	9.1	10.2	141.5
50	214.5	129.0	129.6	7.8	9.8	142.0	100	214.7	130.0	129.4	7.8	10.0	141.2

Gefälschte Banknoten													
id	x1	x2	x3	x4	x5	x6	id	x1	x2	x3	x4	x5	x6
101	214.4	130.1	130.3	9.7	11.7	139.8	151	214.9	130.3	129.9	11.9	10.6	139.8
102	214.9	130.5	130.2	11.0	11.5	139.5	152	214.6	129.9	129.7	11.9	10.1	139.0
103	214.9	130.3	130.1	8.7	11.7	140.2	153	214.6	129.7	129.3	10.4	11.0	139.3
104	215.0	130.4	130.6	9.9	10.9	140.3	154	214.5	130.1	130.1	12.1	10.3	139.4
105	214.7	130.2	130.3	11.8	10.9	139.7	155	214.5	130.3	130.0	11.0	11.5	139.5
106	215.0	130.2	130.2	10.6	10.7	139.9	156	215.1	130.0	130.3	11.6	10.5	139.7
107	215.3	130.3	130.1	9.3	12.1	140.2	157	214.2	129.7	129.6	10.3	11.4	139.5
108	214.8	130.1	130.4	9.8	11.5	139.9	158	214.4	130.1	130.0	11.3	10.7	139.2
109	215.0	130.2	129.9	10.0	11.9	139.4	159	214.8	130.4	130.6	12.5	10.0	139.3
110	215.2	130.6	130.8	10.4	11.2	140.3	160	214.6	130.6	130.1	8.1	12.1	137.9
111	215.2	130.4	130.3	8.0	11.5	139.2	161	215.6	130.1	129.7	7.4	12.2	138.4
112	215.1	130.5	130.3	10.6	11.5	140.1	162	214.9	130.5	130.1	9.9	10.2	138.1
113	215.4	130.7	131.1	9.7	11.8	140.6	163	214.6	130.1	130.0	11.5	10.6	139.5
114	214.9	130.4	129.9	11.4	11.0	139.9	164	214.7	130.1	130.2	11.6	10.9	139.1
115	215.1	130.3	130.0	10.6	10.8	139.7	165	214.3	130.3	130.0	11.4	10.5	139.8
116	215.5	130.4	130.0	8.2	11.2	139.2	166	215.1	130.3	130.6	10.3	12.0	139.7
117	214.7	130.6	130.1	11.8	10.5	139.8	167	216.3	130.7	130.4	10.0	10.1	138.8
118	214.7	130.4	130.1	12.1	10.4	139.9	168	215.6	130.4	130.1	9.6	11.2	138.6
119	214.8	130.5	130.2	11.0	11.0	140.0	169	214.8	129.9	129.8	9.6	12.0	139.6
120	214.4	130.2	129.9	10.1	12.0	139.2	170	214.9	130.0	129.9	11.4	10.9	139.7
121	214.8	130.3	130.4	10.1	12.1	139.6	171	213.9	130.7	130.5	8.7	11.5	137.8
122	215.1	130.6	130.3	12.3	10.2	139.6	172	214.2	130.6	130.4	12.0	10.2	139.6
123	215.3	130.8	131.1	11.6	10.6	140.2	173	214.8	130.5	130.3	11.8	10.5	139.4
124	215.1	130.7	130.4	10.5	11.2	139.7	174	214.8	129.6	130.0	10.4	11.6	139.2
125	214.7	130.5	130.5	9.9	10.3	140.1	175	214.8	130.1	130.0	11.4	10.5	139.6
126	214.9	130.0	130.3	10.2	11.4	139.6	176	214.9	130.4	130.2	11.9	10.7	139.0
127	215.0	130.4	130.4	9.4	11.6	140.2	177	214.3	130.1	130.1	11.6	10.5	139.7
128	215.5	130.7	130.3	10.2	11.8	140.0	178	214.5	130.4	130.0	9.9	12.0	139.6
129	215.1	130.2	130.2	10.1	11.3	140.3	179	214.8	130.5	130.3	10.2	12.1	139.1
130	214.5	130.2	130.6	9.8	12.1	139.9	180	214.5	130.2	130.4	8.2	11.8	137.8
131	214.3	130.2	130.0	10.7	10.5	139.8	181	215.0	130.4	130.1	11.4	10.7	139.1
132	214.5	130.2	129.8	12.3	11.2	139.2	182	214.8	130.6	130.6	8.0	11.4	138.7
133	214.9	130.5	130.2	10.6	11.5	139.9	183	215.0	130.5	130.1	11.0	11.4	139.3
134	214.6	130.2	130.4	10.5	11.8	139.7	184	214.6	130.5	130.4	10.1	11.4	139.3
135	214.2	130.0	130.2	11.0	11.2	139.5	185	214.7	130.2	130.1	10.7	11.1	139.5
136	214.8	130.1	130.1	11.9	11.1	139.5	186	214.7	130.4	130.0	11.5	10.7	139.4
137	214.6	129.8	130.2	10.7	11.1	139.4	187	214.5	130.4	130.0	8.0	12.2	138.5
138	214.9	130.7	130.3	9.3	11.2	138.3	188	214.8	130.0	129.7	11.4	10.6	139.2
139	214.6	130.4	130.4	11.3	10.8	139.8	189	214.8	129.9	130.2	9.6	11.9	139.4
140	214.5	130.5	130.2	11.8	10.2	139.9	190	214.6	130.3	130.2	12.7	9.1	139.2
141	214.8	130.2	130.3	10.0	11.9	139.3	191	215.1	130.2	129.8	10.2	12.0	139.4
142	214.7	130.0	129.4	10.2	11.0	139.2	192	215.4	130.5	130.6	8.8	11.0	138.6
143	214.6	130.2	130.4	11.2	10.7	139.9	193	214.7	130.3	130.2	10.8	11.1	139.2
144	215.0	130.5	130.4	10.6	11.1	139.9	194	215.0	130.5	130.3	9.6	11.0	138.5
145	214.5	129.8	129.8	11.4	10.0	139.3	195	214.9	130.3	130.5	11.6	10.6	139.8
146	214.9	130.6	130.4	11.9	10.5	139.8	196	215.0	130.4	130.3	9.9	12.1	139.6
147	215.0	130.5	130.4	11.4	10.7	139.9	197	215.1	130.3	129.9	10.3	11.5	139.7
148	215.3	130.6	130.3	9.3	11.3	138.1	198	214.8	130.3	130.4	10.6	11.1	140.0
149	214.7	130.2	130.1	10.7	11.0	139.4	199	214.7	130.7	130.8	11.2	11.2	139.4
150	214.9	129.9	130.0	9.9	12.3	139.4	200	214.3	129.9	129.9	10.2	11.5	139.6

Die Klassenzugehörigkeit (echt / gefälscht) ist in der Analyse nicht bekannt und dient später ausschließlich zur Validierung der Clusterstruktur.

1.11.2 Die Analyseergebnisse

Objekt	Länge	Links	Rechts	Unten	Oben	Diagonal	Cluster	Ausreißer	NeuCluster	Distanzmaße
70	214.9	130.2	130.2	8	11.2	139.6	1	0	2	2.60181496
104	215	130.4	130.6	9.9	10.9	140.3	2	0	1	1.90266932
110	215.2	130.6	130.8	10.4	11.2	140.3	2	0	1	2.34632774
123	215.3	130.8	131.1	11.6	10.6	140.2	2	0	1	3.23190137
125	214.7	130.5	130.5	9.9	10.3	140.1	2	0	1	2.35299985

Da wir wissen, dass die ersten 100 Banknoten echt sind und die zweiten 100 Banknoten gefälscht sind, können wir die Ergebnisse der vorangestellten Tabelle verifizieren. Diese fünf Banknoten wurden vom Ward-Algorithmus falsch klassifiziert, weil Objekte mit den Nummern 1-100 echte Banknoten sind und in Cluster 2 der echten Banknoten gehören, Objekte mit den Nummern 101-200 sind gefälschte Banknoten und gehören in Cluster 1. Hier zeigt sich wie wichtig die Verbesserung durch das partitionierende Verfahren k-Means ist. Die Spalte NeuCluster zeigt richtige Zuordnungen nach Partitionierung. In der Spalte NeuCluster sind alle Zuordnungen der gesamten Tabelle richtig.

Clustering mittels Ward- (und k-Mean-Algorithmus)			
Bedingungen für Ward-Algorithmus			
Distanzmaß:	Mahalanobis		
Ausreißertest:	MD²/Chi²		
alpha:	0.99		
Cluster-Suche (CH):	2 .. 10		
k*(optimal):	2		
Anzahl Variabler p:	6		
n (original):	200		
n (ohne Outlier):	191		
Rechenzeit (s):	0.051		
Partitionierung mit k-Means nach Ward-Resultaten			
max. Iterationen:	100	**Hotellings T^2 -Test**	
Toleranz:	0.0000001	T^2:	2290.9896
Iterationen:	2	p-Wert:	0
Umklassifikationen:	6	H0:	verwerfen
SSE (Start):	860.395914		
SSE (Ende):	852.223332		

Der erste Auswertungsblock zeigt 9 Ausreißer = n(Original) – n(ohne Outlier), dies ist nicht verwunderlich. So haben die echten Noten einige Gebrauchsspuren (Knitterfalten), welche zu Messfehlern geführt haben können. Bei den gefälschten Banknoten ist keines Wegs sicher, dass alle von dem gleichen Fälscher stammen. Auch könnte es sich um Produktionsfehler handeln.

Im zweiten Auswertungsblock können wir die erreichte Verbesserung durch die Partitionierung anhand der SSE beurteilen. Nach der Partitionierung ist die Quadratsumme SSE niedriger. Der Hotellings T^2-Test ist hoch signifikant, eine Vorsetzung für eine gute Trennung der Cluster. Sollte das Ergebnis die Nullhypothese bestätigen ist eine Trennung der Cluster nicht mehr möglich. Der Test bezieht sich nicht auf die Cluster, sondern deren Mittelwerte.

Detaillierte CH-Tabelle			
k	**Wk (Within)**	**Bk (Between)**	**CH(k)**
2	860.395914	168.630496	37.0424396
3	759.184548	269.841862	33.4110264
4	675.515135	353.511275	32.6203441
5	616.059966	412.966444	31.1705689
6	569.168697	459.857713	29.8940112
7	527.001905	502.024505	29.2132116
8	497.920447	531.105962	27.8852322
9	471.160544	557.865866	26.9365689
10	447.662791	581.363619	26.117579

Die CH-Tabelle zeigt den größten Wert für zwei Cluster an, deswegen macht es keinen Sinn die Anzahl der Cluster zu erhöhen, weil der CH-Test objektiv und sehr gut ist. Die nächsten Ausgabeblöcke Zentren und Kovarianzen sind wichtig, aber auch abstrakt. Ihre Resultate können besser anhand der Grafiken interpretiert werden.

