Multiregressionsclusterung zur Typisierung von Mensch-Umwelt-Systemen

Carsten Walther

Betreuer:
Matthias Lüdeke

Diplomarbeit
vorgelegt beim Institut für Physik der
Mathematisch-Naturwissenschaftlichen Fakultät
der
Universität Potsdam
Februar 2007

.

Erstgutachter: Prof. Dr. H. J. Schellnhuber
Zweitgutachter: Dr. Udo Schwarz

Inhaltsverzeichnis

Abbildungsverzeichnis

Kapitel 1

Einleitung

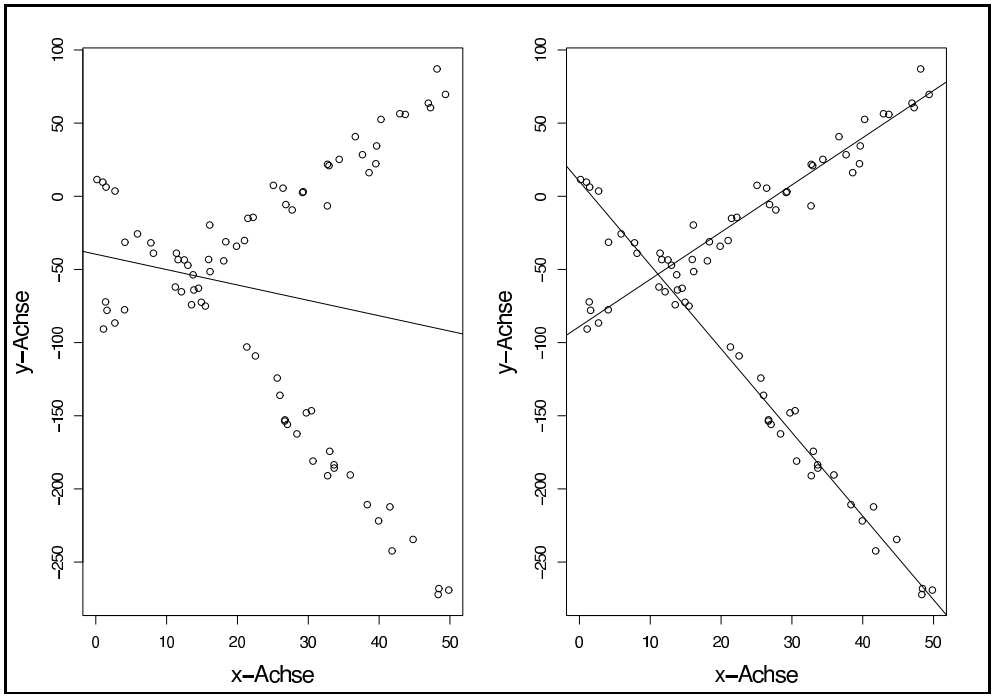

Abb. 1.1: Vergleich der Datenanalyse mittels einfacher Regression und Multiregressionsclusterung.

„Die Informationen in der Welt verdoppeln sich etwa alle 20 Monate“ [Runkler, 2000]. Es müssen große Datenmengen nicht nur gespeichert und verwaltet werden, sondern auf der Suche nach relevanten Informationen auch verarbeitet und aufbereitet. Als Werkzeuge für diese Suche existieren eine Vielzahl von Datenverarbeitungsmethoden wie Korrelations-, Regressions- und Clusteranalyse, Fuzzy-Logik usw., um numerische und nichtnumerische Daten zu verarbeiten, zu filtern, zu visualisieren und zu klassifizieren.
Ist ein $(n \times m)$-dimensionaler Datensatz, der n Objekten m verschiedene Variablen zuordnet, gegeben, kann man ihn auf Zusammenhänge zwischen verschiedenen Variablen untersuchen. Diese können beispielsweise mittels einer Korrelationsanalyse bestimmt werden.
Zur Ermittlung der funktionalen Zusammenhänge verwendet man zum Beispiel die Multiregressionsanalyse. Mit ihr ist es möglich, einen Zusammenhang zwischen einer abhängigen und einer oder mehreren unabhängigen Variablen festzustellen. Bei der Regression ist es jedoch immer nur möglich, *einen* Zusammenhang in den Daten aufzudecken. Sollten, wie in Abb. 1.1, zwei oder mehrere Gruppen von Objekten existieren, in welchen unterschiedliche Zusammenhänge zwischen den Eigenschaften im Datensatz auftreten, wird eine einfache Regression diese nicht finden.

Eine Analysemethode, welche in der Lage ist, Objektgruppen in Daten ausfindig zu machen, ist die Clusteranalyse. Sie sucht Untermengen im Datensatz, deren Datenpunkte sich in ihren Eigenschaften ähnlicher sind als die Datenpunkte aus unterschiedlichen Untermengen. Sie geben Informationen aus dem Datensatz preis, wie z.B. Objekthäufungen bei bestimmten Eigenschaftskombinationen. Aus diesen Gruppen kann man jedoch nicht auf bestimmte Zusammenhänge zwischen den Variablen schließen.
Wird diese Analysemethode mit einer Multiregressionsanalyse kombiniert, lassen sich die oben beschriebenen Objektgruppen mit unterschiedlichen Variablenzusammenhängen ausfindig machen. Die vorliegende Arbeit beschäftigt sich mit dieser Kombination und beleuchtet ihre Vor- und Nachteile.
Die Methode der Multiregressionsclusterung (MRC) kann dann zur Typisierung von Mensch-Umwelt-Systemen verwendet werden, wie sie etwa im Syndromkonzept [Schellnhuber, 1997] angenommen werden.
Am Beispiel des unterschiedlichen Anbauverhaltens einer Bäuerin[1] der Sahelregion und eines industrialisierten Landes wird deutlich, wie zwei gegensätzliche Zusammenhänge innerhalb eines Variablensatzes auftreten können. Solche Aufspaltungen können mittels einer MRC-Analyse aufgedeckt werden. Die Sahelbäuerin lebt typischerweise in Subsistenzwirtschaft. Das von ihr bearbeitete Land ist durch Übernutzung von zunehmender Degradation bedroht. Mit steigendem Einkommen würde sie in der Lage sein, ihr Anbauverhalten so zu steuern, dass der Boden weniger belastet wird und dadurch weniger degradiert. In den industrialisierten Ländern werden Umweltschädigungen eher durch eine kapitalintensive nicht-nachhaltige Nutzung von Böden und Gewässern verursacht. Dort könnte eine Bäuerin mit zunehmendem Einkommen ihre finanziellen Mittel dafür verwenden, den Einsatz von Maschinen und Dünger zu verstärken und dadurch die Böden massiver zu degradieren. Bei einer Gegenüberstellung von Bodendegradation und Einkommen von Bäuerinnen verschiedener Regionen würde sich möglicherweise eine Aufspaltung in zwei Zusammenhänge ergeben.
Die im Verlauf der Arbeit verwendeteten Datensätze sollen beispielhaft die Anwendung der Multiregressionsclusterung bei der Typisierung von Systemen, die die Wechselwirkung zwischen Mensch und Umwelt thematisieren, darstellen. Dabei kommen verschiedene Typen von Daten, wie Zustandsgrößen, Änderungen über einen Zeitraum oder Daten von verschiedenen Zeitpunkten, zur Verwendung.

[1] Der Einfachheit halber wird im Folgenden stets die weibliche Form verwendet.

Die vorliegende Arbeit ist wie folgt aufgebaut.
Im folgenden Kapitel wird der Aufbau und die Funktionsweise der Kombination aus Multiregression und Clusteranalyse beleuchtet. Es werden verschiedene Clusteranalysemethoden vorgestellt und die Multiregressionsclusterung darin eingebettet. Daran anschließend werden kurz verschiedene statistische Methoden eingeführt, welche für die Analysemethode von Bedeutung sind. Der Abschnitt schließt mit der Vorstellung des Problems der Auswahl des geeigneten Modells sowie der zugehörigen Gütekriterien.
Kapitel 3 beschäftigt sich mit der Anwendung der vorgestellten Methode. Dabei werden zuerst Gütekriterien und Algorithmus auf ihre Zuverlässigkeit an synthetischen Daten erprobt. Darauf folgt die Anwendung auf empirischen Daten. Für die Analyse werden weltweite sozioökonomische und Umweltdaten in nationaler Auflösung verwendet. Dabei soll nach Gruppen von Staaten gesucht werden, welche sich in ihren Zusammenhangsmustern unterscheiden. Insbesondere wird angestrebt, aus den Eigenschaften der Cluster Aussagen über die Beziehungen zwischen den Dimensionen der Armut und der Umwelt zu ermöglichen.
Die Arbeit schließt in Kapitel 4 mit einer Bewertung der Methode, Problemen bei der Anwendung sowie einem Ausblick über weitere Variationen und mögliche Optimierungen.

Kapitel 2

Methoden zur Multiregressionsclusterung

2.1 Multiregressionsclusterung

Der Algorithmus der Multiregressionsclusterung nutzt zur Berechnung der Zusammenhänge zwischen mehreren Variablen innerhalb einer Gruppe von Datenpunkten (Cluster) die Abläufe der Multiregressionsanalyse. Um diese Gruppe von Datenpunkten (auch als Objektegruppe benannt) im Datensatz zu finden, ist eine Clusteranalyse integriert. Im Folgenden soll auf die Funktionsweise von Regressionsanalyse und Clusteranalyse eingegangen werden.

2.1.1 Regressionsanalyse

Die Regressionsanalyse untersucht die Struktur der Abhängigkeit zwischen einer abhängigen (Responsevariable oder zu erklärende Variable) und einer oder mehrerer unabhängiger Variablen (Prediktor-Variablen oder erklärende Variablen). Bei nur einer unabhängigen Variable wird das Verfahren als einfache Regression, bei mehreren Unabhängigen als multiple Regression bezeichnet. Im speziellen Fall der linearen Regression wird angenommen, dass eine Variable Y und deren Realisierungen y_i durch eine lineare Kombination anderer Variablen X_j (Realisierungen: x_{ji}) erklärt werden kann.
In folgender Gleichung handelt es sich bei ϵ_i um zufällige Abweichungen vom linearen Zusammenhang. Die Anzahl der Objekte beträgt n und die Anzahl der Variablen m. Bei den β_j handelt es sich um die Regressionskoeffizienten. Sie sind ein Maß für die Abhängigkeit der Responsevariable von der jeweiligen erklärenden Variable. β_0 wird als der Intercept bezeichnet und stellt die Verschiebung der Regressionsgeraden auf der Achse der abhängigen Variable dar.

$$y_i = \beta_0 + \beta_1 x_{1i} + .. + \beta_m x_{mi} + \epsilon_i \quad , \quad i = 1, .., n \tag{2.1}$$

Ziel der Regression ist es, eine lineare Funktion zu finden, für die die nicht erklärten Abweichungen minimal sind. Dies wird als die *Methode der kleinsten Quadrate* bezeichnet. Bei x_{ij} handelt es sich um den Wert der j-ten Variablen von Objekt i. Es wird über alle n Datenpunkte summiert.

$$RSS = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_{1i} + .. + b_m x_{mi}))^2 \rightarrow min! \tag{2.2}$$

RSS steht hier für *residual squared sum*, also die Summe der quadrierten Residuen e_i. Die Residuen sind die nicht erfassten Einflüsse zwischen der zu erklärenden und den erklärenden Variablen. $\hat{y}$ umfasst die durch die Regressionsanalyse gefitteten Werte, also die Werte, welche auf der Regressionsgeraden liegen.

$$Y_i = \hat{Y}_i + e_i \tag{2.3}$$

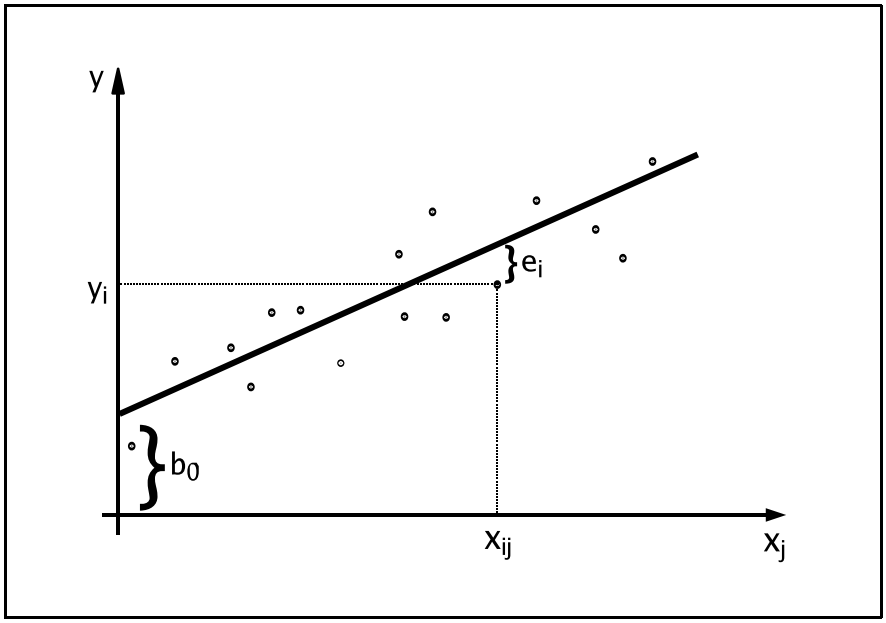

Abb. 2.1: Darstellung einer Punktwolke und der zugehörigen Regressionsgeraden. Bezeichnungen vgl. Gl. 2.1 und 2.3.

Durch partielles Differenzieren von Gleichung 2.2 und Nullsetzen der Ableitungen erhält man ein System von Normalengleichungen:

$$\mathbf{X}^T\mathbf{X}\vec{\beta} = \mathbf{X}^T\vec{y} \quad \text{mit} \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & ... & x_{m1} \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ 1 & x_{1n} & ... & x_{mn} \end{pmatrix}. \tag{2.4}$$

Die Lösungen dieser Normalengleichungen sind die geschätzten Regressionskoeffizienten. $\bar{y}$ ist der Mittelwert der abhängigen Variable und $\bar{x}_j$ der Mittelwert der j-ten erklärenden Variable.

$$\hat{b}_j = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(y_i - \bar{y})}{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2} \qquad \hat{b}_j = \frac{s_{x_j y}}{s_{x_j}^2} \tag{2.5}$$

Wie in Gleichung (2.5) dargestellt, lässt sich jeder Regressionskoeffizient auch als Quotient aus Korrelation zwischen x_j und y ($s_{x_j y}$) sowie der Varianz in x_j ($s_{x_j}^2$) darstellen.
In Abb. 2.1 ist eine lineare Regression beispielhaft dargestellt.
Es gibt verschiedene Methoden um zu prüfen, wie gut sich diese Ergebnisse aus der Stichprobe auf die Grundgesamtheit[1] übertragen lassen.

[1]Die Grundgesamtheit umfasst die Gesamtheit aller möglichen Messwerte (N) einer

Bestimmtheitsmaß R^2

Zur Berechnung dieses 'goodness of fit' wird die Gesamtabweichung (GV) in erklärte (EV) und nicht erklärte Abweichung (NV) aufgespalten.

$$GV \quad = \quad EV \quad + \quad NV \tag{2.6}$$

$$\sum_{i=1}^{n}(y-\bar{y})^2 = \sum_{i=1}^{n}(\hat{y}-\bar{y})^2 + \sum_{i=1}^{n}(\hat{y}-y)^2 \tag{2.7}$$

Bei der nicht erklärten Abweichung handelt es sich um die Residuen. Das Bestimmtheitsmaß gibt dann das Verhältnis von erklärter zu gesamter Abweichung wieder.

$$R^2 = \frac{\sum_{i=1}^{N}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{N}(y_i-\bar{y})^2} \tag{2.8}$$

R^2 liegt zwischen 0 und 1. Je größer R^2, desto höher der Anteil der durch das Modell erklärten Streuung von y.

2.1.2 Clusteranalyse

Die Clusteranalyse sucht in einer heterogenen Menge von Objekten nach homogenen Teilmengen. Im Allgemeinen ist die Zerlegung eines Datensatzes $X = x_1, ..., x_n \subset R^m$ definiert als die Partition von X in $K \in 2, 3, .., n-1$ disjunkte Teilmengen $C_1, .., C_K$ [Runkler, 2000] so, dass

$$X = C_1 \cup ... \cup C_K \tag{2.9}$$

$$C_k \neq \{\} \quad \forall \quad k = 1, .., K \tag{2.10}$$

$$C_k \cap C_{k'} = \{\} \quad \forall \quad k, k' = 1, .., K, k \neq k'. \tag{2.11}$$

Die Objekte innerhalb dieser Teilmengen sollen sich ähnlicher sein als Objekte unterschiedlicher Teilmengen. Diese Ähnlichkeit kann auf unterschiedliche Weise gemessen werden. Nach diesen Ähnlichkeitsmaßen und dem Ablauf des Algorithmus' unterscheidet man verschiedene Verfahren der Clusteranalyse. Hauptsächlich lassen sich die Clusterverfahren in partitionierende und hierarchische Verfahren unterteilen.

Messreihe. Bei einer Stichprobe handelt es sich um eine Teilmenge mit dem Umfang n ($n \leq N$) [Bronstein, 1985].

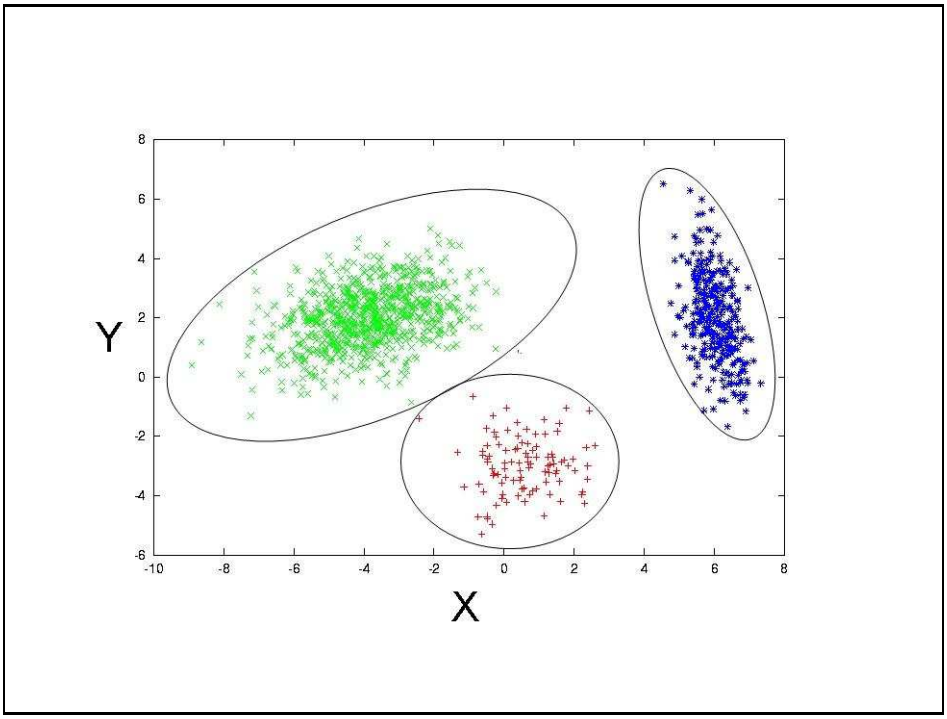

Abb. 2.2: Beispielhafte Darstellung eines Datensatzes mit drei, durch eine partitionierende Clusteranalyse, gefundenen Punktwolken.

Partitionierende Verfahren

Bei den *partitionierenden* Verfahren geht man von einer Startpartitionierung aus, in der durch Zufall oder gezielt die n Objekte einer vorher festgelegten Zahl K von Clustern zugeordnet werden. Die Zuordnung der Objekte zu den Clustern wird über die Partitionsmatrix U definiert.

$$u_{ik} = \begin{cases} 1 \leftrightarrow x_i \in C_k \\ 0 \leftrightarrow x_i \notin C_k \end{cases} \tag{2.12}$$

Die Punkte eines Cluster C_k bilden mit ihrem Schwerpunkt oder ersten Moment ein Clusterzentrum v_k.
Grundsätzlich gilt es nun eine Zielfunktion zu minimieren, welche beispielsweise aus der Summe der quadrierten Abstände zwischen den Objekten $x_i \in C_k$ und den Clusterzentren v_k gebildet wird. Mittels der Partitionsmatrix U werden nur die Abstände zwischen Objekt und zugehörigem Cluster gebildet. Bei Mucha (1992) wird diese Zielfunktion als Interklassenvarianzkriterium V bezeichnet.

$$V = \sum_{k=1}^{K} \sum_{x_i \in C_k} \| x_i - v_k \|^2 = \sum_{k=1}^{K} \sum_{i=1}^{n} u_{ik} \| x_i - v_k \|^2 \tag{2.13}$$

Nun soll durch Wechsel eines Objektes von einem Cluster in ein anderes eine Minimierung der Zielfunktion erreicht werden. Dabei können folgende Vorgehensweisen unterschieden werden.
Bei der *Minimaldistanzmethode* wird der Abstand des Objektes x_i zu jedem Clustermittelpunkt v_k bestimmt. Das Objekt wird in das Cluster mit dem minimalsten Abstand verschoben. Mit der veränderten Partitionsmatrix U verändert sich auch die Matrix der Clusterzentren. Beim *K-Means-Verfahren* werden z.B. nach jeder Änderung von U auch die Zentren neu berechnet. In anderen Verfahren gibt es verschiedene Nebenzyklen, die eine Neuberechnung von V nicht in jedem Iterationsschritt notwendig machen.
Im *Austauschverfahren* werden hingegen die Distanzen eines Objektes x_i zu seinem Clusterzentrum v_k und zu einem anderen Clusterzentrum v'_k verglichen. Sollte der Abstand zum eigenen Clusterzentrum größer sein, wird das Objekt in Cluster k' verschoben. Nun werden alle Zentren neu bestimmt und der nächste Wechsel auf Verbesserung von Gleichung 2.13 getestet.

Allen partitionierenden Verfahren gemein ist eine Abhängigkeit von der oben erwähnten Anfangspartition. Aus diesem Grund sollten die Verfahren mehrmals mit veränderter Anfangspartition durchgeführt werden. Weiterhin ist eine Anwendung der Verfahren auch mit anderen Abstandsmaßen möglich. Beispielsweise kann neben der Euklidische-Metrik auch die
Manhattan-Metrik

$$d_{ij} \equiv \sum_{k=1}^{N} w_k \mid x_{ik} - x_{jk} \mid, \tag{2.14}$$

wobei w_k als Normierungsfaktor beiträgt, oder die
Canberra-Metrik

$$d_{ij} \equiv \sum_{k=1}^{N} \mid x_{ik} - x_{jk} \mid / (\mid x_{ik} \mid + \mid x_{jk} \mid) \tag{2.15}$$

verwendet werden [Gordon, 1999].
Das Abbruchkriterium für die partitionierenden Verfahren kann entweder ein Optimum der gewählten *Zielfunktion* oder ein von der Analytikerin gewähltes Kriterium, wie die maximale Anzahl der Rechenschritte, sein.
Von den bisher beschriebenen Verfahren unterscheidet sich die *Fuzzy*-Methode deutlich in der Art der Zuordnung eines Objektes zu einem Cluster. Hier kann jedes Objekt anteilig mehreren Clustern zugeordnet werden. Die Partitionsmatrix kann also alle Werte im Intervall $u_{ik} \in [0, 1]$ annehmen. Gleichung 2.11 gilt bei dieser Methode nicht mehr.

Hierarchische Verfahren

Beim *hierarchisch agglomerativen* Verfahren wird mit der feinsten Partition (Clusteranzahl entspricht Objektanzahl, $K_0 = n$) begonnen. Es werden in jedem Schritt die Objekte bzw. Objektgruppen mit dem geringsten Abstand zueinander verschmolzen, bis die gröbste Partition (Clusteranzahl = 1) erreicht ist. Bei *hierarchisch divisiven* Verfahren ist der Ablauf umgekehrt.
Die Abstände zwischen den Objekten bzw. Objektgruppen können auf verschiedenste Weise bestimmt werden. Da im Verlauf des Verfahrens die zu verschmelzenden Objektgruppen nicht mehr nur aus einem Objekt bestehen, kann der Abstand auf verschiedene Weisen bestimmt werden. Möglich sind zum Beispiel die Berechnung aus dem Mittelwert aller paarweisen Distanzen (*Average Linkage*), den Minima (*Single Linkage*) oder Maxima (*Complete Linkage*) aller paarweisen Abstände, sowie über die *Minimalvarianzmethode*, bei der Klassen so verschmolzen werden, dass dabei der geringste Zuwachs an Varianz innerhalb einer Klasse verursacht wird. Die Methoden haben verschiedene Vor- und Nachteile, auf die hier jedoch nicht näher eingegangen werden kann. Zusätzlich ist es möglich, die Abstandsberechnung zwischen zwei Objekten oder Objektgruppen ebenfalls mit unterschiedlichen Abstandsmaßen, wie z.B. in Gleichung 2.14 oder 2.15, durchzuführen.
Bei den hierarchischen Verfahren ist die Kenntnis der Clustermenge wie oben dargestellt im Vorhinein nicht notwendig. Der Vorgang des Zusammenfassens der einzelnen Datenpunkte kann in einem sogenannten *Dendogramm* dargestellt werden. Auf der Abszisse sind die Objekte und auf der Ordinate beispielsweise die Abstände der Cluster aufgetragen. Aus dieser Darstellung können Rückschlüsse auf eine geeignete Clusteranzahl gemacht werden.

Nicht-parametrische Verfahren

Gegenüber den bisher vorgestellten *parametrischen* Verfahren, bei denen die grundlegende Struktur der Cluster vorausgesetzt wird, gibt es auch noch *nicht-parametrische* Verfahren, welche weniger Annahmen über die Struktur der Cluster benötigen.
Ein Beispiel dafür ist ein Verfahren, welches auf den physikalischen Eigenschaften eines magnetischen Systems basiert. Durch die dabei jeder Koordinate zusätzlich zugeordneten Spineigenschaften ergeben sich verschiedene selbstorganisierte Phasen, verursacht durch Kopplungen der Spins untereinander. Diese Kopplungen ermöglichen die Ausbildung von sogenannten magnetischen Körnern, welche es ermöglichen, Cluster im Datenraum zu erkennen. Durch variieren eines Parameters des Algorithmus' (physikalische Temperatur) werden sich diese Cluster verbinden und trennen, wodurch die hierarchische Struktur der Daten wiedergegeben wird [Blatt et al., 1996].

2.1.3 Kombination beider Methoden

In der Multiregressionsclusterung (MRC) verschmelzen nun beide oben vorgestellten Analysemethoden. Für die eingehende Clusteranalyse wurden Elemente aus dem partitionierenden Austauschverfahren gewählt. Das heißt, die Objekte werden anfangs zufällig auf eine vorgegebene Zahl von Clustern verteilt und daraufhin der Wert der zu optimierenden Zielfunktion aus dieser Anfangspartition berechnet. Hier setzt nun die Multiregressionsanalyse ein. Die Cluster definieren nicht, wie in der Clusteranalyse üblich, über ihre Abstände zueinander einen Schwerpunkt oder Zentroid, sondern eine Regressionshyperebene. Wie in Abschnitt 2.1 beschrieben, wird diese durch die Minimierung der Residuenquadratsumme (Gl. 2.2), also der summierten quadrierten Abstände zwischen Objekten und Regressionshyperebene, gebildet. Zur verbesserten Anpassung der Regressionshyperebene an die Daten wird eine Zielfunktion über alle Objekte optimiert. Die gewählte Zielfunktion (ZF) ist die Residuenquadratsumme eines jeden Clusters, summiert über alle Cluster:

$$ZF = \sum_{k=1}^{K} \sum_{i \in C_k} [y_i - (b_0 + b_1 x_{1i} + ... + b_m x_{mi})]^2. \tag{2.16}$$

ZF steht hier für Zielfunktion. Die restlichen Bezeichnungen sind identisch mit denen in Abschnitt 2.1.
Mit dem gewählten partitionierenden Austauschverfahren ist es nun möglich zu bestimmen, ob der Wechsel eines Objektes von einem Cluster A in ein Cluster B die Summe der Residuenquadrate der Regressionshyperebenen in A und B vergrößert oder verringert. Es kann also bestimmt werden, ob das *anfitten* der Regressionshyperebenen an die Struktur im untersuchten Datensatz sich durch einen Objektwechsel verbessert oder verschlechtert.
Im folgenden Abschnitt sollen einige Tests vorgestellt werden, mit denen es möglich ist, die Qualität der auf diese Weise bestimmten Regressionscluster abzuschätzen.

2.2 Statistische Tests

Am Ende einer Analyse mittels MRC steht eine Partition, welche die n Objekte den K Clustern zuordnet. Um bewerten zu können, wie gut eine gefundene Partition die Struktur eines Datensatzes wiedergibt und um zu bemessen, wie aussagekräftig beispielsweise ein Regressionsparameter ist, gibt es verschiedene statistische Tests.

2.2.1 Statistische Hypothesen

Statistische Hypothesen sind zu überprüfende Aussagen über die Form oder die Parameterwerte von Zufallsvariablen. Das klassische Verfahren des Hypothesentestens beruht auf der Ablehnung einer trivialen Nullhypothese H_0. Mittels eines statistischen Prüfverfahrens wird darüber entschieden, ob die Nullhypothese H_0 angenommen oder zurückgewiesen wird. Die Nullhypothese nimmt beispielsweise an, dass zwei oder mehr Stichproben aus *einer* Grundgesamtheit stammen. Dies bedeutet, dass die Unterschiede zwischen den Kennwerten (z.B. Mittelwerte) der Stichproben zufällig sind. Die Alternativhypothese geht davon aus, dass die Stichproben aus *verschiedenen* Grundgesamtheiten stammen. Der Unterschied zwischen den Kennwerten der Stichproben wäre somit signifikant. Bei der Entscheidung für oder gegen eine Hypothese besteht das Risiko einer Fehlentscheidung [Sachs, 1984]. Eine Übersicht dieser Entscheidungen und möglicher Fehler ist in Tabelle 2.1 dargestellt. Wichtig ist besonders der Fehler 1. Art oder α-Fehler. Er gibt die Wahrscheinlichkeit an, mit der die Nullhypothese fälschlicherweise abgelehnt wird. Der Fehler 2. Art, oder β-Fehler, gibt an mit welcher Wahrscheinlichkeit die Nullhypothese fälschlicherweise nicht abgelehnt wird.
Es lässt sich kein statistischer Schluss ziehen, welcher mit absoluter Sicherheit gültig ist, weswegen der Grad der Unsicherheit α mit Hilfe einer statistischen Wahrscheinlichkeitsaussage immer anzugeben ist. Die statistische Sicherheit ergibt sich dann aus $S = 1 - \alpha$.

