

Sylvia Winkler

Parallele Koordinaten
Entwicklung einer interaktiven Software

C oordinate

A nalyzing

S tatistical

S oftware

A pplying

T andem

T ransformations

Diplomarbeit

im Studiengang Diplom-Mathematik am Institut für Mathematik
der Universität Augsburg

September 2000

Erstgutachter: Professor Antony Unwin, Ph.D

Zweitgutachter: PD Dr. Adalbert Wilhelm

Inhaltsverzeichnis

1	Einführung	4
1.1	Beitrag dieser Arbeit	4
1.2	Literaturhinweise	6
2	Visualisierung	8
2.1	Graphiken in der Statistik	8
2.2	Darstellungsarten zur Datenanalyse	9
3	Parallele Koordinaten	12
3.1	Zusammenhänge zum Scatterplot – Punkt \leftrightarrow Linie Dualität	12
3.2	Interpretation verschiedener zweidimensionaler Strukturen in einer Parallelen Koordinaten Darstellung	15
3.3	Mehrdimensionale Betrachtung	19
3.4	Ähnlicher Ansatz zur Darstellung multivariater Datensätze	22
3.4.1	Andrews Plot	22
4	Interaktivität	24
4.1	Der Begriff Interaktivität	24
4.2	Anforderungen an interaktive Software	25
5	Entwicklung von CASSATT	27
5.1	Grundlagen zur Softwareentwicklung	27
5.2	Anforderungen an das Systemdesign	28
5.3	Oberflächengestaltung und Klassenstruktur	28
6	CASSATT	30
6.1	Grundlagen	30
6.2	Interaktive Möglichkeiten in Parallelen Koordinaten Darstellungen	32
6.2.1	Darstellungsarten	33
6.2.2	Selektion	35
6.2.3	Abfrage	35

6.2.4	Skala	37
6.2.5	Invertieren einzelner Variablenachsen	38
6.2.6	Umordnung	39
6.2.7	Sortierung	39
6.2.8	Farbe	40
6.2.9	Hiding Plots	42
6.2.10	Toggle1 & Toggle2	42
7	Erweiterte Fähigkeiten	44
7.1	Einfache Selektion	44
7.1.1	Selektionen im Scatterplot, Dotplot und Boxplot	44
7.1.2	Toggle2	45
7.2	Selektionsarten in der Parallelen Koordinaten Darstellung	45
7.2.1	Punktselektion an der Achse	45
7.2.2	Linienselektion mit Hilfe einer Linie (Pinch, Scherung)	46
7.2.3	Linienselektion mit Hilfe einer Dragbox (doppelte Scherung)	46
7.2.4	Linienselektion mit Hilfe eines Winkelintervalls	47
7.3	Translation der Selektionen der Parallelen Koordinaten Darstel- lung auf andere Darstellungen	49
7.3.1	Eindimensionale Selektion	49
7.3.2	Zweidimensionale Selektion	50
7.3.3	Drei- und mehrdimensionale Selektion	58
7.3.4	Selektion bei invertierten Achsen	59
7.4	Selektionssequenzen	59
8	Gruppen in CASSATT	62
8.1	Gruppenerstellung	62
8.2	Gruppenselektion	63
8.3	Gruppen in Parallelen Koordinaten Darstellungen	63
8.4	Gruppeninformation und Selektionsgeschichte	65
9	Beispiel: Zehnkampfdatensatz	69
9.1	Datensatz	69
9.2	Methoden	70
9.3	Analyse	70
9.4	Ausblick	78
9.5	Zusammenfassung	80

10 Zusammenfassung und Ausblick	81
10.1 Zusammenfassung	81
10.2 Ausblick	82
A Shortcuts und Befehle in CASSATT	84
A.1 Datensätze	84
A.2 Shortcuts	84
A.3 Menübefehle	85
Literaturverzeichnis	87

Kapitel 1

Einführung

It has style. Firstly, because all useless details have been eliminated, and all essential details graduated with discernment, these last contributing to an impression of the whole that is clean and vigorous; secondly because of the tone of deep sincerity and life that animates the entire work, then because of the felicitous combination of lines that converge precisely and harmoniously, demonstrating the truth the painter wishes to demonstrate, and finally because of the happy balancing of masses and areas of color, which, in perfect proportion, are also subordinated to the impression of the whole, so that one could change nothing in it, nor add, nor subtract, without diminishing its order, stability, tranquillity, and so to speak the mathematics of the work. These are the qualities that make its style.

Achille Segard, Mary Cassatt: Un Peintre des enfants and des mères, 1913. (zitiert in ...)

1.1 Beitrag dieser Arbeit

Da im Zuge der elektronischen Datenerfassung die Datenerhebung und Datenauswertung immer mehr an Bedeutung gewinnt, muß man häufiger mit multivariaten Datensätzen rechnen. Für verschiedene Voruntersuchungen von Daten eignen sich vor allem graphische Methoden. Es ist relativ einfach mit traditionellen graphischen Methoden einzelne Variablen darzustellen. Schwieriger wird es allerdings, wenn man mehr als 3 Variablen in einer Darstellung untersuchen will. Es ist einzusehen, daß schon bei der Darstellung von nur 3 Variablen, wie in einem dreidimensionalen Scatterplot, Verzerrungen auftreten.

Aus diesem Grund muß eine Darstellungsform gefunden werden, die alle Variablen gleichzeitig und mit möglichst wenig Informationsverlust darstellen kann. Eine weitere wichtige Forderung stellt ebenfalls die Gleichbehandlung dar. Damit ist gemeint, daß beispielsweise keine Achse in den Hintergrund gerückt wird oder

verzerrt dargestellt wird.

Das Problem bei den kartesischen Koordinaten stellt die Orthogonalität dar, da im Raum nur drei derartige Achsen gleichzeitig plaziert werden können. Aus diesem Grund muß man versuchen eine andere Anordnung der Achsen zu finden, um mehr als 3 Achsen darstellen zu können.

Einen Lösungsansatz stellen die Parallelen Koordinaten (Inselberg, 1985 und Wegman, 1990) dar. Inhalt dieser Arbeit soll es sein, die Theorie der Parallelen Koordinaten näher zu erläutern und den Nutzen für die interaktive Datenanalyse multivariater Datensätze darzustellen.

Weiterhin soll eine Software entwickelt werden, die die gefundenen interaktive Methoden in Parallelen Koordinaten Darstellungen verwirklicht. Anhand eines Beispiels soll eine Vorgehensweise beschrieben werden, nach der man eine Datenanalyse eines multivariaten Datensatzes durchführen kann.

Kapitel 2 soll nur einen kurzen Überblick über verschiedene Möglichkeiten der Datenvisualisierung geben, die im Verlauf der Arbeit des öfters erwähnt werden. Weiterhin soll der Übergang auf mehrdimensionale Darstellungsarten geschaffen werden.

Im dritten Kapitel soll die Technik der Parallelen Koordinaten näher beschrieben werden. Es soll versucht werden, diese neue Darstellungsweise verständlich zu machen. Vor allem sollen Zusammenhänge zu anderen Graphiken aufgezeigt werden und die Vorteile dieser Darstellungsart hervorgehoben werden. Es wird hierbei besonderer Wert auf das Erkennen von mehrdimensionalen Strukturen gelegt.

In Kapitel 4 werden kurze Überlegungen zur Interaktivität angestellt. Dabei wird einerseits versucht, den Begriff der Interaktivität zu beschreiben, und andererseits werden verschiedene Anforderungen an interaktive Software zusammengefaßt.

Ein wichtiger Punkt dieser Arbeit ist die Entwicklung der Software CASSATT. Daher werden in Kapitel 5 einige Grundlagen zur Softwareentwicklung dargestellt und erklärt, wie diese in CASSATT verwirklicht werden.

Das darauf folgende Kapitel gibt einen kurzen Überblick über den Aufbau von CASSATT und die grundlegenden interaktiven Methoden. Zusätzlich wird ein Überblick gegeben, welchen Nutzen bestimmte Methoden bei der Datenanalyse besitzen.

In Kapitel 7 werden die erweiterten Fähigkeiten der Software CASSATT erläutert. Dabei wird spezieller auf das Thema Selektion eingegangen. Ebenfalls wird anhand anderer Darstellungen erklärt, was die verschiedenen Selektionen bewirken können.

Am Ende nacheinander ausgeführter Selektionen erhält man meist eine Gruppe. Da dieser Punkt in der Datenanalyse relativ wichtig ist, wird in Kapitel 8 dargelegt, wie man Gruppen erzeugen und in CASSATT damit arbeiten kann. Ebenfalls wird darauf eingegangen, wie man die Informationen über eine Grup-

pe gestalten kann, damit vor allem die Selektion dieser Gruppe nachvollzogen werden kann.

Kapitel 9 zeigt anhand eines Beispiels die Datenanalyse mit Hilfe der Parallelen Koordinaten. Dabei wird vorgestellt, wie die vorher beschriebenen Methoden konkret angewendet und die Ergebnisse interpretiert werden können.

Im letzten Kapitel wird eine Zusammenfassung vorgenommen, sowie ein Ausblick auf mögliche weitere Entwicklungen in Bezug auf CASSATT gegeben.

Im Anhang werden die Grundlagen zur Software noch ergänzt. Es werden daher alle Menü- und Shortcut- Befehle nochmals zusammengefaßt. Auf der beigefügten CD-ROM befindet sich eine lauffähige Version von CASSATT, der komplette Source Code und eine Installationsbeschreibung unter Windows bzw. Macintosh. Ebenfalls findet man dort eine Reihe von Datensätzen, die in der Diplomarbeit verwendet wurden.

1.2 Literaturhinweise

Der Theorieteil dieser Diplomarbeit ist in mehrere Teile unterteilt. Es handelt sich hier bei um

- Parallele Koordinaten
- Interaktivität
- Softwareentwicklung

Daher möchte ich schon an dieser Stelle angeben, welche Quellen ich für diese Themen in meiner Arbeit verwenden und inwiefern diese Literatur mir helfen konnte.

Parallele Koordinaten

Zum Thema Parallele Koordinaten findet man viele Artikel bei A. Inselberg und E. Wegman. A. Inselberg stellt in seinem ersten Artikel "The Plane With Parallel Coordinates" die Technik der Parallelen Koordinaten dar. Seine nachfolgenden Artikel "Don't Panic ... Just Do It Parallel" und "Parallel Coordinates: A Tool For Visualizing Multidimensional Geometry" gehen noch stärker auf die geometrischen Aspekte, wie beispielsweise die Darstellung von Hyperebenen, ein.

E. Wegman geht in seinen Artikeln "High Dimensional Clustering Using Parallel Coordinates", "Hyperdimensional Data Analysis Using Parallel Coordinates" und "Construction Of Line Densities For Parallel Coordinate Plots" ebenfalls auf die geometrischen Zusammenhänge ein, stellt aber zusätzlich noch die Dichteschätzung für Parallele Koordinaten Darstellungen vor. Eine weiterer Artikel, "The Grand Tour In K-Dimensions", zeigt auf, wie Parallele Koordinaten für die Grand Tour geeignet sind.

Weitere Artikel, die sich daneben noch auf spezielle Software beziehen, wären der Artikel von Avidan “ParallAX – A Data Mining Tool Based On Parallel Coordinates” und der Artikel von Bassett “Ibm’s Ibm Fix”.

In den 3 Abhandlungen “Statistische Graphik” von Geßler, “Data Structures for Computational Statistics” von Klinke und “Graphisch gestützte Datenanalyse” von Schnell findet man kurze Kapitel über Parallele Koordinaten Darstellungen. In Wilkinsons “The Grammar Of Graphics” und in “Theorie und Anwendung interaktiver Statistischer Graphik” von Theus wird ebenfalls auf Parallele Koordinaten eingegangen. Jede dieser Arbeiten beschäftigt sich allerdings allgemein mit dem Thema “Statistische Graphiken”, weshalb eher andere Darstellungsmöglichkeiten angegeben und diskutiert werden.

Interaktivität

Im Wesentlichen beziehe ich mich in Kapitel 4 auf die Habilitationsschrift “Interactive Statistical Graphics: The Paradigm of Linked Views” von A. Wilhelm. Er versucht eine Definition für Interaktivität zu geben und geht dabei einerseits auf die Anforderungen an das System und andererseits auf die Anforderungen an die Graphiken ein. Letzteres beschreiben ebenfalls A. Unwin in “Requirements for interactive graphics software for exploratory data analysis” und M. Theus in “Theorie und Anwendung interaktiver Statistischer Graphik”.

Softwareentwicklung

Die Objekt orientierte Programmierung ist zwar schon relativ alt, dennoch ist das Forschungsgebiet der Objekt orientierten Softwareentwicklung erst vor einigen Jahren richtig aufgekommen. Dabei wird großer Wert auf die saubere Analyse und das Design der Software gelegt, um die Wartung einfacher zu gestalten.

Da das Gebiet des “Softwareengineering” noch relativ jung ist, existiert entsprechend wenig gute Literatur. Da die Diplomarbeit sich jedoch mehr auf den statistischen Teil beziehen soll, werde ich das Thema Softwareentwicklung nur kurz anreißen.

Im Buch von H.A. Partsch “Specification and Transformation of Programs” findet man verschiedene Qualitätskriterien für Software. Zum Thema Objekt orientierte Analyse und Design existiert das Werk von C. Larman “Applying UML and Patterns – An Introduction To Object-Oriented Analysis And Design”.

Kapitel 2

Visualisierung

Graphiken sind eine einfache Form, Informationen schnell und übersichtlich weiterzugeben. Da der Informationsfluß immer stärker zunimmt, könnten die Medien ohne Verwendung von Graphiken nur schwer die Masse der Informationen verarbeiten und verständlich präsentieren.