Ausreißeranalyse									
Zentren:	**Global**	**Cluster 1**	**Cluster 2**	**Länge**	**Links**	**Rechts**	**Unten**	**Oben**	**Diagonal**
Ausreißer	**4.100230948**			**2.326347874**					
1	5.09817595	5.2794076	5.09626739	-0.2643259	**2.60331779**	1.509581	**-2.44844415**	**-3.22609479**	0.30315069
5	4.20944597	4.55276557	4.07237412	0.28075088	-1.55943565	0.6941943	1.6189848	**-3.36629693**	-0.89600602
13	4.24414475	4.58085724	4.11265461	0.82582766	1.75678494	**-3.46178145**	-1.07145657	-0.20616753	0.89502598
40	5.58409364	5.86958555	5.45783195	**-2.71717143**	1.20000895	**-4.32579614**	-1.18832187	-1.32107671	0.29387503
160	4.20007962	4.1100651	4.49695555	-0.80940269	1.58831998	-1.02512626	-1.96275037	1.20986027	**-2.73535776**
161	4.78604835	4.81133839	4.95053205	1.91598123	-0.52801214	-0.80659127	-0.68127851	**2.64543271**	**-3.23633721**
167	5.20433227	5.19942201	5.38305918	**3.82374998**	0.72325742	0.03226256	1.00646277	-0.67952039	**-3.17481837**
171	5.34099824	5.23588497	5.609648	**-2.71717143**	**2.35062034**	0.06803528	**-2.72176281**	-0.4728666	**-2.69344738**
180	4.50732843	4.40594949	4.80124092	-1.08194108	0.50551538	1.36217418	-2.03702742	0.72108263	**-3.43930857**

Die letzte Ausgabe ist die Ausreißeranalyse. In der zweiten Zeile sind die verschiedenen Zentren benannt. Die dritte Zeile definiert den Grenzwert für einen Ausreißer von den jeweiligen Zentren. Ab der dritten Zeile werden für jeden Ausreißer die Abweichungen zum jeweiligen Zentrum ausgegeben. Für die Variablen werden signifikante Abweichungen im Fettdruck ausgegeben, dies gibt uns einen Hinweis, welche Variable voraussichtlich für den Ausreißer verantwortlich ist. Auffällig ist, dass alle Ausreißer des Cluster 1 signifikante Abweichungen für die Diagonale produzieren.

Letztendlich kann nur eine Identifikationsanalyse eine besondere Art der Diskriminanzanalyse die Frage, welche Variable für die Abweichungen verantwortlich ist, beantworten. Die Analyse wird in einem folgenden Kapitel zur Diskriminanz- und Identifikationsanalyse fortgesetzt.

Im nächsten Schritt werden wir alle 15 möglichen 2D-Scatterplots darstellen.