	H_0 angenommen	H_0 abgelehnt
H_0 wahr	richtig entschieden	Fehler 1. Art Irrtumswahrscheinlichkeit α
H_0 falsch	Fehler 2. Art Grenzwahrscheinlichkeit β	richtig entschieden

Tabelle 2.1: Entscheidungen in einem Test

Es werden verschiedene Prüfverteilungen bei der Bewertung der Nullhypothese durch statistische Tests verwendet. Dazu gehören unter anderem Normalverteilung, t-Verteilung, χ^2-Verteilung sowie F-Verteilung. Diese werden in den folgenden Abschnitten näher beschrieben und mögliche Anwendungen erläutert.

2.2.2 Prüfverteilungen und Tests

Normalverteilung

Eine Normalverteilung nimmt eine Zufallsvariable immer dort an, wo sie einen Zufallsprozess beschreibt, der aus vielen voneinander unabhängigen Einzelzufallsprozessen besteht (*der Zentrale Grenzwertsatz*). Hat eine Zufallsvariable X eine Dichte der Form

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad (-\infty < x < \infty), \tag{2.17}$$

so heißt X normalverteilt mit den Parametern Mittelwert μ und Standardabweichung σ^2. Man schreibt: $X \sim N[\mu, \sigma^2]$. Bei $\mu = 0$ und $\sigma^2 = 1$ heißt die Verteilung Standardnormalverteilung.

χ^2-Verteilung und χ^2-Anpassungstest

Es seien $X_1, X_2, .., X_n$ n unabhängige standardnormalverteilte Zufallsvariablen. Dann heißt die Verteilung von

$$\chi^2 = X_1^2 + X_2^2 + .. + X_n^2 \tag{2.18}$$

χ^2-Verteilung mit dem Freiheitsgrad[2] n. Die Verteilung geht für $\nu \to \infty$ in die Normalverteilung über.
Diese χ^2-Verteilung kann in einem χ^2-Anpassungstest verwendet werden. Bei diesem handelt es sich um einen Signifikanztest zur Prüfung der Güte der Anpassung einer empirisch gewonnenen Verteilung an eine erwartete Verteilung [George, 2000]. Mit diesem ist es möglich zu testen, ob es sich bei der Grundgesamtheit, aus der eine Verteilung gewonnen wurde, um eine normalverteilte Menge handelt. Dabei werden aus den Differenzen zwischen den Werten einer theoretischen Verteilung ϕ_i und den empirischen Werten f_i eine Teststatistik berechnet, welche bei ausreichender Stichprobengröße χ^2-verteilt ist

$$\chi^2 = \sum_{i=1}^{n} \frac{(f_i - \phi_i)^2}{\phi_i}. \tag{2.19}$$

[2]Die Anzahl der Freiheitsgrade einer Zufallsgröße ist definiert durch die Zahl *frei* verfügbarer Beobachtungen, d.h. dem Stichprobenumfang n minus der Anzahl a aus der Stichprobe geschätzter Parameter $\nu = n - a$ [Sachs, 1984].

Mit dieser Prüfgröße wird getestet, ob die Nullhypothese H_0 gültig ist, nach der die empirische Verteilung sich nicht signifikant von der theoretischen unterscheidet. Mit α als Irrtumswahrscheinlichkeit und ν als Anzahl der Freiheitsgrade wird $\chi^2_{1-\alpha;\nu}$ (1-α-Quantil[3] der χ^2-Verteilung) bestimmt und mit dem berechneten Wert χ^2 verglichen:

$$\chi^2 < \chi^2_{1-\alpha;\nu} \rightarrow H_0 \text{ angenommen} , \tag{2.20}$$

$$\chi^2 \geq \chi^2_{1-\alpha;\nu} \rightarrow H_0 \text{ abgelehnt.} \tag{2.21}$$

Weiterhin ist es möglich, einen gegebenen Datensatz durch einen *Q-Q-Plot* mit einer bestimmten Verteilung zu vergleichen. Dazu werden die empirischen p_i-Quantile des Datensatzes bestimmt und in einem Diagramm gegen die theoretischen p_i-Quantile der Verteilung aufgetragen. Entsprechen die Daten der Verteilung, so ergibt sich eine Gerade mit 45°-Steigung. Das empirische p-Quantil $\Phi(p)$ ist definiert als der kleinste x-Wert eines Datensatzes $x_1, ..., x_n$, für den $F(x) \geq p$ gilt. Dabei gilt, $0 < p \leq 1$ und $F(x)$ ist die empirische Verteilungsfunktion, die die relative Wahrscheinlichkeit angibt, dass ein Wert x_i aus der Zeitreihe $x_1, ..., x_n$ kleiner oder gleich x ist [Schlittgen, 1995].

t-Verteilung und t-Test

Wenn X_N eine normalverteilte Zufallsvariable und X_{χ^2} eine χ^2-verteilte Variable, dann ist

$$T_\nu = \frac{X_N}{\sqrt{\frac{X_{\chi^2}}{\nu}}} \tag{2.22}$$

t-verteilt [Dolić, 2004]. Für verschiedene Freiheitsgrade ν besitzt die Verteilung einen unterschiedlichen Verlauf. Für $\nu \rightarrow \infty$ geht die t-Verteilung in die Normalverteilung über.
Der t-Test ist ein Hypothesentest mit t-verteilter Prüfgröße. In der Praxis sind die Parameter der Grundgesamtheit üblicherweise unbekannt. Bei diesem Test wird die Varianz der Grundgesamtheit über die Varianz der Stichprobe geschätzt.
Es kann beispielsweise die Übereinstimmung der Erwartungswerte zweier normalverteilter Variablen überprüft werden. Der Nachweis bezieht sich darauf, ob die Differenz zwischen den geschätzten Variablen signifikant von Null verschieden ist oder nicht.

[3]Das p%-Quantil ist jene reelle Zahl, für die die kumulierte Verteilungsfunktion den Wert von p% annimmt. Eine Stichprobe aus der Grundgesamtheit ist dann gerade mit p% Wahrscheinlichkeit kleiner gleich dem p%-Quantil, wie in Abb. 2.4 dargestellt.

Weiterhin können die Koeffizienten einer Regressionsanalyse bei normalverteilten Residuen getestet werden. Es wird hierbei überprüft, ob die Nullhypothese gilt, bei der der Parameter $b_j = 0$ ist, also keinen Beitrag zur Erklärung von y liefert. Mit der geschätzten Varianz des j-ten Parameters der Regression $\hat{\sigma}\sqrt{a_{ll}}$ kann die Prüfgröße t_j berechnet werden (2.23). Diese geschätzte Varianz ergibt sich aus der Streuung der Residuen $\hat{\sigma}$ und dem l-ten Diagonalelement von $(X^T X)^{-1}$ (siehe 2.4)

$$t_j = \frac{\hat{b}_j - 0}{\hat{\sigma}\sqrt{a_{ll}}}. \tag{2.23}$$

Der berechnete Wert wird mit den kritischen Werten verglichen, welche für verschiedene Irrtumswahrscheinlichkeiten α und Freiheitsgrade ν bestimmt werden können:

$t_j \leq t_{\alpha,\nu}$	H_0 angenommen, X_j trägt nicht zur Erklärung bei
$t_j > t_{\alpha,\nu}$	H_0 abgelehnt, Beitrag von X_j ist signifikant.

Aus t_j kann auch direkt die Irrtumswahrscheinlichkeit für das Verwerfen der Nullhypothese berechnet werden. Es wird also nicht ein Quantil vorgegeben und überprüft, ob t_j größer oder kleiner ist, sondern das zu t_j gehörende Quantil bestimmt.
Häufig spricht man auch von einem partiellen F-Wert in der Bewertung der Parameter der linearen Regression. Bei diesem handelt es sich um das Quadrat der Prüfgröße t_j. Diese folgt einer F-Verteilung. Eigenschaften einer F-Verteilung und Funktionsweise eines F-Tests werden im nächsten Abschnitt dargestellt.

F-Verteilung und F-Test

Die F-Verteilung setzt sich aus zwei χ^2-verteilten Zufallsvariablen zusammen. Sind beispielsweise S_1^2 und S_2^2 Varianzen zweier unabhängiger Stichproben mit dem Umfang n_1 und n_2 aus zwei normalverteilten Grundgesamtheiten gleicher Varianz, dann ist die Variable

$$F = \frac{S_1^2}{S_2^2} \tag{2.24}$$

F-verteilt. In Abbildung 2.3 ist der Verlauf einer F-Verteilung dargestellt. Sie ist von zwei Freiheitsgraden abhängig. Für $\nu_1 \to \infty$ konvergiert sie in die χ^2-Verteilung und für $\nu_2 = 1$ geht sie in die t-Verteilung über. Die Form der F-Verteilung ist stark von den Samplegrößen abhängig. Die Samplevarianz variiert z.B. mehr von Stichprobe zu Stichprobe, wenn die Anzahl der Beobachtungen klein ist.

Der F-Test prüft, ob die Varianzen aus den voneinander unabhängigen Stichproben aus normalverteilten Gesamtheiten mit *unterschiedlichen* Varianzen gezogen wurden, oder ob sie aus normalverteilten Gesamtheiten mit *gleichen* Varianzen stammen. Dabei wird die F-Verteilung herangezogen, welche die Wahrscheinlichkeit angibt, mit der man ein bestimmtes Verhältnis von Samplevarianzen erhält (Gl. 2.24). Die Werte der Verteilung $F(\alpha|\nu_1;\nu_2)$ geben die Quantile bezüglich augewählter Freiheitsgrade und Wahrscheinlichkeiten an, also den F-Wert, der einen bestimmten Anteil der Grundgesamtheit begrenzt. Ist der aus den Varianzen berechnete F-Wert F_{emp} kleiner als $F(\alpha|\nu_1;\nu_2)$, ist die Homogenität der Varianzen gesichert. α steht hier für die Irrtumswahrscheinlichkeit.

$$F_{emp} < F_{(\alpha|\nu_1,\nu_2)} \tag{2.25}$$

Die Nullhypothese wird in diesem Fall nicht verworfen, da alle existierenden Abweichungen als zufällig angenommen werden müssen. Für den Fall, dass F_{emp} größer als $F_{(\alpha|\nu_1,\nu_2)}$ ist, wird die alternative Hypothese herangezogen, nach der die absolute Differenz zwischen den beiden Varianzen größer ist als für zwei Gesamtheiten mit gleicher Varianz zu erwarten.
Ein Spezialfall ist der F-Test des Bestimmtheitsmaßes(R^2) der linearen Regression, welcher die Signifikanz des gesamten Modells testet. Dabei wird die Nullhypothese, nach der kein Regressionsparameter b_j verschieden von 0 ist, überprüft. Die Alternativhypothese vermutet, dass mindestens ein Parameter b_j signifikant auf die abhängige Variable wirkt.

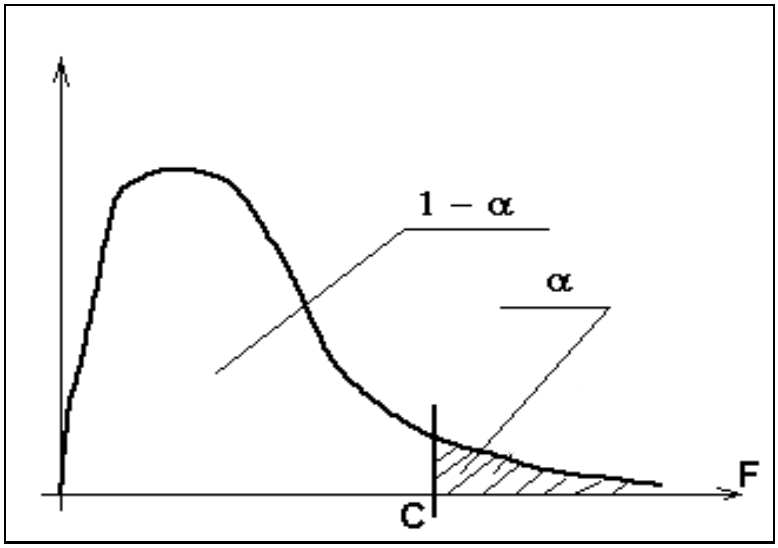

Abb. 2.3: Wahrscheinlichkeitsdichte der F-Verteilung. C steht hier für $F(\alpha|\nu_1;\nu_2)$.

Die Prüfgröße wird aus dem Bestimmtheitsmaß (Gl. 2.8) über

$$F_{emp} = \frac{R^2 \cdot (n - m - 1)}{m \cdot (1 - R^2)} \tag{2.26}$$

berechnet. n ist in Gl. 2.26 die Anzahl der Objekte, über die die Regression vorgenommen wird, und m die Anzahl der Dimensionen.
R^2 ist das Verhältnis von erklärter (EV) und gesamter (GV) Abweichung (siehe 2.8). Aus Gleichung 2.26 ist durch Umformungen zu entnehmen, dass F_{emp} das Verhältnis von erklärter zu nicht erklärter Varianz darstellt. Es ist möglich, F_{emp} mit einem vorgegebenen Quantil der F-Verteilung $F(\alpha|\nu_1;\nu_2)$ zu vergleichen, oder direkt den zugehörigen p-Wert zu berechnen. Er gibt die Irrtumswahrscheinlichkeit für das Verwerfen der Nullhypothese an.
Im Falle einer schlechten Anpassung der Daten durch die Regressionsgerade ist das Verhältnis $F_{emp} = \frac{EV}{NV}$ sehr klein, da beide Varianzen eine ähnliche Größenordnung haben und die Wahrscheinlichkeit für das Auftreten dieses Verhältnisses von Varianzen somit sehr groß ist (Vergleich mit Abb. 2.3). Im anderen Fall der guten Erklärung der Daten durch die Regression ist F_{emp} sehr groß, da die erklärte Varianz viel größer als die nicht erklärte Varianz ist. Die Wahrscheinlichkeit, dass dieses Verhältnis in einer F-Verteilung vorkommt, ist sehr gering.

Vertrauensbereich

Es lässt sich der Vertrauensbereich (Konfidenzintervall) des aus einer Stichprobe vom Umfang n geschätzten Mittelwertes $\hat{x}$ angeben. Hierbei versteht man unter einem $(1-\alpha)$%-Vertrauensbereich ein Intervall, welches so gewählt ist, dass in $(1-\alpha)\%$ aller Schätzungen aus einer Stichprobe der wahre Wert innerhalb des Intervalls liegt. Im Fall einer normalverteilten Zufallsgröße X mit bekanntem Parameter σ^2 lässt sich das Vertrauensintervall ($CI_{\hat{x},1-\alpha}$) mit dem $u_{1-\alpha/2}$ dem $(1-\alpha/2)$%-Quantil der standardisierten Normalverteilung durch

$$CI^1_{\hat{x},1-\alpha} = \hat{x} \pm \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2} \tag{2.27}$$

berechnen. In Abbildung 2.4 ist das Konfidenzintervall mit den zwei symmetrischen Intervallgrenzen dargestellt.
Ist die Streuung σ^2 der Grundgesamtheit nicht bekannt, lässt sich dieses Vertrauensintervall aus der Stichprobenstreuung s^2 und dem Quantil der t-Verteilung bestimmen

$$CI^2_{\hat{x},1-\alpha} = \hat{x} \pm \frac{s}{\sqrt{n}} t_{\alpha/2,n-1}. \tag{2.28}$$

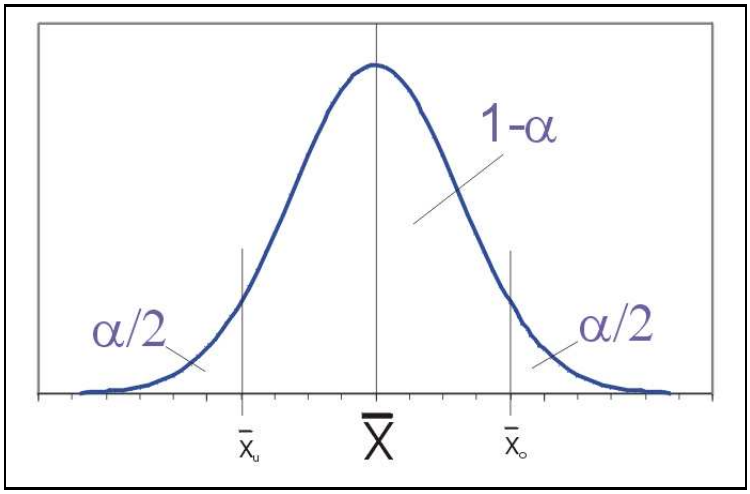

Abb. 2.4: Normalverteilung mit Konfidenzintervall (x_u, x_o) und Konfidenzkoeffizient $1 - \alpha$.

2.3 Modellauswahl

Mittels Multiregressionsclusterung gilt es eine unbekannte Zahl von Clustern in einem m-dimensionalen Datenraum zu finden. Es ist nun folgend zu klären, mit welchen Mitteln die günstige Clusteranzahl sowie die Variablenmenge und -zusammenstellung für die Multiregression mit der größten Erklärungskraft für die Responsevariable zu finden ist.

2.3.1 Variablenauswahl

Es wird eine zu erklärende Variable Y und ein Satz von potentiellen erklärenden Variablen $X_1, ..., X_m$ angenommen. Eine Variablenauswahl ist besonders notwendig bei einem großen Wert von m und bei der Annahme, dass $X_1, ..., X_m$ Variablen beinhaltet, welche nicht zur Erklärung beitragen. Sonst wird der Rechenaufwand unnötig erhöht und die Erklärungskraft des Ergebnisses gemindert. Wenn γ der Index der verschiedenen Modelle ist und q_γ die Anzahl der Variablen im γ-ten Modell darstellt, dann gilt es ein Modell

$$\vec{Y} = X_\gamma \vec{\beta_\gamma} + \vec{e} \tag{2.29}$$

anzupassen und auszuwählen [George, 2000]. Dabei steht X_γ für eine $n \times q_\gamma$ Matrix, deren Spalten zum γ-ten Modell gehören. $\vec{\beta_\gamma}$ ist der Vektor der Regressionskoeffizienten und $\vec{e}$ sind die Residuen. Es existieren verschiedene Methoden zur Auswahl der richtigen Variablen.

'search over all possible subsets' - nimmt alle möglichen Variablenkombinationen und berechnet für jede einen Gütewert, welcher angibt, wie gut sich bei dieser Kombination die Regressiongerade an die Daten anpasst. Nachteil ist der immense Rechenaufwand selbst bei kleinen Werten von p.

'backward elimination' - bei dem mit allen Variablen begonnen und dann fortlaufend immer die schlechteste Variable herausgenommen wird. Als Maß dient der der Variablen zugeordnete t-Wert bzw. partielle F-Wert (dieser gibt an, ob der zusätzliche Einfluss einer Variablen noch signifikant ist).

'forward elimination' - startet mit der am höchsten zu Y korrelierten Variablen und fügt dann Schritt für Schritt weitere Variablen hinzu. Als Kriterium wird meist der partielle F-Wert benutzt.

In der Multiregression *einer* Datengruppe erscheint die Bewertung jeder einzelnen Variablen als unproblematisches und sinnvolles Vorgehen, in der Multiregressionsclusterung ist dies jedoch nicht so einfach, da es mehrere Gruppen von Objekten gibt, die sich durch unterschiedliche Zusammenhänge auszeichnen. In jeder dieser Gruppen, den Clustern, ist die Güte der Regressionsfunktion und die Erklärungskraft der einzelnen Variablen eine andere. Der Schluss vom t-Wert jeder Variablen in jedem Regressionscluster auf die Erklärungskraft der Variablen in der gesamten Partition ist also nicht möglich und kann lediglich als Hilfe dienen, die Bedeutung dieser Variablen abzuschätzen. Es gibt nur die Möglichkeit, für jede Variablenzusammenstellung zu bemessen, wie groß die Güte dieses Modells ist und darüber Rückschlüsse darauf zu ziehen, welche Variablen sich am besten eignen. Die drei Vorgehensweisen können also für die gesamte Suche nach dem richtigen Modell Anwendung finden. Diese schließt die Selektion der richtigen Variablen, deren Zusammenstellung und auch die Menge der im Datensatz vorhandenen Multiregressionscluster ein.

2.3.2 Gütemaße für die Modellwahl

Wie oben beschrieben beschränkt sich die Modellwahl nicht nur darauf, die Anzahl von Cluster zu finden, welche die Zusammenhänge in der Grundgesamtheit am besten darstellt, sondern sie sucht auch nach der günstigen Variablenmenge. Wird für jedes dieser Modelle ein Gütewert angegeben, so erhält man eine Gütematrix folgender Gestalt, wobei die Spalten über alle möglichen nichtleeren Teilmengen von $x_1, .., x_m$ gehen:

$$\begin{pmatrix} Variablenmenge \rightarrow & \{x_1\} & \{x_2\} & ... & \{x_1,x_3\} & ... & \{x_1,..,x_m\} \\ Clusteranzahl \downarrow & & & & & & \\ 1 & G_{1,\{x_1\}} & G_{1,\{x_2\}} & ... & G_{1,\{x_1,x_2\}} & & G_{1,\{x_1,..,x_m\}} \\ 2 & G_{2,\{x_1\}} & G_{2,\{x_2\}} & & G_{2,\{x_1,x_2\}} & & G_{2,\{x_1,..,x_m\}} \\ ... & ... & ... & ... & ... & ... & ... \\ K & G_{K,\{x_1\}} & G_{K,\{x_2\}} & ... & G_{K,\{x_1,x_2\}} & ... & G_{K,\{x_1,..,x_m\}} \end{pmatrix}$$

Die Gütewerte aller Teilmengen können so verglichen und die optimale Variablenmenge und Clusteranzahl ausgewählt werden.
Welche Möglichkeiten gibt es nun, die Erklärungskraft der verschiedenen Partitionen zu bemessen? Naheliegend ist es, auf den Wert der Zielfunktion, die der Summe der Residuenquadrate summiert über alle Cluster entspricht, zurückzugreifen. Problematisch ist hierbei, dass mit der Komplexität des Modells die Anpassung an die Datenpunkte in jedem Fall besser, d.h. die Zielfunktion kleiner wird. Daher wird man zumindest für die Clusteranzahl kein Minimum k_{real} finden.
Wird als Modell-Selektionskriterium das Maß der Anpassungsgüte gewählt, wird stets das umfassendste Modell präferiert. Dies widerspricht jedoch dem *Prinzip der Sparsamkeit*, nachdem möglichst einfache Modelle ausgewählt werden sollen [Schlittgen und Streitberg, 1999].
Das AIC-Kriterium und die verwandten Beziehungen versuchen diesen Nachteil auszugleichen.

Akaike Information Criterion - AIC

Das AIC basiert auf der Loglikelihood-Funktion und hat folgenden Struktur [Maindonald und Braun, 2003]:

$$AIC = -2 \cdot logLikelihood + 2 \cdot \#Parameter. \tag{2.30}$$

Die maximierte Loglikelihood hat nach Schlittgen und Streitberg (1999) folgende Form:

$$l(\vec{y}) = \frac{n}{2}(1 + \ln 2\pi) - \frac{n}{2} \ln \sigma^2, \qquad \sigma^2 := \frac{RSS}{n}. \tag{2.31}$$

Beim RSS in Gleichung 2.31 handelt es sich um die Residuenquadratsumme (Siehe Gleichung 2.2). Nun wird der nicht konstante Teil der Loglikelihood-Funktion so modifiziert, dass *sparsame* Modelle belohnt und *verschwenderische* Modelle bestraft werden [Schlittgen und Streitberg, 1999]. AIC setzt sich also aus der Loglikelihood, einem *fidelity-term* als Maß für die Anpassung des Models an die Daten, und einem *penalty-term*, als Strafterm für die Komplexität des Modells, zusammen.

Nach Einsetzen von Gleichung 2.31 (außer konstante Terme) in Gleichung 2.30 ergibt sich:

$$AIC = n \cdot \log(\frac{RSS}{n}) + 2p. \tag{2.32}$$

n markiert die Anzahl der Beobachtungen und p die Anzahl der Parameter

$$p = K \cdot (m + 1). \tag{2.33}$$

K steht hier für die Clusteranzahl und $m+1$ für die Dimension der erklärenden Variablen bzw. den Intercept-Term. Der konstante Teil fällt heraus, da er keinen Einfluss auf den Verlauf des AIC hat und lediglich eine Verschiebung bewirkt. Ein Nachteil des AIC ist, dass es immer noch zu komplexe Modelle auswählt. Daher existieren noch andere Verfahren, wie das Bayesian Information Criterion (BIC), in dem Modelle mit vielen Parametern stärker bestraft werden.

Bayesian Information Criterion

$$BIC = n \cdot log(\frac{RSS}{n}) + log(n) \cdot p \tag{2.34}$$

Da bei der in Kapitel 3 durchgeführten Bewertung der einzelnen Gütemaße an synthetischen Daten keine eindeutigere Markierung des geeigneten Modells durch BIC gegenüber AIC festgestellt wurde, wird in den Anwendungen dieser Arbeit das Akaike Information Criterion benutzt.

F-Test für die Partition

Eine weitere Methode zur Bewertung von Partitionen bei Clusteranalysen ist das Varianz-Verhältnis aus dem F-Test (siehe Abschnitt 2.2.2). Es basiert auf der Berechnung des Verhältnisses von 'between-groups variance' und 'within-groups variance'. Hier soll ein auf die Multiregressionsclusterung angepasstes, ähnliches Gütemaß Verwendung finden. Dazu berechnen wir die Summe der erklärten Abweichungen und gesamten Abweichungen über alle Cluster der Partition und bestimmen deren Verhältnis zueinander. Damit ist dieses Verhältnis (EV/GV) eine Art *Bestimmtheitsmaß der Partition.* $\hat{y}_{ik}$ in Gleichung 2.35 steht für die berechneten Werte der Regressionsgeraden im Cluster k. Deren Abstand zum jeweiligen Clustermittel $\bar{y}_k$ wird für alle k Cluster bestimmt und aufsummiert.

$$EV/GV = \frac{\sum_{k=1}^{K} \sum_{i \in C_k} (\hat{y}_{ik} - \bar{y}_k)^2}{\sum_{k=1}^{K} \sum_{i \in C_k} (y_{ik} - \bar{y}_k)^2} \tag{2.35}$$

Dieses Gütemaß wird im weiteren als *Erk/GesVarianz* bezeichnet. Der zur Partition gehörende F-Wert, welcher wie in 2.26 aus dem Bestimmtheitsmaß der Partition berechnet wird, kann ebenfalls als Gütemaß genutzt werden. In diese Berechnung gehen zusätzlich die auf dieses multilineare Modell angepassten Freiheitsgrade mit ein. Diese angepassten Freiheitsgrade ν_1 und ν_2 werden wie in der gewöhnlichen Regressionsanalyse bestimmt, jedoch unter Beachtung der Zunahme der Parameter durch die K Multiregressionshyperflächen ($\nu_1 = n - K \cdot m - K \cdot 1$ und $\nu_2 = K \cdot m$).

F-Test der Cluster

Es ist ebenfalls möglich, für jedes einzelne Cluster k einer Partition P ein Bestimmtheitsmaß R_k^2 zu berechnen und aus diesem nach Gl. 2.26 den zugehörigen F-Wert zu berechnen. Wie im Abschnitt 2.2.2 beschrieben, lässt sich zu diesem ein p-Wert angeben, welcher die Wahrscheinlichkeit dafür angibt, das die Ablehnung der Nullhypothese $R_k^2 = 0$ falsch ist.
Als Maß für die ganze Partition ist es nun möglich, den größten aller p-Werte einer Partition auszuwählen und diesen mit den *schlechtesten* Clustern anderer Partitionen zu vergleichen. Zum besseren Vergleich werden die p-Werte logarithmiert. Dieses Gütekriterium wird im folgenden mit *log(max(Ftest))* bezeichnet.
Weiterhin ist es möglich, den Mittelwert der logarithmierten p-Werte aller Cluster einer Partition zu bilden. Dieses Kriterium wird mit *mean(p-Value)* bezeichnet.
Die vorgestellten Kriterien können sowohl für die Auswahl einzelner Variablen wie auch für die Suche nach der richtigen Clusteranzahl benutzt werden. Ihre Qualität soll in Kapitel 3 analysiert und diskutiert werden.

2.4 Der Algorithmus

In diesem Abschnitt soll genauer auf die Umsetzung der Multiregressionsclusterung eingegangen werden. Dafür wird die Methode des Simulated Annealing vorgestellt und der die MRC-Analyse durchführende Algorithmus im Überblick beschrieben.
Wie am Beginn dieses Kapitels erläutert, setzt sich die Multiregressionscluster-Analyse aus einer Multiregressions-Analyse und einer Clusterung zusammen. Mittels eines der partitionierenden Clusterung ähnlichen Verfahrens wird die Zielfunktion (2.16) optimiert. Mit jeder Verbesserung der Zielfunktion durch den Wechsel eines Objektes von Cluster A nach Cluster B nähern sich die Regressionscluster der Struktur des Datensatzes mehr und mehr an.
Ein jedoch bei jeder Optimierung auftretendes Problem sind lokale Minima. Der Algorithmus optimiert nach einer bestimmten Funktion, hier nach der Summe der Residuenquadrate und kann bei der Suche nach dem globalen Minimum in ein lokales Minimum fallen (siehe Abb. 2.5). Da der Algorithmus den Wechsel von Objekten in andere Cluster nur bei Verbesserung der Zielfunktion durchführt, würde er an dieser Stelle abbrechen, da bei keinem Objektwechsel eine Verbesserung eintritt. Die Zielfunktion scheint hier ihr Optimum erreicht zu haben. Mit der Methode des Simulated Annealing wird versucht dieses Problem zu beheben.

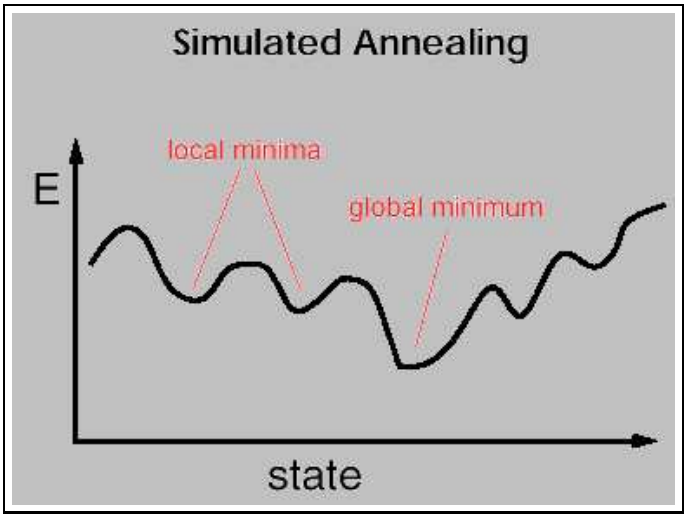

Abb. 2.5: Globales Minimum und lokale Minima. 'E' bezeichnet die Zielfunktion und 'state' symbolisiert die verschiedenen Partitionen.