2.1 Graphiken in der Statistik

In der Statistik werden im Wesentlichen 2 Arten von Graphiken verwendet. Einerseits gibt es die Graphiken zur Datenanalyse und andererseits die Präsentationsgraphiken, welche Jürgen Geßler in seinem Buch “Statistische Graphik” genauer ausführt. Bei den Präsentationsgraphiken steht vor allem die Vermittlung der Information im Vordergrund, wohingegen die Graphiken zur Datenanalyse zum Auffinden von Information verwendet werden.

Ein bedeutender Unterschied dieser zwei Graphiktypen ist das Erscheinungsbild. Ist es bei Präsentationsgraphiken beispielsweise unbedingt notwendig Beschriftungen oder Skalen anzugeben, um die gesamte Information vermitteln zu können, so würde die Angabe dieser Punkte in den Graphiken zur Datenanalyse zur Informationsüberladung führen. Bei der Datenanalyse rücken vielmehr die reinen Daten in den Vordergrund. Spezielle Hilfsmittel, wie die interaktive Abfrage oder Selektion (siehe Kapitel 4), erleichtern das Auffinden verschiedener Strukturen oder Dateneigenschaften.

Da ich mich in der Diplomarbeit vor allem auf die Datenanalyse konzentrieren werde, will ich hier nicht näher auf die Präsentationsgraphiken eingehen. Ich will vielmehr einen kurzen Überblick über verschiedene Darstellungsarten geben, die man bei der Datenanalyse verwendet. Allerdings ist es wichtig zu erwähnen, daß bei der graphischen Datenanalyse die Interaktivität (siehe Kapitel 4) eine wesentliche Rolle spielt. Deshalb möchte ich an dieser Stelle die Arbeit von M. Theus erwähnen, in der sowohl die verschiedenen Graphiken als auch die dazu gehörenden interaktiven Methoden noch näher erklärt werden.

2.2 Darstellungsarten zur Datenanalyse

Um eine gute Darstellungsart auszuwählen, muß man sich vorerst darüber klar werden, was genau dargestellt werden soll. Dabei spielen nicht nur die Variableneigenschaften eine wichtige Rolle, sondern auch auch das Ziel der Datenanalyse (siehe Abb. 2.1).

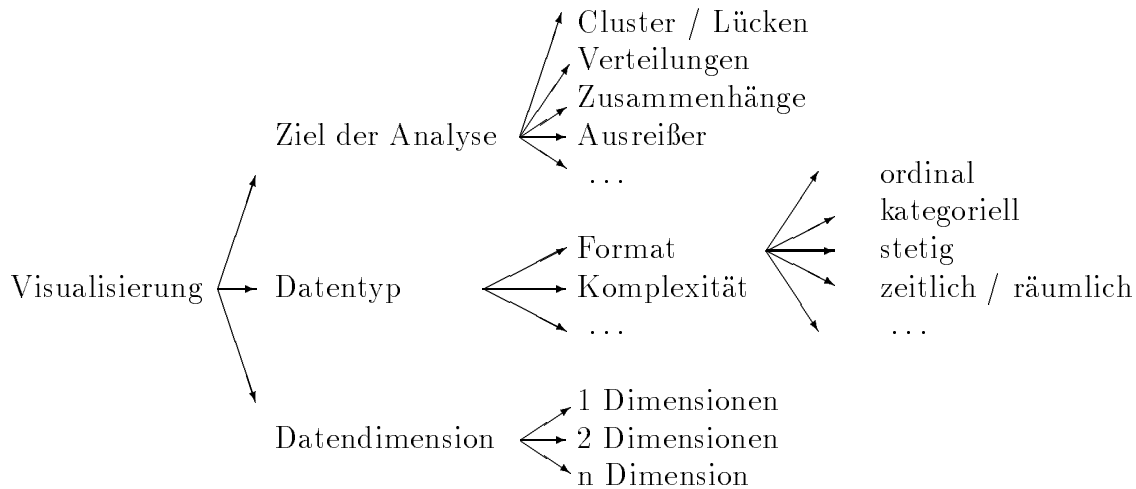


Abbildung 2.1: Überblick über wichtige Kriterien zur Darstellungsart

Vorerst ist es aber von großer Bedeutung, welche und wieviele Variablen, also Dimensionen, dargestellt werden sollen. Ebenso wichtig sind die Eigenschaften dieser Variablen wichtig. Hauptsächlich ist dabei das Datenformat dieser Variablen, wie z.B. kategoriell, ordinal oder stetig, gemeint. Dennoch sollte aber auch die Anzahl der Beobachtungen, also die Komplexität, nicht aus den Augen verloren werden. Als letzten Punkt sollte man sich überlegen, welches Ziel man mit der Darstellung verfolgt. Versucht man eine Verteilung erkennen, so wird man eine andere Graphik wählen als beim Versuch, Ausreißer zu finden.

Zusammenfassend kann man sagen, daß es sinnvoll ist, die graphische Darstellung von der Kombination des Datentyps, der Dimension und des Analysezieles abhängig zu machen.

An dieser Stelle will ich einen kurzen Überblick über die wichtigsten Möglichkeiten zur Datendarstellung geben. Allerdings beschränke ich mich hier nur auf Darstellungen, die ich später auch ansprechen werde. Die wichtigste Einschränkung betrifft das Datenformat. Es werden im Weiteren nur Graphiken mit stetigen Variablen erwähnt, da ich in den folgenden Kapiteln keine kategoriellen oder ordinalen Variablen diskutieren oder anwenden werde. Detailliertere Abhandlungen über die verschiedenen Darstellungsarten findet man bei Klinke oder Theus. Die Möglichkeit der Parallelen Koordinaten Darstellungen werde ich in diesem

Kapitel ebenfalls nicht erwähnen, da diese im Mittelpunkt des nächsten Kapitels stehen.

Eindimensionale Graphiken Eindimensionale Graphiken sind im Normalfall relativ eingängig. So existieren auch je nach Datentyp und Untersuchungsziel verschiedenste Darstellungsarten. Stetige Variablen lassen sich als Boxplot, Dotplot oder auch als Histogramm darstellen (siehe Abb. 2.2).

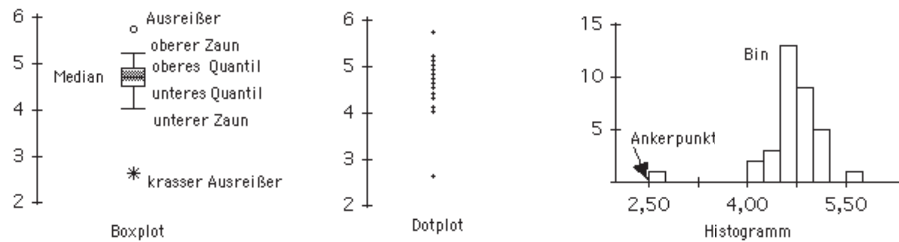
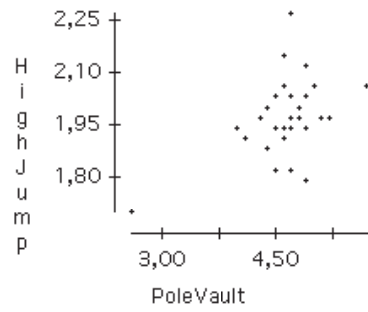
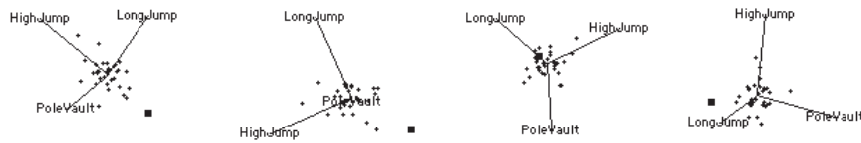


Abbildung 2.2: *Boxplot, Dotplot und Histogramm der Variablen PoleVault aus dem Datensatz aus Kapitel 9*

Jede Darstellung weist dabei ihre eigenen Vor- bzw. Nachteile auf. So erkennt man in einem Boxplot Ausreißer und kann die Verteilung eventuell erahnen. Im Dotplot hingegen kann man bei kleinen Datensätzen schnell Lücken in den Daten finden. Das Histogramm kann bei guter Wahl der Binbreite und des Ankerpunktes einen Einblick in die ungefähre Verteilung der Daten geben. Was hier nicht vergessen werden sollte ist die Tatsache, daß hier von interaktiven Methoden (siehe Kap. 4) Gebrauch gemacht werden sollte. Durch bestimmte Methoden, wie Linking oder Abfrage, kann man so schneller Ergebnisse bei der Datenanalyse erhalten.

Zweidimensionale Graphiken Beim Übergang auf zwei Datendimensionen kann man entweder eindimensionale Graphiken in Verbindung mit der Möglichkeiten des Linkings (siehe Kap. 4) nutzen, oder man verwendet zweidimensionale Graphiken (siehe Abb. 2.3). Die eingängigste Darstellungsart stellt der zweidimensionale Scatterplot dar. Weiterhin gibt es zweidimensionale Histogramme, welche aber durch die Verzerrung einen Informationsverlust mit sich bringen.

Mehrdimensionale Graphiken Beim Versuch drei Dimensionen darzustellen, erweist sich die Erweiterung des Scatterplots zum dreidimensionalen Scatterplots als eine Möglichkeit. Ebenso kann man eine Scatterplotmatrix verwenden. Bei beiden Darstellungsarten kommt man jedoch ohne interaktive Möglichkeiten, wie Highlighting, Linking und für den dreidimensionalen Scatterplot die Rotation, nicht weit.

Abbildung 2.3: *Scatterplot zur zweidimensionalen Darstellung*Abbildung 2.4: *Rotationsdiagramm zur Darstellung von drei Variablen*

Beim Versuch noch mehr Dimensionen darzustellen, kann man nur noch die Scatterplotmatrix erweitern. Andere Ansätze hierzu wären beispielsweise auch noch Chernoff-faces oder Stardiagramme. Bei allen genannten Versuchen stellt der Informationsverlust das größte Problem dar. Aus diesem Grund muß versucht werden, einen anderen Ansatz zur Darstellung von mehreren Dimensionen zu finden. Einen solchen Versuch stellen die Parallelen Koordinaten dar, die im Folgenden genauer erklärt werden sollen.

Kapitel 3

Parallele Koordinaten

Wie schon in der Einleitung erwähnt, dienen die Parallelen Koordinaten dazu, mehrere Variablen gleichzeitig darzustellen. Bei dieser Technik wird jede Variable als eine eigene Achse dargestellt. Diese Achsen werden dann parallel und equidistant nebeneinander angetragen, wobei unterschiedliche Skalierungen gewählt werden können. Von jeder Beobachtung werden nun die Werte jeder Dimension als Punkte auf den jeweiligen Achsen angetragen. Diese Punkte können später zu einer Polylinie verbunden werden (siehe Abb. 3.1).

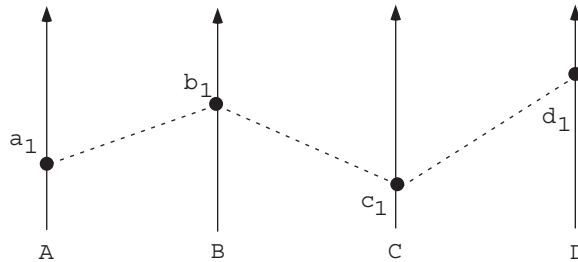


Abbildung 3.1: Darstellung des Vektors (a_1, b_1, c_1, d_1)

3.1 Zusammenhänge zum Scatterplot – Punkt \leftrightarrow Linie Dualität

Parallele Koordinaten weisen in Verbindung mit zweidimensionalen Scatterplots einige sehr schöne Eigenschaften auf. Auffällig ist zuerst die Tatsache, daß eine Punkt \leftrightarrow Linie Dualität vorliegt. Dies bedeutet einerseits, daß ein Punkt in einem Scatterplot als Linie gezeichnet wird, und andererseits, daß eine Linie in einem Scatterplot als ein eindeutiger Punkt in einer Parallelen Koordinaten Darstellung in Erscheinung tritt. Der erste Punkt ist zu erklären, da eine zweidimensionale Parallele Koordinaten Darstellung vorliegt, wird eine einzelne Beobachtung als ei-

ne Linie zwischen den 2 Achsen mit Abstand d , $d > 0$, dargestellt. Die Endpunkte dieser Linie sind die jeweiligen Werte der entsprechenden Dimension.

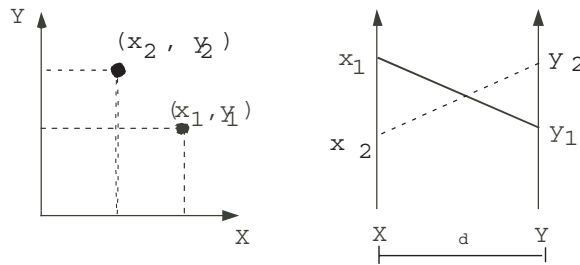


Abbildung 3.2: Darstellung eines Punktes als Linie in einer Parallelen Koordinaten Darstellung

Ausgehend von diesen Informationen läßt sich die zum Punkt $P(x_1, y_1)$ gehörende Geradengleichung folgendermaßen aufstellen:

$$P(x_1, y_1) \quad \Rightarrow \quad f(x) = mx + t \quad \text{mit} \quad m = \frac{y_1 - x_1}{d} \quad \text{und} \quad t = x_1 \quad (3.1)$$

Im anderen Fall betrachtet man eine Linie im Scatterplot. Man kann mit Hilfe der gerade gewonnenen Ergebnisse nachweisen, daß sich alle auf der Linie befindlichen Punkte im Scatterplot in der Parallelen Koordinaten Darstellung, in welcher die Punkte als Linien dargestellt werden, in einem Punkt schneiden werden.