Bild: Länge vs. Links

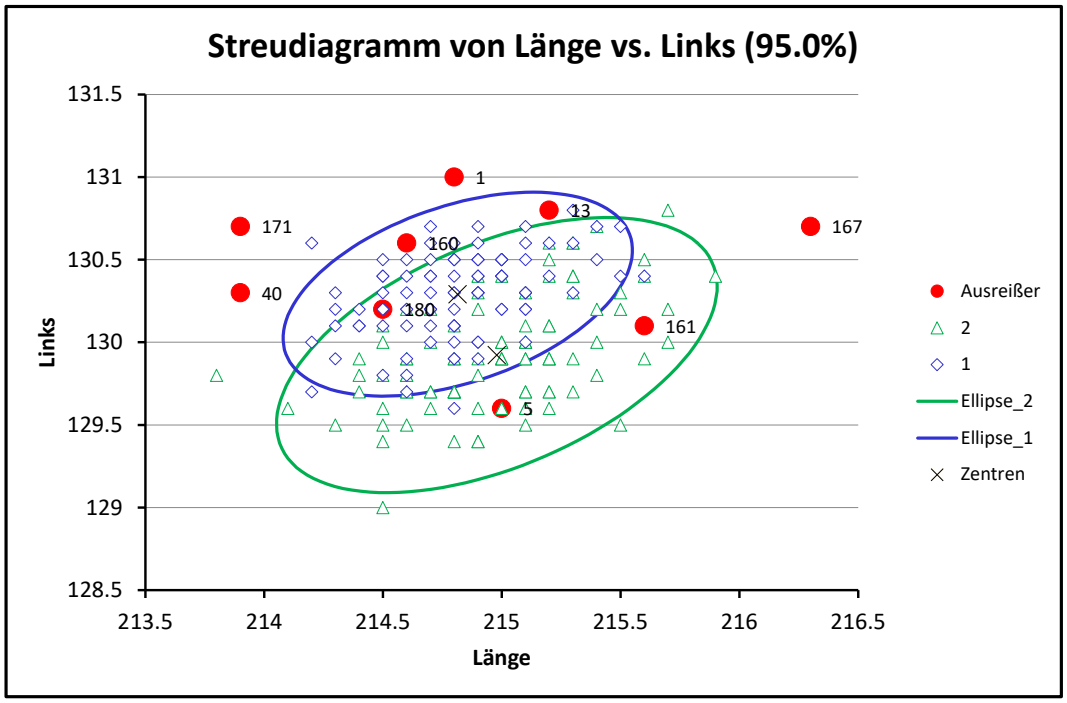


Bild: Länge vs. Rechts

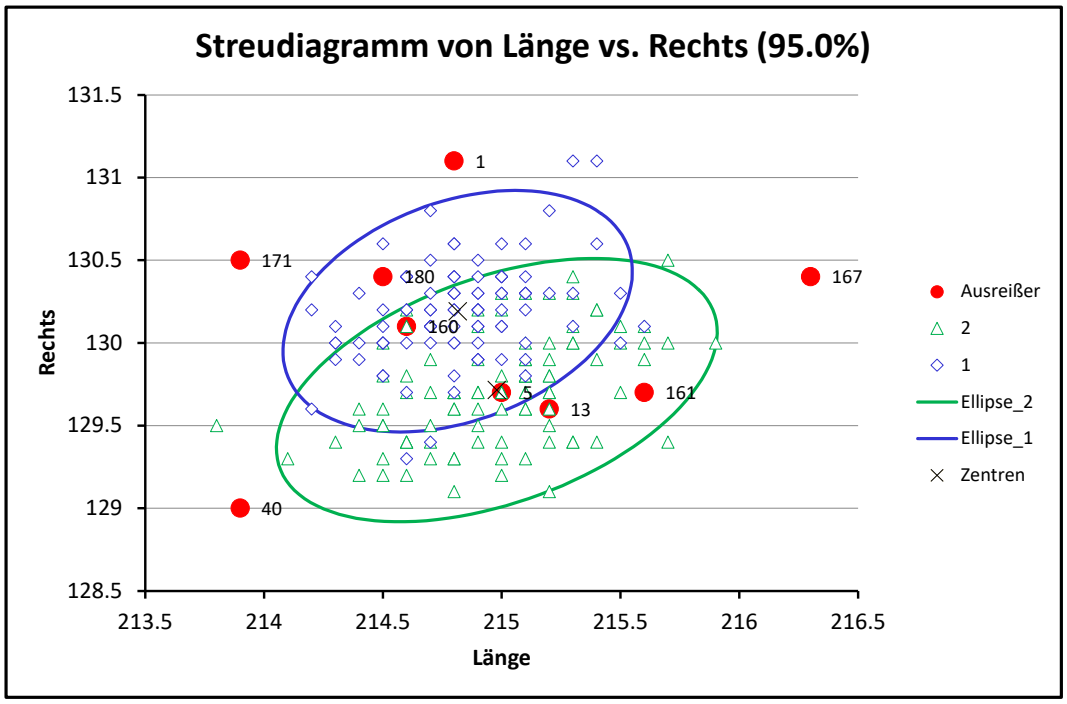


Bild: Länge vs. Unten

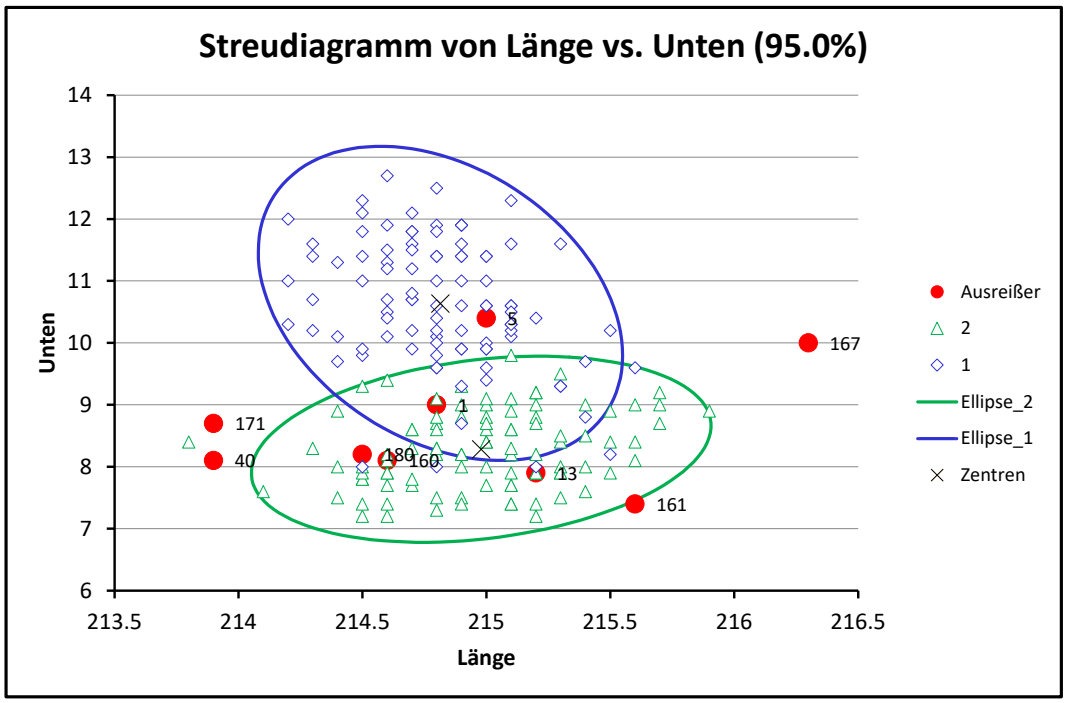


Bild: Länge vs. Oben

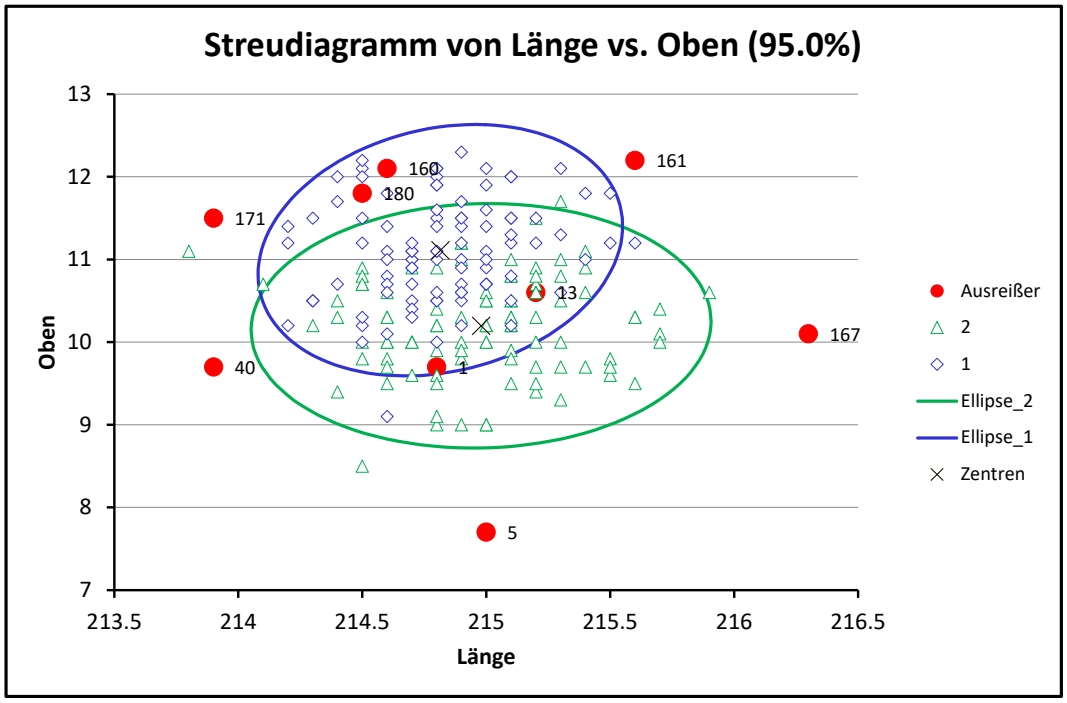


Bild: Länge vs. Diagonale

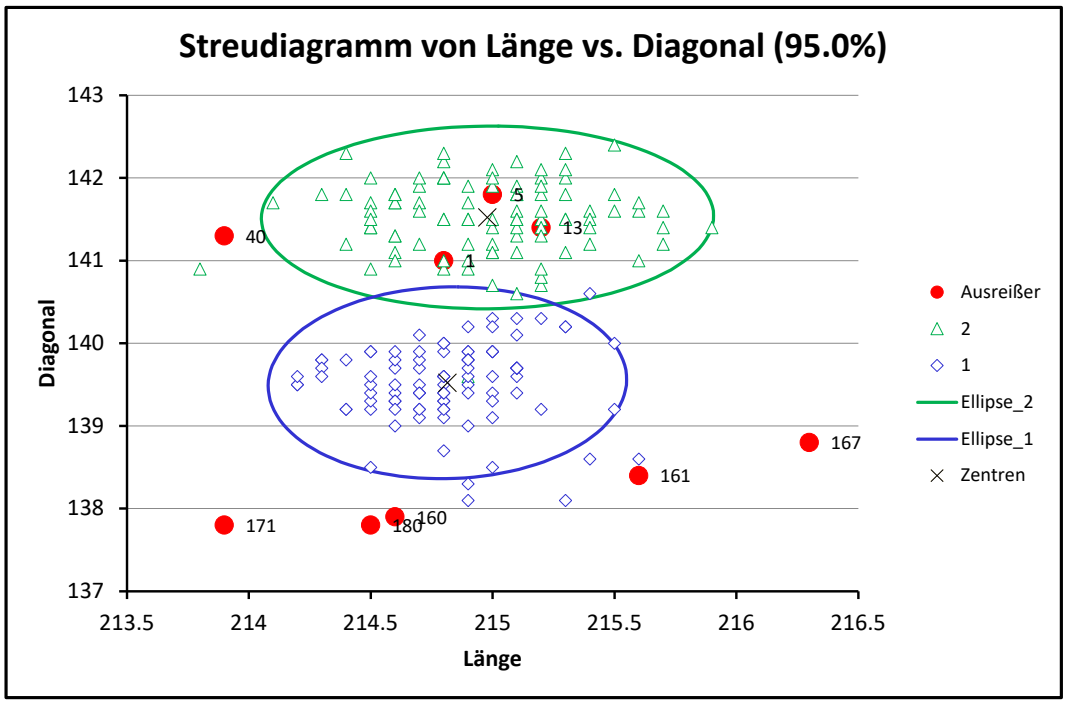


Bild: Links vs. Rechts

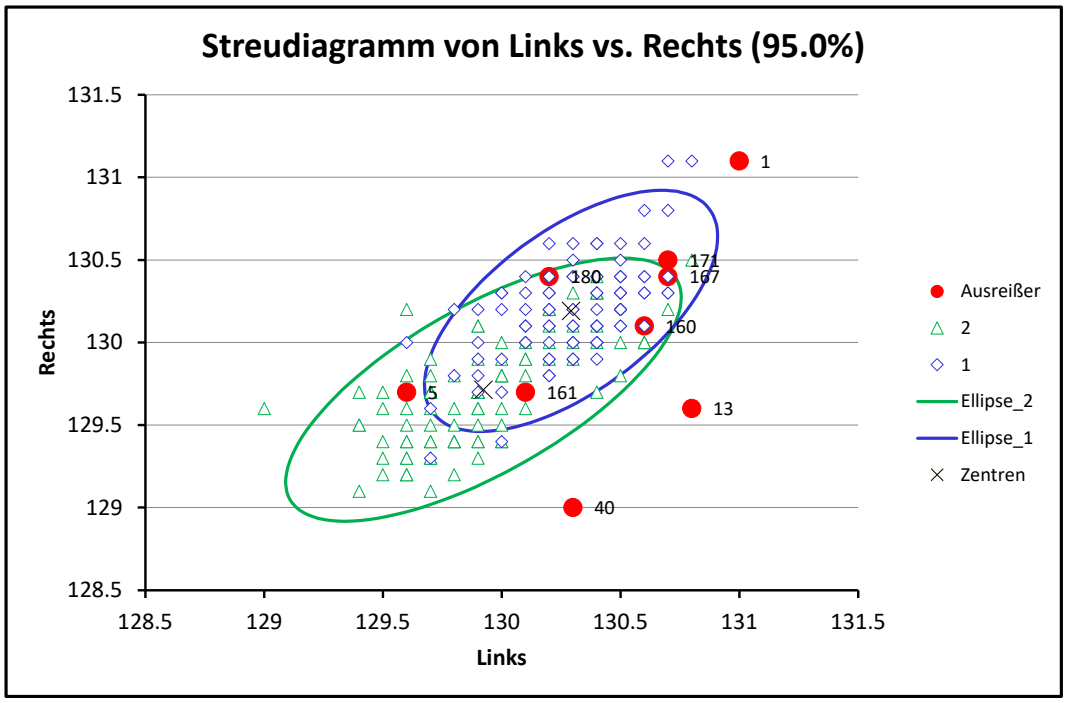


Bild: Links vs. Unten

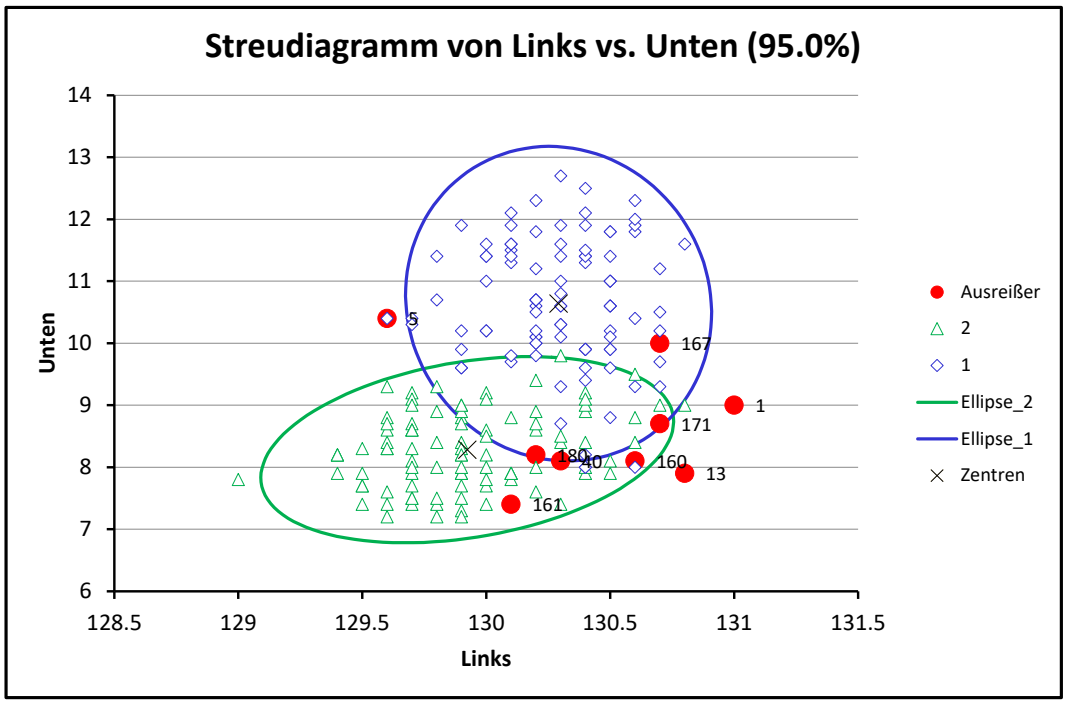


Bild: Links vs. Oben

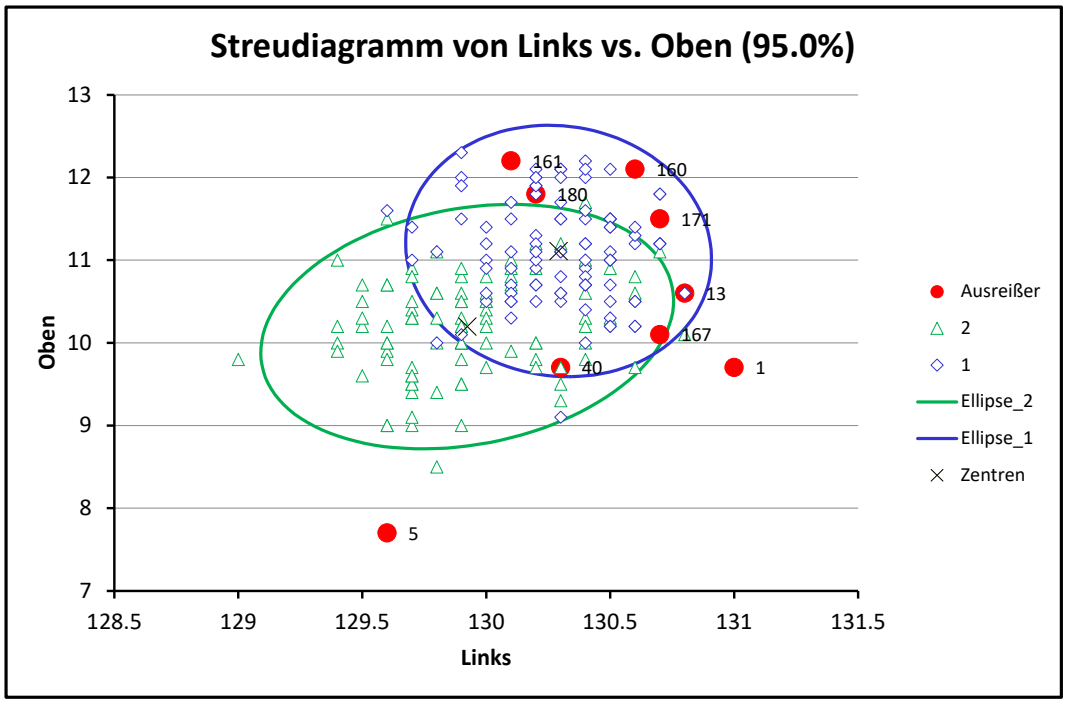


Bild: Links vs. Diagonal

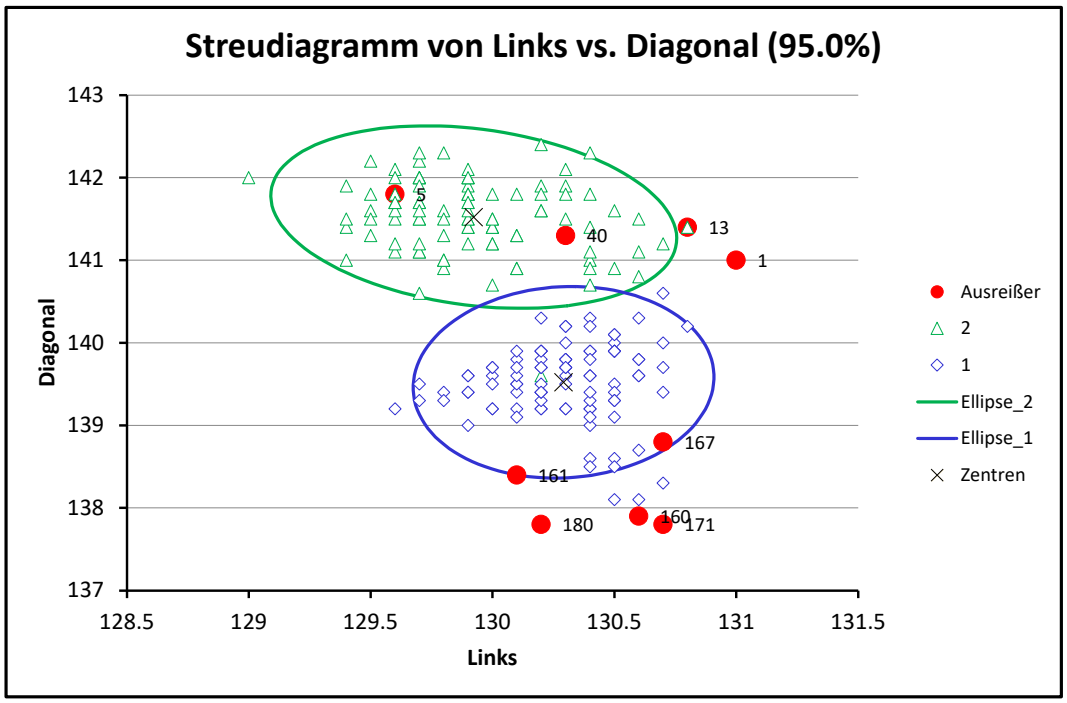


Bild: Rechts vs. Unten

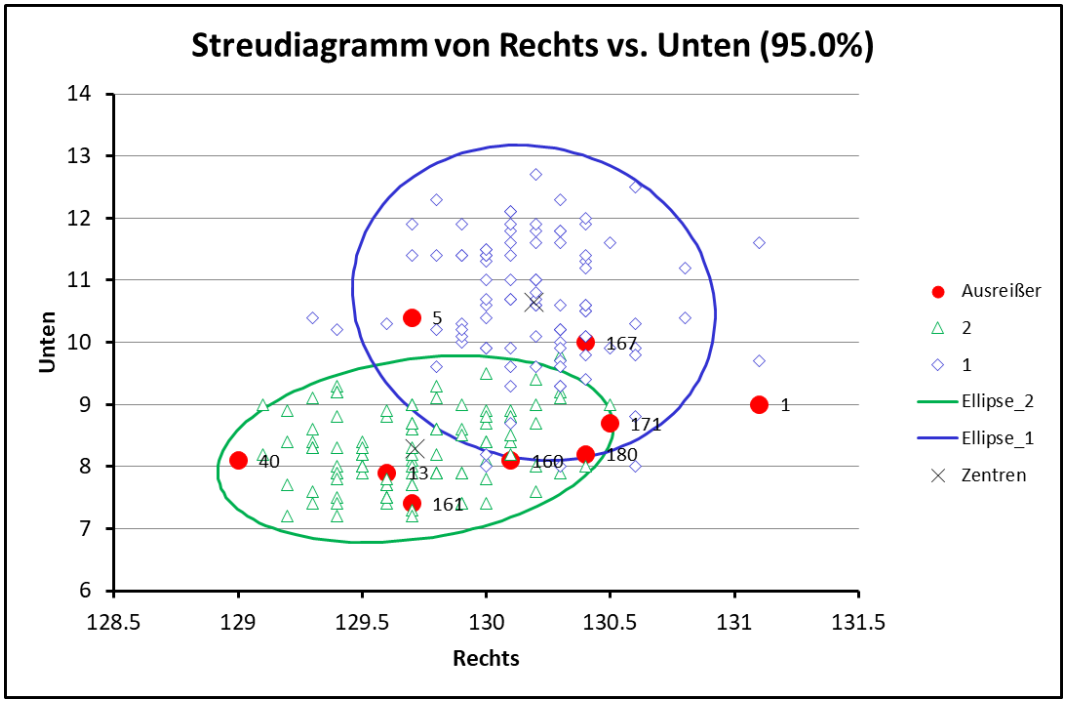


Bild: Rechts vs. Oben

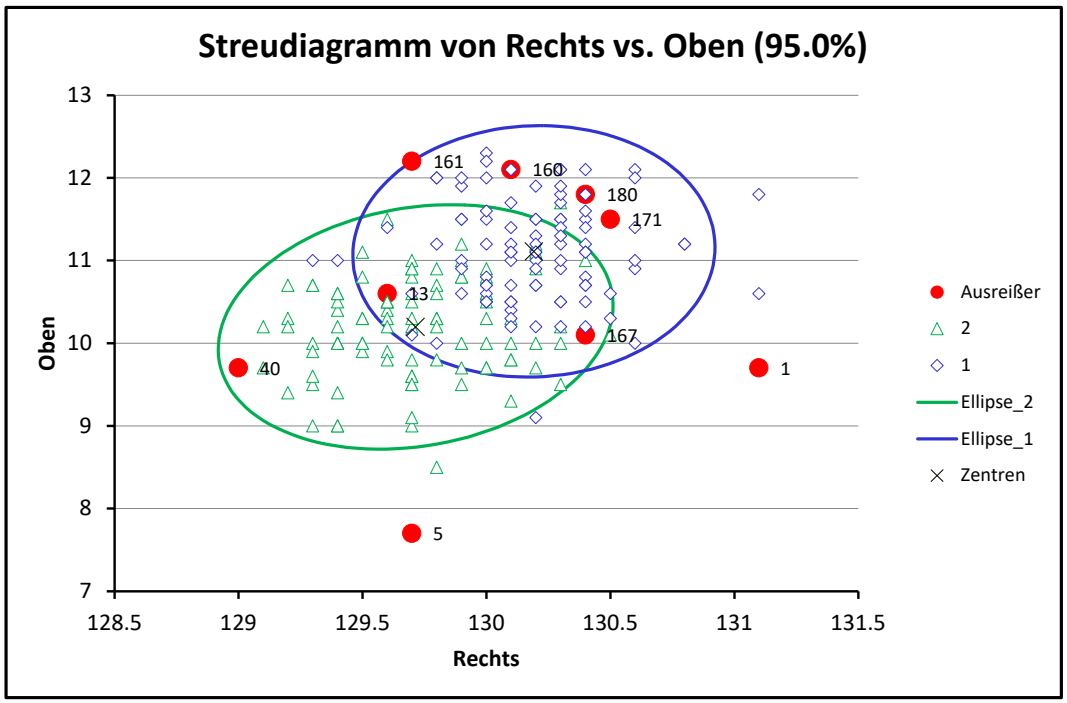


Bild: Rechts vs. Diagonal

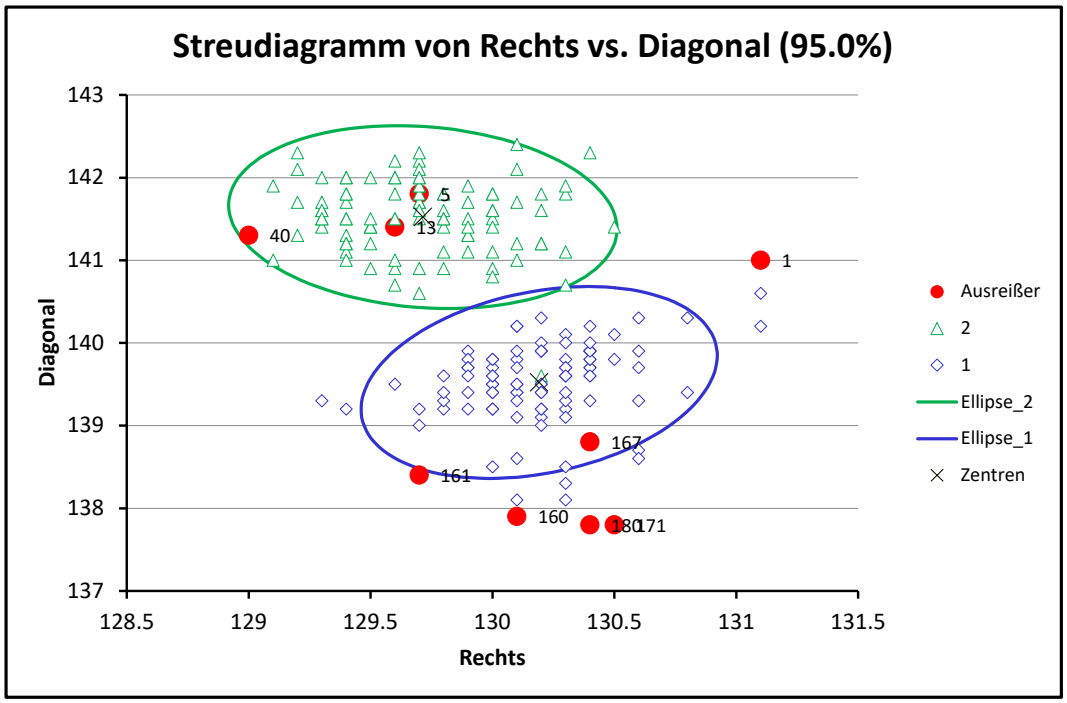


Bild: Unten vs. Oben

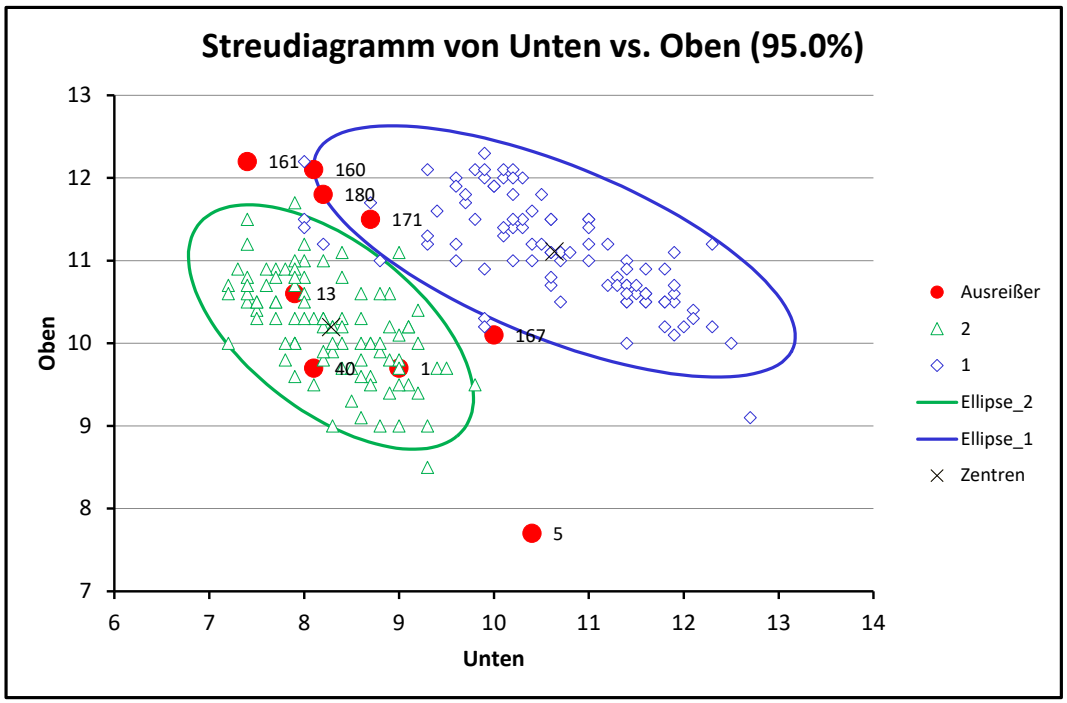


Bild: Unten vs. Diagonale

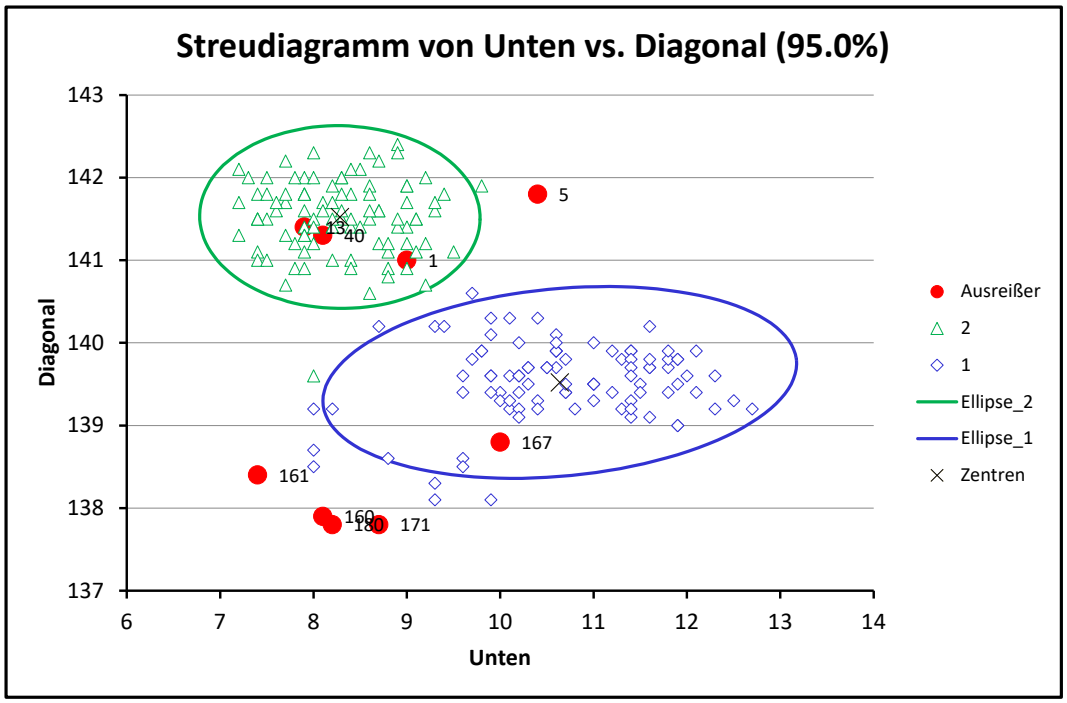
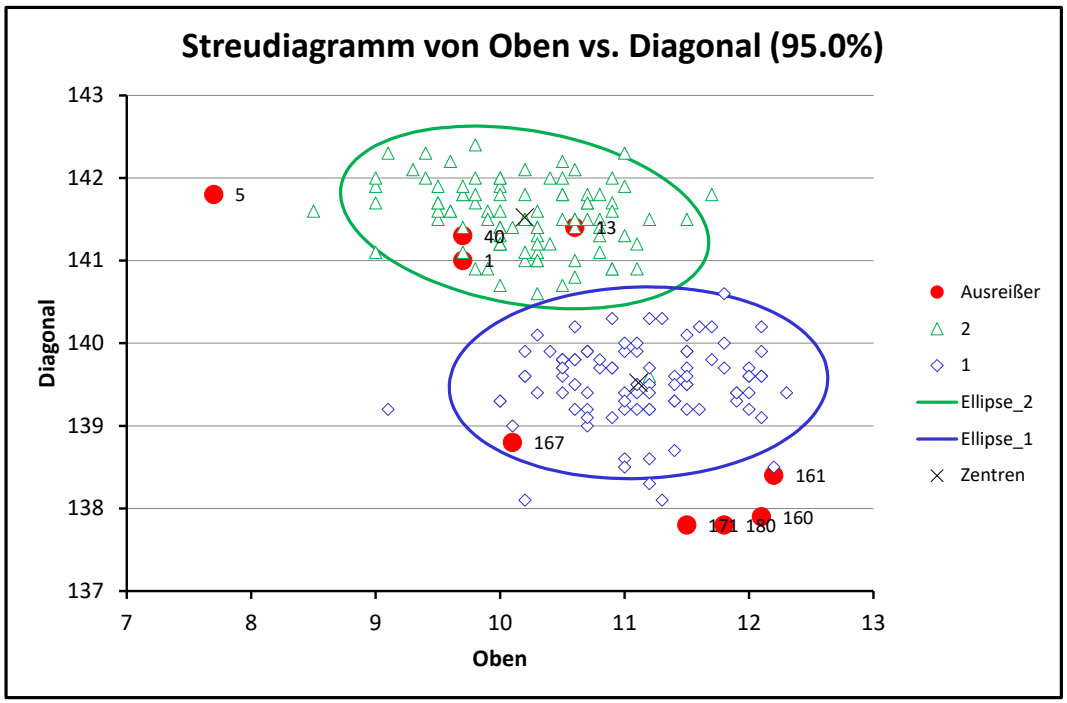


Bild: Oben vs. Diagonale



Die Interpretation 2D-Scatterplots ist nicht einfach, versucht man doch einen mehrdimensionalen Raum anhand verschiedener zweidimensionaler Räume darzustellen. Doch deuten große Abweichungen von Ausreißern zu den Ellipsen auf die verantwortlichen Variablen hin. Auffällig ist auch, das