2.4.1 Simulated Annealing

Diese Methode ähnelt dem physikalischen Prozess des Rekristallisierens von flüssigem Metall im Abkühlungsprozess.
„Zu Beginn besitzt die Schmelze eine hohe Temperatur T und eine ungeordnete Struktur mit der Energie E. Sie wird so langsam abgekühlt, dass sie sich zu jedem Zeitpunkt in einem thermodynamischen Gleichgewicht befindet. Mit fortschreitender Abkühlung des Systems nimmt die Ordnung der inneren Struktur bis zum Grundzustand hin zu. Der Prozess kann als eine adiabatische Zustandsänderung verstanden werden. Ist jedoch die Anfangstemperatur zu niedrig oder der Abkühlungsprozess verläuft zu schnell, kann das System Defekte erzeugen oder in einem metastabilen Zustand erstarren.“ [SNL, 2006]
Man nimmt nun analog an, dass der momentane Zustand des thermodynamischen Systems dem Zustand des kombinatorischen Problems entspricht. Der Energiezustand entspricht dem Wert der Zielfunktion in unserer Berechnung. Bei den metastabilen Zuständen handelt es sich um die lokalen Minima und der Grundzustand des thermodynamischen Systems ist analog zum angestrebten globalen Minimum.
Es sollen nun ähnliche Methoden in der im Multiregressionsclusterung verwendet werden. Die Vor- und Nachteile von zwei Varianten werden hier kurz dargestellt.

Simulated Annealing - konventionell

Bei dieser Variante sind Objektwechsel, bei denen sich die Zielfunktion verschlechtert, während der gesamten Analyse erlaubt. Die Wahrscheinlichkeit für solch einen *verbotenen* Wechsel nimmt während der Analyse ab. Dies kann auf verschiedene Weisen umgesetzt werden. Beispielsweise kann dafür eine Exponentialfunktion mit negativem Exponenten definiert werden. Dieser Exponent setzt sich als Maß für die Dauer der Analyse aus der Anzahl der bisher gewechselten Objekte g, sowie aus der Gesamtanzahl der Objekte n zusammen.

$$h = 1 - exp(-g/n) \tag{2.36}$$

h in Gleichung 2.36 stellt die Hürde dar, welche eine per Zufall aus einer gleichverteilten Menge zwischen 0 und 1 gezogene Zahl überschreiten muss. Für den Fall, dass diese größer ist, wird der *verbotene* Wechsel zugelassen. Die linke Abbildung in 2.6 stellt die resultierende Wahrscheinlichkeit für einen *verbotenen* Wechsel aus Gl. 2.36 dar. In Abbildung 2.6 ist rechts beispielhaft der Verlauf der Zielfunktion über die Anzahl der gewechselten Objekte bei einer Analyse mit konventionellem Simulated Annealing dargestellt.

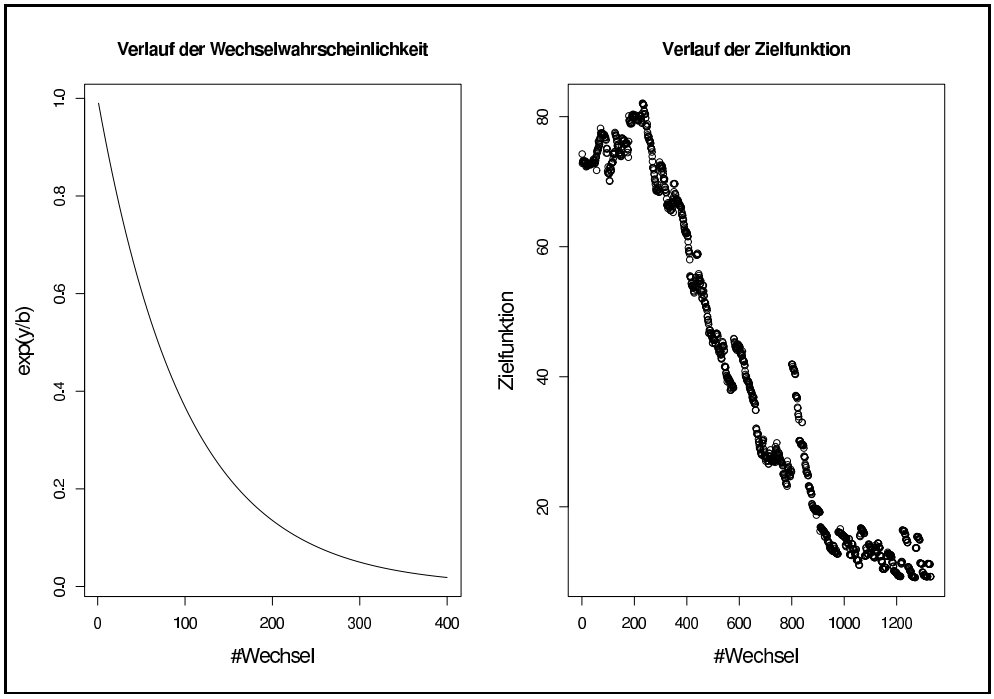

Abb. 2.6: links - Wahrscheinlichkeit für einen Sprung trotz Verschlechterung der Zielfunktion über die Anzahl der bereits gewechselten Objekte. rechts - Verlauf der Zielfunktion über die Anzahl der gewechselten Objekte. Schwankungen sind durch Simulated Annealing verursacht.

Deutlich sind die durch Simulated Annealing verursachten kurzzeitigen Vergrößerungen der Zielfunktion zu erkennen. Eine Änderung in der Häufigkeit der *verbotenen* Wechsel ist, trotz abnehmender Wahrscheinlichkeit, nicht zu erkennen. Dies ist darauf zurückzuführen, dass die Zielfunktion über die Anzahl der bereits gewechselten Objekte dargestellt ist und diese *erlaubten* Wechsel im Verlauf der Analyse immer unwahrscheinlicher werden.
Vorteilhaft an dieser Variante ist das vergleichsweise schnelle Auftreten des besten Minimums. Ein deutlicher Nachteil ist, dass bei nicht eindeutig strukturierten Daten, in denen tiefe lokale Minima existieren, der Algorithmus es nicht schafft, diese wieder zu verlassen. Auch ist es möglich, die Wahrscheinlichkeit für einen *verbotenen* Wechsel über die Variable *Anzahl der nicht gewechselten Objekte* zu bestimmen. Diese Variable misst die Anzahl der Objekte, die ihre Position aufgrund einer daraus folgenden Verschlechterung der Zielfunktion nicht verändert haben. Nähert sich das System nun einem Minimum, wird es für den Algorithmus schwerer einen Wechsel zu finden, bei dem sich die Zielfunktion weiter verbessert - die Variable *Anzahl der nicht gewechselten Objekte* nimmt zu.

Findet ein Wechsel statt, wird diese Variable auf Null zurückgesetzt. Falls sich das System einem Minimum nähert, erhöht sich also die Wahrscheinlichkeit dafür, dass es dieses durch Simulated Annealing wieder verlässt. Dieser Effekt wird für die lokalen Minima angestrebt, muss jedoch beim globalen Minimum vermieden werden. Das ist der große Nachteil dieser Variante. Sie setzt voraus, dass der Anwender eine grobe Vorstellung von der Lage des globalen Minimums hat. Er muss vorher die Grenzen der Anwendung des Simulated Annealings definieren. Diese Grenzen sind natürlich in jedem Datensatz verschieden und nur durch häufige Experimente abzuschätzen.

Simulated Annealing - alternativ

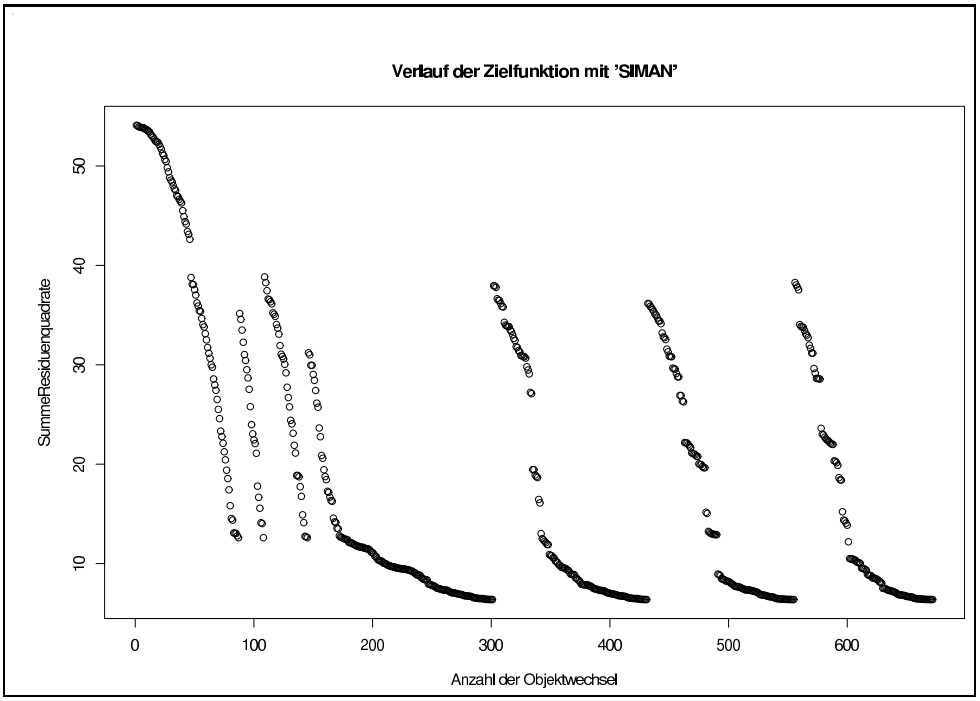

Abb. 2.7: Verlauf der Zielfunktion beim alternativen Simulated Annealing. Das lokale Minimum wird nach mehrmaligem Auftreten als globales angenommen.

Diese Variante des Simulated Annealing weicht etwas von der ursprünglichen Idee ab. Sie zählt ebenfalls die Anzahl der Objekte, die nicht ihre Position gewechselt haben - also wieder ein Maß für die Nähe zu einem Minimum. Die maximal mögliche Anzahl von nicht stattgefundenen Sprüngen setzt sich zusammen aus dem Produkt von Objektanzahl und Clusteranzahl minus 1 (ein Objekt hat nur noch K-1-Cluster zum Wechsel zur Verfügung).

Ist dieser Maximalwert erreicht, reagiert der Algorithmus mit dem zufälligen Wechsel einer vorher definierten Anzahl von Objekten. Dieser Austausch geschieht unabhängig von der dadurch verursachten Veränderung der Zielfunktion. Vorteilhaft gegenüber der vorherigen Variante ist, dass keine Einschränkung auf einen begrenzten Bereich der Zielfunktion notwendig ist. Jedoch muss in gleichem Maße das globale Minimum gegenüber den lokalen Minima gesondert behandelt werden. Dies geschieht durch die Annahme, dass das System nur mit geringer Wahrscheinlichkeit mehrmals in dasselbe Minimum fällt. Lediglich im globalen Minimum würde das System zwangsläufig mehrmals landen. Daher zählt der Algorithmus, wie oft er ein Minimum schon erreicht hat.
Die Wahrscheinlichkeit für das Auftreten des globalen Minimums (p_{GM}) entspricht dem Anteil der Partitionen an der Gesamtmenge der möglichen Partitionen, von denen eine einfache Optimierung der Zielfunktion zum globalen Minimum (GM) führt. Je nach Struktur des Datensatzes kann p_{GM} unterschiedlich ausfallen. Werden die Objekte q-mal neu auf K Cluster verteilt und die Zielfunktion jedes Mal neu optimiert, erhöht das die Wahrscheinlichkeit für das Auffinden des globalen Minimums mit q in folgender Weise. Die Wahrscheinlichkeit dafür, dass bei q-maliger Wiederholung GM mit p_{GM} *mindestens* einmal auftritt, berechnet man über das gegenteilige Ereignis. Dieses lautet: GM tritt bei keinem der q Wiederholungen auf ($\bar{GM}$).

$$p(\bar{GM})^q = (1 - p(GM))^q \tag{2.37}$$

Würde die Wahrscheinlichkeit, bei einem Durchgang auf das Ereignis GM zu treffen, beispielsweise 50% betragen, wäre bei $q = 3$ die Wahrscheinlichkeit dafür, mindestens bei einem Durchgang im globalen Minima zu enden, 87.5%. Bei $q = 4$ wären es bereits 93%. Unter Berücksichtigung der Zunahme der Rechenzeit ist nun abzuwägen, wie groß dieses q und die Anzahl der Objekte, welche neu verteilt werden, gewählt werden. Jedoch sollte dieser Wert nahezu der Gesamtzahl an Objekten entsprechen um zu gewährleisten, dass ein eventuelles lokales Minimum auch verlassen werden kann.
Aufgrund der Annahme, dass die später verwendeten Daten eine eher stark variierende Struktur aufweisen, wird im weiteren nur die alternative Methode zur Verwendung kommen.

2.4.2 Das Bootstrapping-Verfahren

Im Anschluss an eine MRC-Analyse sollen die Ergebnisse mittels des Bootstrapping-Verfahrens auf ihre Robustheit überprüft werden. Im Folgenden wird dieses Verfahren vorgestellt.

Bei Bootstrapping-Verfahren handelt es sich um rechenintensive Methoden der statistischen Analyse zur Bestimmung von Standardfehlern, Konfidenzintervallen oder Signifikanztests [Davison und Hinkley, 1997]. Die Grundidee ist, neue Daten aus den bereits vorhandenen Daten auszuwürfeln. Dies kann direkt oder über ein gefittetes Modell geschehen. Mittels dieser neuen Datensätze können dann die gesuchten Variablitäten von Größen bestimmt werden, ohne dass auf langwierige und auf Näherungen basierende analytische Berechnungen zurückgegriffen werden muss.
Für den hier angestrebten Nachweis der Stabilität der MRC-Ergebnisse werden zuerst die Eigenschaften der Regressionscluster näher betrachtet. Man bestimmt Mittelwerte $\hat{x_k}$ und Varianzen s_k^2 der Residuen e_k in jedem der K Cluster. Nun wird angenommen, die Verteilung der Residuen spiegelt die Verteilung einer normalverteilten Grundgesamtheit wider.

$$\hat{x_k} = \mu_k \text{ bzw. } s_k^2 = {\sigma_k}^2$$

Es werden dann neue Sample aus einer normalverteilten Grundgesamtheit mit den Eigenschaften $\hat{x_k}$ bzw. s_k^2 gezogen. Die Menge der neu gezogenen Sample entspricht der Objektanzahl in den ursprünglichen Clustern. Es existiert nun ein neu generierter Datensatz mit statistisch dem Ursprungsdatensatz äquivalenten Eigenschaften. Dieser wird im Weiteren als *resample*-Datensatz bezeichnet. Von diesem *resample*-Datensatz können per Regression die Eigenschaften der einzelnen *resample*-Cluster bestimmt werden. Der *resample*-Datensatz wird nun in gleicher Weise wie der Ursprungsdatensatz der MRC-Analyse mit dem vorher bestimmten K unterzogen. Dieser Prozess des *resamplings* und der darauf folgenden MRC-Analyse wird R-mal wiederholt.
Im Anschluss können die Eigenschaften der *resample*-Cluster und der MRC-Cluster in einem Histogramm verglichen werden. Weiterhin werden die normalverteilten Anstiege und Intercepts der *resample*-Cluster im Mittel den Eigenschaften der Ursprungscluster entsprechen. Der Vergleich der Standardabweichung der Parameter der Ursprungscluster mit denen der MRC-Cluster, sowie der Vergleich der Mittelwerte der Verteilungen der Parameter der *resample*-Cluster mit den Verteilungen der Parameter der MRC-Cluster, ermöglicht eine Einschätzung der Stabilität der gefundenen Partition.
Mittels Standardabweichungen der Verteilungen der MRC-Cluster kann ein Vertrauensintervall für die Regressionsparameter angegeben werden. Diese Intervallgrenzen können entweder wie in Gleichung 2.28 bestimmt oder durch Aufsummieren bis zur gewählten Irrtumswahrscheinlichkeit an den Rändern der Verteilungen der Parameter erhalten werden.

2.4.3 Drehung des Koordinatensystems

Es ist möglich, dass die Lage eines Clusters im Raum ungünstig für die Anwendung des Algorithmus ist. Dies ist der Fall, wenn ein Cluster nahezu parallel zur y-Achse liegt, da dann die bei der Regression berechneten Residuen, also die Abstände in y-Richtung, besonders groß sind. Um das Auffinden des globalen Minimums zu ermöglichen, ist es notwendig, die Daten im Koordinatensystem durch Anwendung einer Drehmatrix zu drehen.

$$M = \begin{pmatrix} 1 & & & & & 0 \\ & \ldots & & & & \\ & & \cos\alpha & & \sin\alpha & \\ & & & \ldots & & \\ & & -\sin\alpha & & \cos\alpha & \\ & & & & & \ldots \\ 0 & & & & & 1 \end{pmatrix} \tag{2.38}$$

Abb. 2.8: Verallgemeinerte Drehmatrix. Im 2-d Fall verbleiben nur die 4 Terme im Zentrum der Matrix.

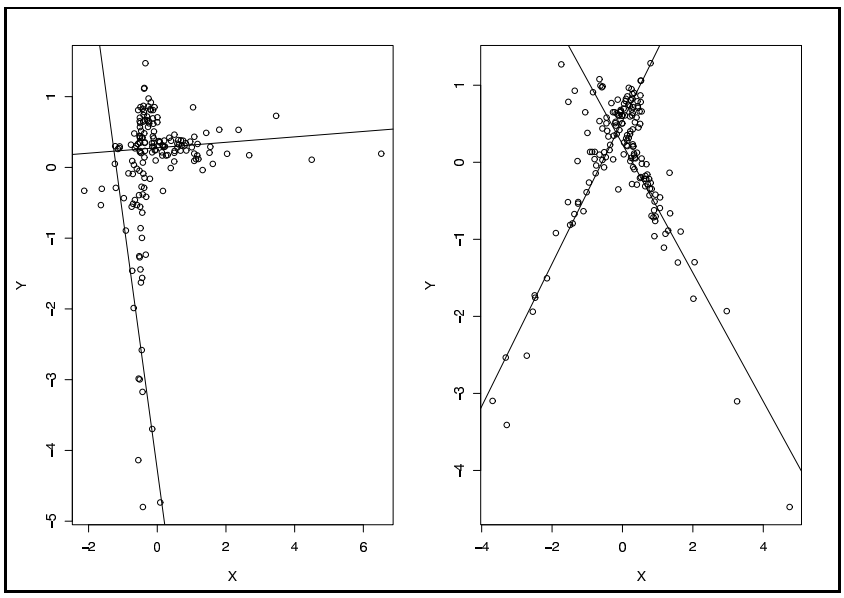

Abb. 2.9: Drehung eines Beispieldatensatzes. Die scheinbar unterschiedlichen Winkel zwischen den Regressionsgeraden beruhen lediglich auf der Darstellungsart.

In Abbildung 2.9 ist beispielhaft ein Datensatz dargestellt, der verdeutlicht, wie notwendig eine Drehung des Datensatzes sein kann. An den eingezeichneten Regressionslinien für die zwei leicht identifizierbaren Cluster ist gut zu erkennen, ob der Algorithmus diese gefunden hat oder nicht. Im linken Bild ist zu sehen, dass der Algorithmus wegen der Steilheit der Objektwolke nicht in der Lage ist, diese zu bestimmen. Auf die Bewertung mittels einer Gütefunktion soll hier noch nicht eingegangen werden (siehe Kapitel 3). Die Drehung aller Datenpunkte um einen wählbaren Winkel α stellt kein Problem dar. Jedoch die Rücktransformation der berechneten Parameter der Regressionsgeraden ist etwas umständlicher und soll hier kurz erläutert werden. Als Beschreibung der Regressionsgeraden werden die Schnittpunkte mit den Koordinatenachsen ausgewählt (hier zweidimensional !).

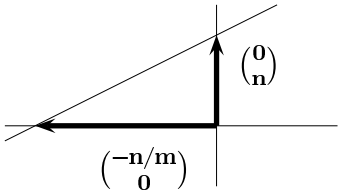

Abb. 2.10: Vor der Drehung

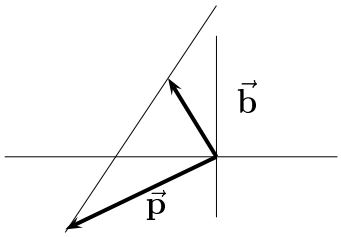

Abb. 2.11: Nach der Drehung

Diese beiden Vektoren werden mittels Drehmatrix zu $\vec{b}$ und $\vec{p}$ gedreht (Siehe Abb. 2.10 und 2.11) . Da die Datenmatrix zu Beginn der Berechnung auf den Mittelwert 0 und die Standardabweichung 1 normiert wurde, müssen diese Werte jetzt mit $y_N \cdot sd(y) + mean(y)$ wieder zurücknormiert werden. Man erhält dabei für $\vec{b}$ und $\vec{p}$:

$$\begin{pmatrix} b_x \\ b_y \end{pmatrix} = \begin{pmatrix} n \cdot \sin\alpha \cdot sd(x) + mean(x) \\ n \cdot \cos\alpha \cdot sd(y) + mean(y) \end{pmatrix} \tag{2.39}$$

und

$$\begin{pmatrix} p_x \\ p_y \end{pmatrix} = \begin{pmatrix} (-n/m) \cdot \cos\alpha \cdot sd(x) + mean(x) \\ (+n/m) \cdot \sin\alpha \cdot sd(y) + mean(y) \end{pmatrix}. \tag{2.40}$$

Aus diesen beiden Vektoren des Rohdatenraumes muss nun noch die zugehörige Gerade berechnet werden. Wir haben einen Punkt und einen Vektor gegeben und können damit die Gerade angeben als: $g(\vec{x}) = \vec{p} + r \cdot (\vec{b} - \vec{p})$. Die x- bzw. y-Komponente dieser Gerade wird Null gesetzt und in die jeweils andere Komponente eingesetzt. Dies ermöglicht die Berechnung der Schnittpunkte mit der x- bzw. y-Achse. Mit $n = y_0$ und $m = -y_0/x_0$ können wir den Verlauf des Regressionsclusters innerhalb der Rohdaten angeben.

2.4.4 Ablaufschema

Für die Umsetzung des Algorithmus, der Datenaufbereitung im Vorfeld, sowie der graphischen Darstellung der Ergebnisse wurde die Open-Source-Programmiersprache R verwendet [www.r-project.org]. Sie wurde für die statistische Analyse und Darstellung entwickelt. Viele Prozesse können durch die Benutzung einer Vielzahl implementierter statistischer und graphischer Methoden vereinfacht werden.
In Abbildung 2.12 ist eine Übersicht zum Ablauf und zur Funktionsweise des Algorithmus dargestellt. Die Darstellung bezieht sich auf die Verwendung des alternativen Simulated Annealing, da dieses hauptsächlich zur Anwendung kam.
Der Algorithmus beginnt mit dem Einlesen der Datenmatrix, welche vorher aus verschiedenen Datenquellen zusammengestellt wurde. Die Zeilen der Matrix entsprechen den n Objekten, die m Spalten beinhalten die verschiedenen Variablen. Es wird eine Anfangspartition erstellt, indem die Objekte zufallsgeneriert auf die vorher festgelegte Clusteranzahl verteilt werden. Der erste Wert der Zielfunktion wird berechnet. Nun wird für jedes Objekt geprüft, ob der Wechsel in ein anderes Cluster eine Verbesserung der Zielfunktion bewirkt. Ist dies der Fall, findet der Wechsel statt und der Algorithmus geht zum nächsten Objekt bzw. Cluster über. Ist dies nicht der Fall, wird die Zahl der hintereinander nicht gewechselten Objekte bestimmt. Diese Menge definiert ein Minimum, wenn keines der Objekte in einer Berechnungsrunde gewechselt wurde. Handelt es sich nicht um ein Minimum, geht der Algorithmus zum Wechsel zum nächsten Cluster über. Bringt keiner der K Cluster eine Verbesserung der Zielfunktion mit sich, wird zum nächsten Objekt gewechselt. Wenn es sich um ein lokales Minimum handelt wird daraufhin eine bestimmte Anzahl von Objekten per Zufall auf die Cluster verteilt und der Algorithmus beginnt wieder beim Berechnen der Zielfunktion.
Erst wenn ein lokales Minimum mit einer vorher festgelegten Häufigkeit wieder erreicht wurde, wird es als globales Minimum angesehen und die Berechnung ist beendet. Weiterhin kann zur Einschränkung der Rechenzeit die Anzahl der Objektwechsel begrenzt werden.

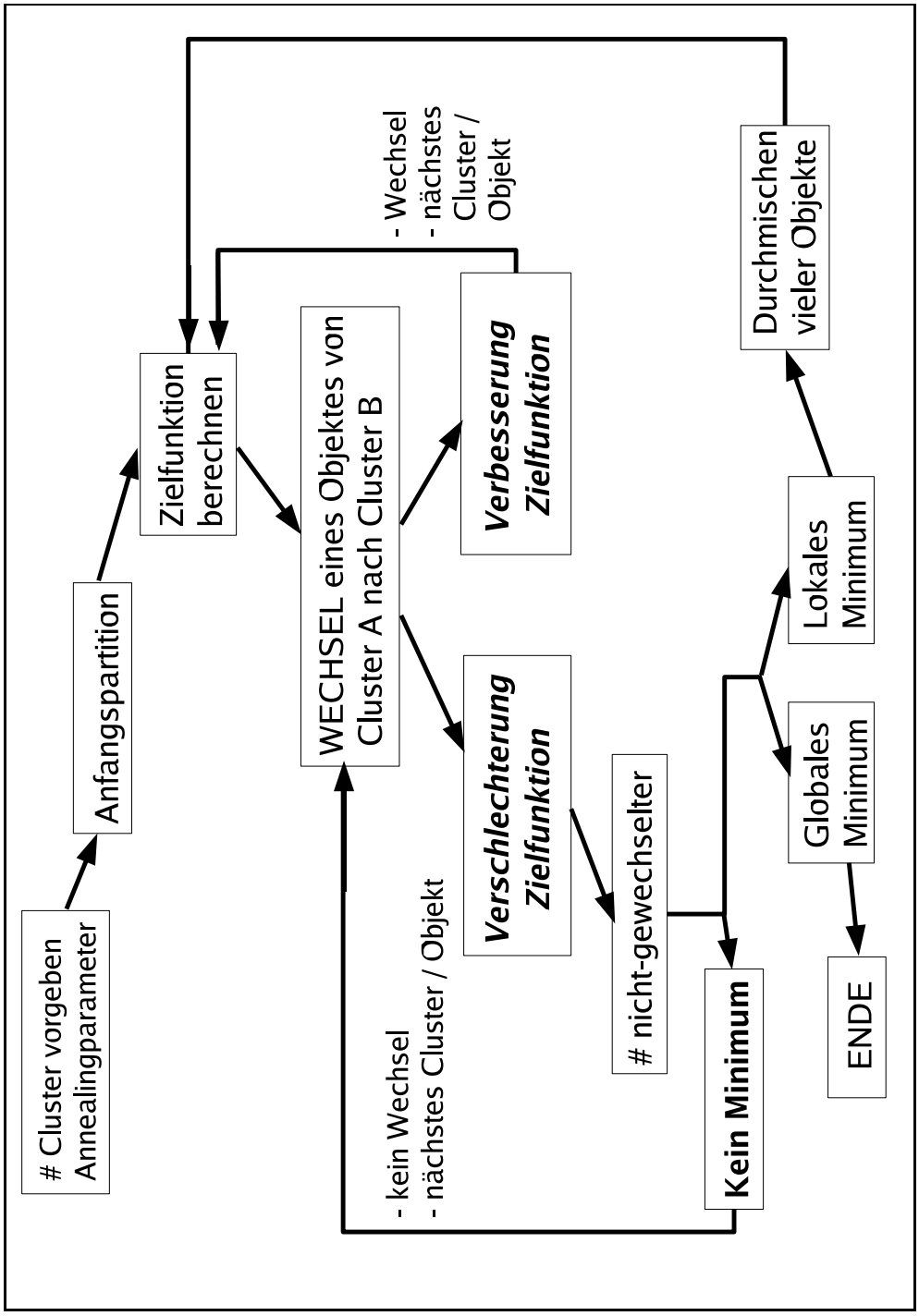

Abb. 2.12: Ablaufschema

2.5 Datenvorbearbeitung

$$V_{ij} = \begin{pmatrix} y_1 & x_{11} & ... & x_{1m} \\ y_2 & x_{21} & ... & x_{2m} \\ ... & ... & ... & ... \\ y_n & x_{n1} & ... & x_{nm} \end{pmatrix} \tag{2.41}$$

Zu Beginn einer Datenanalyse ist es notwendig, sich mit der Verteilung und dem strukturellen Aufbau der Daten zu beschäftigen. Der Datensatz ist in einer $(n \times m)$-Matrix zusammengefasst. Wie in Abbildung 2.41 zu sehen, besteht die Matrix V_{ij} aus n Zeilen, welche den n Objekten entsprechen, sowie m Spalten. Diese stehen für die m Eigenschaften der Objekte.
Eine erste Möglichkeit für einen guten Überblick bietet die Projektionsdarstellung. Sie stellt alle Variablen aus einem mehrdimensionalen Datensatz paarweise zweidimensional dar. Bei empirischen Daten wird das Fehlen einiger Datenpunkte unvermeidlich sein. Es bestünde die Möglichkeit, diese Lücken durch Mittelung der umgebenden Werte zu füllen. Da es sich bei den später verwendeten Objekten um Länder handelt, würde dies eine Glättung der Werte über Ländergrenzen hinweg bedeuten, obwohl gerade dort sehr große Schwankungen zu erwarten sind. Daher erscheint dieses Vorgehen bei unseren Daten unangebracht. Die Beobachtungen mit Fehlstellen werden vollständig herausgenommen, eine Zeile in der Datenmatrix wird also gelöscht.