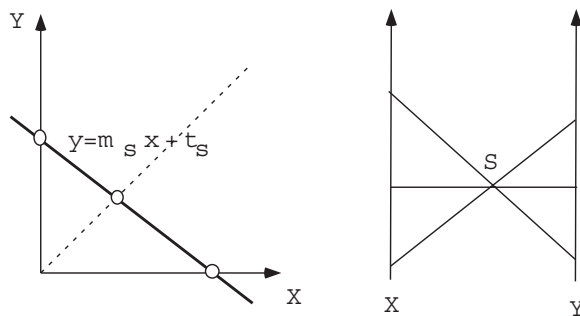


Abbildung 3.3: Darstellung einer Geraden als Punkt in einer Parallelen Koordinaten Darstellung

Gegeben sei Gerade $g(x) = m_s x + t_s$ mit beliebigen

Punkten $P_1(x_1, m_s x_1 + t_s)$ und $P_2(x_2, m_s x_2 + t_s) \in g$

⇒

Schnittpunkt $S\left(\frac{d}{(1-m_s)}, \frac{t_s}{(1-m_s)}\right)$, für $m \neq 1$ und S unabhängig von P_1 und P_2

(3.2)

Begründung:

$$P_1 \Rightarrow f_1(x) = \frac{(m_s x_1 + t_s) - x_1}{d} x + x_1 \text{ aus (3.1)}$$

$$P_2 \Rightarrow f_2(x) = \frac{(m_s x_2 + t_s) - x_2}{d} x + x_2 \text{ aus (3.1)}$$

Berechnet man nun den Schnittpunkt der beiden Geraden, also $f_1(x) = f_2(x)$

⇒

$$x = \frac{d}{(1-m_s)}$$

$$y = f_1\left(\frac{d}{(1-m_s)}\right) = \frac{t_s}{(1-m_s)}$$

Für $m=1$ erhält man keinen Schnittpunkt, da sich in diesem Fall nur waagerechte, parallele Linien in einer Parallelen Koordinaten Darstellung ergeben. Aus den Koordinatenwerten kann man erkennen, daß die Ordinate des Schnittpunktes S genau dem Koordinatenwert des Schnittpunktes T der Gerade g mit der Winkelhalbierenden im Scatterplot entspricht. Die Abszisse hängt bei festem d nur von der Steigung m_s ab. Daher kann man sich nun näher überlegen, wo sich der Schnittpunkt bei unterschiedlichen Steigungswerten m_s befindet (siehe Abb. 3.4).

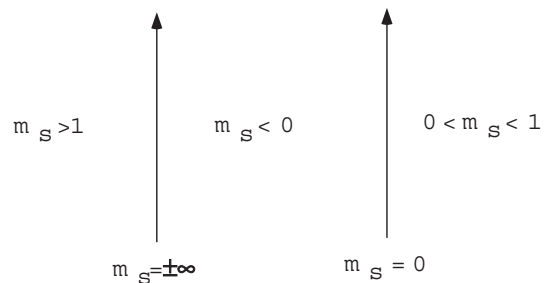


Abbildung 3.4: Schnittpunktpositionen bei unterschiedlichen Steigungen der Geraden des Scatterplots

3.2 Interpretation verschiedener zweidimensionaler Strukturen in einer Parallelen Koordinaten Darstellung

Aufgrund der Ergebnisse des vorherigen Abschnittes kann man nun auch unterschiedliche Strukturen, die in einer Parallelen Koordinaten Darstellung auftreten, leichter interpretieren.

Korrelation Erkennt man in einer Parallelen Koordinaten Darstellung beispielsweise einen Knotenpunkt, so kann man davon ausgehen, daß ein negativer linearer Zusammenhang vorliegt. Je stärker dieser Knoten zu einem Punkt verengt ist, umso höher ist diese negative Korrelation zwischen den dargestellten Variablen. Nimmt im Gegensatz dazu der Einschnitt zwischen den Achsen ab, so nimmt auch die vorhandene Korrelation ab (siehe Abb. 3.5).

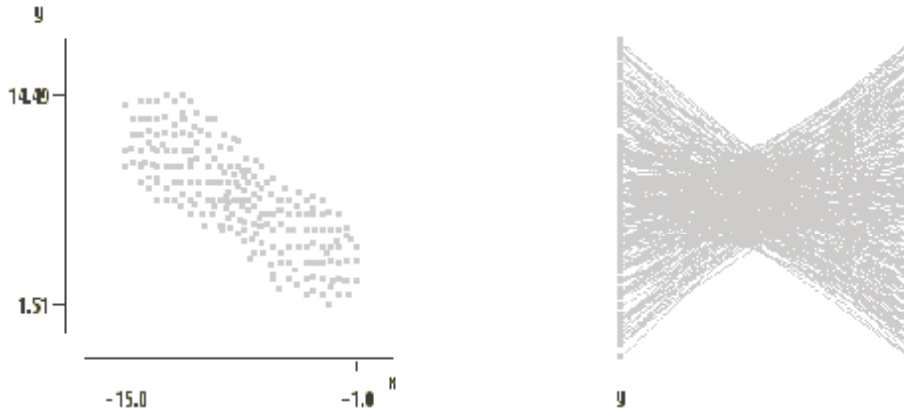


Abbildung 3.5: vorliegende negative Korrelation

Im Fall einer vorliegenden positiven Korrelation, kann man im Normalfall eher parallele Linien sehen, als einen Knotenpunkt. Dennoch ist es sehr schwer, eine vorliegende positive Korrelation zu erkennen (siehe Abb. 3.6).

Durch einen einfachen Trick kann man diese Korrelation dennoch in einer Parallelen Koordinaten Darstellung sichtbar machen. Man muß einfach eine der beiden Achsen invertieren, also umdrehen.

Die verwendete Transformation, $T(x) = \min\text{Value}(X) + \max\text{Value}(X) - x$, bewirkt, daß eine vorhandene positive Korrelation als negative Korrelation dargestellt wird. Diese kann nun ebenfalls in einer Parallelen Koordinaten Darstellung erkannt werden (siehe Abb. 3.7).

Liegt keine Korrelation vor, so erkennt man, daß ungefähr ein Drittel der Linien als parallel verlaufende Linien, ein weiteres Drittel als steigende Linien und das letzte Drittel als fallende Linien zu erkennen sind (siehe Abb. 3.8). Nach der

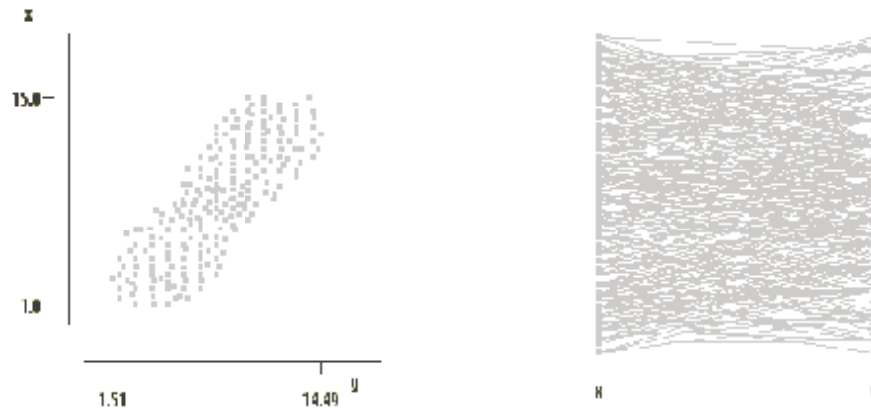


Abbildung 3.6: vorliegende positive Korrelation

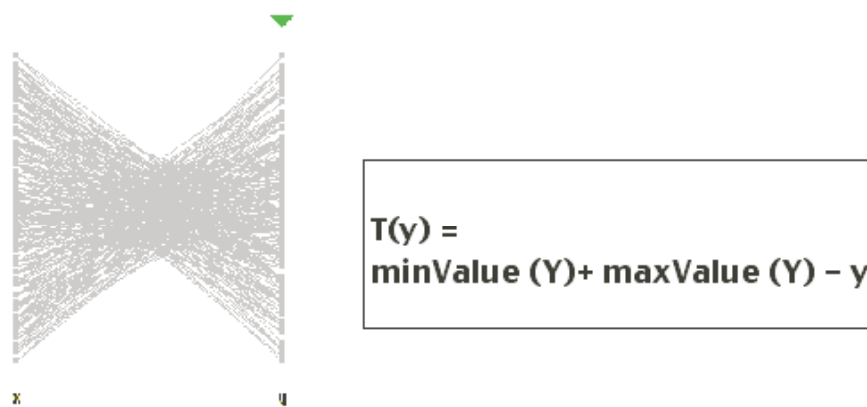


Abbildung 3.7: Transformation um negative Korrelation zu erhalten

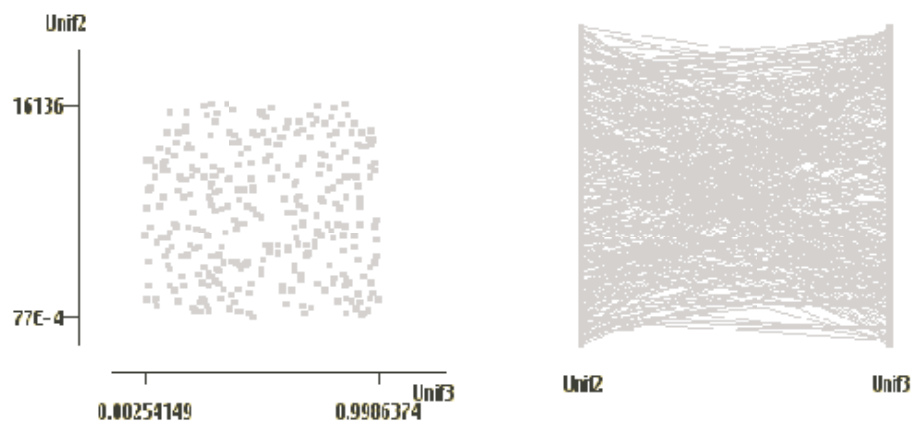


Abbildung 3.8: Erscheinungsbild zweier unkorrelierter Variablen

Inversion einer Achse erkennt man keine wesentlichen Änderungen im Erscheinungsbild der Darstellung.

Cluster Ein Cluster ist im Normalfall ein relativ unkorrelierter Punkthaufen, der sich von anderen Punktanhäufungen abhebt. Ein solches Cluster kann man in einer Parallelen Koordinaten Darstellung als zusammenhängendes Linienband erkennen. Ein derartiges Band weist zusätzlich eine unkorrelierte Struktur auf, d.h. es besitzt Linien mit unterschiedlichen Steigungen und keine einzelnen Knoten.

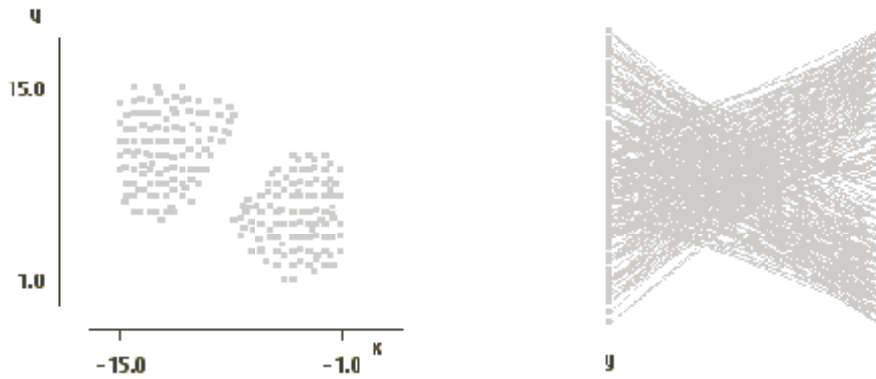


Abbildung 3.9: *Cluster*

Wie auch bei linearen Zusammenhängen kann man Cluster, die mit fallender Tendenz aufeinander folgen, relativ gut erkennen. Die einzelnen Bänder werden sich in einem dicken Knoten schneiden, aber immer noch als zusammenhängende Gruppen erkennbar bleiben (siehe Abb. 3.9).

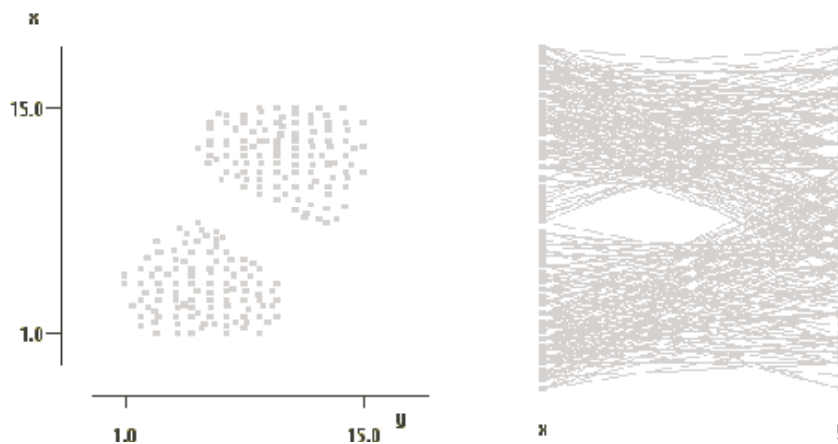


Abbildung 3.10: *Cluster mit monoton steigender Tendenz*

Anders verhält es sich mit Clustern, die mit steigender Tendenz auftreten. Analog zur Korrelation ist hier wiederum die Inversion einer Achse notwendig, um die Präsenz solcher Strukturen sichtbar zu machen.