die Größen der Ellipsen (ein Maß für die Streuung) für den Cluster 2 (echte Banknoten) meist größer ist als die von Cluster 1 (gefälschte Banknoten). Dies ist wohl die Gebrauchsspuren der echten Banknoten und der damit einhergehenden Messunsicherheit zurückzuführen. Außerdem ist erkennbar, welche Variablen einen großen Einfluss auf die Trennfunktion der Cluster haben. Wenn die Ellipsen nur eine kleine oder gar keine Schnittmenge ergeben ist der Varableneinfluss für die Trennfunktion groß, umgekehrt gilt, dass bei großen Schnittmengen kein oder nur ein geringer Einfluss festgestellt werden kann. Demnach ist der Einfluss der Länge gering, Links und Rechts haben einen Einfluss, den größten Einfluss kann man für die Variablen Unten, Oben und Diagonale erwarten.

Da die Clusteranalyse ein strukturbildendes Verfahren ist, sind fast alle Aussagen nur Vermutungen und Hypothesen. Damit die Hypothesen geprüft werden können bedarf es strukturprüfender Verfahren. Für die Clusterprüfung und Trennfunktion ist dies die Diskriminanzanalyse, welche wir in OQM-Stat im wichtigen Zweigruppenfall mittels multipler Regressionsanalyse durchführen können. Dazu sind die Daten entsprechend aufzubereiten. Für Prüfung neuer Objekte oder Ausreißer liefert die Identifikationsprüfung entsprechende Ergebnisse.