2.5.1 Standardisierung

In mehrdimensionalen Datensätzen haben einzelne Komponenten häufig sehr unterschiedliche Wertebereiche [Runkler, 2000]. Um jeder Variablen dasselbe Gewicht zu geben, müssen die Daten vor der Analyse standardisiert werden. Dies ist auf verschiedene Weisen möglich, z.B. durch die z-Transformation oder die μ-σ-Standardisierung. In dieser Arbeit wird nur letztere verwendet. Dafür werden spalten- bzw. variablenweise die Daten gemittelt ($\bar{x}_j$) und die Varianzen s_j berechnet. Mit diesen Größen werden die i Datenpunkte einheitlich auf den Mittelwert 0 und die Varianz 1 transformiert.

$$x_{ij}^{Norm} = \frac{x_{ij} - \bar{x}_j}{s_j} \tag{2.42}$$

2.5.2 Ausreißer

Gibt es im Datensatz Ausreißer, das heißt Punkte, die einen ungewöhnlich großen Abstand zum Rest der Punkte aufweisen, stellt sich die Frage, ob diese auf Messfehler zurückzuführen sind oder ob sie wichtige Informationen für die Analyse beinhalten.
Um dies festzustellen kann man z.B. auf die *2-Sigma-Regel* zurückgreifen. Sie klassifiziert jede Beobachtung als Ausreißer, sobald mindestens eine Komponente um mehr als das zweifache der Standardabweichung vom Mittelwert abweicht [Runkler, 2003]. Weiterhin kann man durch die Darstellung der einzelnen Komponenten in einem *Histogramm* auf ungewöhnliche Werte aufmerksam werden. Mit diesem ist es möglich, eine grob aufgelöste Darstellung der Häufigkeitsverteilung einer Variablen zu erhalten.
Nach einer *Plausibilitätskontrolle* der gefundenen Messwerte kann entschieden werden, ob diese Punkte weiter in die Analyse eingehen. Neben dem vollständigen Herausnehmen der entsprechenden Beobachtung gibt es andere Möglichkeiten Ausreißer zu behandeln. Beispielsweise Ersetzen durch den globalen Mittelwert, durch die nächsten Nachbarn, durch die Minimal- oder Maximalwerte (ohne den oder die Ausreißer) der betroffenen Komponente u.a. [Runkler, 2000]. In dieser Arbeit wird im Folgenden mit als Fehlwert erkannten Ausreißern so verfahren, dass die gesamte Beobachtung aus dem Datensatz entfernt wird.

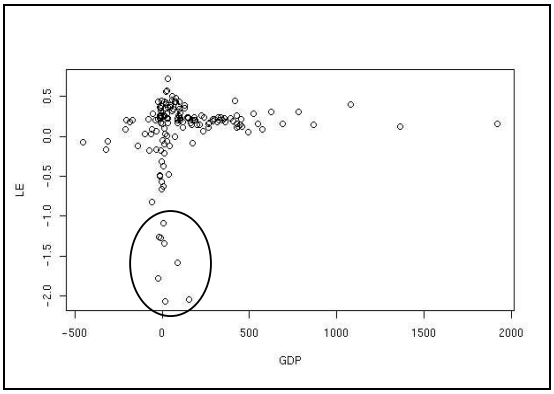

Abb. 2.13: HIV-Staaten. Dargestellt sind Änderung der Lebenserwartung über Änderung des GDP/Kopf.

Staat	Reihenfolge HIV-Rate
Botswana	2
Kenya	17
Lesotho	3
Namibia	6
SouthAfrica	5
Swaziland	1
Zambia	7
Zimbabwe	4

Tab. 2.2: Rangordnung nach HIV-Rate.

Hier soll beispielhaft eine Plausibilitätsanalyse von auffälligen Werten in einem Datensatz durchgeführt werden. Bei der zweidimensionalen Darstellung handelt es sich um die Änderung der Lebenserwartung in den 90-er Jahren über die Änderung des Bruttoinlandsproduktes in den 90-ern.

Wie in Abbildung 2.13 gut zu erkennen, existiert eine Gruppe von Staaten mit einem auffallend negativen Bevölkerungswachstum. Handelt es sich um falsche Messwerte oder können die starken Abweichungen zu den restlichen Daten inhaltlich begründet werden? In Tabelle 2.2 sind die Staaten innerhalb der Ellipse mit der Reihenfolge der Rate der HIV-Betroffenen in dem jeweiligen Land aufgelistet.
Es ist gut zu erkennen, dass sämtliche Staaten mit auffällig stark fallender Lebenserwartung zu den am stärksten von HIV betroffenen Ländern zählen. Es gibt also eine sinnvolle Erklärung für die große Abweichung dieser Werte. Wie weiter mit diesen Werten verfahren wird, hängt vom speziellen Fall der Analyse ab. Auf die Staaten mit besonders hohem Bruttoinlandsprodukt, welche ebenfalls näher untersucht werden müssten, soll hier nicht weiter eingegangen werden.

Im folgenden Kapitel wird die Methode der Multiregressionsclusterung an synthetischen und realen Daten angewendet.

Kapitel 3

Anwendung der Methoden

3.1 Anwendung auf synthetische Daten

Bevor der Algorithmus auf reale Daten angewandt wird, wird die Funktionalität sowohl des Algorithmus selbst als auch der Gütemaße an synthetisch erzeugten Daten getestet. Der Vorteil dieser Daten ist die bekannte Struktur, die es ermöglicht, Ergebnisse einer Analyse objektiv zu bewerten.

3.1.1 Erstellen der synthetischen Daten

Zur Generierung der Daten wird eine vorher festgelegte Zahl (n) von Punkten aus einer gleichverteilten Menge bestimmt. Dies wird für jede der m erklärenden Dimensionen $(X_1, X_2, .., X_m)$ sowie für jedes Cluster wiederholt. Nun werden für jedes Cluster k (k=1..K) die Punkte nach folgendem Berechnungsmuster bestimmt:

$$y_k = b_{0k} + b_{1k}x_{1k} + .. + b_{mk}x_{mk} + e_k. \tag{3.1}$$

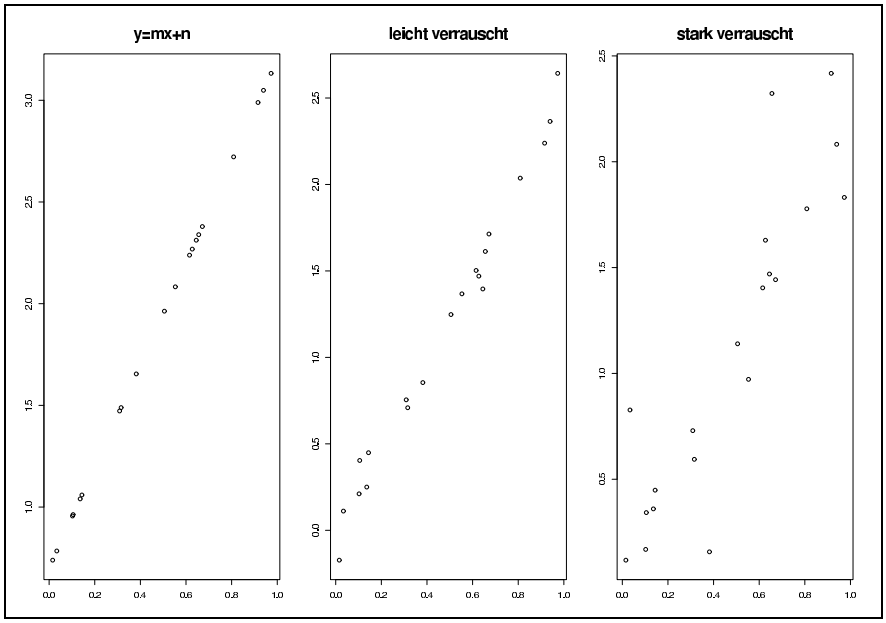

Abb. 3.1: links - Generierte Gerade; mitte - Schwaches aufaddiertes Rauschen; rechts - Starkes aufaddiertes Rauschen; Die Beschriftungen der x- und y-Achsen wurde der Übersicht halber weggelassen.

Die $b_{ij}'s$ sind die wählbaren Paramter einer Hyperfläche. Beim letzten Term handelt es sich um einen Vektor, welcher die Stärke des Verrauschens der Hyperfläche definiert.

Die Elemente des Vektors werden einer normalverteilten Menge mit dem Mittelwert $\mu = 0$ und einer wählbaren Standardabweichung σ^2 entnommen. Die bestimmten Datenreihen werden wie in Abbildung 2.41 als eine $(n \times m)$-Matrix aufgefasst.
In Abbildung 3.1[1] ist beispielhaft die Entstehung eines solchen synthetischen Clusters dargestellt.

3.1.2 Prüfen der Gütemaße

Wie oben bereits erwähnt, liegt ein grundlegendes Problem der Analyse in der Wahl angemessener Gütemaße. Nur wenn deren Einsatz sich an den künstlichen Daten bewährt hat, kann er an Daten mit unbekannter Struktur erfolgen. Die Prüfung verläuft auf drei verschiedenen Ebenen.
Anwendung auf einen ...
...zufallsverteilten 2-d Datensatz
...schwach verrauschten 2-d Datensatz mit multilinearer Struktur
...stark verrauschten 2-d Datensatz mit multilinearer Struktur

Datensatz ohne Struktur

Bei diesem Datensatz sollten die Gütefunktionen bei der Verteilung der Datenpunkte auf Partitionen mit von 1 bis 10 zunehmender Clusterzahl einen Verlauf aufweisen, welcher keine Clusteranzahl präferiert. Mit jedem hinzugenommenen Cluster wird im Mittel der Abstand aller Punkte zu den Regressionsgeraden geringer und die Erklärungskraft der Partitionen selbst in einem Datensatz ohne Struktur besser. Da die Anzahl der Datenpunkte auch in diesem Datensatz beschränkt ist (200 Datenpunkte), würde es auch eine Clusteranzahl K geben, bei der die vollständige Varianz erklärt ist. Dies ist spätestens dann der Fall, wenn jeweils 2 Datenpunkte in einem Regressionscluster liegen.
Die Analyse wurde für die Partitionen mit der Clusteranzahl 1 bis 10 bei zehnmal neu ausgewürfeltem Rohdatensatz durchgeführt. Die sich daraus ergebenden 100 Gütewerte sind in jeder der 5 Gütegrafiken in Abbildung 3.2 eingetragen. Die fünf verwendeten Gütemaße wurden in Abschnitt 2.3.2 näher vorgestellt.
In Abbildung 3.2(1) ist beispielhaft einer der zehn zweidimensionalen zufallsverteilten Datensätze dargestellt. Abbildung 3.2(2) stellt den Verlauf des Anteils der erklärten gegenüber der gesamten Varianz dar.

[1] Im Großteil der folgenden Darstellungen werden die Beschriftungen der x- und y-Achsen aus Gründen der Übersichtlichkeit weggelassen.

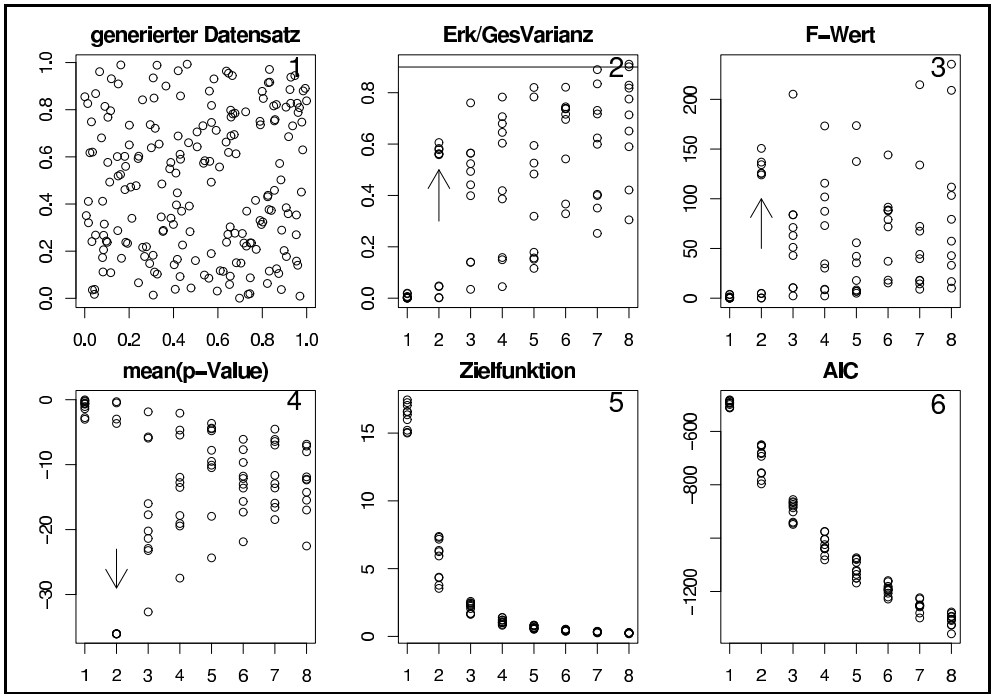

Abb. 3.2: Beispieldatensatz und Gütekriterien für Clusterzahlen K=1..8 des 2-d Datensatzes ohne Struktur. Die Ergebnisse für die gekreuzten Cluster sind mit Pfeilen markiert.

Die Abbildung 3.2(3) und Abbildung 3.2(4) geben die Ergebnisse der Signifikanzanalyse wieder. Alle drei Gütefunktionen haben keinen eindeutigen Verlauf und es wird wie erwartet keine Clusteranzahl präferiert. Auffällig ist in Bild 3.2(4) jedoch das bei $K = 2$ auftretende Minimum. Dieses lässt sich darauf zurückführen, dass bei einer nahezu gleichverteilten Datenwolke die Erklärungskraft bei gekreuzten Regressionsgeraden vergleichsweise gut ist. Dieser Umstand ist auch in 3.2(2) bei $K = 2$, allerdings etwas schwächer, auszumachen. In Abbildung 3.2(5) ist die Zielfunktion selbst dargestellt. Sie hat einen stetig fallenden Verlauf. Hier ist deutlich zu erkennen, wie eine Zunahme an Clustern eine Verbesserung der Anpassung mit sich bringt. Die Abweichungen, verursacht durch unterschiedliche Datensätze, sind gering. Beim Akaike Information Criterion ist der Verlauf durch die zunehmende Erklärungskraft und die verbesserte Zielfunktion ebenfalls stetig fallend. Die Zuverlässigkeit der Analysemethode hängt, wie die folgende Untersuchung zeigen wird, von der Stärke des statistischen Rauschens ab, welches über die linearen Strukturen gelegt wurde. Daher werden Analysen bei unterschiedlichen Rauschstärken gemacht.

Datensatz mit schwach verrauschter multilinearer Struktur

Zunächst wird die Multiregressionscluster-Analyse auf einen zweidimensionalen Datensatz angewendet, dessen Datenpunkte sich in vier gut trennbare Cluster aufteilen lassen. Das aufaddierte Rauschen ist hier also gering gewählt.
Ein Beispieldatensatz und die Gütekriterien sind in Abbildung 3.3 dargestellt. Abbildung 3.3(1) zeigt den analysierten Datensatz nur beispielhaft, da die MRC-Analyse zehnmalig mit neu ausgewürfeltem Datensatz durchgeführt wurde. In Abbildung 3.3(2) folgt der Anteil der mit den K Clustern erklärten Varianz an der gesamten Varianz des Datensatzes. Die hier und im weiteren bei 0.9 eingefügte Linie soll die Marke, ab welcher 90 % der Varianz erklärt sind, leichter erkennbar machen. Es ist zu sehen, dass diese Marke bei einigen Durchgängen schon bei 3 Clustern leicht, bei 4 jedoch deutlich überschritten wird. Beim F-Wert in Abbildung 3.3(3) ist ebenfalls eine deutliche Stufe von $K = 3$ zu $K = 4$ zu erkennen. Die mittleren Signifikanzwerte der Cluster in Abbildung 3.3(4) zeigen bei $K = 4$ die geringsten Schwankungen zwischen den zehn Durchgängen.

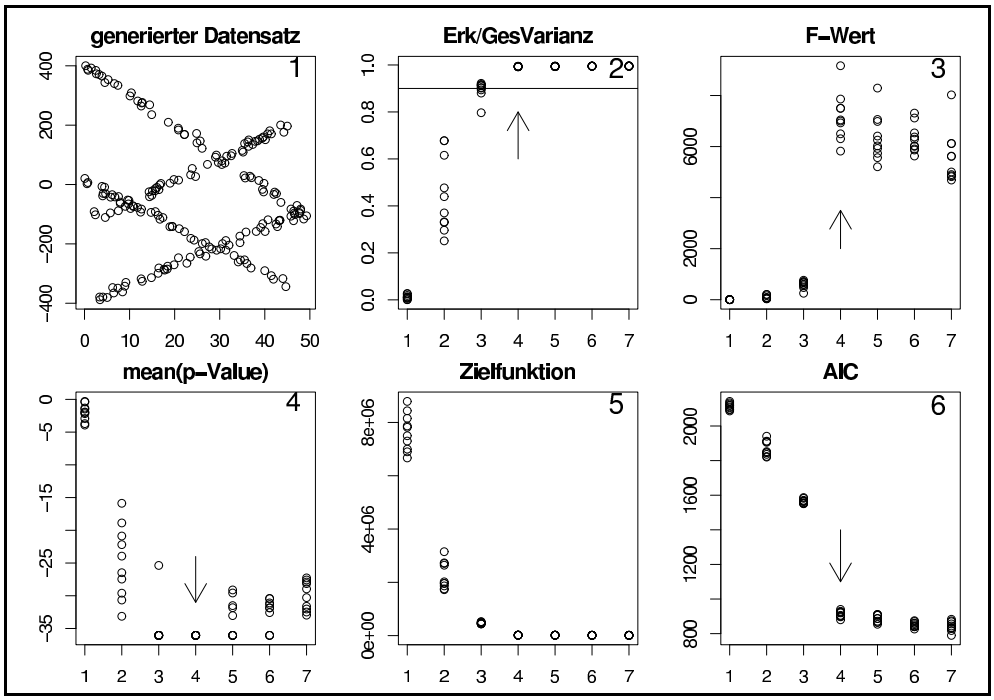

Abb. 3.3: Beispieldatensatz und Gütekriterien für Clusterzahlen K=1..8 vom 2-d Datensatz mit 4 Clustern und schwachem Rauschen.

Bei $K = 3$ bis $K = 6$ endeten einige Durchgänge mit einem gleich guten Ergebnis. Abbildung 3.3(5) stellt den Verlauf der Zielfunktion dar, welcher höchstens über eine abnehmende Änderungsrate auf eine Clusterzahl schließen lassen würde. Ab $K = 4$ ist diese Änderungsrate, im Vergleich zur vorherigen Änderung, sehr klein.
Das *Akaike Information Criterium* in Abbildung 3.3(6) stärkt die Annahme von $K = 4$ als günstige Partitionsgröße mit einer deutlichen Stufe zwischen $K = 3$ und $K = 4$.
Es ist also festzustellen, dass die Gütemaße es bei diesem leicht verrauschten Datensatz gut ermöglichen, auf die im Datensatz steckende Struktur zu schließen. Ein Vergleich der Parameter der durch den MRC-Algorithmus gefundenen Regressionsgeraden mit den vorher festgelegten belegt die Übereinstimmung. Auch hat sich bei Versuchen mit veränderten Parametern der Geraden, anderen Clusteranzahlen aber ähnlich niedrigen Rauschstärken ein vergleichbares Ergebnis gezeigt.

Datensatz mit stark verrauschter multilinearer Struktur

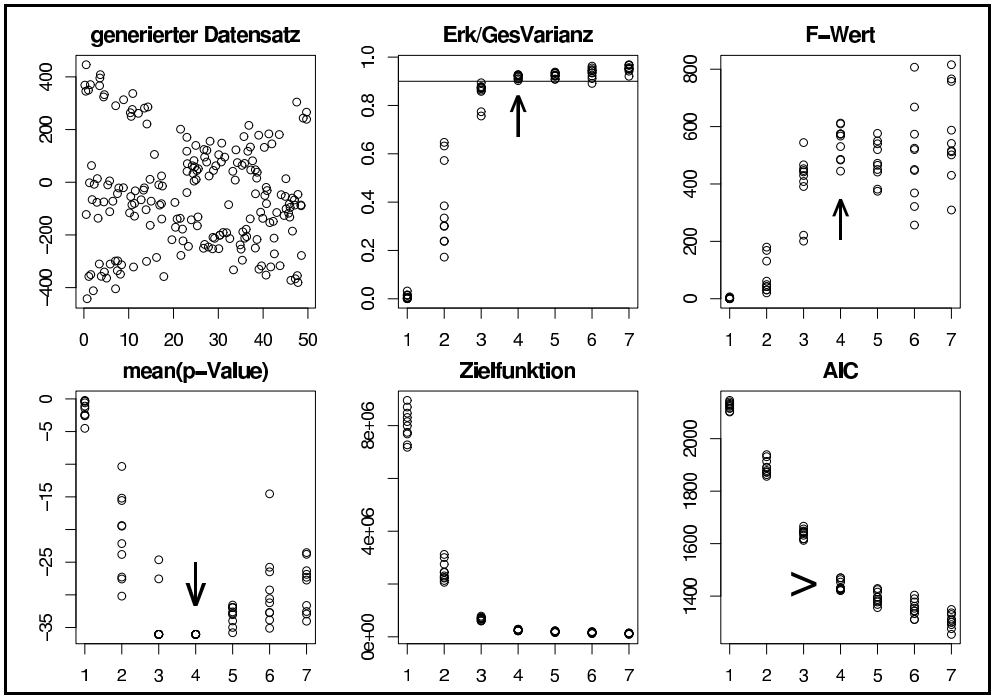

Abb. 3.4: Beispieldatensatz und Gütemaße zur Analyse eines stark verrauschten Datensatzes.

Nun soll ein Datensatz untersucht werden, der in den Eigenschaften der Cluster (Anstieg, Intercept) dem Datensatz in Abbildung 3.3 entspricht. Der Datensatz besteht ebenfalls aus 200 Objekten. Das aufaddierte Rauschen ist jedoch um den Faktor 4 vergrößert worden. Die Analyse wurde von Clusteranzahl 1 bis 7 durchgeführt.
In Abbildung 3.4 ist im ersten Bild wieder beispielhaft der Datensatz aus einer der zehn Wiederholungen mit neu erstelltem Datensatz dargestellt. In folgender Abbildung für die erklärte Varianz sieht man, dass bei $K = 4$ über 90% der Varianz erklärt sind. In Abbildung 3.4(3) gibt es für den F-Wert der vierten Partition ein lokales Maximum. Die mittleren Signifikanzlevel (Abb. 3.4(4)) sind wiederum bei $K = 4$ über die zehn Wiederholungen sehr stabil. Das Minimum dieser Gütefunktion wird auch bei $K = 3$ und $K = 5$ erreicht. Die Zielfunktion verlangsamt ihren Abfall ab der vierten Partition deutlich. Beim *Akaike Information Criterium* gibt es bei dieser Rauschstärke keinen deutlichen Sprung mehr. Lediglich ein *Knick* im Verlauf ist zu erkennen, der ebenfalls die vierte Partition markiert.

Datensatz mit variierender Varianz

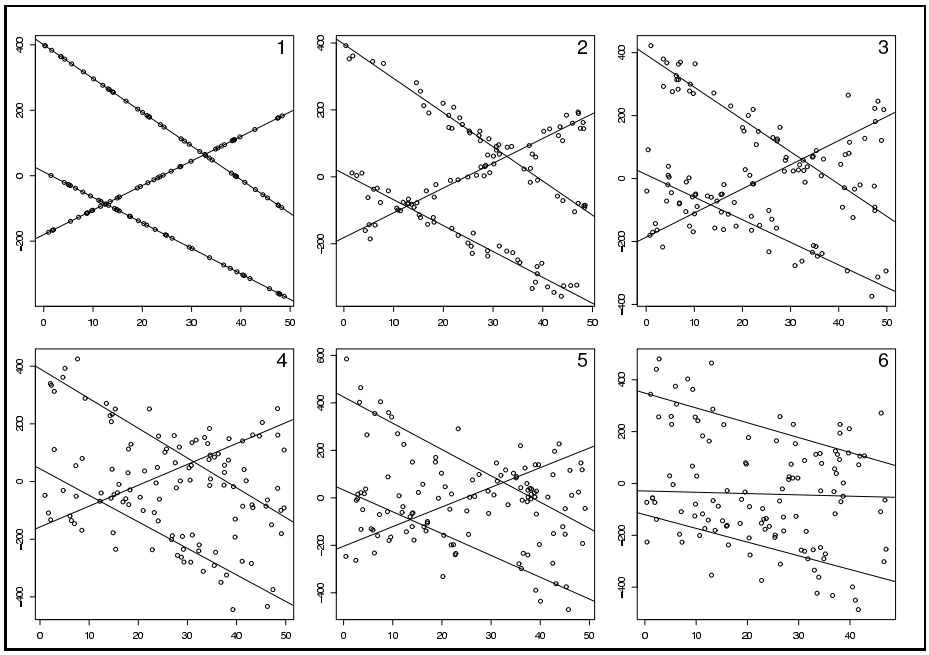

Abb. 3.5: Generierte Daten mit in 6 Etappen ansteigender Varianz von sehr gering bis sehr stark. Zusätzlich sind die durch die MRC-Analyse bei vorgegebener Clusteranzahl $K = 3$ gefundenen Regressionsgeraden eingetragen.

Für die letzte Analyse an synthetischen Daten wurde ein Datensatz generiert, der drei Cluster aufweist. Die Standardabweichung des Rauschens (siehe 3.1.1) nimmt in 6 Etappen von sehr schwach bis sehr stark zu (Standardabweichung des Rauschens: ($sd_{i=1..6}$=0;20;40;60;80;120). Mit diesen 6 Datensätzen wurde die Multiregressionsclusterung bei einer vorgegebenen Clusteranzahl von drei durchgeführt. In Abbildung 3.5 ist zu erkennen, wie die drei Cluster trotz deutlicher Verstärkung des Rauschens noch erkannt werden. Erst in der letzten Darstellung stimmen die gefundenen Regressionsgeraden nicht mehr mit den ursprünglich definierten überein. Jedoch bereits im Abbildung 3.5(5) ist die Struktur des Datensatzes nicht mehr eindeutig mit dem bloßen Auge zu erkennen. Der Algorithmus ordnet jedoch immer noch die drei Cluster richtig zu.

In diesem Kapitel wurden die in 2.3.2 vorgestellten Gütemaße an synthetisch erzeugten Daten getestet. Sie sollen der Analytikerin bei der Wahl der geeigneten Clusteranzahl helfen. Wichtig zu bemerken ist, dass eine eindeutige Aussage mit einem der Gütemaße nicht zu treffen ist. Es können nur aus dem Verlauf mehrerer Gütemaße Schlussfolgerungen auf die bevorzugte Partition gemacht werden.
Häufig tritt auch der Fall ein, dass ein Hinzufügen eines Clusters keine qualitativ neue Gruppe bildet, da nur zwei Cluster parallel laufend ein altes Cluster aufteilen. Dies kann nur eine genaue Untersuchung der Parameter der Regressionscluster aufdecken.
Welches die *wirkliche* Struktur im Datensatz ist, kann selbst bei bekannten Parametern der generierten Daten nicht eindeutig gesagt werden. Eine ursprünglich zugrundegelegte Struktur mit vorgegebener Zahl von Clustern kann durch das aufaddierte Rauschen seine Struktur soweit verloren haben, dass eine andere Partition mit neuer Clusterzahl optimal ist.
Für die spätere Analyse ist auch die richtige Einstellung der Parameter der Analyse wichtig. Wie in Abschnitt 2.4 näher erläutert, gibt es Möglichkeiten, das *Simulated Annealing* zu verstärken. Dies erleichtert das Auffinden der geeigneten Partition auch bei verrauschteren Daten, verlängert jedoch auch die Rechenzeit.
Abschließend bleibt zu bemerken, dass die drei Gütemaße *EV/GV, F-Wert* und *mean(p-Value)* besonders aussagekräftig sind. Sie ermöglichen, zumindest bei nicht zu starkem Rauschen, das Auffinden der generierten Cluster (siehe Abschnitt 3.1.2 und 3.1.2). Die Aussagekraft der Zielfunktion ist allgemein nur gering. Das *AIC* erweist sich ebenfalls als sehr nützlich. Bei geringem Rauschen zeigt sich bei der vorher definierten Clusteranzahl ein Absatz und bei stärkerem Rauschen ein Abknicken im Verlaufes des *AIC*.

Ein weiteres Merkmal für eine geeignete Partition hat sich bei beiden Datensätzen mit Struktur gezeigt. Bei mehrmaligem Wiederholen der Analyse mit unterschiedlicher Anfangspartition sind die Gütemaße bei der ursprünglich gewählten Clusteranzahl am stabilsten gewesen. Insbesondere beim mittleren Signifikanzlevel in Abbildung 3.3(4) und Abbildung 3.4(4) ist diese Stabilität auffallend.
Im nächsten Abschnitt werden die an synthetischen Daten erprobten Methoden und verwendeten Gütemaße auf empirische Daten angewandt.

3.2 Anwendung auf empirische Daten

Zur Anwendung der Multiregressionscluster-Analyse wurden unterschiedliche empirische Datensätze ausgewählt. Bei der Auswahl wurde nach folgenden Kriterien vorgegangen:

Verfügbarkeit der Daten
Die Daten sollten weltweit mindestens auf Länderebene vorhanden sein. Um eine sinnvolle Analyse von Zusammenhängen zu betreiben ist es notwendig, eine nicht zu geringe Anzahl von Datenpunkten in die Untersuchung einfließen zu lassen. Waren die räumlichen oder zeitlichen Lücken in den Datensätzen zu groß, fielen die entsprechenden Größen als potentiell zu untersuchenden Variablen heraus.

inhaltliche Gesichtspunkte
Weiterhin mussten die Daten eine sinnvolle Interpretation eines potentiellen Zusammenhangs zulassen. Viele Daten bilden in der Gegenüberstellung lediglich Hypersphären aus, bei denen eine *Multiregressionsclusterung* zwar eine von der Güte her betrachtet bessere Zuordnung der Datenpunkte zulässt als bei einer gewöhnlichen Multiregression, jedoch erscheinen die Cluster als nur schlecht voneinander getrennt und schwer interpretierbar.
Es mussten also Datenmatrizen zusammengestellt werden, in denen trennbare und sinnvolle Zusammenhänge vermutet werden und deren einzelne Variablen ausreichend viele Datenpunkte aufweisen.