Hier kann noch erwähnt werden, daß ein geschultes Auge die Existenz derartiger Formationen auch ohne Transformationen erkennen kann (siehe Abb. 3.10). Dennoch werden solche Methoden in der Regel zur Überprüfung der Hypothesen angewendet.

Ausreißer Ein weiterer erwähnenswerter Punkt stellen Ausreißer dar. Diese sind besonders schnell in Parallelen Koordinaten Darstellungen zu identifizieren. Obwohl ich in diesem Abschnitt vorerst nur auf den zweidimensionalen Fall eingehen, kann ich schon einmal vorweg nehmen, daß es besonders einfach ist, mehrdimensionale Ausreißer in Parallelen Koordinaten zu entdecken. Bei Ausreißern einzelner Dimensionen findet man die entsprechenden Werte entweder am oberen oder unterem Rand des zugehörigen Dotplots. Besitzt ein einzelner Fall extreme Werte in zwei Dimensionen, so erkennt man dies entweder durch eine Linie, die sich am oberen oder unteren Ende der Parallelen Koordinaten Darstellungen befindet, oder durch eine Linie, die quer von oben nach unten bzw. entgegengesetzt verläuft. Mit Hilfe der vorgeschlagenen Inversion läßt sich der letzte Fall wiederum in den ersten Fall transformieren.

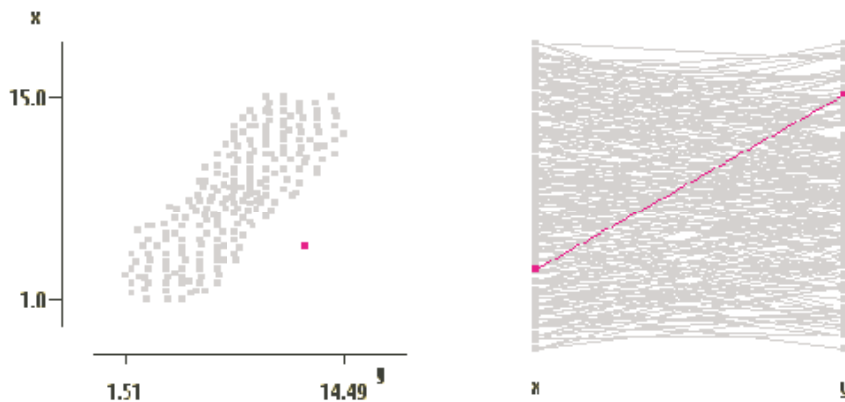


Abbildung 3.11: *Zweidimensionaler Ausreißer*

Ausreißer, die keine extremen Werte besitzen, muß man auf andere Weise auffinden. Als sehr hilfreich erweist es sich, wenn man einen Haupttrend der Linien ausmachen kann. Gibt es einzelne Fälle, deren Linienzüge außerhalb dieser Hauptlinien liegen, kann man behaupten, daß es sich um Ausreißer handelt. Zusätzlich sollte man noch eine der Achsen invertieren, da man bei einer solchen Invertierung auch noch weitere Fälle finden kann, die man als Ausreißer bezeichnen könnte. Wie auch schon bei anderen Strukturen, kann man diese Invertierung

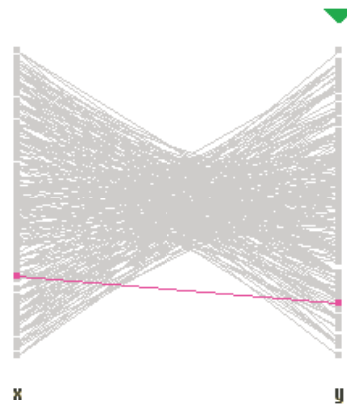


Abbildung 3.12: *Ausreißer in einer transformierten Parallelen Koordinaten Darstellung*

auch als Überprüfung anwenden. Fällt immer noch der schon vorher gefundene Fall als Sonderfall auf, so war die vorherige Vermutung richtig.

3.3 Mehrdimensionale Betrachtung

Die Ergebnisse aus dem vorherigen Abschnitt kann man auf höherdimensionale Zusammenhänge erweitern.

Wie schon ausführlich erklärt wurde, wird ein mehrdimensionaler Punkt in einer Parallelen Koordinaten Darstellung als eine Polylinie dargestellt. Interessanter wird es allerdings, wenn man sich das Erscheinungsbild anderer geometrischer Formen überlegt. Relativ wichtig erscheint für die Interpretation von Korrelationen eine n -dimensionale Linie.

Eine solche n -dimensionale Linie kann als Serie von $(n - 1)$ linearen Gleichungen der Form

$$x_i = m_i x_{(i-1)} + b_i, \text{ mit } i = 2 \dots n,$$

dargestellt werden.

Aus diesem Grund läßt sich eine n -dimensionale Linie in einer Parallelen Koordinaten Darstellung erkennen, wenn man die Ergebnisse aus dem vorherigen Abschnitt auf mehrere Dimensionen überträgt. Man erkennt demnach immer zwischen zwei benachbarten Achsen zweidimensionale lineare Strukturen, also Schnittpunkte bzw. Linien, die sich außerhalb der Achsen schneiden (siehe Abb. 3.13).

Man kann erkennen, daß es bezüglich der linearen Zusammenhänge analog zur Punkt \leftrightarrow Linie Dualität im zweidimensionalen Raum eine Linie \rightarrow Punkte Beziehung gibt.

Ebenfalls interessant werden noch andere geometrische Strukturen, wie z.B. Hyperebenen, die bei Wegman abgehandelt werden. Man muß allerdings im Bezug

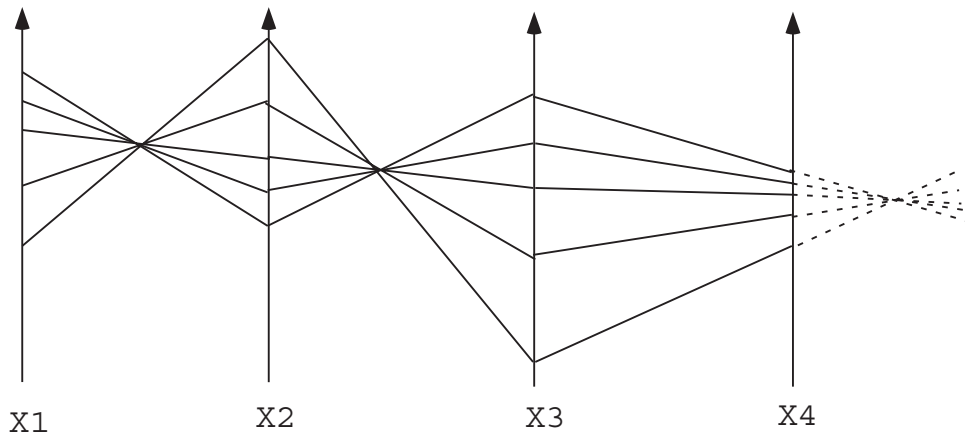


Abbildung 3.13: Darstellung einer vierdimensionalen Geraden

auf diese Strukturen erwähnen, daß dabei die menschliche Vorstellungskraft an ihre Grenzen kommt, weshalb ich auch nicht näher darauf eingehen möchte.

Korrelation Bei der Suche nach mehrdimensionalen Korrelationen befindet man sich, wie auch im zweidimensionalen Fall, in der Situation, daß man das Verhalten von linearen Strukturen schon kennt. Aus diesem Grund kann man die Ergebnisse des vorherigen Abschnittes direkt anwenden und mehrdimensionale Korrelationen als einzelne zweidimensionale Korrelation in einer Parallelen Koordinaten Darstellung erkennen. So weiß man, daß sich bei einem negativen Zusammenhang ein Knoten ergeben muß und im Gegensatz dazu bei einem positiven Zusammenhang parallele Linien zeigen. Letzteres kann man wieder durch die Inversion einer benachbarten Achse als Knoten erkennbar machen. Auch hier weisen enge Knoten auf starke Zusammenhänge und lockere Knoten auf weniger starke Zusammenhänge hin.

Cluster Auch bei den Clustern helfen die Ergebnisse aus den zweidimensionalen Untersuchungen weiter. Cluster erkennt man auch im mehrdimensionalen Raum als breite unkorrelierte Bänder. Diese kann man je nach Zusammenhang als Knoten, als parallel liegende oder getrennte Bänder erkennen. Als ein sehr hilfreiches Instrument erweist sich in einer Parallelen Koordinaten Darstellung das Einfärben solcher Untergruppen mit Hilfe der Selektion, da man über mehrere Variablenachsen hinweg eine derartige Gruppe verfolgen will. Die Parallele Koordinaten Darstellung eignet sich jedoch nicht, um reine dreidimensionale Cluster zu finden. Man kann diese Darstellungsart lediglich dazu verwenden, ein schon gefundenes Cluster einzufärben und dessen Eigenschaften zu überprüfen.

Ausreißer Noch weit interessanter erscheint mir aber das Vorhandensein von Ausreißern. Wie schon im vorherigen Abschnitt angedeutet, kann man in einer

Parallelen Koordinaten Darstellung äußerst schnell und einfach Ausreißer aufspüren. Besonders schnell kann man Ausreißer sehen, die in einzelnen Dimensionen Extremwerte aufweisen (siehe Abb. 3.14).

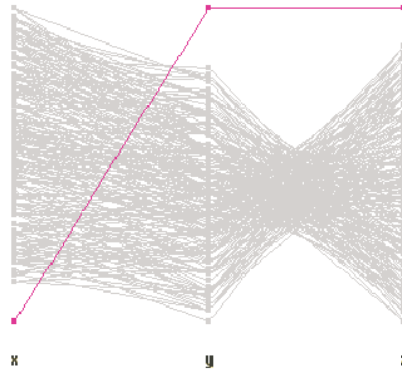


Abbildung 3.14: *Ausreißer in 3 Dimensionen*

Ebenfalls als mehrdimensionale Ausreißer zählen Individuen, deren Linienabschnitte einen anderen Verlauf besitzen als die Hauptgruppe. Vor allem kann man hier verschiedene interaktive Methoden, wie Winkelselektion oder die Inversion der Achsen (siehe Kapitel 6 bzw. 7), gut zum Einsatz bringen, um eine Darstellung zu erhalten, in der man einen solchen Ausreißer erkennen kann (siehe Abb. 3.15).

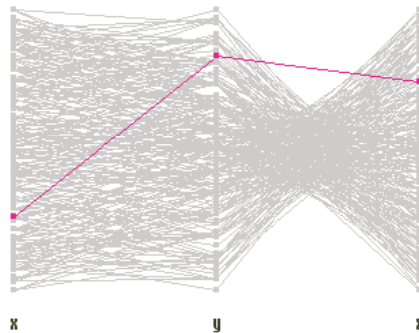


Abbildung 3.15: *mehrdimensionaler Ausreißer in einer Parallelen Koordinaten Darstellung*

Wie bei den Clustern muß hier erwähnt werden, daß nur das Auffinden von Ausreißern gelingt, die auch schon in niedrigeren Dimensionen Ausreißer darstellen. Punkte, die nur im mehrdimensionalen Fall als Ausreißer zu erkennen sind, kann man in der Parallelen Koordinaten Darstellung nur überprüfen.

3.4 Ähnlicher Ansatz zur Darstellung multivariater Datensätze

3.4.1 Andrews Plot

Es gibt eine weitere Darstellung für multivariate Datensätze, die mit den Parallelen Koordinaten sehr verwandt zu sein scheint. Bei dieser Darstellungsart wird für jede n -dimensionale Beobachtung (x_1, x_2, \dots, x_n) eine eigene Funktion

$$f(t) = \frac{x_1}{\sqrt{2}} + x_2 \sin t + x_3 \cos t + x_4 \sin 2t + x_5 \cos 2t + \dots$$

gezeichnet, wobei $-\pi \leq t \leq \pi$. Somit entspricht also jede Beobachtung einer Linie im Andrews Plot. Diese Linien haben durch die Sinusfunktionen eher welliges Verhalten. Wie auch in einer Parallelen Koordinaten Darstellung besitzen ähnliche Fälle auch ähnliche Strukturen, weshalb man auch hier gut Cluster erkennen kann (siehe Abb. 3.16 und 3.17).

Hier muß noch erwähnt werden, daß die Andrews Plots nicht die gewünschte Gleichbehandlung der einzelnen Variablen erfüllen. Einerseits ist diese Darstellung gegen unterschiedliche Wertebereiche der dargestellten Variablen empfindlich und andererseits spielt die Reihenfolge der Variablen eine wesentlich größere Rolle als bei den Parallelen Koordinaten Darstellungen. Die Variablen, die zuerst in die Funktion eingehen, bestimmen im Wesentlichen den Verlauf dieser Kurven.