Deshalb ist der nächste Schritt, die Diskriminanzanalyse und Identifikationsanalyse im Zweigruppenfall darzulegen und ihre Anwendung am Beispiel der Banknoten zu demonstrieren.

2 Diskriminanz- und Identifikationsanalyse im Zweigruppenfall

Die Diskriminanzanalyse dient der strukturprüfenden Auswertung bereits gebildeter Gruppen (Cluster) oder bestehender Gruppen. Im Zweigruppenfall wird untersucht, ob sich zwei Gruppen statistisch trennen lassen und wie gut diese Trennung ist. Außerdem wird die Frage beantwortet, wie gut neue Beobachtungen einer Gruppe zugeordnet werden können, und mit welcher Wahrscheinlichkeit eine Beobachtung zu Gruppe 1 oder Gruppe 2 gehört.

Die Identifikationsanalyse bei der ein Referenzcluster gegen eine neue Beobachtung (Objekt) geprüft wird. Auch können Ausreißer verifiziert werden. Die Identifikationsanalyse ist eine spezielle Diskrimianzanlyse im Zweigruppenfall, wobei der Referenzcluster unter beliebigen Clustern ausgewählt werden.

In OQM-Stat wird die Diskriminanzanalyse im Zweigruppenfall nicht über klassische Fisher-Diskriminanzfunktionen, sondern über eine äquivalente Formulierung mittels multipler Regression realisiert. Dies erlaubt eine einheitliche mathematische Behandlung und einen klaren probabilistischen Zugang.