3.2.1 Datenquellen

Als Datenquelle nutzen wir die *World Development Indicators* der Weltbank [WDI, 2001]. Ein Vorteil dieser Datenquelle ist, dass hier die Zeitreihen von mehr als 500 Indikatoren für eine feste Ländermenge zusammengestellt sind und diese sich auch gegenüber anderen Quellen als weitaus umfangreicher bezüglich der verfügbaren Zeiträume herausgestellt haben. Die CD-Rom wird jährlich von der Weltbank herausgegeben und enthält Daten für 208 Staaten von 1960 bis 2002. Es ist möglich, die Daten in verschiedenen Formaten zu exportieren, anschließend zu bearbeiten und für die Analyse in R [www.r-project.org] einzulesen.
Weiterhin werden Daten aus dem *World Development Report* genutzt, welcher im Netz frei zugänglich ist [WDR]. Diese Daten werden seit 1990 herausgegeben und sollen es ermöglichen, einen Einblick in die Entwicklung von Armut und Wohlstand über die bloße Angabe des Pro-Kopf-Einkommens hinaus zu erlangen.

Bei der Auswahl weiterer exemplarischer Datensätze wurden ernährungsspezifische Daten ausgewählt, da diese als ein wichtiger Indikator für Armut und Wohlstand angesehen werden können. Die Ertragsdaten ausgewählter Nutzpflanzen wurden aus der im Internet frei zugänglichen statistischen Datenbasis der *Food and Agricultural Organization* [FAOSTAT] entnommen.
Es liegt nahe, dass die Erträge der Nutzpflanzen in starkem Maße von klimatischen Gegebenheiten beeinflusst werden. Daher wurden in einem weiteren Schritt Klima- bzw. Wetterdaten in die Analyse einbezogen. Bei diesen handelt es sich um Daten vom *Climate Research Unit*, welche am PIK auf ein räumliches Raster interpoliert wurden. Sie liegen nun mit einer Auflösung von $0.5° \times 0.5°$ für die monatlichen Mittelwerte von Temperatur und den monatlichen Niederschlägen vor. Zur Analyse können Datensätze von 1951 bis 2000 herangezogen werden.
Eine Übersicht über die in den Beispielen verwendeten Daten gibt Tabelle 3.1. Bei diesen kann es sich lediglich um eine beispielhafte Auswahl an Daten zur Anwendungen des MRC-Verfahrens handeln. Eine systematische und automatisierte Durchforstung der genannten Datenquellen würde den Rahmen dieser Diplomarbeit sprengen.

Datum	Einheit	Auflösung	Zeitraum	Quelle	Abk.
HIV-Rate	Anzahl	pro Land	2003	CIA	
Niederschlag, monthly	mm	0.5×0.5	1961-2000	CRU	P
Temperatur, monthly mean	mm	0.5×0.5	1961-2000	CRU	T
Kindersterblichkeit	pro 1000	pro Land	2002	WDR	IMR
Unterernährung bei Kindern	%	pro Land	1995-2002	WDR	ChMal
Human Development Index	Index	pro Land	2002	WDR	HDI
GDP pro Kopf	Dollar	pro Land	2002	WDR	
Düngemittel (Fertilizer)	hg/ha	pro Land	1961-1999	WDI	F
Traktor (Mechanisierung)	#/ha	pro Land	1961-1999	WDI	M
Anteil Anbaufläche	%	pro Land	1961-1999	WDI	L
Sorghumertrag	hg/ha	pro Land	1961-2000	FAOSTAT	
Weizenertrag	hg/ha	pro Land	1961-2000	FAOSTAT	
Getreideertrag	hg/ha	pro Land	1961-2000	FAOSTAT	Y

Tab. 3.1: Verwendete Daten mit Aufbau und Quelle. Genauere Angaben zur Quelle befinden sich im Literaturverzeichnis.

3.2.2 Kindersterblichkeit versus Unterernährung

Als erstes soll unsere Analysemethode auf einen zweidimensionalen Datensatz, bestehend aus Kindersterblichkeit und Unterernährung bei Kindern, angewandt werden. Die Kindersterblichkeit (Infant Mortality Rate - IMR) ist als Anzahl verstorbener Kinder zwischen Geburt und 5. Lebensjahr pro 1000 Lebendgeborenen im Jahr 2002 angegeben. Die Unterernährung bei Kindern (Children under weight for age - ChMal) ist in Prozent der Kinder bis 5 Jahren mit Untergewicht im Zeitraum 1995 bis 2002, angegeben. Beide Datensätze liegen für 134 Staaten in Südamerika, Afrika und Asien vor.
Zuerst wurden beide Datensätze mittels eines Histogrammes auf auffällige Werte untersucht. Wie den beiden Verteilungen in Abbildung 3.6 zu entnehmen ist, gibt es jedoch keine Ausreißer unter den Datenpunkten.

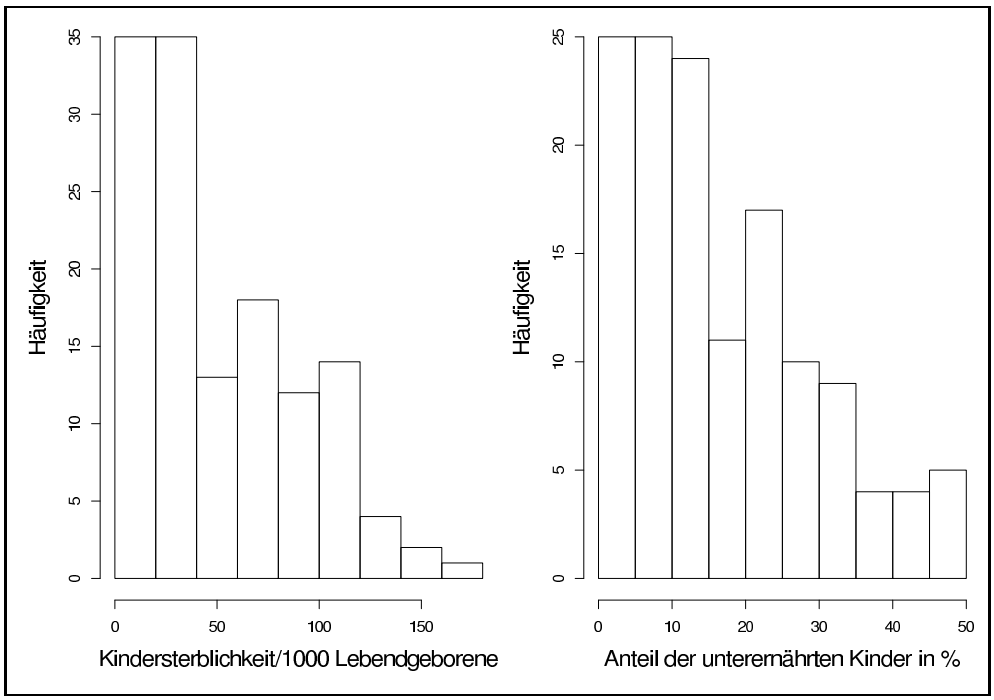

Abb. 3.6: Histogramme der Datensätze Child Malnutrition sowie Infant Mortality Rate.

Der Vorteil eines zweidimensionalen Datensatzes, wie in Abbildung 3.7 dargestellt, ist die Möglichkeit einer direkten optischen Inspektion. Wie zu erwarten, besteht insgesamt eine positive Korrelation zwischen der mangelhaften Ernährung bei Kindern und ihrer Sterblichkeit.

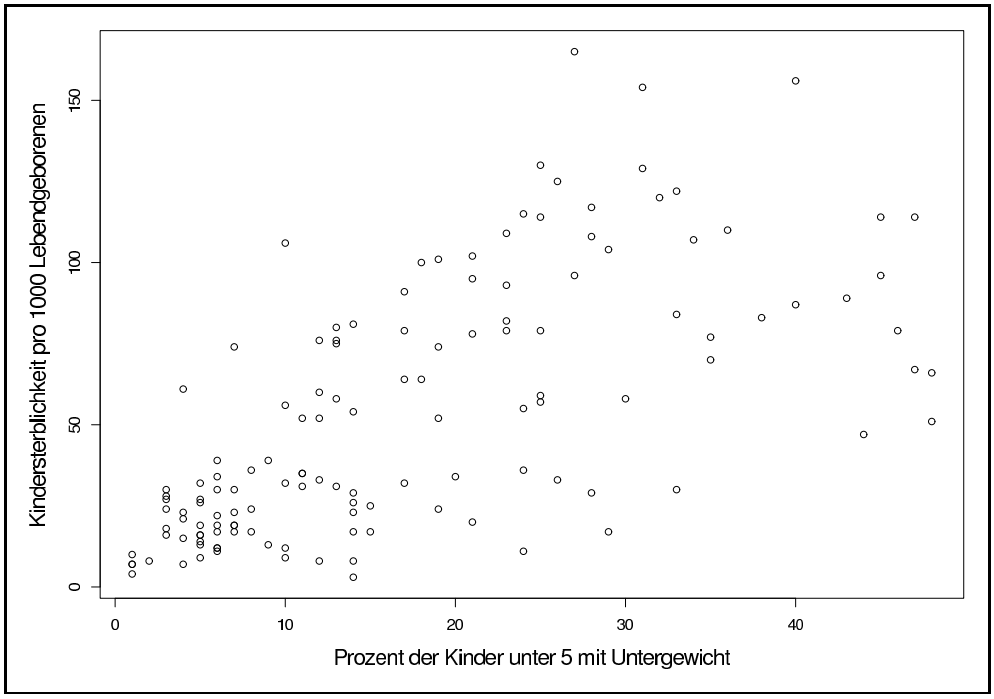

Abb. 3.7: Zusammenhang zwischen Unterernährung und Sterblichkeit bei Kindern.

Bei genauerem Hinsehen scheinen sich jedoch zwei Gruppen von Ländern abzuzeichnen, die durch unterschiedliche lineare Zusammenhänge der beiden untersuchten Variablen gekennzeichnet sind. Im nächsten Abschnitt folgt mittels der Gütekriterien eine Analyse der geeigneten Clusteranzahl.

Optimale Clusteranzahl

In Abbildung 3.8 sind verschiedene Gütemaße für die Partitionen mit der Clusteranzahl von 1 bis 8 dargestellt. Für diese und die folgenden Analysen wurden die Gütemaße ausgewählt, welche bereits in Kapitel 3.1.2 auf ihre Funktion an synthetischen Daten getestet wurden. Zusätzlich kommt das Gütemaß *log(max(Ftest))* zur Verwendung, dessen Zweckmäßigkeit sich erst im späteren Verlauf der Arbeit bewiesen hat. Seine Funktionsweise wird in Abschnitt 2.3.2 erläutert.
Die Analyse wurde zehnmal mit veränderter Anfangspartition wiederholt und diese mehrfachen Ergebnisse in jedem Gütemaß den Partitionen zugeordnet. Der Anteil der erklärten Varianz von 0.9, wird erst bei einer Clusteranzahl von 5 überschritten. Auch ist beim F-Wert ein markanter Sprung von P_4 (Partition mit $K = 4$) auf P_5 zu bemerken.

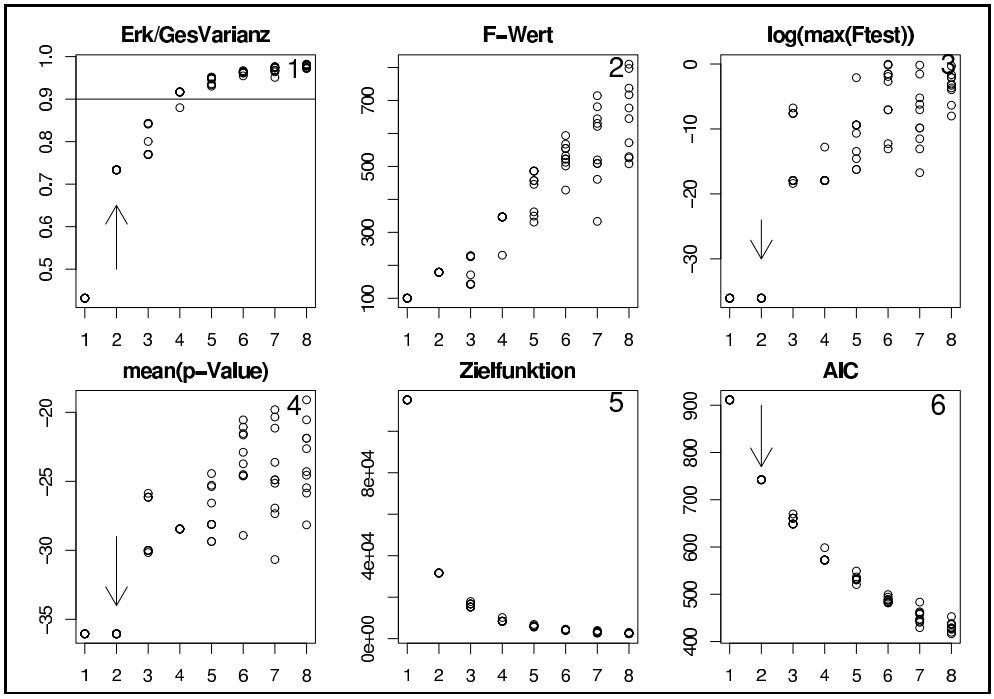

Abb. 3.8: Gütemaße zur Analyse vom Datensatz Unterernährung und Sterblichkeit bei Kindern.

Trotzdem erscheint diese Menge von Objektgruppen besonders in der graphischen Darstellung als zu groß. Es werden dabei hauptsächlich Cluster nochmals geteilt, wodurch keine qualitativ neue Gruppe entsteht, die Güte jedoch verbessert wird. Dem Prinzip der Sparsamkeit dienend würden wir die Clusteranzahl $K = 2$ bevorzugen. Sie lässt sich auch mit leichten Abstrichen durch die Gütemaße belegen. Der Anteil der erklärten Varianz macht seinen größten Sprung von P_1 auf P_2. Der Logarithmus vom Signifikanzwert des schlechtesten Clusters der Partition ist bei P_1 und P_2 am kleinsten, also am besten, und der mittlere p-Wert der Partition ist ebenfalls bei diesen beiden Partitionen optimal.
Auch bei der Zielfunktion und beim AIC ist der Sprung von der ersten zur zweiten Partition am auffälligsten. Aus diesen Gründen würden wir neben der ersten optischen Analyse auch bei der Analyse der Gütemaße auf die Partition mit der Clusteranzahl *zwei* schließen.

Statistische Prüfung der Robustheit des Ergebnisses

Nun soll das in Abschnitt 2.4.2 vorgestellte Verfahren zur statistischen Analyse des Ergebnisses zur Anwendung kommen. Dafür werden die Eigenschaften der beiden Cluster näher betrachtet.

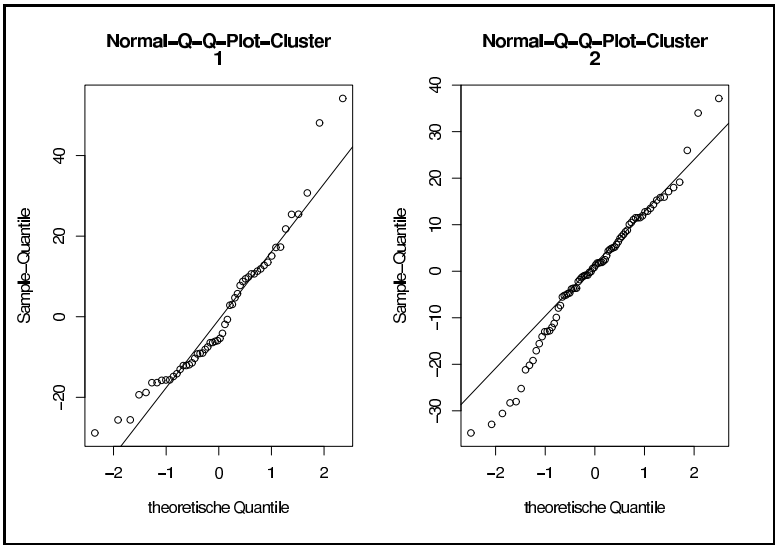

Abb. 3.9: Q-Q-Plot der Residuen aus C1 (links) und C2 (rechts).

Abbildung 3.9 zeigt den *Q-Q-Plot* (siehe Abschnitt 2.2) der Residuen der beiden Cluster. Die Abweichungen der beiden Datensätze von der 45°-Geraden sind relativ gering und wir nehmen daher die Residuen als normalverteilt an. Wie beschrieben werden aus den Mittelwerten und der Varianz der Residuen neue, aber statistisch äquivalente Daten gewonnen. Der so generierte Datensatz wird wieder per Multiregressionsclusterung untersucht. Es werden zwei Cluster und die entsprechenden Anstiege bestimmt. Diese Datengenerierung und -untersuchung wurde 1500-mal wiederholt. Die Ergebnisse sind in Abbildung 3.10 dargestellt.
Die Abbildungen 3.10(1,4) stellen die Verteilung der Anstiege der Regressionsgeraden dar, welche per einfacher Regression aus den neu generierten Punkten bestimmt werden können. Die Verteilungen beider Cluster haben die Struktur einer Normalverteilung. Es werden Mittelwerte und Standardabweichungen bestimmt und diese Normalverteilungen in Abbildung 3.10(3,6) eingetragen. Die Abbildungen 3.10(2,5) stellen die Verteilungen der Anstiege dar, welche durch die Multiregressionsclusterung aus den generierten Daten bestimmt wurden. Man erkennt deutlich, dass es zwei Zustände gibt, welche mit unterschiedlicher Häufigkeit eintreffen. Bei Cluster 1 (C1) fällt in 83,6% der Fälle der Anstieg unter den ersten Peak, welcher einen Mittelwert von 3,12 hat. Der Anstieg des zweiten Clusters (C2) fällt in 83,9% der Wiederholungen in den Peak mit dem mittleren Anstieg 1,68. In einem statistisch äquivalenten Datensatz findet der Algorithmus diese Zuordnung in zwei Cluster in über vier Fünftel der Fälle.

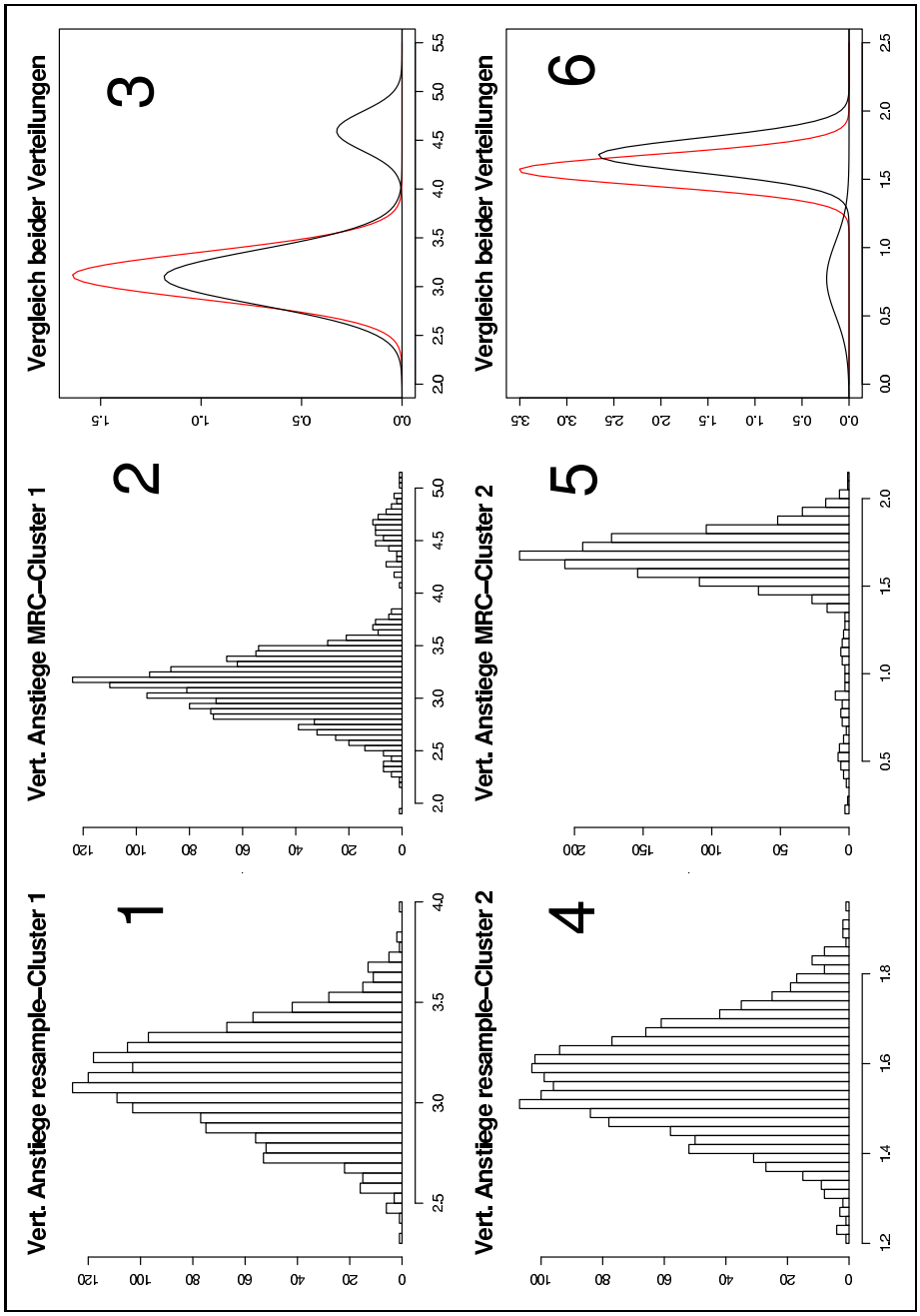

Abb. 3.10: Statistische Analyse nach 1500-facher Sampleauswahl. Zum Vergleich sind die Verteilungen der resample-Anstiege (rot) und die Verteilungen der MRC-Anstiege (schwarz) noch einmal in Abbildung (3) und (6) dargestellt.

Die zwei Peaks in beiden Clustern wurden jeweils getrennt auf ihre Eigenschaften untersucht. Dafür wurde im 1. Cluster der Anstieg 4.0 und im 2. Cluster der Anstieg 1.25 als Trennwert genommen. Die aus den Kennwerten der Verteilungen gewonnenen Gaußverteilungen wurden mit dem Anteil der unter dem jeweiligen Peak befindlichen Ereignisse normiert. Das heißt, dass beispielsweise in Cluster 1 der Hauptcluster mit 0.836 multipliziert wurde, da er 836 von 1000 Ereignissen vereint. In Abbildung 3.10(2,5) sind nun die idealisierten Gaußverteilungen sowohl der generierten als auch der mit MRC ermittelten Anstiege zum direkten Vergleich eingetragen. Bis auf den zweiten selteneren stabilen Zustand überlagern sich die Anstiege der resample und MRC-Cluster von Cluster 1 ohne Abweichung. Im zweiten Cluster gibt es eine leichte Verschiebung des MRC-Anstieges gegenüber dem resample Anstieg. Diese lässt sich wahrscheinlich auf Fehler bei der Trennung der beiden Peaks zurückführen.
Zuletzt werden die Vertrauensgrenzen durch Abzählen aus den Verteilungen entnommen. Wir interessieren uns für das $\alpha = 95\%$ Intervall. Dafür müssen, da 1500 Wiederholungen vorgenommen wurden, 75 Ereignisse abgezogen werden. Das entspricht 37.5 Ereignissen am rechten und linken Rand der jeweiligen Verteilung. Da nach den ganzzahligen Ereignissen erst die nächstfolgenden Ereignisse als Intervallränder ausgewählt werden, soll hier der nicht ganzzahlige Teil abgerundet werden. Die Ergebnisse sehen wie folgt aus:

$$\bar{\hat{b}}_{C1} = 3.12 \qquad 2.52 \leq \beta_{C1} \geq 3.61 \tag{3.2}$$

$$\bar{\hat{b}}_{C2} = 1.68 \qquad 1.44 \leq \beta_{C2} \geq 1.93. \tag{3.3}$$

Die angegebenen Vertrauensintervalle geben an, zwischen welchen Werten sich der Parameter (β_{Ci}) der Grundgesamtheit mit 95%er Wahrscheinlichkeit befindet.
Mit der Bootstrapping-Methode konnte gezeigt werden, dass die Einteilung des Datensatzes in zwei Cluster gerechtfertigt ist. Bis auf etwa ein Fünftel der Wiederholungen wurde diese Partition wiedergefunden. Diese abweichenden Zustände werden durch Objekte im unteren Bereich des Datensatzes verursacht, in dem die Regressionscluster sehr nahe beeinander liegen und die Variationen beim Bootstrapping dazu führen können, dass Länder unterschiedlich zugeordnet werden. Diese Zustände beschreiben keine qualitativ anderen Zusammenhänge, jedoch ist der Anstieg in Cluster 2 weitaus flacher und in Cluster 1 steiler.

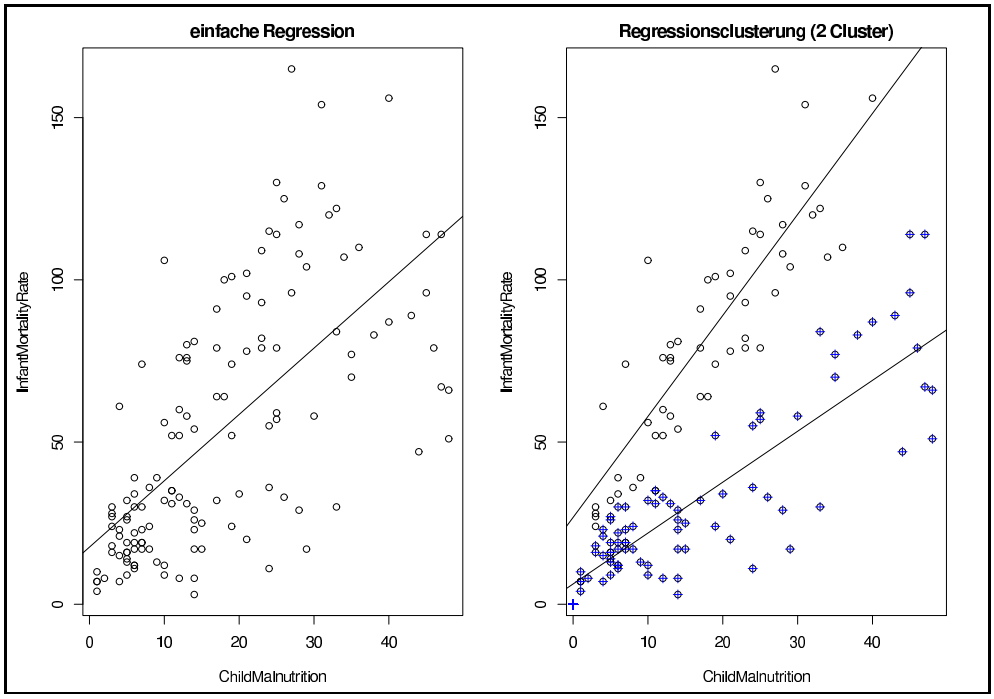

Abb. 3.11: Vergleich der Datenanalyse mittels einfacher Regression und Multiregressionsclusterung. In der rechten Grafik sind die Objekte der einzelnen Cluster mit unterschiedlichen Symbolen markiert (C1-gekreuzte Kreise, C2-Kreise).

Inhaltliche Analyse

Ist eine inhaltlich sinnvolle Interpretation der gefundenen Partition möglich? Um dies zu untersuchen ist es notwendig, sich die Ergebnisse der Regressionsclusterung genauer anzusehen. Welche Länderpunkte befinden sich in welchem der beiden Cluster? Lässt sich begründen, warum eine Ländergruppe einen stärkeren Anstieg in der Sterblichkeit aufweist? Welche Eigenschaften haben die Länder einer Gruppe noch gemein?
Abbildung 3.11 soll verdeutlichen, wie sich eine einfache Regression von einer Regressionsclusterung an diesen Daten unterscheidet. Es ist gut zu erkennen, dass die zwei Regressionsgeraden die Zusammenhänge in den Daten besser wiedergeben.
Die Anstiege der Regressionsgeraden in beiden Analysen unterscheiden sich deutlich. Bei der einfachen Regression lautet der Zusammenhang zwischen beiden Größen: $\text{IMR} = 17.62 + 2.04 \cdot \text{ChMal}$
Bei der Regressionsclusterung lauten die Zusammenhänge in beiden Clustern: $\text{IMR} = 6.25 + 1.57 \cdot \text{ChMal}$ sowie $\text{IMR} = 26.77 + 3.11 \cdot \text{ChMal}$.

In einem Cluster nimmt die Kindersterblichkeit in stärkerem Maße mit schlechter werdender Ernährung zu.
In Abbildung 3.12 sind die Länder der beiden Cluster farblich unterschieden auf der Weltkarte dargestellt. Bei den grün gekennzeichneten Ländern handelt es sich um das Cluster mit dem steileren Anstieg (C2) der Regressionsgeraden. Hier wirkt sich Unterernährung noch stärker auf die Kindersterblichkeit aus als in Ländern mit roter Markierung (C1). Die grünen Flecken auf der Weltkarte füllen den afrikanischen Kontinent fast vollständig aus. Einige Länder in Asien sowie in Südamerika sind ebenfalls grün. Die grau hinterlegten Flächen in der Darstellung stehen für fehlende Daten. Die Zugehörigkeit fast des gesamten afrikanischen Kontinents zu C1 lässt vermuten, dass dieses Cluster viele weniger entwickelte Staaten beinhaltet. Dies lässt sich bestätigen, wenn man den Mittelwert der HDI- (Human Development Index) und GDP-pro-Kopf-Werte bildet. Dabei lassen sich deutliche Unterschiede zwischen den beiden Gruppen ausmachen. Im ersten Cluster beträgt der Mittelwert vom HDI = 0.53 und vom GDP = 3104 $/Kopf. Diese unterscheiden sich deutlich von den Mittelwerten des zweiten Clusters: HDI = 0.71 und GDP = 6991 $/Kopf.