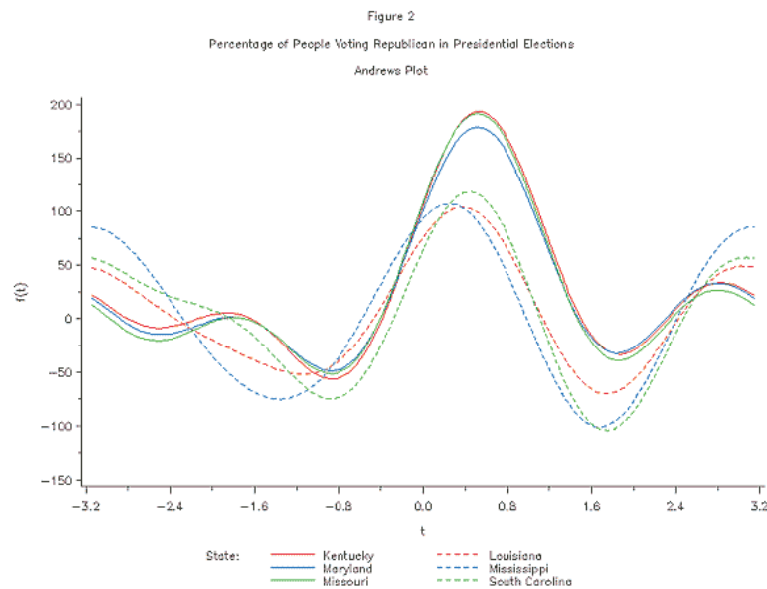
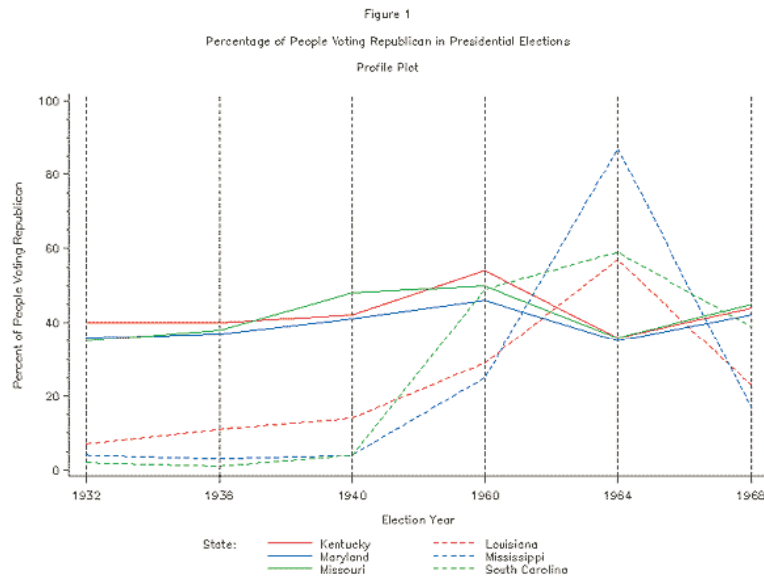


Abbildung 3.16: Andrews Kurve der Datentabelle 3.1

In den Abbildungen 3.16 und 3.17 werden, einerseits als Andrews Plot und andererseits als Parallele Koordinaten Darstellung, die Daten der Präsidenten-

Abbildung 3.17: *Parallele Koordinaten Darstellung der Datentabelle 3.1*

State	1932	1936	1940	1960	1964	1968
Missouri	35	38	48	50	36	45
Maryland	36	37	41	46	35	42
Kentucky	40	40	42	54	36	44
Louisiana	7	11	14	29	57	23
Mississippi	4	3	4	25	87	17
South Carolina	2	1	4	49	59	39

Tabelle 3.1: *Prozentualer Anteil der Wähler der Republikaner bei verschiedenen Präsidentenwahlen*

wahlen aus verschiedenen Jahren dargestellt (vgl. Tabelle 3.1, Schwenke & Fergen).

Man erkennt in beiden Darstellungen, daß verschiedene Untergruppen zu sehen sind. In der Parallelen Koordinaten Darstellung kann man jedoch noch weit mehr erkennen. Beispielsweise kann man behaupten, daß das Jahr 1964 ein besonderes Jahr war, da eine vollkommen andere Anordnung der einzelnen Fälle vorliegt. Was man zusätzlich bemerken sollte ist die Tatsache, daß der Andrews Plot zwar in dieser Form die gleichen Untergruppen liefert, aber bei einer anderen Variablenreihenfolge vielleicht zu anderen Schlüssen gekommen wäre. Ebenfalls besitzen hier die Variablen zufällig schon den selben Wertebereich, da es sich um Prozentwerte handelt.

Zum Abschluß möchte ich daher bemerken, daß der Andrews Plot zwar im Bereich der multivariaten Darstellungen angewendet wird, jedoch im Vergleich zu den Parallelen Koordinaten wesentlich weniger Informationen vermittelt.

Kapitel 4

Interaktivität

4.1 Der Begriff Interaktivität

Der Begriff der “Interaktivität” ist vom Begriff “Interaktion” abgeleitet. Dieser Begriff stammt wiederum vom lateinischen Wörtern “inter - actio”, was soviel wie “miteinander Handeln” bedeutet, wird aber oft mit “Wechselwirkung” übersetzt.

Das Psychologie - Fachgebärdenlexikon liefert je nach Themengebiet verschiedene Definitionen für den Begriff “Interaktion”.

So verstehe man in der Biologie unter Interaktion die wechselseitige Beeinflussung verschiedener Teilsysteme. In der Statistik hingegen bedeute Interaktion den gemeinsamen Effekt, den zwei oder mehrere unabhängige Variablen über ihre Einzeleffekte hinaus auf eine abhängige Variable haben. Weiterhin bezeichne man in der Sozialpsychologie die durch Kommunikation vermittelte gegenseitige Beeinflussung von Personen im Hinblick auf ihr Verhalten, ihr Handeln oder ihre Einstellungen als Interaktion.

Allerdings wird hier kein Hinweis auf den Begriff Interaktion im Bezug auf Computersysteme gegeben.

Robert Schmidtner gibt im Osinet im Bereich Psychologie eine genauere Erklärung zu den Begriffen “Interaktion” und “Interaktivität”.

So verstehe man in den Sozialwissenschaften unter dem Begriff Interaktion die gegenseitige Beeinflussung, die wechselseitige Abhängigkeit, und das “Miteinander - in - Verbindung - treten” zwischen Individuen und sozialen Gebilden.

Der Begriff der “Interaktivität” sei als abgeleiteter Begriff zu verstehen. Er beschreibe in Bezug auf Computersysteme die Eigenschaften von Software, welche dem Benutzer eine Reihe von Eingriffs- und Steuermöglichkeiten eröffnen. Als konstitutiv für die Interaktivität eines Computerprogrammes werden die aktive Rolle des Benutzers und die Freiheitsgrade der Auswahl betrachtet. Im Idealfall komme es zu einer wechselnden Dialog - Initiative von Mensch und Maschine.

In Bezug auf statistische Software entspricht das Vorhandensein von interaktiven Methoden der Möglichkeit, mit den Daten interagieren und die dadurch

verursachten Veränderungen sofort wahrnehmen zu können.

Welche Anforderungen aber an eine interaktive statistische Software gestellt werden, soll nun im Weiteren diskutiert werden.

4.2 Anforderungen an interaktive Software

Obwohl der Begriff der Interaktivität in Bezug auf statistische Software äußerst häufig in der Literatur vorkommt, findet man erstaunlicherweise nur sehr selten Kommentare darüber, was “Interaktivität” in dieser Beziehung bedeutet.

M. Theus versucht in seiner Dissertation eine Zusammenfassung der wichtigsten interaktiven Methoden zu geben. Er beschreibt in seiner Arbeit verschiedene Grundelemente interaktiver Graphiken. Dazu gehören verschiedene Methoden

- Highlight – die hervorgehobene Darstellung von Daten in Graphiken und Tabellen.
- Linking – die Methode, Veränderungen, die in irgendeiner Form auf den Daten durchgeführt werden, auch in allen anderen Darstellungen, die von diesen Daten abhängen, durchzuführen.
- Abfrage – die Möglichkeit, in statistischen Graphiken Werte oder Gruppen von Werten bzw. Statistiken abzufragen.
- Warnungen – Informationen über verschiedene Dinge, die ein Benutzer übersehen kann und die deshalb vom System bereit gestellt werden müssen.

Daneben gibt er auch noch verschiedene Anforderungen, die an Hard- und Software gestellt werden müssen, in seiner Arbeit an. Allerdings werden diese Anforderungen nicht im Rahmen der Interaktivität aufgelistet, weshalb ich diese hier nicht zusätzlich aufführen will.

A. Unwin gibt beispielsweise eine etwas abgeänderte Zusammenfassung von Anforderungen, die eine interaktive statistische Software erfüllen muß. Er beschreibt verschiedene Methoden und gibt ebenfalls Beispiele an, die deren Bedeutung unterstreichen. Die erwähnten Methoden sind:

- direkte Abfrage
- Zoomen
- Variation der Darstellung, z.B. Reskalierung
- Verschiedene Ansichten
- Gruppierungen
- Linking

- Selektion mit Linking

Während A. Unwin und M. Theus hauptsächlich auf die interaktiven Methoden in Graphiken hinweisen, versucht A. Wilhelm eine feste Definition für die Interaktivität zu finden.

Dabei kommt er auch zu dem Entschluß, der Interaktivität verschiedene Schichten zuzuordnen, die jeweils durch die vorliegende graphische Schicht bestimmt werden. Ebenso stellt er nicht nur an die Graphiken sondern an das komplette System Design bestimmte Anforderungen, die für die Interaktivität wichtig sind. Ich möchte aber dennoch nur kurz die wichtigsten graphischen Befehle zusammenstellen, auf die ein System schnell reagieren können muß.

- Skalierung – Möglichkeit, flexibel die Skala zu ändern.
- Abfrage – Möglichkeit, Information bereit zu halten, ohne daß die Graphik überladen wirkt.
- Selektion – Möglichkeit, Untergruppen auszuwählen und genauer zu fokussieren, um eventuelle Strukturen in den Daten zu finden.
- Projektionsansichten – Möglichkeit mit Hilfe dimensionsreduzierender Techniken multidimensionale Strukturen sichtbar zu machen.
- Linking – Möglichkeit, Selektionen durch interne Verbindungen der Graphiken in allen Darstellungen zu zeigen.

Dabei ist mit dem Begriff der Interaktivität nicht nur die Bearbeitung der Daten gemeint, sondern vorallem auch die schnelle Reaktion des Systems auf die Benutzeranfragen.

Welche interaktiven Möglichkeiten in den Parallelen Koordinaten Darstellungen von Bedeutung sind, wird in den Kapiteln 6, 7 und 8 näher erläutert.

Kapitel 5

Entwicklung von CASSATT

Mit den Parallelen Koordinaten steht eine Darstellungsart zur Verfügung, mit deren Hilfe mehrdimensionale Datensätze in nur einer Darstellung gezeichnet werden können. Es erscheint sinnvoll, zur Erstellung von Parallelen Koordinaten Darstellungen ein Softwarepaket zu verwenden. Leider sind entweder bestimmte Methoden, die nur für Parallele Koordinaten sinnvoll sind, nur spärlich implementiert, oder der Zugang zu den Softwarepaketen ist relativ schwer. Weiterhin ist es im Normalfall nicht gegeben, daß man eigene Ideen mit Hilfe dieser Pakete in die Realität umsetzen kann.

Aus diesen Gründen wurde der Entschluß gefaßt, eine Software zu entwickeln, die besonderen Wert auf Parallele Koordinaten Darstellungen und deren Bearbeitung legt. Es soll dabei darauf geachtet werden, daß neue Ideen umgesetzt und bei Bedarf geändert werden können.

5.1 Grundlagen zur Softwareentwicklung

Bevor eine Software implementiert werden kann, müssen einige Vorüberlegungen angestellt werden. Hier werde ich nun einerseits auf die Anforderungen eingehen, die eine gute Software erfüllen muß, und andererseits auf die Spezifikation, also auf die spezielle Leistungsbeschreibung, der zu entwickelnden Software.

Zu den Anforderungen gehören hauptsächlich die Standardanforderungen an gute Software. Offensichtliche Merkmale sind hier die Zuverlässigkeit, die Benutzerfreundlichkeit und die Korrektheit. Noch wichtigere Punkte – vor allem in Bezug auf die Wartung und Pflege – stellen die Punkte Portabilität, leichte Erweiterbarkeit und leichte Änderbarkeit dar. Die Portabilität kann durch Verwendung einer plattformunabhängigen Programmiersprache erreicht werden. Durch die Benutzung einer objektorientierten Programmiersprache können Änderungen und Erweiterungen ohne große Eingriffe in den schon vorhandenen Programmcode durchgeführt werden. Allerdings muß hier erwähnt werden, daß dies zwar für die Entwickler und Programmierer dieser Software zu bewerkstelligen ist, andere

Programmierer zur Erweiterung jedoch eine gute Dokumentation der vorhandenen Klassenbibliotheken benötigen, um die vorhandenen Methoden und Klassen auch verwenden zu können.

Um nun näher die Leistungsanforderungen zu spezifizieren, muß man sich klar werden, was die Hauptanwendung dieser Software sein soll. Da es sich um eine graphische Software handelt, die auch Interaktivität (siehe Kapitel 4) aufweisen soll, sind im Wesentlichen drei Gesichtspunkte zu betrachten. Neben dem Systemdesign muß einerseits eine benutzerfreundliche und intuitive Oberfläche gestaltet werden und andererseits eine Klassenstruktur mit den dazugehörigen Methoden und Attributen entworfen werden.

5.2 Anforderungen an das Systemdesign

Da die Software ein interaktives Programm werden soll, ist es von Bedeutung schon von Anfang an verschiedene Anforderungen an das Systemdesign zu stellen. Wilhelm hat die Anforderungen an interaktive Statistik in einzelnen Punkten zusammengefaßt:

- Unabhängigkeit
- Konsistenz
- Flexibel
- Erweiterbar
- Eignung für Benutzer-Interaktion
- Abdeckung vorhandener Implementationen
- unabhängig von einer speziellen Implementierung

Vor allem Punkte wie Unabhängigkeit, Erweiterbarkeit und Interaktivität, müssen schon in der Entwicklungsphase mit eingeplant werden. Um zusätzlich ein Programm zu haben, welches plattformunabhängig ist, wird die gesamte Implementierung in Java gehalten. Desweiteren ist bei Verwendung dieser Programmiersprache von Vorteil, daß man so ein einfaches GUI (Graphische Benutzer Schnittstelle) zur Verfügung hat, um die Oberflächengestaltung durchzuführen.