2.1 Datenaufbereitung

Aus der Clusteranalyse werden folgende Spalten verwendet:

- Objekt (ID)
- Variablen: $x_1, x_2, \ldots, x_p$
- Clusterzuordnung (Cluster oder NeuCluster)
- Ausreißerkennzeichnung

Die Objekt (ID) ist notwendig um die Objekte nach der Diskiminanzanalyse bzgl. Ausreißern und großen Residuen, sowie großen Einflüssen auf die Diskriminanzfunktion zu bewerten. Da die multiple Regression eine eigene ID erzeugt, muss die Objekt-Spalte kopiert und in die Ausgabe der multiplen Regression manuell eingefügt werden. Dies ist nicht zwingend erforderlich, macht aber einfacher die Objekte zu identifizieren. Die Variablen werden zur Berechnung der Diskriminanzfunktion benötigt. Dies gilt auch für die Clusterzuordnung die verwendet wird um die Anzahl von Objekten jeden Clusters zu ermitteln und aus diesen die Diskriminanzwerte der Gruppen nach der Fisher-Codierung festzulegen. Die Spalte der Ausreißer wird zur Bereinigung des Datensatzes benötigt. Vor der Diskriminanzanalyse erfolgt automatisch:

- Entfernung aller Ausreißer
- Reduktion auf genau zwei Cluster

Erzeugung einer neuen Datenmatrix mit:

- der Objekt (ID)
- den Variablen $x_1, \ldots, x_p$
- einer zusätzlichen Diskriminanzvariable

Die Datenmatrix aus Variablen plus Diskriminanzvariable kann natürlich auch direkt in Excel erstellt werden, eine zuvor durchgeführte Clusteranalyse ist nur erforderlich, wenn keine zwei

Gruppen existieren. Zur Regression wird eine künstliche Zielvariable D definiert. Standardmäßig wird die Fisher-Codierung verwendet: Für zwei Gruppen mit Umfängen n_1 und n_2:

$$D = \begin{cases} c_1 = \dfrac{n_2}{n_1 + n_2} & \text{für Gruppe 1} \\ c_2 = \dfrac{-n_1}{n_1 + n_2} & \text{für Gruppe 2} \end{cases}$$

Eigenschaften dieser Codierung:

- gewichtete Zentrierung bei ungleichen Gruppengrößen
- bei $n_1 = n_2$: symmetrisch um 0
- Summe aller Diskriminanzwerte = 0
- Addition der Beträge von den Gruppenmittelwerten ergibt das Bestimmtheitsmaß R^2
- maximiert die Trennschärfe (äquivalent zur Fisher-Diskriminanz)

Alternativ können auch einfache Codierungen (z. B. 1 und 2) verwendet werden, jedoch ist die Fisher-Codierung statistisch optimal.

2.2 Trennfunktion (Diskriminanzfunktion)

Der Datensatz wird in zwei Stufen, jeweils mit Kopfzeile eingelesen, zuerst die Variablen gefolgt von der Spalte in der die Diskriminanzwerte D stehen. Die multiple Regression mit D als abhängiger Variabler liefert die Diskriminanzfunktion:

$$D = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$$

mit: b_0 = Achsenabschnitt der Diskriminanzfunktion
b_i = Diskriminanzkoeffizienten

Diese Funktion ist die Trennfunktion zwischen den beiden Gruppen. Jede Beobachtung erhält einen Diskriminanzwert d_{ij}, diese Werte stehen in der Spalte „geschätzt“. Als kritischer Schwellenwert dient der Mittelpunkt zwischen den Gruppenzentren der Diskriminanzwerte:

$$D_{krit} = \frac{\bar{D}_1 + \bar{D}_2}{2}$$

Zuordnung: wenn $d_{ij} > D_{\text{krit}} \rightarrow$ Gruppe 1
wenn $d_{ij} \leq D_{krit} \rightarrow$ Gruppe 2

Der Wert Dkrit wird auch als Trennmaß TM bezeichnet. Dies entspricht der linearen Entscheidungsgrenze. Für jede Beobachtung wird geprüft: die tatsächliche Gruppenzugehörigkeit, die zugeordnete Gruppe nach Trennfunktion. Daraus entsteht eine Konfusionsmatrix:

	vorhergesagt für	
	Gruppe1	Gruppe2
Gruppe1:	richtig	falsch
Gruppe2:	falsch	richtig

Daraus ergeben sich die Kennzahlen für die Trefferquote, die Fehlklassifikationsrate und die Sensitivität / Spezifität. Die Diskriminanzwerte der beiden Gruppen werden als normalverteilt angenommen:

$$D \sim N(\mu_1, \sigma_1^2), \quad D \sim N(\mu_2, \sigma_2^2)$$

Für einen neuen Wert d werden berechnet:

$$P(G_1 \mid d) = \frac{f_1(d)}{f_1(d) + f_2(d)}, \quad P(G_2 \mid d) = \frac{f_2(d)}{f_1(d) + f_2(d)}$$

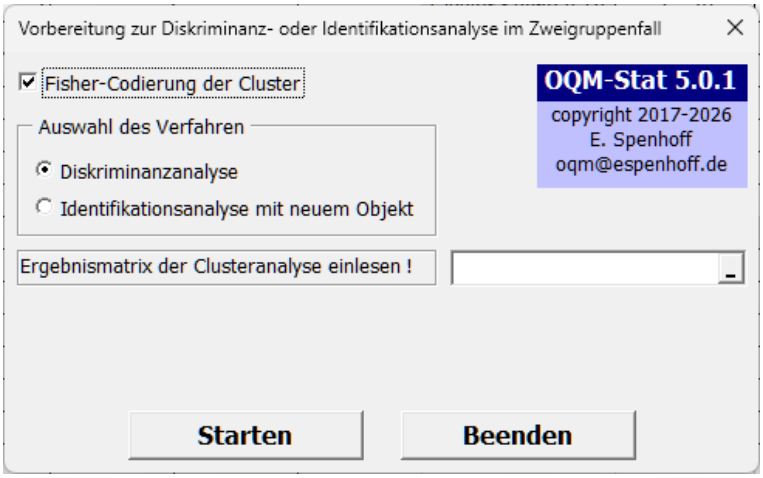
mit den Dichtefunktionen f_1, f_2. Die Zuordnung erfolgt nach der höheren Wahrscheinlichkeit.

2.2.1 Analyseaufbereitung der Banknoten

Ausgehend von den Ergebnissen der Clusteranalyse, welche wir als Auszug darstellen:

Objekt	Länge	Links	Rechts	Unten	Oben	Diagonal	Cluster	Ausreißer	NeuCluster
1	214.8	131	131.1	9	9.7	141		1	
2	214.6	129.7	129.7	8.1	9.5	141.7	2	0	2
3	214.8	129.7	129.7	8.7	9.6	142.2	2	0	2
4	214.8	129.7	129.6	7.5	10.4	142	2	0	2
5	215	129.6	129.7	10.4	7.7	141.8		1	
6	215.7	130.8	130.5	9	10.1	141.4	2	0	2

Nun können wir einen Menüpunkt „Prep. Diskriminanzanalyse“ in OQM-Stat aufrufen und müssen die gezeigten Daten komplett einlesen. Das Menü zeigt die folgende Grafik:



Nach dem Einlesen der Daten und dem Drücken des Startbutton erhalten wir eine bereinigte Datei, welche für die Diskriminanzanalyse (multiple Regression) genutzt werden kann. Das Ergebnis zeigt folgende Tabelle (Auszug).

Objekt	Länge	Links	Rechts	Unten	Oben	Diagonal	D
2	214.6	129.7	129.7	8.1	9.5	141.7	-0.5
3	214.8	129.7	129.7	8.7	9.6	142.2	-0.5
4	214.8	129.7	129.6	7.5	10.4	142	-0.5
6	215.7	130.8	130.5	9	10.1	141.4	-0.5

2.2.2 Die Diskriminanzanalyse der Banknoten

Die multiple Regression bestimmt Koeffizienten b_0, b_1, …, b_p durch Minimierung der Fehlerquadratsumme:

$$\sum_{i=1}^{n}\left(D_i - \hat{D}_i\right)^2$$

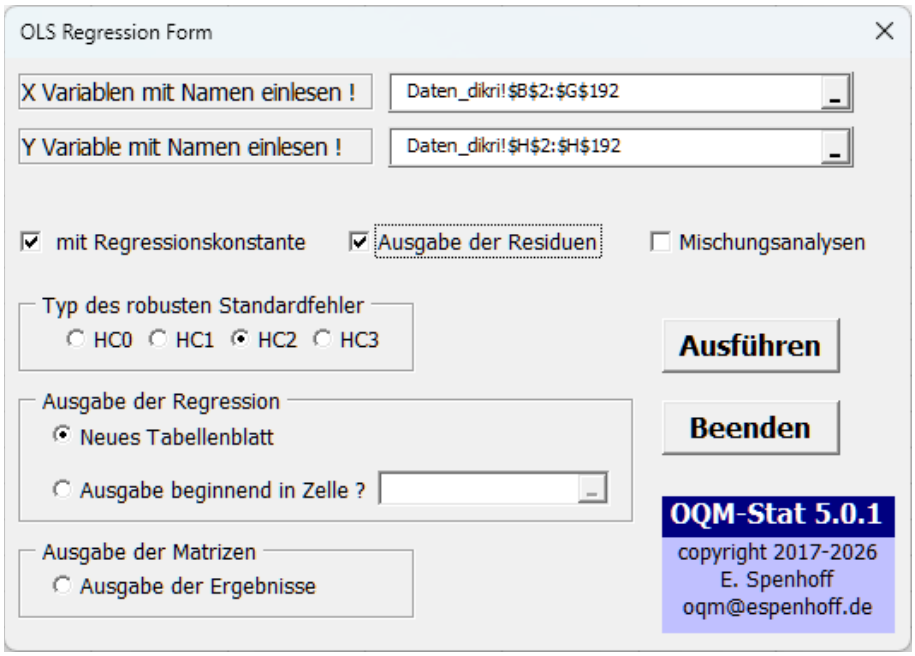
mit

$$\hat{D}_i = b_0 + \sum_{j=1}^{p} b_j x_{ij}$$

Dies ist äquivalent zur Lösung des Normalgleichungssystems:

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{D}$$

Wir lesen die aufbereiteten Daten im Menü der multiplen Regression ein und dabei müssen die Checkboxen „mit Regressionkonstante“ und „Ausgabe der Residuen“ aktiviert sein. Mit Anklicken des Button „Ausführen“ startet die Diskriminanzanalyse.