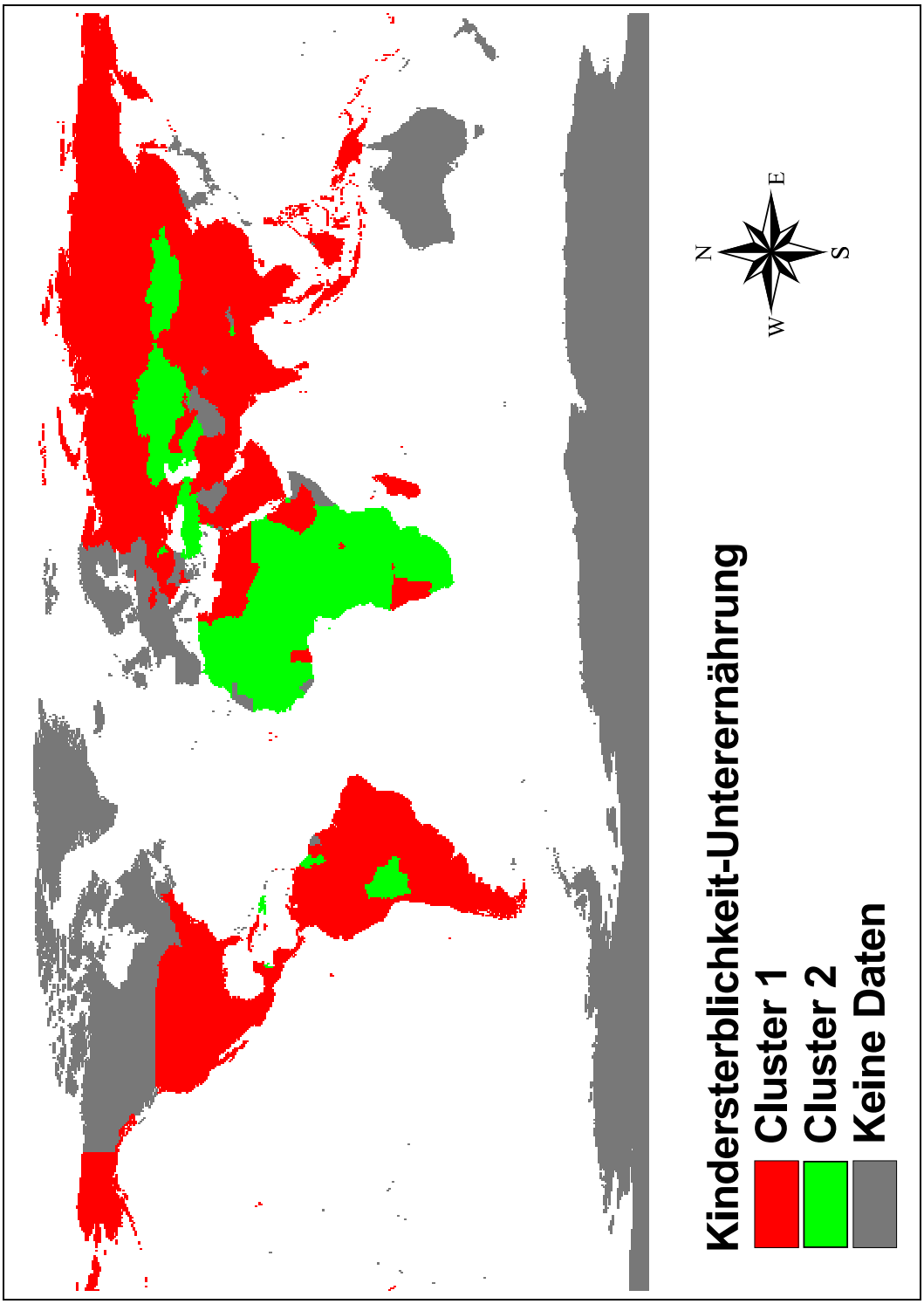

Abb. 3.12: Weltkarte mit farblicher Unterscheidung beider Cluster.

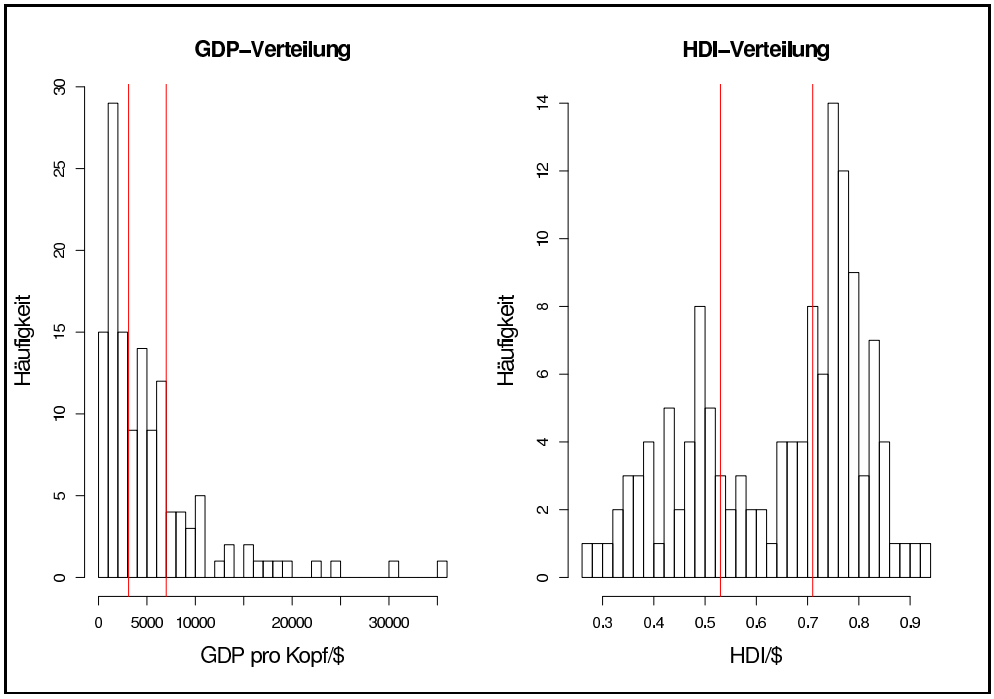

Abb. 3.13: Verteilung von Bruttoinlandsprodukt (GDP) und Human Development Index (HDI) der an der Analyse beteiligten Staaten. Bei den eingetragenen waagerechten roten Linien handelt es sich um die Mittelwerte der beiden Cluster. (HDI(C1) = 0.53, HDI(C2) = 0.71, GDP/Kopf(C1) = 3104 $, GDP/Kopf(C2) = 6991 $)

In Abbildung 3.13 sind die Verteilungen der Größen GDP/Kopf und HDI aus den Ländern der Analyse dargestellt. Insbesondere beim HDI kann man erkennen, wie sehr sich beide Gruppen in ihrer Entwicklung unterscheiden. Zur Bestätigung der Signifikanz des Unterschiedes beider Gruppenmittelwerte wurde ein t-Test (siehe Abschnitt 2.2) durchgeführt.
Beim HDI beträgt die Irrtumswahrscheinlichkeit $p_{HDI} \simeq 5 \times 10^{-11}$ und beim GDP/Kopf $p_{GDP} \simeq 4 \times 10^{-5}$. Die beiden Mittelwerte in beiden Clustern weichen also statistisch signifikant voneinander ab.
Was sind die Gründe für die unterschiedlichen Zusammenhänge zwischen Unterernährung und Kindersterblichkeit in den Gruppen unterschiedlichen Wohlstandes ?
Es ist bereits belegt worden, dass sich die beiden Ländergruppen deutlich in ihrer ökonomischen Entwicklung unterscheiden. Das nachweislich geringere durchschnittliche Einkommen der Menschen in Cluster 2 korreliert mit anderen Indikatorgrößen der Entwicklung [Ravallion, 1994].

Dazu zählen beispielsweise die Verfügbarkeit von sauberem Trinkwasser und Sanitäranlagen, der Zugang zu Bildungsmaßnahmen und die Anzahl von Ärzten pro Einwohner. An diesen Größen wird deutlich, dass sich eine nicht ausreichende Ernährung eines Kindes in einem ärmeren Land viel verheerender auswirken kann. Ein unterernährtes Kind hat weitaus schlechtere Chancen auf einen Arzt, ihm fehlen zusätzlich sauberes Trinkwasser und den Eltern die Möglichkeiten, auf Folgen der Mangelernährung adäquat zu reagieren. Der größere Wert des Intercepts in Cluster 2 unterstützt dies Interpretation. Selbst bei ausreichender Ernährung können die oben genannten Faktoren, wie fehlende ärztliche Versorgung zu einem frühzeitigen Tod des Kindes führen.

3.2.3 Landwirtschaftliche Erträge versus Wetter

Als weiteres Beispiel werden Ertragsdaten von vielerorts verwendeten landwirtschaftlichen Nutzpflanzen Wetterdaten gegenübergestellt. Die Daten für die Erträge wurden der statistischen Datenbasis der *Food and Agricultural Organication of the United Nations [FAOSTAT]* entnommen. Für ca. 200 Länder liegen dort Ertragsdaten von 1961 bis 2005 für verschiedenste Anbaupflanzen vor. Es wurden die Ertragsdaten der Nutzpflanzen Sorghum und Weizen zwischen 1961 und 2000 ausgewählt, da hier ausreichend viele Daten vorhanden waren und nach einer optischen Inspektion der Struktur der Datensätze eine sinnvolle Anwendung der vorgestellten Analysemethode möglich erschien.
Wir erwarten, dass die Erträge von Sorghum und Weizen in unterschiedlichen Regionen der Welt auf unterschiedliche Weise vom Einflussfaktor Niederschlag abhängen. Aufgrund von ähnlichen Gegebenheiten in bestimmten Ländergruppen sollten sich abgrenzbare Gruppen von Staaten mit ähnlichen Zusammenhängen ausbilden. Beispielsweise könnte in einigen Regionen der Ertrag mit zunehmendem Niederschlag ebenfalls zunehmen, da dieses zusätzliche Wasserangebot die Wachstumsbedingungen der Pflanze verbessert. In anderen Regionen zeigt sich möglicherweise der umgekehrte Zusammenhang, da zusätzlicher Niederschlag zu übermäßiger Bodenfeuchte führt und dadurch negativ auf das Wachstum wirkt. Weiterhin wird eine erhöhte Niederschlagsmenge mit stärkerer Bewölkung und daher geringerer Sonnenscheindauer einhergehen. Dies kann sich je nach Pflanze ebenfalls kontraproduktiv auf den Ertrag auswirken.

Aufbereitung der Wetterdaten

Die Temperatur- und Niederschlagsdaten sind, wie oben beschrieben, auf einem $0.5° \times 0.5°$-Raster vorhanden. Weiterhin liegen für jedes Jahr und jeden der 67420 Punkte zwölf Monatsmittel (Temperatur) bzw. Monatssummen (Niederschlag) vor.
Da das Pflanzenwachstum hauptsächlich von den Wetterbedingungen während der Vegetationsperiode bestimmt wird, ist lediglich eine Betrachtung der Vegetationsphase von Interesse. Diese wurde aus Cassel-Gintz et.al. (1997) entnommen, und aus jedem Jahr in jedem Rasterelement nur die entsprechende Vegetationsphase extrahiert. Die Niederschlagsmengen wurden auf Angaben pro Monat gemittelt.
Es ist weiterhin notwendig, die Daten räumlich zu mitteln, da die Ertragsdaten nur auf Länderebene vorliegen. Dabei kam eine Zuordnungsmatrix zur Anwendung, welche den 67420 Punkten eine Länderzugehörigkeit zuordnet.

Alle Datenpunkte aus einem Land wurden zu einem dem jeweiligen Land und Jahr zugeordneten Wert gemittelt. Niederschlags- und Temperaturwerte liegen für 187 Länder vor.

Sorghumertrag versus Niederschlag

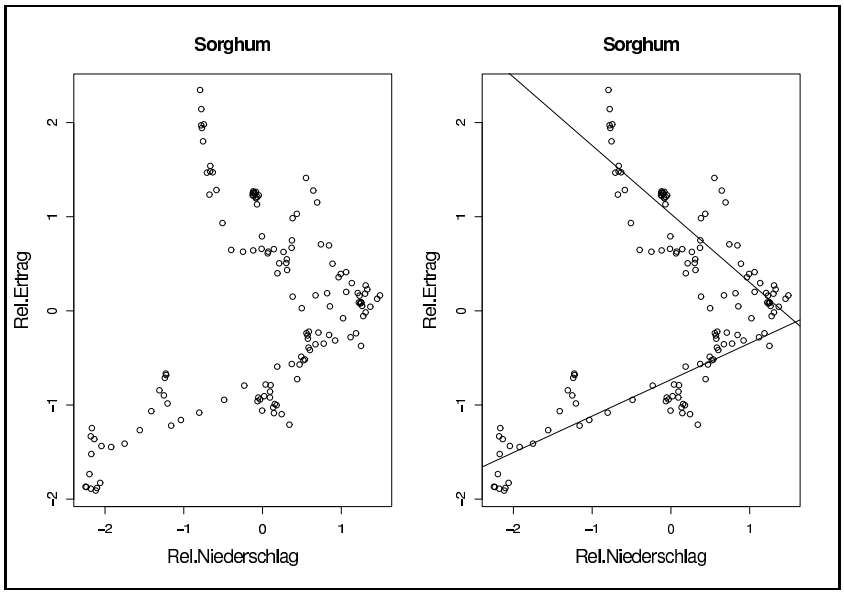

Abb. 3.14: Sorghumdatensatz und der Verlauf der Regressionsgeraden bei K=2.

Sorghum oder auch Hirse ist eines der wichtigsten Getreide in Westafrika. Weltweit ist die Anbaufläche die fünftgrößte unter den Getreidesorten (Produktion 2004: 49 Mio t [FAOSTAT]).
Die Ertragsdaten von Sorghum liegen für 63 Länder vor. Sie wurden mittels eines laufenden Dreijahresmittels geglättet und auf ihren Anfangswert normiert. Nachdem die Länder ausgewählt wurden, in denen der Einfluss des Niederschlages signifikant ist, blieben 140 Datenpunkte übrig. Dabei handelt es sich um die Länder Benin, Burkina Faso, Niger und Sudan mit jeweils 35 Zeitpunkten. Der Datensatz ist in Abbildung 3.14 dargestellt.

Optimale Clusteranzahl

Der Datensatz wurde einer MRC-Analyse unterzogen. Es wurde nach den Gruppengrößen von 1 bis 8 gesucht. Zusätzlich wurde die Analyse fünfmal mit neuer Anfangspartition wiederholt. Die zugehörigen Gütewerte sind in Abbildung 3.15 dargestellt.
Der Anteil der *erklärten Varianz* macht einen starken Sprung von der Partition mit einem zu zwei Clustern, überschreitet jedoch nicht die Marke von $R^2 = 0.9$. Ein ebenso markanter Sprung ist bei der *Zielfunktion* und dem *AIC* zwischen P_1 (Partition mit $K = 1$) und P_2 zu bemerken. Der *mittlere p-Wert* hat bei P_2 und P_3 ein Minimum zu verzeichnen. Die Schlussfolgerung ist, dass die Gütewerte die Partition P_2 präferieren.
Eine graphische Überprüfung (siehe Abbildung 3.14) bestätigt, dass es sich nicht um den trivialen Fall der gekreuzten Cluster (wie in Abschnitt 3.1.2 beschrieben) in einer Punktwolke handelt. Die Regressionsgeraden lauten:

$$Y^1 = -0.73 + 0.39 \cdot P \tag{3.4}$$

$$Y^2 = 1.03 - 0.72 \cdot P. \tag{3.5}$$

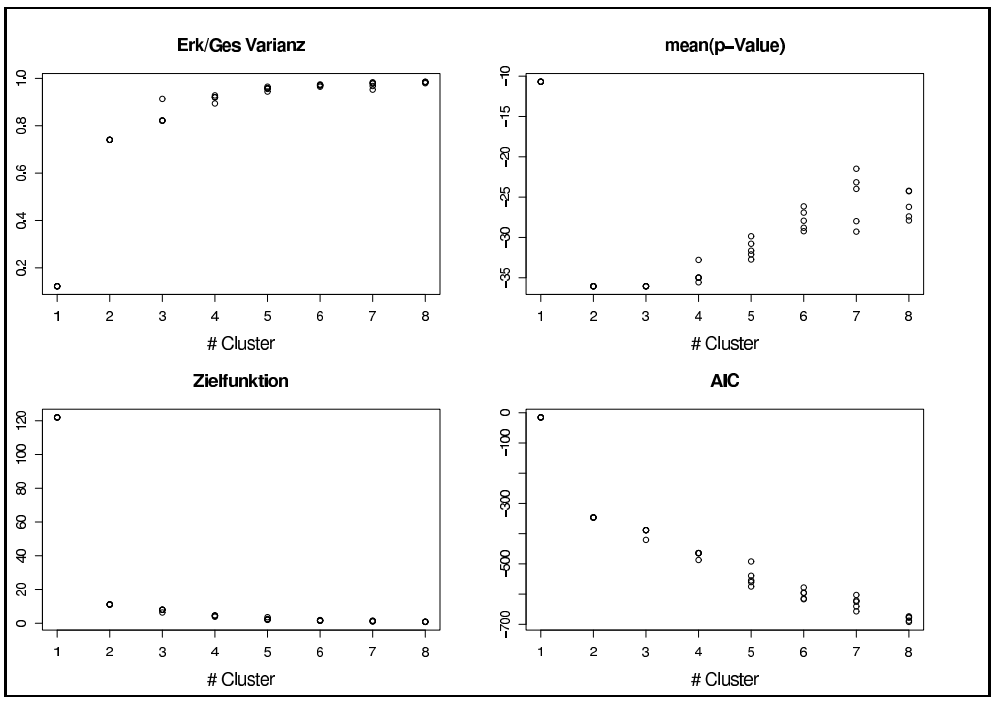

Abb. 3.15: Darstellung der Gütewerte für die Partitionen von K=1..8.

Statistische Prüfung der Robustheit des Ergebnisses

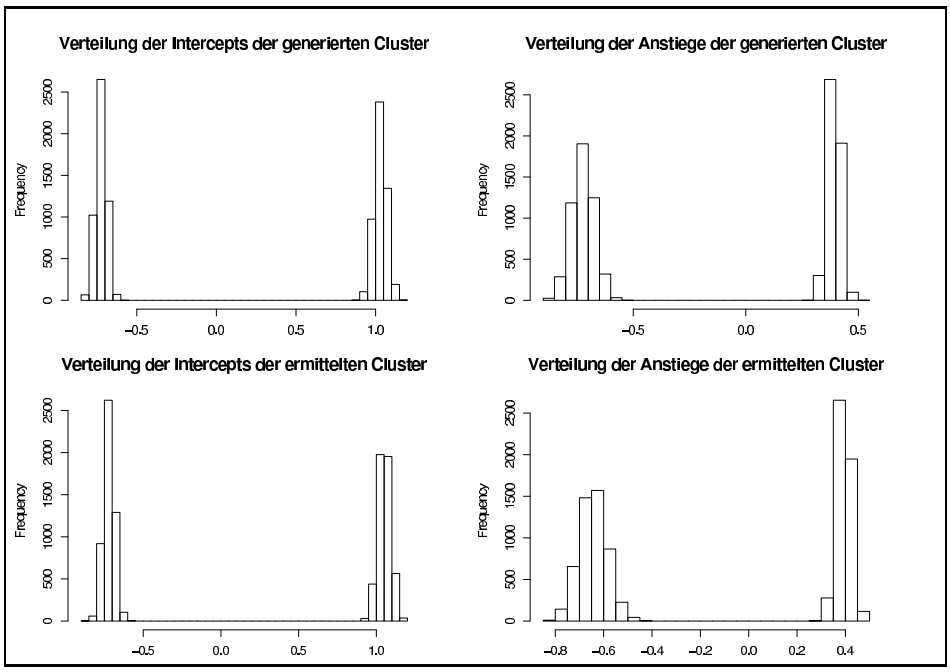

Abb. 3.16: Verteilungen der resample- und MRC-Parameter nach 5000 Wiederholungen.

Nun wird die gefundene Partition wie in 3.2.2 einer Stabilitätsanalyse unterzogen. Die Normalverteilung der Residuen wurde mittels eines Q-Q-Plots überprüft. Die Abweichungen von der 45°-Geraden waren minimal.
Das Ergebnis ist in Abbildung 3.16 zusammengefasst. Die Darstellung soll lediglich übersichtsweise darstellen, wie die Verteilungen der Intercepts und Anstiege der Regressionsgeraden der beiden Cluster der Partition nach 5000 Wiederholungen aussehen. Beim Vergleich der Lage der Verteilungen in den oberen Darstellungen mit denen in den unteren, ist zu erkennen, dass die aus den Orginalclustern generierten Objektgruppen gut durch die MRC-Analyse wiedergefunden werden.
Der Mittelwert der Anstiege und deren Vertrauensintervall beträgt:

$$\bar{\hat{b}}_{C1} = -0.64 \qquad -0.76 \leq \beta_{C1} \geq -0.53 \tag{3.6}$$

$$\bar{\hat{b}}_{C2} = 0.39 \qquad 0.34 \leq \beta_{C2} \geq 0.45. \tag{3.7}$$

Im Mittel nehmen die Anstiege der beiden Cluster den Wert $\bar{b}_{Ck}$ an und der Parameter der Grundgesamtheit (β_{Ck}) befindet sich mit 95%ger Wahrscheinlichkeit im angegebenen Intervall.

Inhaltliche Analyse

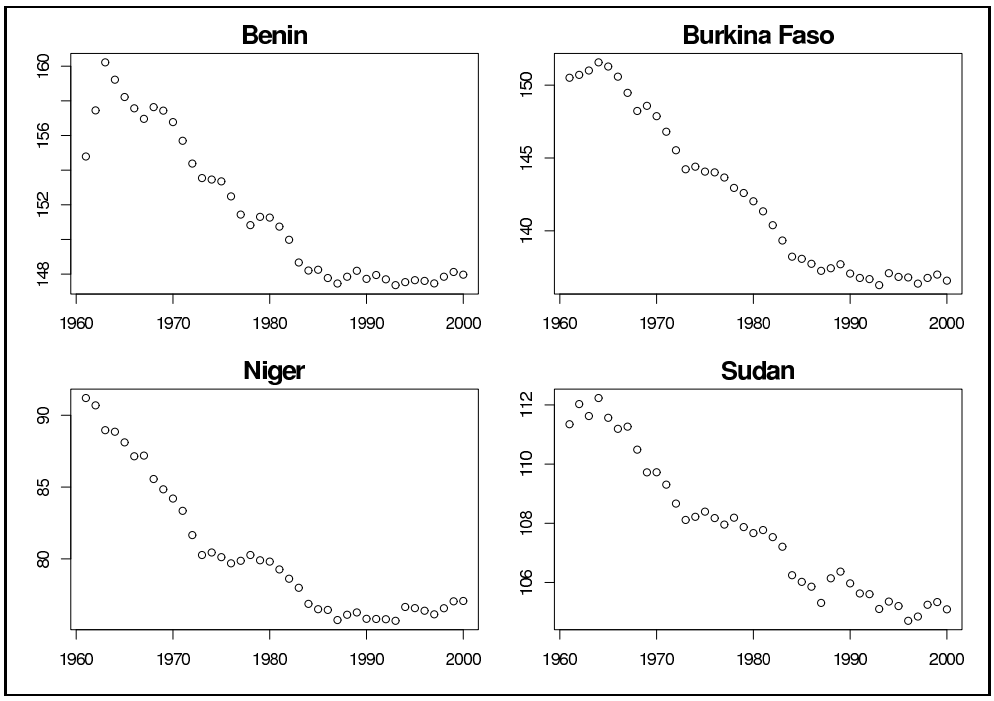

Abb. 3.17: Entwicklung der Niederschlagsmengen in der Vegetationsperiode von 1961-2000.

Die Datenpunkte der vier Länder teilen sich wie folgt auf: Benin und Burkina Faso gehören zum Cluster 1. Hier nehmen die Erträge für Sorghum mit zunehmender Niederschlagsmenge ab. Im Niger und Sudan, welche in Cluster 2 liegen, nehmen dagegen die Erträge mit dem Niederschlag zu. Worauf kann diese unterschiedliche Entwicklung in den beiden Ländergruppen zurückgeführt werden? Die Niederschlagsmengen nehmen in allen vier Staatsgebieten ab (siehe Abb. 3.17). Dies kann also keine Teilung der Datenpunkte verursacht haben. Die Trennung liegt also trivialer Weise nur in einer unterschiedlichen Entwicklung des Ertrages begründet. Wodurch lässt sich diese erklären?
Bei der Betrachtung verschiedener anderer Größen fällt folgendes auf:
Aufgrund der geographischen Lage befinden sich die Anbaugebiete im Niger und Sudan größtenteils in ariden bis semi-ariden, also trockeneren Gebieten als die Anbaugebiete in Benin und Burkina Faso, welche sich in semi-ariden bis semi-humiden Gebieten befinden. Eine Abnahme der Niederschläge ist daher in den feuchteren Gebieten Benins und Burkinas noch nicht so einschränkend für den Anbau von Sorghum wie in Niger und Sudan.

Allen vier Ländern gemein ist eine hohe Abhängigkeit der Bevölkerung von der Landwirtschaft, also eine große Zahl von Kleinbauern, die in Subsistenzwirtschaft leben. Speziell bei Sorghum handelt es sich um eine häufig von Kleinbauern zum Eigenbedarf angebaute Nutzpflanze. Allgemein werden südlich der Sahara 80% des Grobgetreides als direktes Nahrungsmittel[2] verwendet. Vorteilhaft beim kleinflächigen Anbau ist die relativ schnelle Anpassungsfähigkeit gegenüber äußeren Einflussfaktoren wie z.B. veränderten Niederschlagsmengen. Dem Kleinbauern sollte es unter idealen Bedingungen möglich sein, schnell auf eine andere den klimatischen Bedingungen besser angepasste Nutzpflanze umzusteigen. Je nach Pflanze kann diese dann weiterhin zum Eigenbedarf verwendet oder als *Cash Crop* auf Märkten verkauft und vom erwirtschafteten Geld Nahrungsmittel gekauft werden.
Es ist vorstellbar, dass diese *idealen* Bedingungen insbesondere im Sudan nicht gegeben sind. Das Land befindet sich seit Jahrzehnten in einem instabilen Zustand, es herrschen Bürgerkriege, Dürren und es existiert seit Jahrzehnten keine stabile Regierung. Gegenwärtig hat die UN einen Teil des Landes, die westlichen Darfur Regionen, als *unsafe for humanitarian operations*, als sogenannte *no-go areas* erklärt. Dies führt dazu, dass Arbeitskräfte durch Tod, Verletzung, Flucht oder Krankheit aufgrund fehlender medizinischer Versorgung bei der Feldarbeit fehlen. Auch ist der Verkauf von *Cash Crops* auf Märkten vielleicht nicht möglich, da die Wege zu unsicher oder die Infrastruktur zerstört ist. Unter diesen ungünstigen Bedingungen ist eine Kleinbäuerin kaum in der Lage, schnell auf sich verändernde äußere Einflussfaktoren zu reagieren.
Jedoch ist zu bemerken, dass auch die anderen Länder, wenn auch nicht so stark wie im Sudan, geprägt sind von verschiedensten politischen Spannungen.
Ein weiterer entscheidender Einflussfaktor für die Erträge von Nutzpflanzen ist die technische Entwicklung und die damit einhergehende Intensivierung der Anbaumethoden. Als Indikator für diese zunehmende Intensivierung kann die Verwendung von Düngemittel je Hektar herangezogen werden. Bei Betrachtung dieser Größe zeigt sich eine den Clustern ähnliche Aufteilung der vier Länder. In Burkina und Benin hat die Verwendung von Fertilizer in den letzten Jahrzehnten stetig zugenommen. Hingegen stagnieren im Sudan und im Niger die verwendeten Düngemengen je Hektar [FAOSTAT].
Welche der hier erwähnten Einflüsse zur Erklärung der Entwicklung der Erträge beitragen und in welchem Maße dies geschieht, könnte durch eine ausführlichere Multiregressionsclusterung mit mehr Indikatorgrößen ermittelt werden. Dieser Schritt wurde im auf Seite 69 folgenden Datensatz *Änderung des Flächenertrages von Getreide* gemacht.

[2]Dies ist ein gravierender Unterschied zu den durchschnittlich 60%, die weltweit nicht für den menschlichen Verzehr sondern als Tierfutter verwendet werden.

Weizenertrag versus Niederschlag

Weizen stellt in vielen Ländern ein Grundnahrungsmittel dar. Nach Mais ist er das am zweithäufigsten angebaute Getreide weltweit (Produktion 2004: Mais-703 Mio t; Weizen-614 Mio t [FAOSTAT]). Er ist an trockene und warme Sommer angepasst. Moderne Sorten erlauben einen Anbau in kühleren Klimazonen. In der FAO-Datenbank befinden sich für 77 Staaten Ertragsdaten für Weizen. Die Datensätze für Niederschlag und Ertrag werden auf ihren Anfangswert von 1961 normiert.
Die Erträge des Weizens unterliegen mit Sicherheit vielen Einflussfaktoren. Da diese Analyse lediglich beispielhaft die Beziehung vom Ertrag zum Niederschlag untersucht, werden jene Länder ausgewählt, in denen der Einfluss des Niederschlages signifikant ist (R^2>0.7). Bei den in Abb. 3.19 dargestellten Werten handelt sich um 240 Relativwerte von Niederschlag und Weizenertrag in den Ländern: Österreich, Indien, Mauretanien, Schweiz, Türkei und Großbritannien.

Optimale Clusteranzahl

Eine erste optische Inspektion des Datensatzes ließ vermuten, dass der Algorithmus die beiden zu vermutenden Cluster, aufgrund ihres starken Anstiegs nicht ermitteln würde. Daher wurde eine Drehung des Datensatzes um $\alpha = 90°$ durchgeführt (siehe Abschnitt 2.4.3).
Nun wird eine Multiregressionsclusteranalyse durchgeführt. Dabei wird nach allen Partitionen mit einer Clusteranzahl von 1 bis 8 gesucht. Eine Analyse der Gütemaße wird Indizien für die geeignete Partitionsgröße liefern.
Im Gegensatz zur Analyse der Sorghumdaten erbringt die Analyse der Weizendaten ein nicht so eindeutiges Ergebnis in Bezug auf die Clusteranzahl. Es gibt zwar auch einen Sprung bei der *erklärten Varianz* sowie bei der *Zielfunktion*, jedoch ist der Sprung beim *AIC* von P_1 (Partition mit K=1) zu P_2 nicht so eindeutig von späteren Absätzen zu trennen. Beim *mittleren p-Wert* ist das Minimum bei P_1 zu finden. Dies ist vermutlich darauf zurückzuführen, dass ein Cluster weitaus kleiner als das andere ist und eine Abweichung dieser Punkte von einem einzelnen Cluster nicht so schwer wiegt.
In Abbildung 3.19 sind die im gedrehten Datensatz (links) ermittelten Regressionsgeraden eingetragen. Im rechten Bild wurde der Datensatz und die ermittelten Geraden wieder zurücktransformiert.
Die Regressionsgeraden lauten (gedreht (Y^k_{ged}) und zurücktransformiert (Y^k)):

$$Y^1 = -1.63 - 0.58 \cdot P \qquad Y^1_{ged} = -2.82 + 1.73 \cdot P \tag{3.8}$$

$$Y^1 = 0.28 + 0.68 \cdot P \qquad Y^2_{ged} = -0.40 - 1.47 \cdot P. \tag{3.9}$$

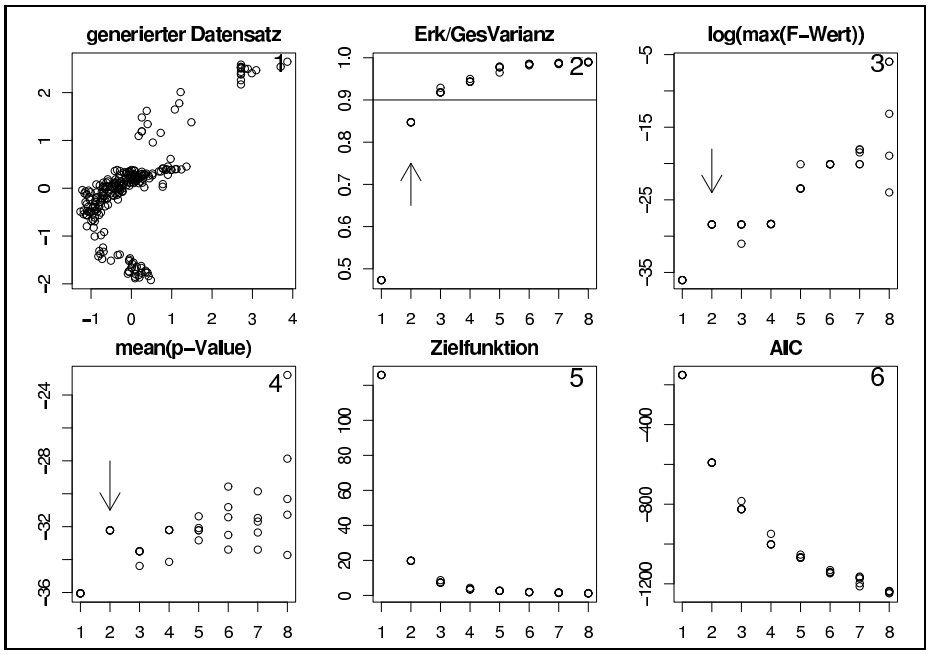

Abb. 3.18: Darstellung des gedrehten Datensatzes und der Gütewerte für die Partitionen von K=1..8.