5.3 Oberflächengestaltung und Klassenstruktur

Es ist verständlich, daß die Oberflächengestaltung einer Software eine besonders wichtige Rolle darstellt. Die Grundlagen dazu findet zusammengefaßt man bei R. Baecker (Baecker, 1995).

Da CASSATT eine reine Forschungssoftware ist, wurde die Oberfläche einfach und benutzerfreundlich erstellt. Wie die Oberfläche genau aussieht, wird im nächsten Kapitel beschrieben.

Einen weiteren wesentlichen Punkt der Softwareentwicklung stellt die Klassenstruktur dar. Dabei will ich nur kurz erwähnen, daß CASSATT so gestaltet ist, daß Erweiterungen und neue Ideen ohne großen Aufwand durchgeführt werden können. Es würde allerdings den Rahmen dieser Arbeit sprengen, wenn ich näher auf alle Klassen und die zugehörigen Methoden und Attribute eingehen würde.

Kapitel 6

CASSATT

Wie im vorherigen Kapitel erklärt wurde, ist CASSATT entwickelt worden, um Datenanalyse mit Hilfe der Parallelen Koordinaten betreiben zu können. Dabei ist CASSATT zu einer Software herangewachsen, die weit mehr, als ursprünglich geplant, zu leisten vermag.

6.1 Grundlagen

Startet man nun das Programm, so wird man aufgefordert, eine Datei, den Datensatz, auszuwählen. Kurz darauf wird der Datensatz eingelesen und das Hauptfenster erscheint im Zentrum des Bildschirms (siehe Abb. 6.1).

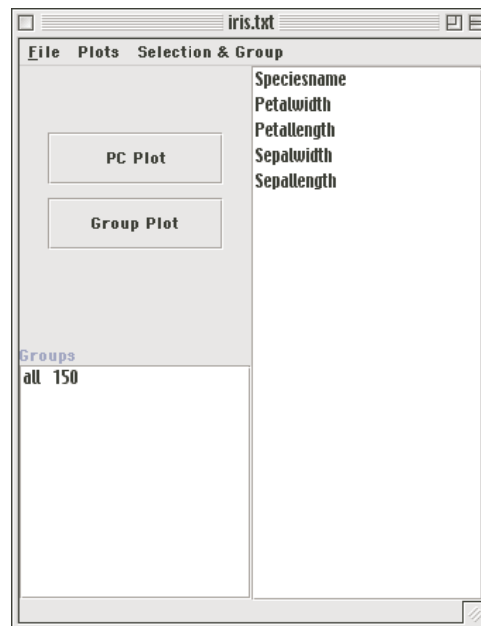


Abbildung 6.1: Hauptfenster mit Variablenliste (rechts) und Gruppenliste (links)

Ausgehend von diesem Hauptfenster lassen sich alle wesentlichen Aktionen, wie z.B. das Öffnen und das Schließen verschiedener Datensätze oder die Erzeugung der Darstellungen, über das Menü steuern. Es enthält einerseits eine Liste mit den Variablennamen und andererseits eine Liste mit den Namen und der Individuenanzahl der vorhandenen Gruppen.

Eine standardmäßig erzeugte Gruppe bei CASSATT ist die Menge aller vorkommenden Individuen. Man erkennt daher in der Gruppenliste eine Gruppe namens “all” und deren Individuenzahl, in diesem Fall die Zahl aller Beobachtungen des Datensatzes. Was genau aber eine Gruppe ausmacht, und wie man eine solche bei CASSATT erzeugen kann, wird in Kapitel 8 näher erläutert.

Als weiteres wichtiges Fenster erscheint das Toolfenster am oberen Bildschirmrand (siehe Abb. 6.2). Dieses Fenster enthält 4 Auswahlknöpfe für den jeweils gewünschten Selektionsmodus und einen Knopf zum Erzeugen von Gruppen. Auf die Wirkungsweise dieser Knöpfe werde ich aber später noch genauer eingehen (siehe Kapitel 7 und Kapitel 8).



Abbildung 6.2: *Toolfenster*

Eher unauffällig am linken Bildschirmrand befindet sich ein kleines Fenster mit einer Liste, in der zu Beginn nur der Name des momentanen Datensatzes enthalten ist (siehe Abb. 6.3).

Bei der Bearbeitung eines einzelnen Datensatzes spielt dieses Fenster, das Auswahlfenster, keine große Rolle. Will man jedoch mehrere Datensätzen gleichzeitig bearbeiten, so erweist es sich doch als sinnvoll, wie die folgende Erklärung zeigt.



Abbildung 6.3: *Auswahlfenster*

Für jeden Datensatz befindet sich ein Hauptfenster und eventuell zugehörige Graphiken auf dem Bildschirm. Um zwischen diesen Datensätzen wechseln zu können, muß man nur in dem Auswahlfenster den entsprechenden Datensatz anklicken, wodurch das zugehörige Hauptfenster mit allen entsprechenden Graphiken in den Vordergrund gesetzt wird.

Damit wären gegenwärtig alle wichtigen Fenster erklärt, so daß näher auf den praktischen Teil eingegangen werden kann.

Zu den grundlegenden Fähigkeiten, die CASSATT aufweist, gehört das Erstellen von verschiedenen Darstellungen. Dabei kann man neben den Parallelen Koordinaten Darstellungen auch Dotplots, Boxplots oder Scatterplots erstellen. Diese erzeugt man, indem man im Hauptfenster die gewünschten Variablen auswählt und im Menüpunkt "Plots" die entsprechende Darstellungsart anklickt. Zusätzlich existiert für die Parallelen Koordinaten Darstellungen neben dem Menübefehl die Möglichkeit, über den Button "PC Plot" im Hauptfenster diese Darstellungsart zu erstellen.

In CASSATT können in einer Parallelen Koordinaten Darstellung auch eine bzw. mehrere Gruppen gleichzeitig gezeichnet werden. Diese besondere Form der Parallelen Koordinaten Darstellung wird als Parallele Gruppen Darstellung bezeichnet und kann durch zusätzliche Selektion der gewünschten Gruppen im Hauptfenster und Drücken des "Group Plot Buttons" angefertigt werden.

Da die einzelnen Darstellungen in eigenen Fenstern erstellt wurden, kann man mit diesen Graphiken alles machen, was mit normalen Standardfenstern des jeweiligen Betriebssystems möglich ist. So können die einzelnen Darstellungen unabhängig voneinander vergrößert, verkleinert, verschoben oder auch minimiert werden.

Eine große Bedeutung spielt in CASSATT die Interaktivität (siehe Kapitel 4). Dies bedeutet, daß in jeder Darstellung die grundlegenden interaktiven Möglichkeiten, wie beispielsweise Selektion oder Abfragen, vorhanden sind. Daneben ist ebenso das Linking implementiert, so daß alle Darstellungen intern miteinander in Verbindung stehen und sich jede Selektionsänderung auf alle Graphiken auswirkt.

In den Parallelen Koordinaten Darstellungen sind neben den interaktiven Standardmethoden zusätzlich noch viele neue interaktive Ideen verwirklicht worden. Welche Methoden das genau sind und für welche Zwecke diese besonders gut geeignet sind, wird in den folgenden Abschnitten näher erklärt.

6.2 Interaktive Möglichkeiten in Parallelen Koordinaten Darstellungen

Ein wichtiger Bestandteil der Interaktivität ist das einfache und schnelle Abändern einer Darstellung. Um nun aber derartige Änderungen zügig durchzuführen, ist es ratsam mit Mausbefehlen oder Shortcuts, also Tastenkombinationen, zu arbeiten. In CASSATT sind daher nahezu alle interaktiven Möglichkeiten (interactive Modifiers), die in einer Parallelen Koordinaten Darstellung ausgeführt werden können, als Mausbefehle oder Shortcuts implementiert.

Grundlegende Methoden sind zusätzlich noch im Menü zu finden, was zwar

Redundanz bedeutet, aber dennoch oft hilfreich sein kann. Da sich im Menü damit hauptsächlich redundante Methoden befinden, gibt es die Möglichkeit, dieses Menü ein- bzw. ausblenden. Dies vermeidet auch zusätzliche Informationsüberladung in der Darstellung.

Die implementierten Shortcuts und Befehle sind am Ende dieser Arbeit nochmals als Liste zusammengefaßt und befinden sich zusätzlich als Hilfe im Menü jeder Parallelen Koordinaten Darstellung.

6.2.1 Darstellungsarten

Standardmäßig erscheint eine Parallele Koordinaten Darstellung als Paralleler Dotplot, was bedeutet, daß Dotplots der einzelnen Variablen parallel nebeneinander gezeichnet werden, und noch nicht, wie in Kapitel 3 erwähnt, die zusammengehörigen Punkte miteinander verbunden werden (siehe Abb. 6.4). Dies ist vor allem deshalb von Vorteil, da man beim Betrachten durch die Menge von Linien nicht verwirrt wird. Dadurch hat man anfangs die Möglichkeit, Ausreißer, Lücken oder besondere Strukturen innerhalb einzelner Variablen auszumachen.

Zusätzlich besteht die Möglichkeit, die Parallele Koordinaten Darstellung so abzuändern, daß man einen Parallelen Boxplot erhält (siehe Abb. 6.4). Man kann dies durch entsprechende Shortcuts erreichen. Ebenso lassen sich Boxplots und Dotplots gleichzeitig überlagern. Dadurch kann man sowohl die Vorteile des Dotplots als auch die des Boxplots nutzen.

Um nun die entsprechenden Linien hinzuzufügen, muß man wiederum nur einen Shortcut anwenden. Dies bewirkt folglich, daß durch die einzelnen zusammengehörigen Punkte jeweils eine Polylinie gezeichnet wird (siehe Abb. 6.5). Diese Linien können sehr aussagekräftig in Bezug auf Datenstrukturen oder Zusammenhänge zwischen Variablen sein. Welche Datenstrukturen man mit Hilfe dieser Linien entdecken kann, wurde bereits ausführlich in Kapitel 3 erklärt.

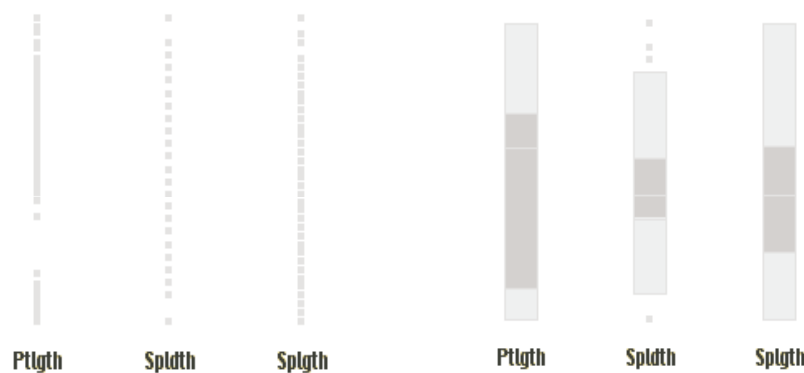


Abbildung 6.4: Darstellungsarten der Parallelen Koordinaten

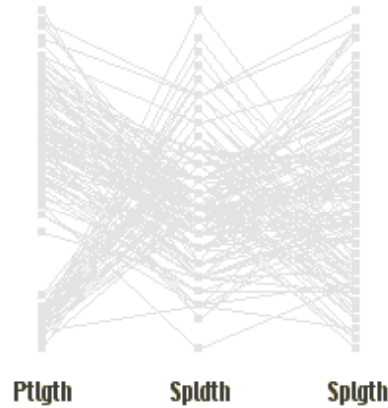


Abbildung 6.5: *Parallele Koordinaten in der Liniendarstellung*

Um einzelne Individuen besser erkennen zu können, ist es oft gewünscht, die Punkte größer darstellen zu können. Man muß aber einräumen, daß zu große Punkte auch häufig den Überblick über die Linien verfälschen können. Daher besteht auch die Möglichkeit die Punktgröße zu variieren. Sind Linien eingeblendet, so können die Punkte auch ganz ausgeblendet werden.

Eine andere Größe, die oft geändert werden muß ist der Achsenabstand. Diesen kann man beliebig mit Hilfe eines Shortcuts vergrößern bzw. verkleinern.