Wir kopieren die Objektspalte aus der bereinigten Datei und fügen diese Werte in die Spalte „Nr.“ der Residuenausgabe ein. Das Ergebnis zeigt der Auszug folgender Ergebnisse:

Objekt	beobachtet	geschätzt	Residual	Std.Residual	Hebel	stud. Res	del. Res	Cooks D	Dffits	MD^2
2	-0.50000000	-0.58977805	0.08977805	0.68525033	0.03127148	0.69622275	0.69523927	0.00223535	0.12491296	4.91557324
3	-0.50000000	-0.60445649	0.10445649	0.79728670	0.02452945	0.80724886	0.80647742	0.00234094	0.12788778	3.64132922
4	-0.50000000	-0.62779025	0.12779025	0.97538663	0.01999352	0.98528603	0.98520697	0.00282935	0.14072058	2.78403826
6	-0.50000000	-0.34814464	-0.15185536	-1.15906882	0.05838823	-1.19446460	-1.19586743	0.01263869	-0.29778976	10.04063897
7	-0.50000000	-0.58521209	0.08521209	0.65039964	0.06639552	0.67312982	0.67212074	0.00460337	0.17924012	11.55401647
8	-0.50000000	-0.60584831	0.10584831	0.80791010	0.03877398	0.82404377	0.82331814	0.00391307	0.16535796	6.33354511
9	-0.50000000	-0.37332376	-0.12667624	-0.96688373	0.05383367	-0.99400943	-0.99397681	0.00803100	-0.23709341	9.17982585
10	-0.50000000	-0.15603273	-0.34396727	-2.62540444	0.02559293	-2.65965927	-2.70518157	0.02654199	-0.43841580	3.84232658
11	-0.50000000	-0.32513161	-0.17486839	-1.33472069	0.05768102	-1.37496431	-1.37834056	0.01653179	-0.34101544	9.90697621
12	-0.50000000	-0.61337049	0.11337049	0.86532475	0.03523127	0.88098298	0.88044165	0.00404895	0.16824925	5.66397357

Die Spalten „Objekt, beobachtet und geschätzt“ sind die wichtigsten Spalten, auch für die spezielle Ausgabe der Diskriminanzanalyse. Alle weiteren Spalten dienen der Beurteilung einzelner Objekte, die letzte Spalte ist die Malahanobis-Distanz und zeigt, ob irgendwelche Objekte Ausreißer sind. Cook’s D ist die wichtigste Maßzahl zur Bestimmung sogenannter *einflussreicher Beobachtungen*, während die DFFITS den Einfluss auf das angepasste Modell anzeigen. Die hauptsächliche Ausgabe der Analyse entspricht der multiplen Regression.

Regressionsstatistiken für D

			Ursache	SSQ	FG	MSS	F-Stat	p-Wert
Anz. Beob.:	190		TSS	47.50000000	189	0.25132275		
Anz. fehl. Beob.:	0		RSS	44.35881942	6	7.39313657	430.71194364	0.00000000
R^2:	0.93386988		Error	3.14118058	183	0.01716492		
St.Abw. Error:	0.13101497		LoF	---	---	---	---	
AIC_ols:	-765.46081886		pure Error	---	---	---		
BIC_ols:	-742.73165036							

Variable	Koeffizient	SE	SE (HC2)	t_SE	t_HC2	p_SE	p_HC2	VIF
Konstante	27.87833281	6.58563537	7.39205892	4.23320321	3.77138942	**0.00001821**	0.00010950	
Länge	-0.03036790	0.03211821	0.03685717	-0.94550388	-0.82393452	0.17282421	0.20552465	1.41210361
Links	-0.11884019	0.04583350	0.04868530	-2.59286771	-2.44098696	**0.00514366**	0.00779953	2.89433963
Rechts	0.16244973	0.04084699	0.04596566	3.97703042	3.53415424	**0.00005020**	0.00025901	2.84171583
Unten	0.13846782	0.01070477	0.01137715	12.93514746	12.17069946	**0.00000000**	0.00000000	2.67613178
Oben	0.14571235	0.01738523	0.01765932	8.38138941	8.25129882	**0.00000000**	0.00000000	1.96535335
Diagonal	-0.21251358	0.01592547	0.01681256	-13.34425969	-12.64017102	**0.00000000**	0.00000000	3.51021795

Prüfung auf Heteroskedastizität			Prüfung auf Normalität		
BP-Test:	1.31929610	0.25071842	AD-Test:	0.44568399	0.28266382

Die Diskriminanzfunktion ergibt sich aus der multiplen Regression:

für das Beispiel gilt:

$$\hat{D} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_4 x_4 + b_5 x_5 + b_6 x_6$$

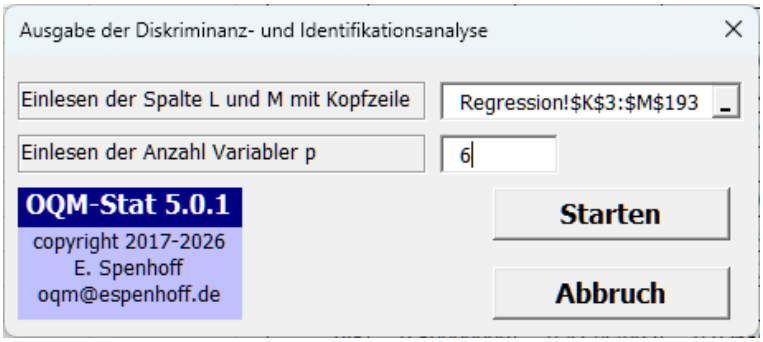
und allgemein gilt:

$$\hat{D} = b_0 + \sum_{j=1}^{p} b_j x_j$$

Die Koeffizienten *bj* werden aus den Banknotendaten geschätzt.

Jede Variable trägt mit einem Gewicht zur Trennung bei. Große Beträge von b_j zeigen besonders trennwirksame Merkmale. In der Praxis erweist sich häufig die Bilddiagonale x_6 als dominant.

Rechts in der ANOVA sehen wir, dass die Nullhypothese abgelehnt wird. Wir haben eine gesicherte Diskriminanzfunktion. Das Bestimmtheitsmaß R^2 ist hoch. Die Varianzen der Residuen sind homogen und die Residuen normalverteilt. Die Variable „Länge“ ist nicht signifikant, trägt zur Trennung der Cluster nur wenig bei. Alle anderen Variablen sind für die Trennung der Cluster bedeutsam. Zur spezifischen Ausgabe der Diskriminanzanalyse müssen wir ein weiteres Menü aufrufen.



Nach der Eingabe der Spalten „Objekt, beobachtet und geschätzt“ sowie der Anzahl Variabler p, starten wir die spezielle Ausgabe der Diskriminanzanalyse.

Diskriminanzanalyse für D

	Gruppe1	Gruppe2	gezählte Zuordnung			wahrscheinliche Zuordnung		
Gruppenwerte D:	-0.5	0.5		richtig	falsch		D > TM	D < TM
Priors:	0.5	0.5	Gruppe1:	95	0	Gruppe1:	9.27248E-05	0.999907275
Trennmaß:	**0.00000000**		Gruppe2:	95	0	Gruppe2:	0.999907275	9.27248E-05

Name	Umfang	Mittelwert	Varianz	St.Abw.				
Gruppe1:	95	-0.466934941	0.01785787	0.133633342				
Gruppe2:	95	0.466934941	0.013349087	0.115538248				
Summe:	190	0	0.015603478	0.124913884				
F-Test:	F_ratio:	1.337759682	F_tab (.95):	1.406395056	nicht signifikant			

In diesem ersten Teil der Ausgabe ist das Trennmaß TM berechnet, es werden die Zuordnung gezählt und auf Basis einer Normalverteilung der Anteil der richtigen bzw. falschen Zuordnungen berechnet. Das Trennmaß berechnet sich für den Fall gleicher Varianzen nach

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$$TM = \frac{m_1 + m_2}{2} + \frac{s_p^2}{m_1 - m_2} \ln\left(\frac{\pi_1}{\pi_2}\right)$$

und für den Fall ungleicher Varianzen nach

$$a = \frac{1}{\sigma_2^2} - \frac{1}{\sigma_1^2}$$

$$b = -2\frac{\mu_2}{\sigma_2^2} + 2\frac{\mu_1}{\sigma_1^2}$$

$$c = \frac{\mu_2^2}{\sigma_2^2} - \frac{\mu_1^2}{\sigma_1^2} - 2ln\left(\frac{\pi_2\sigma_1}{\pi_1\sigma_2}\right)$$

$$x1,2 = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \quad \text{TM der Wert zwischen } \mu_1 \text{ und } \mu_2$$

Deswegen wird der F-Test mit den Varianzen durchgeführt, um im günstigen Fall die Gleichheit der Varianzen annehmen zu dürfen und die einfachen Formel zu nutzen.

Der zweite Teil (Auszug) ist die wahrscheinliche Zuordnung einzelner Objekte zu den Clustern.

Zuordnung nach Wahrscheinlichkeiten (Detail)

Objekt	D beobachtet	D geschätzt	P(G1|Dhat)	P(G2|Dhat)	Zuordnung	richtig?
2	-0.5	-0.58977805	1.00000000	0.00000000	G1	1
3	-0.5	-0.60445649	1.00000000	0.00000000	G1	1
4	-0.5	-0.62779025	1.00000000	0.00000000	G1	1
6	-0.5	-0.34814464	1.00000000	0.00000000	G1	1
7	-0.5	-0.58521209	1.00000000	0.00000000	G1	1
8	-0.5	-0.60584831	1.00000000	0.00000000	G1	1
9	-0.5	-0.37332376	1.00000000	0.00000000	G1	1
10	-0.5	-0.15603273	0.99991204	0.00008796	G1	1

Die Zuordnungen berechnen sich für jede Gruppe:

$$fk\left(\hat{D}\right) = \phi\left(\frac{\hat{D} - \mu_k}{\sigma_k}\right)$$

und dann Bayes:

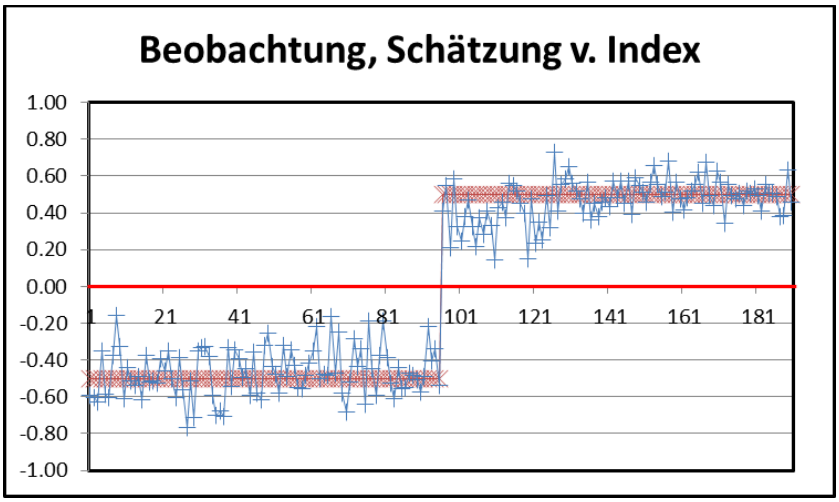
$$P\left(G_k \mid \hat{D}\right) = \frac{f_k\left(\hat{D}\right)\cdot\pi_k}{\sum_j f_j\left(\hat{D}\right)\cdot\pi_j} \text{ mit } \pi_k = \frac{n_k}{n}$$

2.2.3 Zusammenfassung

Die Diskriminanzanalyse beantwortet die folgenden Fragen:

- **Sind die Cluster tatsächlich trennbar?**

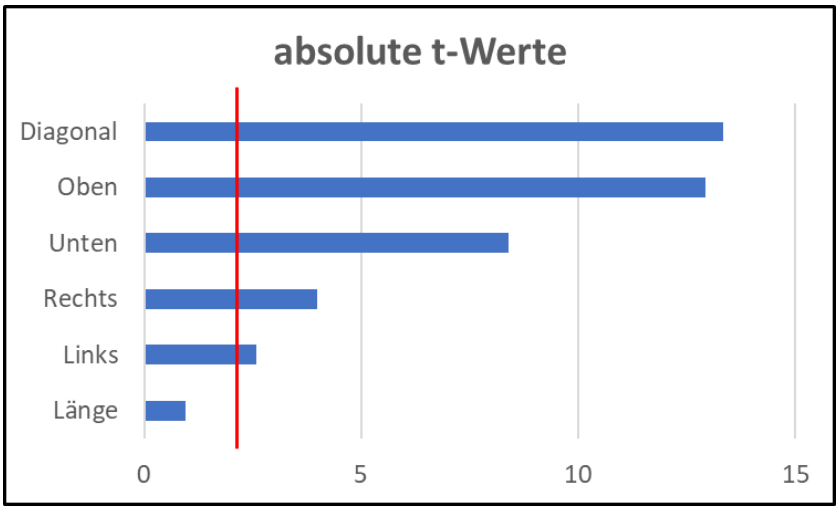
Die folgende Grafik bestätigt das gute Ergebnis des F-Test der ANOVA F=430.71194364 und einem p-Wert = 0.00000000. Dies Ergebnis deckt sich mit dem Hotelling T^2-Test in der Clusteranalyse.



Die Grafik zeigt die perfekte Trennung der Cluster und auch kein einzelner Wert überschreitet das Trennmaß.

- **Welche Variablen tragen zur Trennung bei?**

Eine weitere Grafik mit den t_SE beantwortet diese Frage.



Nur die Variable Länge ist nicht signifikant, alle anderen Variablen tragen zur Trennung der Cluster echte und gefälschte Banknoten bei.

- **Wie zuverlässig ist die Zuordnung?**

Die Tabelle der Zuordnungen zeigt ein eindeutiges Ergebnis: kein Objekt wurde falsch zugeordnet. Aber man sollte berücksichtigen das 10 Objekte als Ausreißer entfernt wurden. Fragen zu diesem Sachverhalt können mit der Identifikationsanalyse und 5 neuen Banknoten beantwortet werden. Eine wirklich sichere Aussage zu diesem Sachverhalt ist nur mit einem zweiten Datensatz, welcher unter gleichen Bedingungen vermessen wurde, möglich.

- **Wie sicher ist die Identifikation neuer Objekte?**

Um diese Frage zu beantworten berechnen wir die Diskriminanzwerte aller zehn Ausreißer plus fünf neuer falscher Banknoten. Das Ergebnis dieser Bewertung:

Objekt	Länge	Links	Rechts	Unten	Oben	Diagonal	D	Banknote	Zuordnung
1	214.8	131	131.1	9	9.7	141	-0.22039125	echt	richtig
5	215	129.6	129.7	10.4	7.7	141.8	-0.55509880	echt	richtig
13	215.2	130.8	129.6	7.9	10.6	141.4	-0.55862387	echt	richtig
40	213.9	130.3	129	8.1	9.7	141.3	-0.63939154	echt	richtig
70	214.9	130.2	130.2	8	11.2	139.6	0.10305909	echt	FALSCH
103	214.9	130.3	130.1	8.7	11.7	140.2	0.11720560	gefälscht	richtig
161	215.6	130.1	129.7	7.4	12.2	138.4	0.33010867	gefälscht	richtig
167	216.3	130.7	130.4	10	10.1	138.8	0.32027679	gefälscht	richtig
171	213.9	130.7	130.5	8.7	11.5	137.8	0.64590742	gefälscht	richtig
180	214.5	130.2	130.4	8.2	11.8	137.8	0.64534161	gefälscht	richtig
201	215.2	130.4	130.1	10.1	11.6	139.8	0.36050035	gefälscht	richtig
202	214.9	130.4	130.3	10.7	11.6	139.6	0.52768407	gefälscht	richtig
203	215.3	130.4	130.4	7.7	12	139.9	0.11090930	gefälscht	richtig
204	215.1	130.6	130.6	9.2	11	138.8	0.42145909	gefälscht	richtig
205	214.9	130.5	130.2	8.4	11.6	138.4	0.42145909	gefälscht	richtig

Bis auf die echte Banknote 70, welche als „gefälscht“ zugeordnet wurde, waren alle anderen Banknoten korrekt klassifiziert. Dieses ist einem standardisierten Residuum von 4.1 zur Mitte des echten Clusters geschuldet. Wahrscheinlich bedingt durch die Gebrauchsspuren und damit ein hergehenden Messfehlern. Eine echte Banknote falsch zu klassifizieren hat keine großen Konsequenzen.

Die Diskriminanzanalyse bildet damit das strukturprüfende Gegenstück zur strukturbildenden Clusteranalyse.

2.3 Identifikationsanalyse (neues Objekt prüfen)

Bei der Identifikationsanalyse wird ein Referenz-Cluster oder auch mehrere ausgewählt und jedem Referenz-Cluster wird ein neuer Objektvektor x_{neu} hinzugefügt. Wir haben eine spezielle Diskriminanzanalyse, welche als Identfikationsanalyse bezeichnet wird. Der Datensatz besteht aus dem Referenz-Cluster und einem Cluster nur aus einem Objekt bestehenden Datensatz.

Ziel der Analyse ist zu prüfen, ob dieses neue Objekt zum Referenz-Cluster gehört oder nicht gehört. Dieses entscheidet die Prüfung des Diskriminanzmodell mit dem globalen F-Test. Ist das Ergebnis nicht signifikant, dann wird die Nullhypothese bestätigt und das neue Objekt gehört zur Referenz. Lautet dagegen das Ergebnis signifikant, wird die Nullhypothese abgelehnt und das neue Objekt gehört nicht zum Referenz-Cluster. In diesem Fall möchte man wissen, welche Variablen zu einem signifikanten Ergebnis geführt haben. Dies sieht man an der Signifikanzprüfung der Diskriminazkoeffizienten.

Will man ein Objekt (Ausreißer, einzelnes neues Objekt) beurteilen, dann wird man wie folgt Vorgehen: Die Messwerte x_1, …, x_p werden in den bestehenden Referenzdatensatz eingefügt. Die Fisher-Codierung wird angepasst:

- bestehende Gruppe: $c1=1/(n+1)$
- neue Beobachtung: $c2=-n/(n+1)$

Die Regressionsfunktion wird mit dem ergänzten Objekt neu berechnet. Anhand des globalen F-Test kann nun entschieden werden, ob das Objekt zur Referenz gehört oder nicht.

- Nullhypothese H_0 angenommen: Unterschied zwischen neuem Objekt und Referenz ist zufällig.
- Nullhypothese H_0 abgelehnt: Unterschied zwischen neuem Objekt und Referenz ist signifikant.

	echt				gefälscht	
Objekt	F-Test	p-Wert	MD^2	p-Wert MD	F-Test	p-Wert
1	5.1729880	0.0001325	24.3080550	0.0004583	8.8619188	0.0000001
5	3.1837360	0.0070692	16.6122661	0.0108190	21.2706265	0.0000000
13	2.9005438	0.0124993	15.3762952	0.0175235	15.3000786	0.0000000
40	5.1252686	0.0001455	24.1414024	0.0004919	24.1774206	0.0000000
70	4.2194827	0.0008785	20.8197740	0.0019765	2.3230800	0.0394610
103	3.2463919	0.0062303	16.8804918	0.0097326	1.6914613	0.1323014
161	11.3259647	0.0000000	40.7028759	0.0000003	3.3146923	0.0054285
167	9.9995435	0.0000000	37.8555115	0.0000012	5.4538279	0.0000765
171	16.2781142	0.0000000	49.1880778	0.0000000	5.6179482	0.0000556
180	15.4501640	0.0000000	47.9626359	0.0000000	3.3055206	0.0055299
201	7.4199676	0.0000019	31.3461022	0.0000218	0.5576375	0.7628209
202	10.9057812	0.0000000	39.8326459	0.0000005	0.2605826	0.9536891
203	3.9548907	0.0014940	19.7890402	0.0030192	1.9354937	0.0836623
204	8.4704539	0.0000003	34.1708793	0.0000062	1.1053400	0.3656499
205	9.6804446	0.0000000	37.1245359	0.0000017	1.2651133	0.2815898

Zur Interpretation, die ersten fünf Banknoten sind echte Banknoten welche als Ausreißer in der Clusteranalyse gefunden wurden. Alle diese Banknoten wurden als Ausreißer bestätigt, sie gehören weder zu den echten noch zu den gefälschten Banknoten. Die Banknote 70 würde bei einer Klassifikation falsch zugeordnet werden. Die Ursache ist wahrscheinlich ein Messfehler in der Diagonalen x_6.

Die weiteren fünf Banknoten sind gefälschte Banknoten, wobei die Banknote 103 kein Ausreißer ist und richtig zugeordnet wird. Die restlichen Banknoten können nicht zu geordnet werden und sind somit als Ausreißer bestätigt.

Die letzen fünf Banknoten sind neue gefälschte Banknoten, welche alle richtig zugeordnet werden konnten.

Ergänzend muss gesagt werden, man kann ein neues Objekt auch mit Hilfe der Mahalanobis-Distanz MD^2 bewerten, signifikante Werte werden rot ausgegeben.

2.4 Zusammenfassung der Ergebnisse

Im Banknotenbeispiel zeigt sich typischerweise:

- Sehr klare Trennung der Diskriminanzwerte.
- Geringe Überlappung der Verteilungen.
- Hohe Trefferquote.
- Eine bis drei Variablen dominieren die Trennung.

Die Analyse bestätigt somit: Die Clusteranalyse hat eine reale, physikalisch interpretierbare Klassenstruktur entdeckt. Doch erst die Kombination aus:

- Clusteranalyse (strukturbildend)
- Diskriminanzanalyse (strukturprüfend)
- Identifikationsanalyse (neue Objekte)

ermöglicht eine objektive Qualitätskontrolle, automatische Klassifikation und statistisch abgesicherte Entscheidungen. Im Fall der Banknoten heißt das: Eine neue Banknote kann anhand weniger geometrischer Messgrößen mit quantifizierter Sicherheit als echt oder gefälscht identifiziert werden. Dieses Beispiel zeigt den vollständigen Analysezyklus:

- Explorative Clusteranalyse
- Bestätigung durch Diskriminanzanalyse
- Ableitung einer Trennfunktion
- Wahrscheinlichkeitsbasierte Klassifikation
- Identifikation neuer Objekte

Damit wird aus einer rein beschreibenden Analyse ein entscheidungsfähiges statistisches Verfahren.