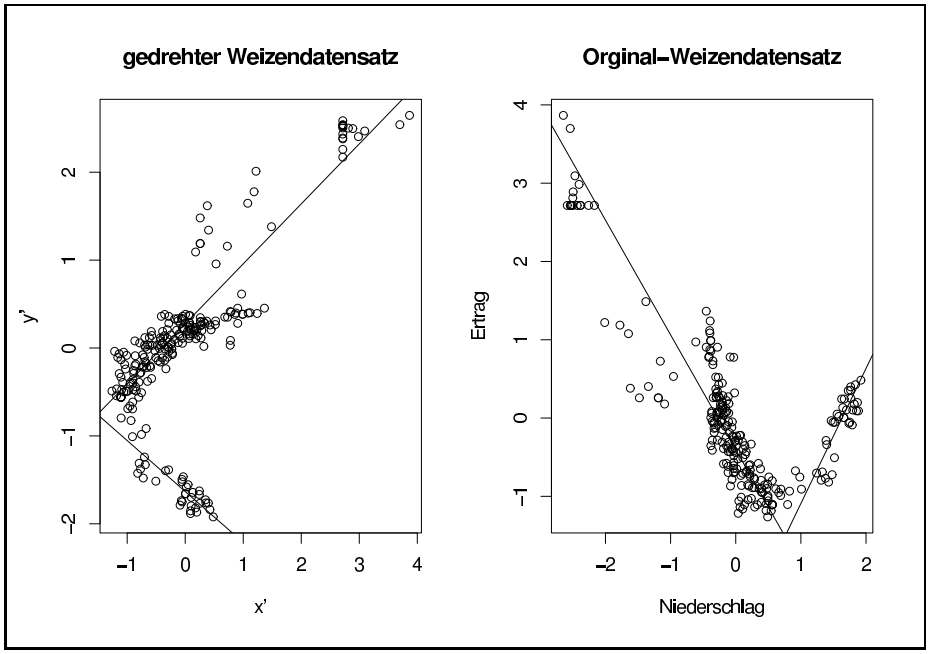

Abb. 3.19: Rel.Weizenertrag über Rel.Niederschlag (links: um 90° gedreht); Es sind zusätzlich die ermittelten Regressionsgeraden eingetragen.

Statistische Prüfung der Robustheit des Ergebnisses

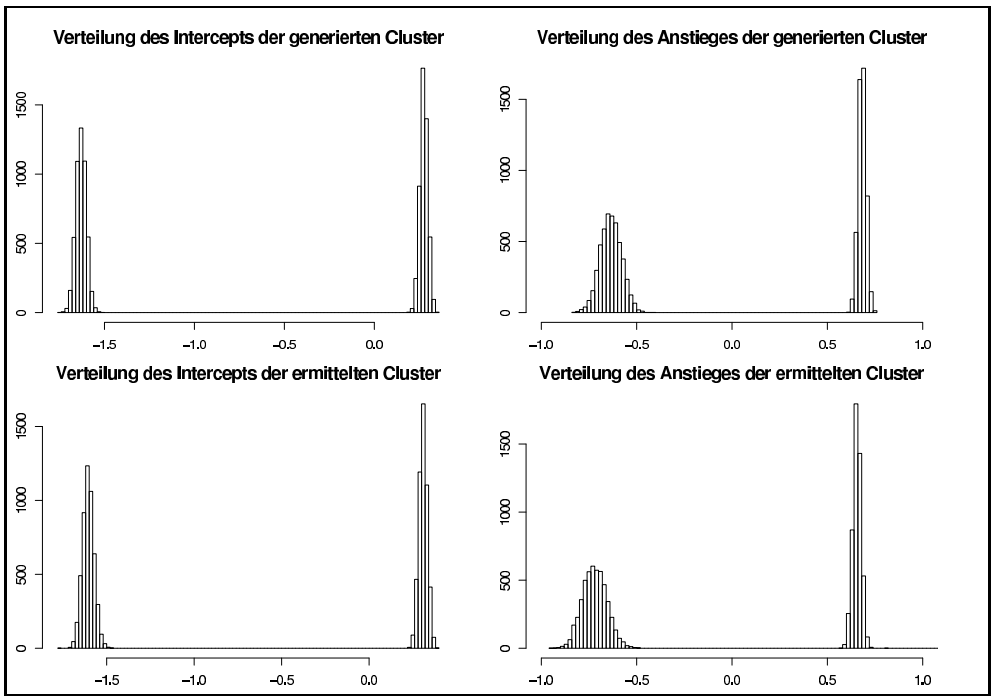

Abb. 3.20: Verteilungen der resample- und MRC-Parameter nach 5000 Wiederholungen.

Auch bei diesem Beispiel waren die Abweichungen der Residuen in jedem Cluster von der 45°-Geraden im Q-Q-Plot minimal. Es werden daher normalverteilte Residuen angenommen. Abbildung 3.20 soll lediglich einen Vergleich der Parameter der generierten mit den ermittelten Clustern ermöglichen. Die Tatsache, dass es für jeden Parameter eines Clusters nur einen Peak in der Häufigkeitsverteilung gibt, zeigt, wie stabil die Zuordnung der Objekte zu den Clustern ist. Zu hundert Prozent findet der Algorithmus die Partition in ihrer grundlegenden Gestalt wieder. Lediglich einige Punkte am Schnittpunkt beider Regressionscluster werden systematisch in der MRC-Analyse anders als beim Generieren zugeordnet. Dadurch ergibt sich für das kleinere Cluster (C1) ein steilerer Anstieg[3] von $b_{C1}^{MRC} = -0.72$ gegenüber $b_{C1}^{res} = -0.64$. Die Anstiege des größeren Clusters (C2) sind dagegen kaum verändert, da die größere Anzahl von Objekten den Cluster träger gegenüber Veränderungen durch einige wenige Objektwechsel macht.

[3]Die Kürzel *res* und *MRC* stehen für die Kennzeichnung der Ergebnisse aus den neu generierten Daten (res) und der Multiregressionscluster-Analyse an diesen neugenerierten Daten (MRC)

Hier verschiebt sich der Anstieg nur von $b_{C2}^{res} = 0.68$ zu $b_{C2}^{MRC} = 0.66$. Zur Bestimmung des 95%-Vertrauensintervalls müssen auf beiden Seiten 125 Ereignisse der Analyse abgezählt werden. Daraus ergeben sich folgende Grenzen für den Mittelwert der Grundgesamtheit:

$$-0.84 \leq \beta_{C1} \geq -0.59 \tag{3.10}$$

$$0.61 \leq \beta_{C2} \geq 0.70. \tag{3.11}$$

Innerhalb dieser Intervalle befinden sich die wahren Anstiege der beiden Cluster, mit einer Irrtumswahrscheinlichkeit von $\alpha = 5\%$. In Abbildung 3.19 ist der Datensatz relativer Weizenertrag über relative Niederschlagsmenge dargestellt. Es ist der um 90° gedrehte Datensatz, welcher in die Analyse einging sowie der ursprüngliche Datensatz eingetragen. Zusätzlich wurden die beiden Regressionscluster in beide Darstellungen eingefügt.

Inhaltliche Analyse

Nun soll im nächsten Schritt diskutiert werden, wie sinnvoll diese Klassifizierung der Länder unter inhaltlichen Gesichtspunkten erscheint. Dazu werden zuerst die Länder den Clustern zugeordnet und in einer Tabelle mit den zugehörigen Klasseneigenschaften dargestellt.

Land	Cluster 1	Cluster 2	
	$Y^1 \sim +1.7 \cdot P$	$Y^2 \sim -1.5 \cdot P$	
Österreich	-	x	
Thailand	-	x	
Mauretanien	-	x	
Schweiz	-	x	
Türkei	x	-	
Grossbritanien	-	x	

Tabelle 3.2: Zuordnungstabelle der beiden Cluster des Weizendatensatzes.

Das größere Cluster, in dem der Ertrag von Weizen mit zunehmender Niederschlagsmenge abnimmt, wird von fünf Ländern belegt. Lediglich die Türkei bildet das zweite fallende Cluster. Bei Betrachtung der Veränderung der Niederschlagsmengen der letzten 40 Jahre ist diese Zweiteilung innerhalb dieser Länder auch zu finden. Die über die Landesfläche und die Vegetationsperiode gemittelte Niederschlagsmenge nimmt über den Zeitraum lediglich in der Türkei zu. In den anderen fünf Ländern nimmt diese Größe hingegen ab. Diese Zunahme der Regenmenge wirkt sich mit Sicherheit positiv auf den Flächenertrag von Weizen und anderen Nutzpflanzen aus. Mit zunehmender Trockenheit müßten sich die Erträge dann jedoch auch verringern.

Da in den anderen Gebieten der Zusammenhang jedoch genau umgekeht ist, kann dies nicht die einzige Kraft sein, welche auf die Erträge wirkt. Die Erträge nehmen in den letzten Jahren in nahezu allen Ländern der Welt durch intensivierte Anbaubedingungen zu. Zu diesen Maßnahmen zählen der Einsatz von Düngemitteln, Pestiziden sowie eine verstärkte Mechanisierung der landwirtschaftlichen Methoden.
Die unterschiedliche Veränderung der Niederschlagsmengen in den sechs Ländern verursacht die Aufspaltung im Verlauf des Flächenertrages. Mit Sicherheit gibt es neben den Niederschlägen weitere Ursachen für die Entwicklung des Ertrages vom Weizen. Welche dies sind kann nur eine tiefergehende Analyse mit einer größeren Zahl von erklärenden Variablen liefern. Die Aufteilung der Datenpunkte in diesem Datenraum in zwei Cluster scheint jedoch gerechtfertigt.
Im folgenden Datensatz werden nun mehrere unabhängige Variablen zur Erklärung der Änderung des Ertrages einer Nutzpflanze verwendet.

Änderung des Flächenertrages von Getreide

In der folgenden Analyse werden die relativen jährlichen Änderungsraten der Variablen Ertrag pro Hektar (Y), Temperatur (T), Niederschlag (P), Düngemitteleinsatz pro Hektar (F), Mechanisierung pro Hektar (M) und Landwirtschaftliche Nutzfläche (L) verwendet.

$$\Delta Y_{CerealToT} = f(\Delta T, \Delta P, \Delta F, \Delta M, \Delta L)$$

Diese wurden per Regression über die Zeit von 1961 bis 2000 aus den absoluten Daten berechnet.
Zur Überprüfung der Anpassung jeder einzelnen Regression zur Trendbestimmung wurden die Bestimmtheitsmaße (2.8) untersucht. 20% der Regressionen konnten nur schlecht ($R^2 < 0.2$) an die Daten angepasst werden. Die entsprechenden Länder verblieben trotzdem weiterhin in der Analyse, da es sich in allen Fällen nur um eine oder zwei der sechs Trendbestimmungen (sechs Variablen!) pro Land handelte.
Die berechneten Anstiege wurden mit dem Mittel der jeweiligen Variablen über denselben Zeitraum normiert. In Abbildung 3.21 sind die Daten durch Projektionen auf die entsprechenden Koordinatenebenen dargestellt.
Mittels der Abbildung 3.21 konnten die Daten gut nach Ausreißern untersucht werden. Bei dem mit einem Kreis markierten auffälligen Wert handelt es sich um die Temperaturänderung in der Schweiz. Beim Nachprüfen konnte dieser hohe Wert nicht als Fehlwert interpretiert werden und wird daher in der Analyse belassen. Weitere Ausreißer sind nicht zu erkennen.

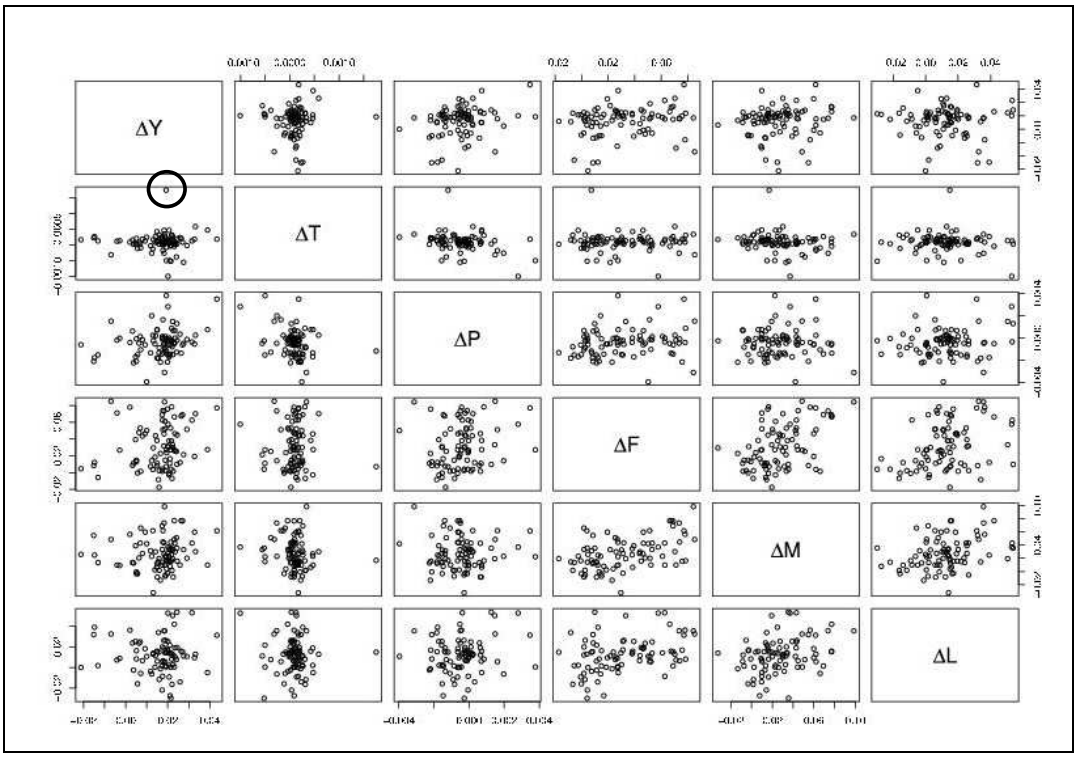

Abb. 3.21: Projektionsdarstellung der Matrix bestehend aus abhängiger Variable *Änderung des Ertrages von Cerealien Total* und den fünf erklärenden Variablen *Änderung von Niederschlag, Temperatur, Düngemittel, Mechanisierung und Anbaufläche.*

In den 85 Staaten wird nun nach Ländergruppen und deren Zusammenhänge zwischen der zu erklärenden Variable *Änderung des Ertrages aller Getreidearten* und den oben erwähnten unabhängigen Variablen gesucht. Die Indikatoren für die einzelnen Größen sind in Tabelle 3.1 eingetragen[4].

Optimale Clusteranzahl

Die Analyse der optimalen Clusteranzahl wird mittels einer Multiregressionsclusterung für die Partitionen mit der Clusteranzahl 1 bis 7 durchgeführt. Die Ergebnisse von fünf wiederholten Durchgängen mit veränderter Anfangspartition sind in Abbildung 3.22 dargestellt.

[4]Als Indikator für die Mechanisierung wird die Anzahl der Traktoren pro Hektar Anbaufläche verwendet.

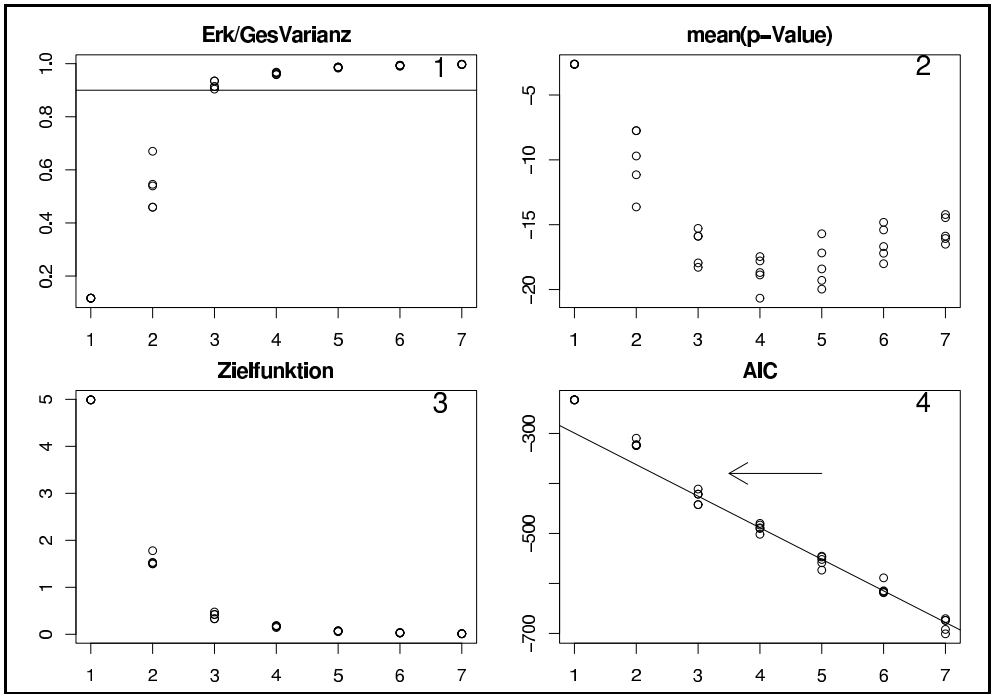

Abb. 3.22: Darstellung der Gütewerte für die Partitionen von K=1..7.

Im Abbildung 3.22(1) ist der Anteil der erklärten an der gesamten Varianz dargestellt. Bei P_3 (Partition mit $K = 3$) werden 90% erklärte Varianz überschritten. In 3.22(2) ist das Mittel der Irrtumswahrscheinlichkeiten (p-Value) der einzelnen Cluster aufgetragen. Hier ist deutlich ein Minimum bei P_4 zu erkennen. Die Abbildungen (3) und (4) stellen die Zielfunktion, also die Summe der Residuenquadrate und das *AIC* dar. In Abbildung 3.22(4) ist zusätzlich eine Regressionsgerade, welche aus den Werten der Partitionen P_3 bis P_7 berechnet wurde, eingetragen. Mit Hilfe dieser Geraden wird der Absatz im Verlauf der *AIC*-Werte von $K = 2$ auf $K = 3$ deutlich.
In diesem Fall ist die Analyse der geeigneten Clusteranzahl nicht eindeutig. Das Minimum beim Gütemaß *mean(p-Value)* spricht für die Partition P_4. Jedoch deutet der Absatz im Verlauf der *AIC*-Werte auf eine Partition mit drei Clustern. Bei beiden Partitionen sind mehr als 90% der Varianz erklärt. Die im folgenden Abschnitt angewendete Bootstrap-Analyse wird die auch vom Prinzip der Sparsamkeit geforderte Partition P_3 bestätigen. Bei der Bootstrap-Analyse stellte sich die Partition P_4 als nicht stabil heraus.
Tabelle 3.3 gibt detailierte Informationen über die Paramter der Cluster in der Partition P_3 und ihre zugehörigen Signifikanzwerte.

	$\hat{b}_0$	$\hat{b}_{\Delta T}$	$\hat{b}_{\Delta P}$	$\hat{b}_{\Delta F}$	$\hat{b}_{\Delta M}$	$\hat{b}_{\Delta L}$	R^2	p-Val
Cluster 1	0.47	-0.09	-0.03	-0.17	0.16	0.11	0.43	2.7e-03
Signifikanz	+++	.		+++	++			
Cluster 2	0.30	2.06	0.36	-0.53	0.76	-0.55	0.94	5.8e-12
Signifikanz	+++	+++	+++	+++	+++	+++		
Cluster 3	0.01	-3.34	0.32	1.01	-0.05	0.54	0.96	9.6e-11
Signifikanz		+++	+++	+++		+++		

Tab. 3.3: Parameter der drei Cluster und ihre Signifikanzwerte aus den partiellen F-Werten. Signifikanzlevel: +++ p<0.001, ++ p<0.01, + p<0.05, . p<0.1. *p-Val* in der letzten Spalte gibt den Signifikanzwert aus dem F-Test des Bestimmtheitsmaßes des gesamten Regressionsclusters an.

Alle Werte haben in zwei von den drei Clustern eine signifikant erklärende Bedeutung. Lediglich die Änderung der Düngemenge trägt in allen drei Clustern bedeutend zur Ertragsänderung bei. Auch sind das Bestimmtheitsmaß und der p-Wert für jedes Regressionscluster mit angegeben. Hierbei wird deutlich, wie sehr das erste Cluster (C1) in seiner Güte von den beiden anderen abweicht. Die Irrtumswahrscheinlichkeit bei C1 beträgt 0.3%. Die Anzahl der Objekte in den Clustern beträgt $C1 = 37$, $C2 = 27$ und $C3 = 21$.

Statistische Prüfung der Robustheit des Ergebnisses

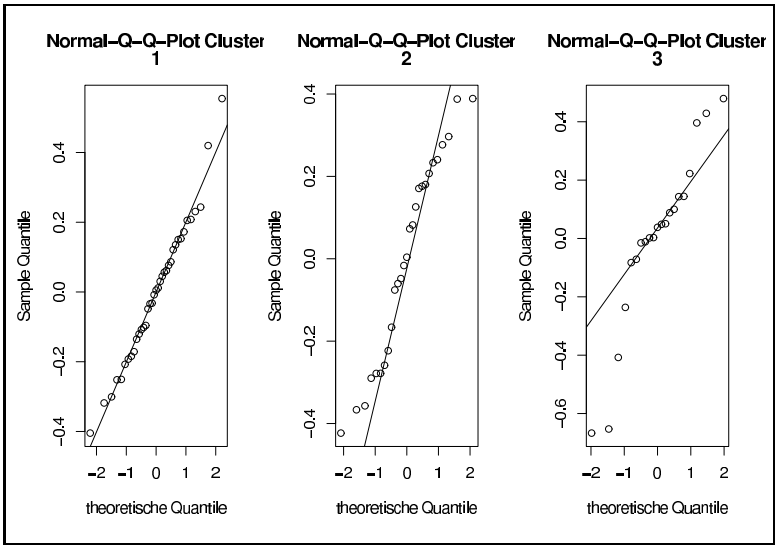

Abb. 3.23: Q-Q-Plot der Residuen aus C1 (links), C2 (mitte) und C3 (rechts).

Zu Beginn der statistischen Prüfung werden die Residuen der drei Datensätze durch einen *Q-Q-Plot* mit einer Normalverteilung verglichen. In Abbildung 3.23 wird deutlich, dass im ersten und zweiten Cluster die Abweichungen zur 45°-Geraden nur gering sind während in Cluster drei die Abweichungen bei den Randwerten stark zunehmen. In diesem Fall wäre es möglich, anstatt neue Bootstrap-Werte aus einer Gaußverteilung zu ziehen, diese per Zufall aus der alten Datenmenge zu entnehmen. Damit würde diese nichtgaußsche Verteilung kompensiert.
In Abbildung 3.24 sind, wie bei den vorherigen Datensätzen, die Ergebnisse des Bootstrapping angegeben, mit welchem die Stabilität des Ergebnisses der MRC-Analyse überprüft wurde (siehe Abschnitt 2.4.2). Da der Datensatz in dieser Analyse aus fünf unabhängigen Variablen besteht, sind fünf Verteilungen jeweils für die Anstiege jeder Dimension nebeneinander dargestellt ($\hat{b}_{\Delta T}, \hat{b}_{\Delta P}, \hat{b}_{\Delta F}, \hat{b}_{\Delta M}, \hat{b}_{\Delta L}$). Während die obere Reihe die *resample*-Parameter darstellt, sind in der unteren die durch die MRC-Analyse gefundenen MRC-Parameter dargestellt. In jedem Einzelbild befinden sich die überlagerten Verteilungen der Parameter von drei Clustern. Der Vergleich der Mittelwerte der resample-Parameter mit den Parametern der entsprechenden Partition (siehe Tab. 3.3) zeigt, dass die aus den neu generierten Daten gebildeten Regressionsgeraden den Ursprungsdaten in ihren Parametern gleichen. Ein weiterer Vergleich der unteren Abbildungen mit den oberen Abbildungen, also von resample- und MRC-Parametern, belegt ebenfalls eine gute Übereinstimmung. Das heißt, die Cluster der Partition werden unter Bootstrap-Bedingungen gut durch die MRC-Analyse wiedergefunden.
Die Mittelwerte und Vertrauensintervalle der Verteilungen der fünf Anstiege in den drei Clustern lassen sich nur schwer bestimmen, da die Verteilungen von ΔP, ΔM und ΔL stark und bei ΔF teilweise überlappen. Es werden daher nur beispielhaft die Mittelwerte und Vertrauensintervalle für ΔT bestimmt. In den Gleichungen 3.12-3.14 sind die Mittelwerte $\bar{b}_{Ck}$ der Verteilungen der Anstiege in der Variable ΔT dargestellt. Weiterhin sind die Vertrauensintervalle der Mittelwerte β_{Ck} der Grundgesamtheit angegeben.

$$\bar{\hat{b}}_{C1} = -0.07 \qquad -0.17 \leq \beta_{C1} \geq 0.04 \tag{3.12}$$

$$\bar{\hat{b}}_{C2} = 2.07 \qquad 1.66 \leq \beta_{C2} \geq 2.40 \tag{3.13}$$

$$\bar{\hat{b}}_{C3} = -3.23 \qquad -4.14 \leq \beta_{C3} \geq -1.44 \tag{3.14}$$

Beim Vergleich mit Tabelle 3.3 ist zu erkennen, dass der Mittelwert der ΔT-Anstiege für C2 sehr gut übereinstimmt. Bei C3 gibt es eine leichte Verschiebung und bei C1 eine prozentual größere Verschiebung des Mittelwertes.

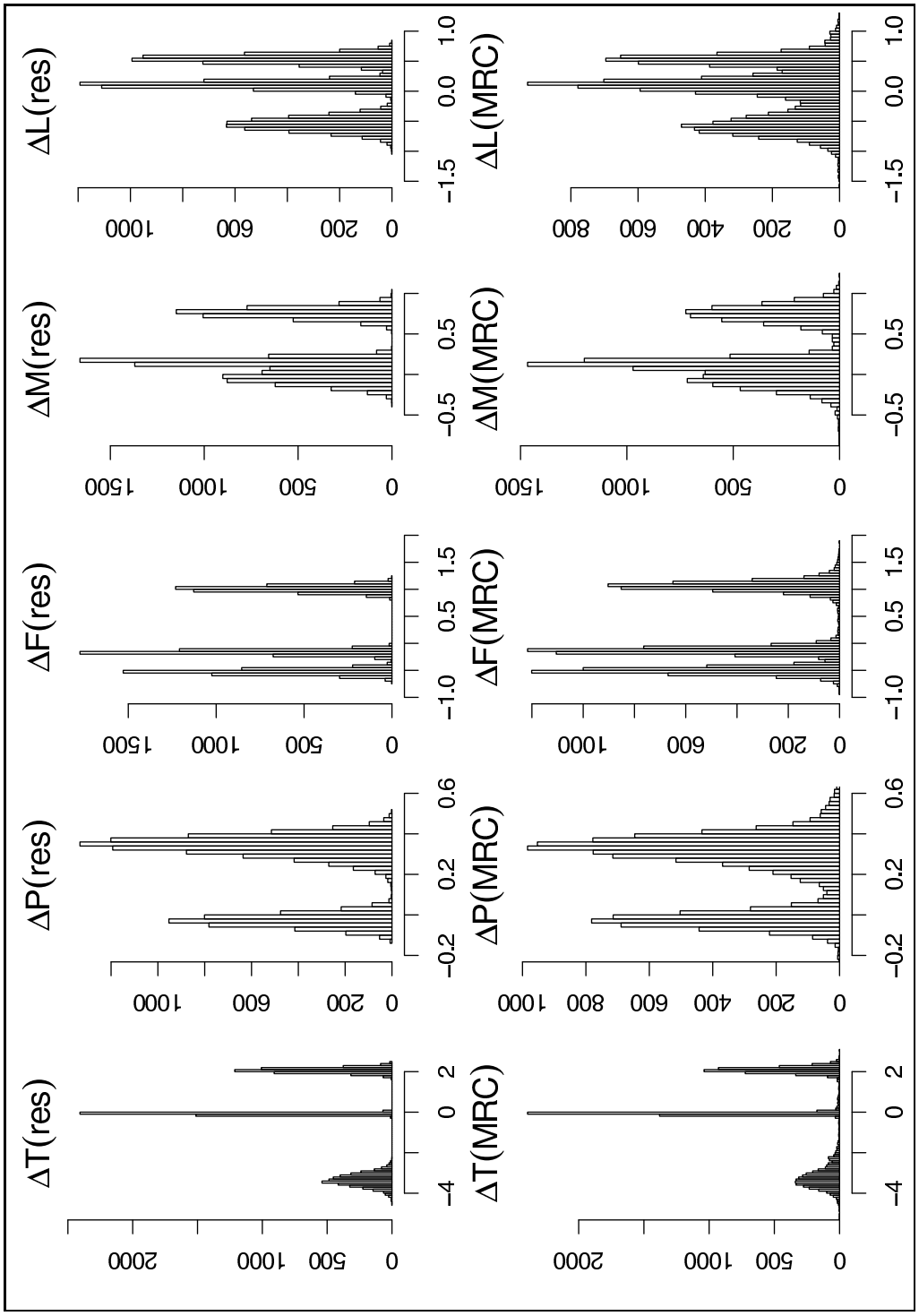

Abb. 3.24: Verteilungen der resample- und MRC-Anstiege nach 4000 Wiederholungen. In den Überschriften der einzelnen Grafiken sind die Variablen der geschätzten Anstiege eingetragen.