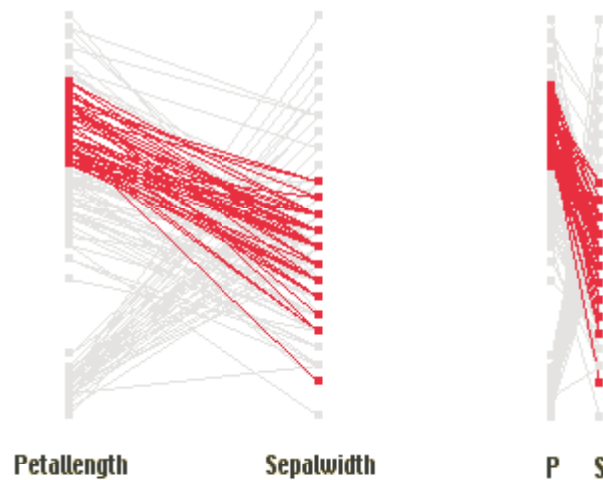


Abbildung 6.6: *Parallele Koordinaten mit unterschiedlichen Achsenabständen*

Um einen gesamten Überblick über die Daten zu erhalten, werden die Achsen aller ausgewählten Variablen standardmäßig in der Graphik so eng nebeneinander gesetzt, daß alle Variablen sichtbar sind. Dies führt bei einer großen Anzahl

von Variablen folglich zu einem sehr geringen Achsenabstand, was aber bei der standardmäßigen Darstellung mit Dotplots für die Individuendarstellungen keine Einschränkungen mit sich bringt. Will man aber eine Darstellung mit Linien erzeugen, so sollte man den Achsenabstand vergrößern, um die Variablen und die Linien deutlich erkennen zu können. Hier sollte allerdings erwähnt werden, daß es bei der Verringerung des Achsenabstands unter Umständen zu einer Überlagerung der Variablennamen kommen kann. Es werden zwar die Namen bei zu geringem Achsenabstand entsprechend gekürzt, dennoch besitzen die Variablen eine bestimmte Mindestlänge.

Hinzu kommt, daß die Linien je nach Abstand der Achsen unterschiedliche Winkel aufweisen, was bewirkt, daß die einzelnen Linienstrukturen je nach Winkelverteilung zwischen den Variablen unterschiedlich gut erkennbar sind. Liegen hauptsächlich parallele Linien vor, so kann der Achsenabstand relativ gering ausfallen. Befindet sich aber ein Knoten zwischen den Achsen, so kann man erst bei einem größeren Achsenabstand einzelne Linienzüge erkennen (siehe Abb. 6.6).

6.2.2 Selektion

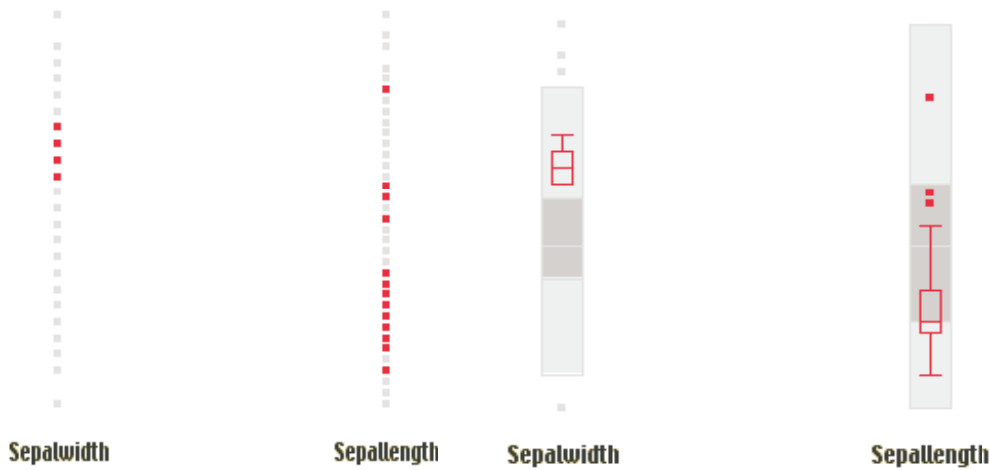
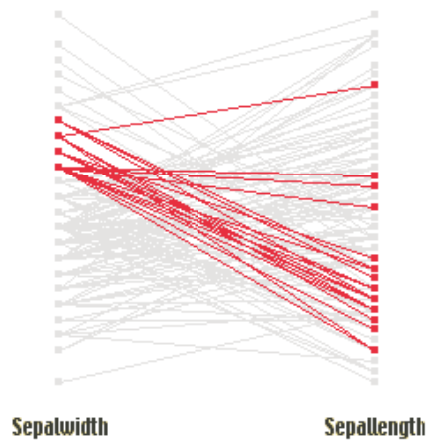
Mittlerweile wurde schon desöfteren von der Selektion einzelner Beobachtungen gesprochen. Diese Methode beschreibt im Wesentlichen das Markieren bestimmter Individuen. Eine solche Markierung wird "Highlight" genannt und im Normalfall farbig dargestellt. Um konsistent zu bleiben, werden die Individuen je nach Darstellungsart anders markiert. So werden beispielsweise die ausgewählten Punkte in einem Boxplot ebenfalls durch einen Boxplot dargestellt, in Dotplots die zugehörigen Punkte eingefärbt (siehe Abb. 6.7 und 6.8). In Parallelen Koordinaten Darstellungen, in denen die Linien eingezeichnet sind, werden die Polylinien der ausgewählten Beobachtungen farbig gezeichnet.

In interaktiven Systemen ist die Selektion stark mit dem Begriff "Linking" verknüpft". Das bedeutet im Wesentlichen nur, daß die einzelnen Graphiken miteinander verknüpft sind, so daß jede Selektionsänderung sofort an alle Graphiken weitergeleitet wird und dort graphisch umgesetzt wird.

Eine solche Selektion kann man je nach Darstellungsart auf verschiedene Arten erreichen. Da dieser Themenbereich aber sehr komplex ist, werde ich im Kapitel 7 näher auf die verschiedenen Selektionsmöglichkeiten eingehen.

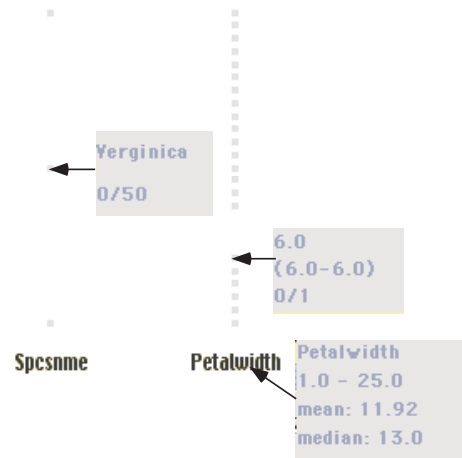
6.2.3 Abfrage

In CASSATT erhält man überall dort Information, wo man dies erwarten kann. Mit Hilfe der "Alt"-Taste und Betätigung der Maus kann man damit prinzipiell alle Objekte, die in der Darstellung vorhanden sind, abfragen. Diese Abfragen sind in CASSATT kontextsensitiv, was bedeutet, daß die Information je nach Objekt unterschiedlich ausfällt.

Abbildung 6.7: *Selektionen in unterschiedlichen Darstellungen*Abbildung 6.8: *Selektion in der Liniendarstellung*

Ein einfacher Punkt spiegelt dabei den exakten Wert und den Wertebereich der Individuen, die von dem gezeichneten Punkt überdeckt werden, wider. Letzteres wird vor allem bei geringer Auflösung relativ wichtig, da in diesem Fall ein gezeichneter Punkt einen größeren Wertebereich darstellt. Zusätzlich erhält man die Information, wie viele Punkte an dieser Stelle vorhanden sind und wie viele davon selektiert sind. Im Gegensatz dazu erhält man bei der Abfrage eines Variablennamens die grundlegenden Informationen über diese Variable, Mittelwert, Median, Minimum und Maximum (siehe Abb. 6.9).

Diese Art der Informationsgewinnung hat vor allem den Vorteil, daß man alle Werte abfragen kann und somit keine Skala mehr braucht, die die Graphik mit

Abbildung 6.9: *verschiedene Abfragemöglichkeiten*

Information anfüllt. Man erhält also nur dort Information, wo man diese benötigt. Ebenso kann man sich schnell und problemlos einen Überblick über grundlegende Statistiken verschaffen.

Neben dieser Standardabfrage besteht grundsätzlich die Möglichkeit, unterschiedliche Abfrageebenen zu schaffen. Dies würde bedeuten, daß man noch ergänzende Informationen zu dem entsprechenden Objekt bekommen kann, indem man beispielsweise neben der “Alt”-Taste noch die Umschalttaste betätigt.

6.2.4 Skala

Im allgemeinen haben verschiedene Variablen nicht immer die gleiche Einheit. Aus diesem Grund wird in CASSATT standardmäßig die Parallele Koordinaten Darstellung mit standardisierten Variablen erstellt. Dabei werden jeweils die Minima und die Maxima der einzelnen Variablen auf gleicher Höhe angetragen und die restlichen Punkte dazwischen im richtigen Verhältnis eingezeichnet.

Eine Gleichskalierung standardmäßig einzuführen wäre sehr unbefriedigend, wenn man bedenkt, daß man bei sehr unterschiedlichen Skalen sehr schnell nur noch wenig erkennen kann. So kann es vorkommen, daß der Pixelabstand zwischen Minimum und Maximum einzelner Variablen bei entsprechender Gesamtskala in der Graphik so gering ausfällt, daß die Bildschirmauflösung die Punkte nicht mehr einzeln darstellen kann (siehe Abb. 6.10).

In manchen Fällen will man dennoch Gebrauch einer einheitlichen Skala machen. Diese Umstellung kann man im Menü durchführen. Eine Gleichskalierung ist aber nur dann sinnvoll, wenn man Daten hat, die gut miteinander verglichen werden können.

Man sollte sich aber bewußt sein, daß es in Bezug auf Skalierung noch weit mehr Möglichkeiten gibt, die jedoch nur teilweise in CASSATT verwirklicht sind.

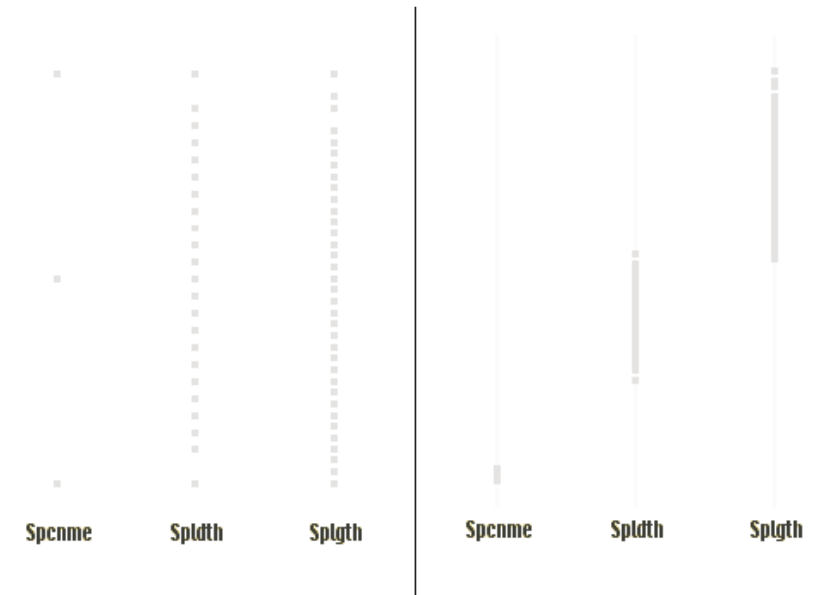


Abbildung 6.10: *Parallele Koordinaten Darstellung standardisiert und mit gleicher Skala*

So kann man beispielsweise über das Menü eine Transformation durchführen, so daß die Mittelwerte aller Variablen auf gleicher Höhe liegen. Die restliche Skalierung der einzelnen Variablen hängt im Folgenden dann von der Standardabweichung der jeweiligen Variablen ab. Dabei kann es bei einzelnen Punkten vorkommen, daß sie außerhalb des sichtbaren Bereichs liegen.

6.2.5 Invertieren einzelner Variablenachsen

Eine andere spezielle Skalierung ist das Invertieren einzelner Variablenachsen. In jeder Parallelen Koordinaten Darstellung können zwei nebeneinander stehende Variablen gut miteinander verglichen werden. In Kapitel 3 wurde schon die Inversion einzelner Variablenachsen und der Nutzen, den man daraus ziehen kann besprochen. Diese Inversion führt im Wesentlichen dazu, daß eine positive Korrelation in eine negative Korrelation umgewandelt wird, was in einer parallelen Koordinaten Darstellung gut erkannt werden kann (siehe Abb. 3.6 und 3.7 in Kapitel 3).

Diese Methode ist ebenfalls dazu geeignet einzelne Linien zu selektieren. Einerseits kann man so stark korrelierte Linien einfach selektieren, indem man bei vorliegender negativer Korrelation den Knoten selektiert und bei positiver Korrelation eine Achse invertiert und ebenso wie bei einer negativen Korrelation vorgeht. Andererseits lassen sich mit Hilfe dieser Methode auch einfach Ausreißer selektieren, da diese mit Hilfe der Inversion einzelner Achsen auch entdeckt werden können.

In CASSATT kann man diese einfache, aber sehr nützliche Aktion erreichen, indem man rechts unter der zu invertierenden Variablen die Maus betätigt, während der Cursor einen schrägen Doppelpfeil zeigt. Ist die Achse invertiert, erscheint automatisch eine grüne Pfeilspitze über der entsprechenden Achse. Klickt man in dem beschriebenen Bereich ein weiteres Mal die Maus, so wird der Achse wieder ihre ursprüngliche Orientierung zugewiesen.

6.2.6 Umordnung

Da in einer Parallelen Koordinaten Darstellung stets nur bei direkten Nachbarn die Möglichkeit besteht, Zusammenhänge abzulesen, ist es einleuchtend, daß man sich eine Methode wünscht, mit Hilfe derer man schnell und einfach die einzelnen Achsen umordnen oder vertauschen kann.

In CASSATT ist daher der intuitive Ansatz implementiert, daß man einzelne Variablenachsen durch “Drag & Drop” verschieben kann. Es ist daher möglich eine Variablenachse mit der Maus am Namen zu packen und die Variablenachse zwischen die Variablenachsen zu bewegen, wo man diese haben möchte. Die gezogene Variablenachse wird also an der Position eingefügt, an der sich diese beim Loslassen der Maustaste befunden hat. Es ist jedoch eine Besonderheit zu beachten. Zieht man eine Variablenachse genau über eine andere Variablenachse, so wird letztere blau eingefärbt. Dies bedeutet, daß man die Stelle erreicht hat, an der man die beiden Achsen vertauschen kann.