Das Vertrauensintervall für $\alpha = 0.05$ ist bei C3 im Vergleich zum Mittelwert sehr unsymmetrisch. Möglicherweise überlagern sich hier zwei Verteilungen, die darauf hindeuten, dass es zwei nahe beieinanderliegende stabile Zustände für den Anstieg dieses Clusters gibt. Bei C1 tritt sogar der Fall auf, dass der Parameter der Grundgesamtheit innerhalb des Vertrauensintervalls zwei qualitativ unterschiedliche Werte annehmen kann. Der Anstieg dieses Clusters in ΔT könnte auch positiv sein. Da dieser Parameter, wie aus Tabelle 3.4 zu entnehmen, keinen hohen Signifikanzwert hat, wirkt sich dies nicht auf die Robustheit des Ergebnisses aus.

Inhaltliche Analyse

Um eine Übersicht zu erhalten, wie die 85 Länder, welche in die Analyse eingegangen sind, auf die drei Cluster aufgeteilt sind, wurden die Staaten und ihre Cluster farblich auf einer Weltkarte Abbildung 3.25 markiert. Eine offensichtliche geographische Einteilung der Länder gemäß ihrer Clusterzugehörigkeit lässt sich nicht ausmachen. Es wurde daher eine weitere Untersuchung der Eigenschaften der drei Cluster vorgenommen. Tabelle 3.4 zeigt die clusterweise gemittelten Werte der einzelnen in der MRC verwendeten Variablen. In Tabelle 3.5 sind zusätzlich Clustermittelwerte von weiteren Eigenschaften der Objekte aufgelistet. Eine solche Betrachtung zur Interpretation nähert sich wieder einer eher traditionellen Clusterung an, trägt aber in diesem Fall zur Plausibilisierung einiger Eigenschaften der MRC-Cluster bei.

	# Länder	$\Delta\bar{Y}$	$\Delta\bar{P}$	$\Delta\bar{T}$	$\Delta\bar{F}$	$\Delta\bar{M}$	$\Delta\bar{L}$
Cluster 1	37	0.020	- 3.5e-04	6.4e-05	0.029	0.023	0.010
Cluster 2	27	0.014	- 7.1e-04	10.0e-05	0.032	0.029	0.012
Cluster 3	21	0.012	- 2.5e-04	12.0e-05	0.037	0.035	0.018

Tab. 3.4: Mittelwerte der Änderungen der sechs in die MRC-Analyse eingegangenen Variablen im jeweiligen Cluster. Beispielsweise beträgt die Änderung des Ertrages in Cluster 1 im Mittel über alle Länder des Clusters 2%/Jahr.

	Y 100g/ha	T °C	P mm	F 100g/ha	M #/ha	L %	GDP/Kopf $/Kopf
Cluster 1	29700	17.9	111	1525	4.9	3.9	8895
Cluster 2	21100	19.7	105	882	1.8	3.0	5267
Cluster 3	21500	22.7	135	1075	3.1	2.1	4520

Tab. 3.5: Mittelwerte von Eigenschaften der Staaten im jeweiligen Cluster (1990).

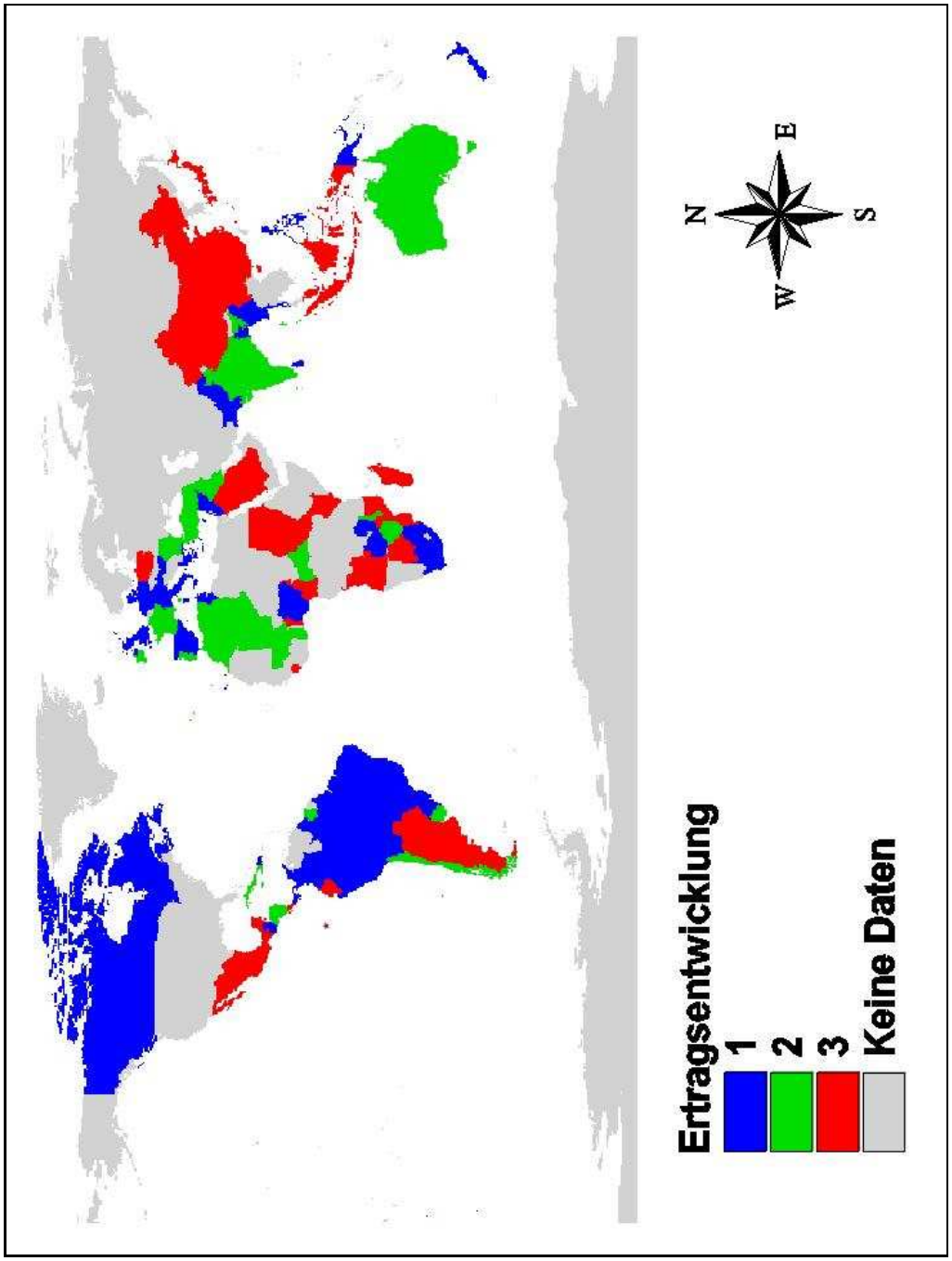

Abb. 3.25: Weltkarte mit farblicher Unterscheidung der drei Cluster.

Die in den Tabellen 3.4 und 3.5 aufgelisteten Eigenschaften lassen einige grundsätzliche Unterschiede der drei Cluster erkennen.

Typisierung der drei Cluster

C1 Das Getreide-Ertragsmittel (Y) ist in C1 um ca. 1/3 höher als in C2 bzw. C3. Da auch die Mechanisierung (M) und der Fertilizereinsatz (F) in C1 sich im Mittel deutlich über den beiden anderen Clustern befinden, liegt die Vermutung nahe, dass es sich bei einem Großteil der Länder dieses Clusters um entwickeltere Staaten mit einer industrialisierten, intensiven Landwirtschaft handelt. Die niedrigsten Änderungen in der Düngemittelmenge (ΔF) und der Mechanisierung (ΔM) sprechen für eine Sättigung im Bereich der Technisierung der Landwirtschaft in diesen Regionen. Geeignete Beispiele für diese Gruppe sind Kanada, Deutschland und Großbritannien.

C2 Dieses Cluster zeichnet sich durch die niedrigsten Werte für den Einsatz von Traktoren und Dünger aus. Die zugehörigen Länder sind im Vegetationsperiodenmittel die trockensten und nach Tabelle 3.4 auch die Länder mit der stärksten Abnahme der Niederschlagsmenge (ΔP) pro Jahr. Der Abfall der Regenmenge liegt in C2 um 170% über dem von C3 und ist doppelt so groß wie in C1. Diese Gruppe zeichnet sich durch eine weniger technisierte Landwirtschaft (M) aus. Beispielhaft sind Staaten wie Indien, Burkina Faso und Zimbabwe.

C3 Beim dritten Cluster ist das Mittel der Temperaturen (T) mit 23°C am höchsten und mit 135mm Regen pro Monat in der Vegetationsperiode am feuchtesten. Die Zunahme der Technisierung in der Landwirtschaft (ΔM) ist am stärksten. Ebenfalls ist die Zunahme der Anbaufläche (ΔL) in C3 um 80% stärker als in C1 und 50% über der Zunahme in C2. Jedoch ist der absolute Anteil der landwirtschaftlichen Nutzfläche (L) in C3 auch weitaus geringer als in C2 und C1. Dieses Cluster umfasst Staaten mit moderat technisierter Landwirtschaft und einer dynamischen Entwicklung in feuchten und warmen Regionen der Erde wie z.B. China und Indonesien.

Eine Übersicht aller signifikanten Zusammenhänge ist in Tabelle 3.6 angefertigt.

$\Delta Y \sim$	ΔT	ΔP	ΔF	ΔM	ΔL
C1			$-$	$+$	
C2	$++$	$++$	$--$	$++$	$--$
C3	$--$	$++$	$++$		$++$

Tab. 3.6: Die signifikanten Zusammenhänge in den Clustern. $(++, --)$ stehen für starke Zusammenhänge. $(+, -)$ für schwache Zusammenhänge (Vergleich nur in einer Variablen, da Regressionskoeffizienten nicht standardisierten).

Deutung der Parameter

ΔT Im Mittel hat die Temperatur in allen drei Clustern im Untersuchungszeitraum zugenommen. Dies spiegelt den weltweiten Trend zum Temperaturanstieg, maßgeblich mitverursacht durch einen verstärkten Treibhauseffekt, wider. In C3 führt eine Schwächung dieser Temperatursteigerung zu einer verstärkten Ertragssteigerung. Dies erscheint plausibel, da eine starke Zunahme der Temperatur zu einer verstärkten Verdunstung führt, welche sich negativ auf das Wachstum der Getreidepflanzen auswirkt. Aus diesem Grund eher kontraintuitiv ist der Zusammenhang in C2, wonach bei verstärkter Zunahme der Temperatur mit einer weiteren Ertragssteigerung zu rechnen ist. Besonders unter dem Aspekt, dass die Länder aus C2 zu den am stärksten von zunehmender Trockenheit betroffenen zählen.

ΔP Im Clustermittel nehmen die Niederschläge in allen drei Clustern ab. Vermutlich ist dies ebenfalls als eine Auswirkung des globalen Klimawandels zu sehen. Die nur in C2 und C3 signifikanten Zusammenhänge, nach denen mit zunehmenden Niederschlagsänderungen, also zusätzlichem Regen, die Ertragsänderungen zunehmen, sind ebenfalls einsichtig. Solange dies nicht zu übermäßiger Bodenfeuchte führt, wirken sich vermehrte Niederschläge positiv auf das Wachstum der Nutzpflanze aus.

ΔF Bei verstärkter Düngemittelzunahme nimmt im Cluster der dynamischen Staaten C3 die Ertragsrate ebenfalls zu. In den Clustern C2 und C1 jedoch wirkt sich diese Zunahme der Intensivierung der Landwirtschaft eher negativ auf die Ertragsentwicklung aus. In den entwickelten Staaten aus C1 lässt sich dies möglicherweise mit einem Sättigungseffekt durch Überdüngung der Böden erklären. Da die Länder in C2 zu den trockensten gehören und wie oben erwähnt hier auch die größte Abnahme an Niederschlägen zu verzeichnen ist, ist anzunehmen, dass der Mangel an Wasser den positiven Effekt von mehr Düngemittel limitiert.

ΔM Eine forcierte Mechanisierung wirkt sich signifikant nur auf die beiden Cluster C1 und C2 aus; auf beide jedoch positiv. In den entwickelten Staaten bewirkt diese Intensivierung eine lediglich moderate Zunahme der Erträge. Dies kann auf den bereits hohen Grad an Mechanisierung in C1 zurückgeführt werden. Dieser könnte sich, wie beim Dünger, ab einem bestimmten Level durch eine Übernutzung des Bodens, z.B. durch eine verstärkte Bodenverdichtung, negativ auf die Qualität des Bodens und damit auf die Erträge der Nutzpflanzen auswirken. Durch den geringen Grad an Intensivierung in C2 wirkt sich eine Zunahme von ΔM hier deutlich positiv auf die Ertragszunahme aus.

ΔL Für diese Variable gibt es wieder nur für C2 und C3 signifikante Zusammenhänge. Dabei sind die Wirkungen von einer verstärkten Ausweitung der Nutzfläche in beiden Cluster entgegengesetzt. In C3 bewirkt eine forcierte Ausweitung bei einem Nutzflächenanteil (1990) von 2.3% eine positive Änderung der Ertragsentwicklung. Beim zweiten Cluster mit einem Nutzflächenanteil von 3.0% bewirkt sie eine Schwächung der Ertragsentwicklung. Wir vermuten, dass ein Ausweiten der Nutzflächen in diesem Cluster nur auf Gebiete mit Böden schlechterer Qualität möglich ist.

Kapitel 4

Auswertung

4.1 Zusammenfassung und Diskussion

Im Verlauf dieser Arbeit konnte gezeigt werden, dass die hier vorgeschlagene Multiregressionsclusterung (MRC) in der Lage ist, an synthetischen und empirischen Datensätzen eine Gruppierung von Datenpunkten nach ihren funktionalen Zusammenhängen durchzuführen. Dabei wird für entsprechende Datenstrukturen eine bessere Beschreibung erreicht als dies mit einer einfachen Multiregressionsanalyse möglich wäre.
Wie in jeder Clusteranalyse stellt sich auch in der MRC das Problem der Identifikation der optimalen Clusteranzahl. In Abschnitt 3.1 wurde hierzu die Wirkungsweise unterschiedlicher Gütemaße vorgestellt. Es wurde deutlich, dass eine eindeutige Aussage zur optimalen Anzahl der Cluster mit nur einem der Gütemaße nicht zu treffen ist. Vielmehr müssen viele Informationen aus mehreren Gütemaßen und eine inhaltliche Bewertung der geeigneten Partitionen zusammenfließen. Auch wurde in diesem Abschnitt deutlich, wie verschiedene Parameter des Algorithmus auf das Ergebnis wirken können.
Die Methode hat jedoch auch ihre Grenzen. Bei Daten, die eine multilineare Struktur enthalten, jedoch zu stark verrauscht sind, lässt sich keine eindeutige Aussage über das bevorzugte Modell abgeben. Jedoch bleibt die Frage offen, ob eine Struktur im Datensatz einfach nicht existiert oder nicht gefunden wurde.
Ein Problem in der Optimierung, so auch in der MRC-Analyse, sind lokale Minima. Im Abschnitt 2.4.1 wird das Simulated Annealing beschrieben, mit welchem die Wahrscheinlichkeit, das globale Minimum zu finden erhöht werden kann. Es wurden eine konventionelle und eine alternative Methode vorgestellt.
Im zweiten Teil des Kapitels *Anwendung der Methode* wurde die Multiregressionsclusterung auf empirische Daten angewandt. Dabei wurden verschiedene Datensätze analysiert.
Im ersten Datensatz zu *Sterblichkeit und Unterernährung bei Kindern* wurden Zustandsgrößen einander gegenübergestellt. Es konnte gezeigt werden, dass eine Einteilung in zwei Regressionscluster besonders geeignet ist und sich sinnvoll interpretieren lässt. Als besonders aussagekräftig haben sich hier die Signifikanzwerte *log(max(Ftest))* und *mean(p-Value)* gezeigt. Eine nähere Untersuchung der beiden Ländergruppen konnte darlegen, dass es sich bei den analysierten Ländern um zwei Typen unterschiedlicher ökonomischer Entwicklung handelt. Aufgrund des unterschiedlichen mittleren Einkommens in den beiden Gruppen war auf Unterschiede bei anderen Indikatoren der Entwicklung wie Zugang zu Trinkwasser oder Sanitäranlagen zu schließen.

Diese wiederum werden als Ursache dafür angenommen, dass in den Ländern des ökonomisch weniger entwickelten Multiregressionsclusters die Menschen weniger Möglichkeiten haben, auf Folgen einer Mangelernährung des Kindes zu reagieren und diese Unterernährung daher zu einer höheren Sterblichkeit bei Kindern führt.
Die beiden folgenden Datensätze waren ebenfalls zweidimensional, jedoch aus Daten verschiedener Zeitpunkte zusammengesetzt. Dabei wurden die Erträge von Sorghum bzw. Weizen der Entwicklung der Niederschläge in der Vegetationsperiode ausgewählter Länder gegenübergestellt. Es ergab sich eine klare Struktur aus zwei Clustern. In beiden Datensätzen wurde deutlich, dass eine Aufspaltung der Länder in zwei Typen durch unterschiedliche sozioökonomische Faktoren verursacht wurde. Die beiden Cluster im Datensatz *Sorghumertrag und Niederschlag* werden vermutlich durch Unterschiede in der Intensivierung der Anbaumethoden verursacht. Diese beispielhaften Anwendungen machten deutlich, dass nur eine Analyse mit mehr Einflussfaktoren eine umfassende Erklärung für die Ertragsentwicklung liefern würde.
Daher wurden im letzten Beispiel die Änderungen von fünf unabhängigen Variablen über 40 Jahre dazu verwendet, die Änderung von Getreideerträgen in diesem Zeitraum zu erklären. Besonders aussagekräftig waren für das Auffinden der geeigneten Clusteranzahl die Gütemaße *Erk/GesVarianz* und *mean(p-Value)*. Die drei Cluster der bestimmten Partition konnten aufgrund der Eigenschaften der Länder typisiert und die zugehörigen Parameter, also die Zusammenhänge zwischen den Variablen, sinnvoll mit dieser Typisierung in Verbindung gebracht werden. Die drei Ländertypen zeichneten sich durch einen unterschiedlichen Grad der ökonomischen Entwicklung, der Intensivierung ihrer Landwirtschaft sowie der Dynamik ihrer Entwicklung aus.
Als wirkungsvolles Werkzeug bei der Ermittlung der statistischen Signifikanz der Ergebnisse erwies sich das *Bootstrapping*-Verfahren.

4.2 Probleme und Ausblick

Ein grundlegendes Problem stellt die Annahme dar, dass empirische Daten auf Länderebene repräsentativ für die in diesem Land lebende Bevölkerung sind. Über alle sozialen Bevölkerungsschichten, ethnischen Gruppen, geographischen und infrastrukturellen Unterschiede wird hier gemittelt und so für alle eine zumindest ähnliche Situation vorausgesetzt. „Shifting scales from groups of individuals or regions to the nation requires aggregation and generalization, such that some losers will not be identified when the country is considered a winner“ [O'Brien et.al, 2000].

Regionen und Bevölkerungsteile, in denen bestimmte Mechanismen eine bedeutende Rolle spielen und die gesuchten Zusammenhänge deutlich erkennbar sein sollten, werden so mit anderen Regionen und Bevölkerungsteilen vermischt und typische funktionale Zusammenhänge damit auf dieser Ebene schwerer identifizierbar. Für eine klarere Identifizierung dieser funktionalen Zusammenhänge bedarf es Daten mit feinerer geographischer und sozialer Auflösung. Da dies weltweit nahezu unerreichbar scheint, sollte sich die Analyse auf bestimmte Gebiete beschränken, in denen die Datenlage ausreichend ist.

Weiterhin stellt sich die Frage inwieweit die Zuordnung eines Objektes zu einem Cluster eine Aussage über eine zeitliche Dauer der Zugehörigkeit zulässt. Es wäre sinnvoll, an diese Untersuchung eine Analyse anzuschliessen die die Stabilität von Clusterzugehörigkeiten über die Zeit untersucht.

Der vorgestellte Multiregressionscluster-Algorithmus verwendet als zu optimierende Zielfunktion die Summe der quadrierten Residuen. Diese stellen jedoch lediglich den Abstand in der Dimension der zu erklärenden Variablen dar. Befindet sich ein Regressionscluster nahezu parallel zu einer Ebene der erklärenden Variablen, werden die Abstände zwischen Punkten und Regressionsgerade unverhältnismäßig groß und der Algorithmus ist unter Umständen nicht in der Lage, das Regressionscluster zu finden. Bisher wurde der Datensatz gedreht, um diesen Umstand zu verhindern. Mit dem Hesse-Abstand sollte es möglich sein, auch ohne Drehung dieses Problem zu beheben. Der Abstand zwischen einem Punkt und einer Ebene im Raum kann über die Hessesche Normalenform bestimmt werden. Dafür wird der Normalenvektor der Ebene $\vec{n_0}$, ein beliebiger Ortsvektor der Ebene $\vec{b_0}$ und der Ortsvektor des Punktes $\vec{p_0}$, dessen Abstand zur Ebene berechnet werden soll, verwendet.

$$\mid \vec{n_0} \cdot (\vec{p_0} - \vec{b_0}) \mid = d \tag{4.1}$$

Es müsste überprüft werden, inwieweit aus den neuen *Hesse*-Residuen (siehe 4.1) über ein der *Methode der kleinsten Quadrate*-ähnliches Verfahren die Regressionsgerade direkt bestimmt werden kann oder ob mit dieser Form der Abstandsberechnung lediglich die Zuordnung der Objekte zu Clustern verbessert und die eigentliche Berechnung der Regressionshyperebene unverändert bleibt. Besonders geeignet für diese modifizierte Zielfunktion erscheint das Minimaldistanzverfahren, welches jedes Objekt ins nächstliegende Cluster legt.

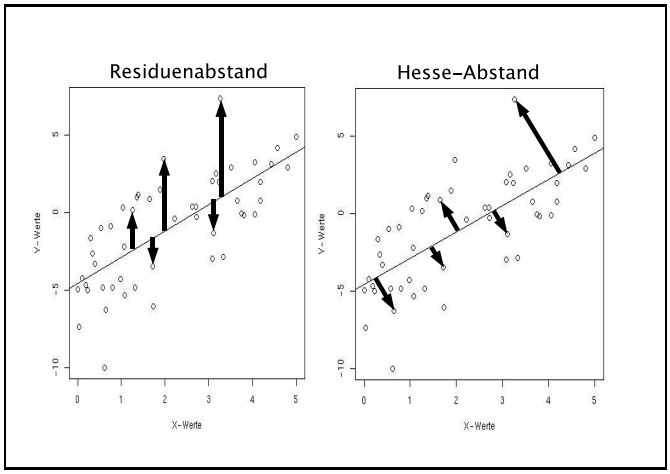

Abbildung 4.1: Gewöhnliche Residuen und Hesseabstand.

Weiterhin hat sich im Laufe der Untersuchung abgezeichnet, dass bei zukünftigen Anwendungen eine Simulated Annealing Version, die sowohl konventionelles als auch alternatives Vorgehen koppelt, sich als sehr wirkungsvoll herausstellen könnte. Zu Beginn der Analyse eines Datensatzes muss der Algorithmus q-mal in den Bereich eines Minimums gelangen, nachdem viele Objekte getauscht wurden. Damit ist der Algorithmus mit großer Wahrscheinlichkeit in der Nähe des globalen Minimums und kann nun mit dem konventionellen Simulated Annealing, also einzelnen *unerlaubten* Objektwechseln, das globale Minimum in kurzer Zeit finden. Abschließend lässt sich sagen, dass je aufwändiger und zeitintensiver eine Rechnung, desto wahrscheinlicher das Auffinden des globalen Minimums.

.

Eidesstattliche Erklärung

Ich erkläre, daß ich die vorliegende Arbeit selbständig, ohne fremde Hilfe angefertigt und nur die in den beigefügten Verzeichnissen angegebenen Hilfsmittel verwendet habe. Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

Potsdam, den 12.Februar 2007

Carsten Walther

Danksagung

Ich möchte an dieser Stelle Dr. Matthias Lüdeke, Oli Walkenhorst, Diana Sietz, Henning Rust, Malaak Kallache, Dr. Gerhard Petschel-Held († 09.09.2005) und den anderen Menschen aus der Integrierten Systemanalyse am Potsdam Institut für Klimafolgenforschung für ihre Unterstützung, ihre Vorschläge und die Diskussionen danken. Weiterhin gilt mein Dank Reyko Schachtschneider, Berit Kitzing sowie meinen Freunden und meinen Eltern für Ihre Geduld und Ratschläge.

Literaturverzeichnis

[1] Backhaus; Erichson; Plinke; Weiber *Multivariate Analysemethoden - Eine anwendungsorientierte Einführung*, Springer, Heidelberg (1980)

[2] Blatt et al. *Superparamagnetic clustering of data* Phys. Rev. Lett., 76, 3251-3254 (1996)

[3] Bollwien; Auerbach *Stochastische Prozeßmodellierung* Fachbuchverl., Leipzig (1982)

[4] Bronstein; Semendjajew; et al. *Taschenbuch der Mathematik*, Teubner, Stuttgart (1985)

[5] Cassel-Gintz; Lüdeke; Petschel-Held; Reusswig; Plöchl; Lammel; Schellnhuber *Fuzzy logic based global assessment of the marginality of agricultural land use* Climate Research 8 (2): 135-150 (1997)

[6] CIA *World Factbook* www.cia.gov/cia/publications/factbook/index.html (01.12.2006)

[7] CRU - Climate Research Unit, www.cru.uea.ac.uk/ (verwendete Daten stammen vom PIK)

[8] Dasgupta, *An Inquiry into Well-Being and Destitution*, Clarendon Press, Oxford (1995)

[9] Davison; Hinkley *Bootstrap Methods and their Application*, Cambridge University Press, Cambridge series in Statistical and Probabilistic Mathematics, Cambridge (1997)

[10] Deleeuw, *Introduction to Akaike (1973) information mining*, Springer, London (1992)

[11] Dolić *Statistik mit R - Einführung für Wirtschafts- und Sozialwissenschaftler*, Oldenbourg, München (2004)

[12] FAOSTAT - *Statistische Datenbasis der Food and Agricultural Organication of the United Nations* www.faostat.fao.org (1.12.2006)

[13] Förster; Rönz *Methoden der Korrelations- und Regressionsanalyse*, Die Wirtschaft, Berlin (1979)

[14] George, *The Variable Selection Problem* Journal of the American Statistical Association, 95 (2000)

[15] Glaser, *Varianzanalyse*, Fischer, Stuttgart (1978)

[16] Gordon, *Classification*, Chapman & Hall, New York (1999)

[17] Kaufmann; Rousseeuw *Finding Groups in Data: An introduction to clusteranalysis*, Wiley, New York (1990)

[18] Kirkpatrick; Gelatt; Vecchi, *Optimization by Simulated Annealing*, Science 220, 4598, 671-680 (1983)

[19] Lohse, *Prüfstatistik - Ein programmierter Lehrgang*, Fachbuchverl., Leipzig (1982)

[20] Mucha, *Clusteranalyse mit Mikrocomputern*, Akademie Verlag, Berlin (1992)

[21] Maindonald; Braun *Data Analysis and Graphics Using R - An Example based Approach*, Cambridge series in Statistical and Probabilistic Mathematics, Cambridge University Press, Cambridge (2003)

[22] O´Brien; Leichenko, *Double Exposure: assessing the impacts of climate change within the context of economic globalization*, Global Environmental Change, 10, 3, 221-232 (2000)

[23] Petschel-Held; Lüdeke; Schellnhuber et al., *Syndromes of global change: a qualitative modelling approach to assist global environmental managment*, Environmental Modeling and Assessment, 4, 4, Springer (1999)

[24] Petschel-Held; Lüdeke; Schellnhuber *The Syndromes Approach to Scaling Describing Global Change on an Intermediate Functional Scale*, Integrated Assessment 3, 2-3, 201-219 (2005)

[25] Ravallion, *Poverty Comparisons*, Harwood Academic Publ., Chur (1994)

[26] Ribot et al. *Climate Variability, Climate Change and Social Vulnerability in the Semi-Arid Tropics*, Global Environmental Change 7, 1, 82-83 (1997)

[27] Runkler, *Information Mining* Springer, Braunschweig, Wiesbaden (2000)

[28] Rust, *Modellselektion und Parameterschätzung in dynamischen Systemen* unv. Diplomarbeit, Freiburg (2001)

[29] Sachs, *Angewandte Statistik*, Springer, Berlin (1984)

[30] SNL-Sandia National Laboratories, *Simulated Annealing* www.cs.sandia.gov/opt/survey/sa.html (01.12.2006)

[31] Schellnhuber et al., *Syndromes of Global Change*, GAIA, 6, 1, 19-34 (1997)

[32] Schlittgen; Streitberg *Zeitreihenanalyse* Oldenbourg, München, Wien (1999)

[33] Schlittgen, *Einführung in die Statistik - Analyse und Modellierung von Daten*, Lehr und Handbücher der Statistik, Oldenbourg, München, Wien, 5. Auflage (1995)

[34] Schmutzer, *Grundlagen der Theoretischen Physik*, Wissenschaftsverlag, Wien, Zürich (1989)

[35] Sen; Srivastava, *Regression Analysis - Theory, Methods and Applications*, Springer, New York (1997)

[36] Vogel, *Probleme und Verfahren der numerischen Klassifikation*, Vandenhoeck & Ruprecht, Göttingen (1975)

[37] Voss, *Nichtlineare statistische methoden zur Datenanalyse*, Doktorarbeit, Potsdam (1999)

[38] WDI - World Development Indicators, publiziert von der Weltbank, CD-Rom (2001)

[39] WDR - World Development Report www.worldbank.org/ (01.12.2006)

[40] www.r-project.org - Free Software Foundation (FSF), Wien (01.12.2006)