Mit dieser Methode ist ein Instrument geschaffen, mit dem man flexibel alle gewünschten Kombinationen erstellen und analysieren kann. Es wäre allerdings zu überlegen, ob man zusätzlich noch eine automatische Umordnung der Achsen einführt, um alle Kombinationen ansehen zu können.

Um alle paarweise Vergleiche anstellen zu können, muß man insgesamt $\binom{n}{2}$, also $\frac{n(n-1)}{2}$, Paarkombinationen betrachten. Jede Anordnung aller n Achsen in einer Parallelen Koordinaten Darstellung könnte idealerweise $(n-1)$ neue Paarkombinationen enthalten. Mit einer geeigneten automatischen Methode könnte man durch $\lceil \frac{n(n-1)}{2(n-1)} \rceil$ unterschiedliche Anordnungen, also in $\frac{n}{2}$ für gerade n bzw. $\frac{(n+1)}{2}$ für ungerade n , alle paarweisen Kombinationen erzielen.

6.2.7 Sortierung

Eine ganz andere Art der Umordnung stellt die automatische Sortierung nach bestimmten Statistiken dar. Oft will man nicht nur direkte Zusammenhänge erkennen. Bei manchen Datensätzen, die beispielsweise Bewertungen widerspiegeln, will man eher Vergleiche anstellen. Dies kann man einerseits durch die schon angesprochenen Parallelen Boxplots bzw. Dotplots erreichen, wenn man dabei eine gemeinsame Achsenskalierung eingestellt hat. Bei einer großen Anzahl von Variablen ergibt sich allerdings im Normalfall eine relativ unübersichtliche Parallele

Koordinaten Darstellung, so daß es sinnvoll wird, die Variablen nach verschiedenen Statistiken zu sortieren.

In CASSATT kann man nach verschiedenen Orientierungswerten, wie Mittelwert, Median, Interquartilsabstand, Minimum oder Maximum, ordnen (siehe Abb. 6.11). Eine solche Sortierung läßt sich über den Menüpunkt “Order data”. Die Variablen werden dann nach den bestimmten Orientierungswerten geordnet, wobei aber jeweils die Lage dieses Wertes auf der Darstellung als Maßstab zählt. Der Ansatz, “What you see is what you get!”, ist in CASSATT verwirklicht, was dazu führt, daß sich je nach gewählter Skalierung eine andere Reihenfolge ergibt.

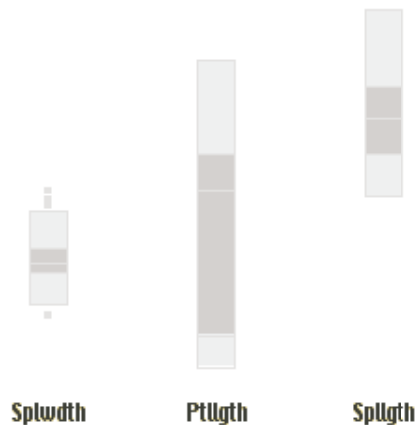


Abbildung 6.11: *Sortierung nach dem Median*

In CASSATT ist eine weitere Sortierung implementiert. Die Variablen können hierbei nach den Orientierungswerten der selektierten Punkte geordnet werden (siehe Abb. 6.12). Im Wesentlichen kann man diese Sortierung genau wie die erste Sortierung erreichen, indem man den Menüpunkt “Sort Selected” auswählt.

Die jeweiligen Sortierungen erfolgen immer aufsteigend. Will man für bestimmte Zwecke die absteigende Reihenfolge erhalten, so kann man nach der Sortierung den Menüpunkt “Reverse” auswählen.

Diese Art der Sortierung kann dem Benutzer also schnell einen vergleichenden Überblick über Daten verschaffen. Besonders mit Hilfe der Sortierung nach Selektierten Punkten kann man schnell einzelne Untergruppen in ein und derselben Darstellung vergleichen.

6.2.8 Farbe

Wie schon bei der Selektion angedeutet wurde, kann man einzelne Beobachtungen in CASSATT farbig kennzeichnen. Dies führt unweigerlich zu einem Problem: Welche Farbkombinationen von selektierten und nicht selektierten Punkten sollte am besten verwendet werden. Es spielen dabei nicht nur die Einstellungen des

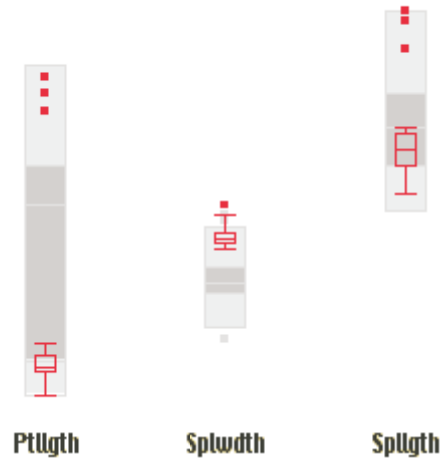


Abbildung 6.12: *Sortierung der selektierten Fälle nach dem Median*

jeweiligen Bildschirmes, also Kontrast, Helligkeit und Farbtiefe, eine wichtige Rolle sondern auch die Auswahl der Farben (siehe Abb. 6.13).

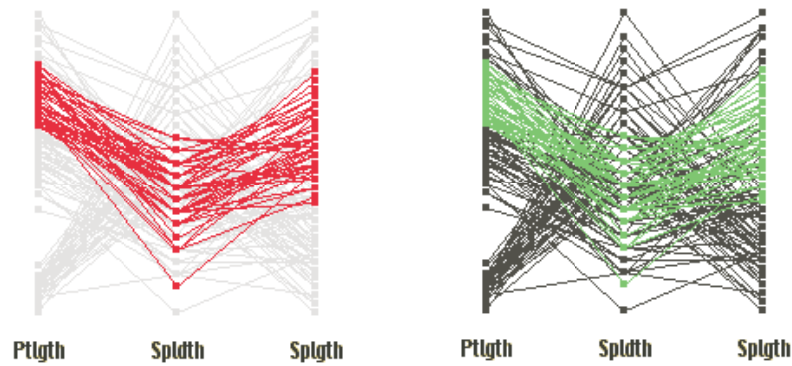


Abbildung 6.13: *unterschiedliche Farbzusammenstellungen*

Eine eingängige Lösungsvariante stellt hier die freie Wahl dieser Farben dar, auch wenn relativ kontrastreiche Farbeinstellungen als Standardeinstellung verwendet werden. In CASSATT gehört auch diese Möglichkeit zu den interaktiven Fähigkeiten in Parallelen Koordinaten Darstellungen, da man das Ändern der einzelnen Farben direkt in den Graphiken über einen Menüpunkt erreichen kann. Je nachdem ob man die Selektionsfarbe oder die Linienfarbe ändern will, muß man den Menüpunkt “Color Selection” bzw. “Color Background” auswählen. In einem Dialogfenster kann man daraufhin die Farbe auswählen, die dann global für sämtliche Darstellungen angewendet werden, so daß sich jeder Benutzer seine individuellen Farbzusammenstellungen selbst wählen kann.

6.2.9 Hiding Plots

Eine ganz besondere nur temporäre Farbzusammenstellung erhält man durch ausblenden bestimmter Individuen. Dabei entstehen sogenannte Hiding Plots, was bedeutet, daß selektierte bzw. nicht selektierte Punkte die Farbe des Hintergrundes erhalten. Die Möglichkeit der Selektion bzw. Deselektion bleibt aber noch erhalten. Besonders nützlich erweist sich diese Methode, wenn man versuchsweise einzelne Untergruppen genauer betrachten will, aber nicht den komplizierten Weg der Gruppenerstellung gehen will (siehe Kapitel 8). Parallele Koordinaten Darstellungen kann man in Hiding Plots umwandeln, indem man den Menüpunkt “hide selected” bzw. “hide not selected” betätigt. Mit dem Menüpunkt “show all” wird dies wieder rückgängig gemacht.

6.2.10 Toggle1 & Toggle2

Die Selektion bringt noch weitere Probleme mit sich. Dies hängt aber besonders mit der graphischen Darstellung der Individuen zusammen. Liegen zwei Datenpunkte genau übereinander, so kann man nur den Punkt erkennen, der als letztes gezeichnet wurde.

Da die selektierten Punkte im Standardfall nach den nicht selektierten Punkten gezeichnet werden, kann man nun nicht immer erkennen, ob unter Punkten, die die Selektionsfarbe besitzen, noch andere nicht selektierte Punkte liegen. Eine Möglichkeit besteht hier in der schon erwähnten Abfrage der einzelnen Punkte. Allerdings kann die Anzahl der zu betrachtenden Fälle ziemlich groß werden, so daß eine einfachere und eingängigere Methode angebracht wäre.

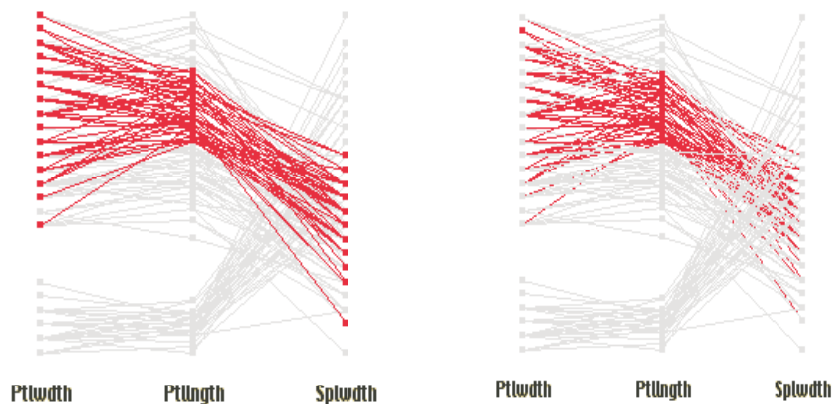


Abbildung 6.14: *Wirkungsweise des Toggle1*

In CASSATT gibt es daher das Toggle1. Dabei werden nur über einen gewünschten Zeitraum, die selektierten Punkte zuerst gezeichnet, so daß man alle nicht-selektierten Punkte sofort erkennen kann (siehe Abb. 6.14). Ausgelöst wird

dieser Effekt durch einen Shortcut. Läßt man die Tasten wieder los, erscheint sofort wieder das gewohnte Bild.

Wie an der Überschrift dieses Abschnittes schon zu erkennen ist, gibt es eine weitere Art des Toggles, das Toggle2, welches nicht nur einer kurzzeitigen optischen Veränderung der Darstellung sondern einer echten Selektion entspricht. Es werden also durch einen einfachen Shortcut alle selektierten Individuen deselektiert und umgekehrt. Geeignet ist diese Art der Selektion, wenn die gewünschte Auswahl von Datenpunkten komplizierter auszuwählen ist als die zugehörige inverse Menge.

Eine weitere Anwendung stellt der Fall dar, daß man zwei Gruppen, A und das Komplement von A, miteinander vergleichen will. So kann man mit Hilfe des Toggle2 zwischen den Gruppen hin und her wechseln, um eventuelle Punktüberschneidungen oder Strukturen zu erkennen.

Kapitel 7

Erweiterte Fähigkeiten

7.1 Einfache Selektion

Eine wichtige Fähigkeit, die interaktive Graphiken aufweisen sollten, ist die mit Linking verknüpfte Selektion. Je nach Darstellungsart eignen sich hierbei nur bestimmte Selektionsmechanismen. So ist es relativ unsinnig, in einem Scatterplot Individuen mit Hilfe eines Striches zu selektieren, was aber wiederum bei Linien in einer Parallelen Koordinaten Darstellung eine sehr vernünftige Selektionsweise darstellen kann. Im Folgenden möchte ich nun näher auf die einzelnen Selektionsmöglichkeiten in den einzelnen Darstellungsarten eingehen und deren Nutzen zur Datenanalyse kurz erklären.

7.1.1 Selektionen im Scatterplot, Dotplot und Boxplot

Betrachtet man nun Darstellungen, wie Scatterplots oder Dotplots, so ist es verständlich, daß man mit Punktdarstellungen der Individuen zu tun hat. Einen Boxplot kann man ebenfalls in diese Gruppe einordnen, weil dieser eigentlich ein Dotplot ist, dessen Punkte hinter der Box verborgen sind.

Da Punkte aber im Normalfall keine räumliche Ausbreitung besitzen, ist es nur möglich diese mit Hilfe einer Fläche auszuwählen. Die einfachste Fläche, die man mit einer Maus auf dem Bildschirm erzeugen kann, ist ein Rechteck, eine sogenannte Dragbox. Mit dieser Selektionsart kann man sehr genau einzelne Punkte oder auch Gruppierungen selektieren, da im Normalfall die erzeugte Dragbox auch während der Auswahl zu sehen ist.

Eine weitere Selektionsart wäre auch eine Selektion mit Hilfe eines “Messers”, welches erlaubt, ganze Intervalle einer Achse zu selektieren. Diese Art der Selektion ist bei Dotplots oder Boxplots äußerst sinnvoll, bei Scatterplots hingegen kann man damit Punktgruppen nur schlecht auswählen. Andere Möglichkeiten wären auch ein Kreis oder ein Lasso. Vor allem beim Lasso muß man noch erwähnen, daß es sehr schwierig ist, eine geometrische Definition für diese Selektion anzugeben, was eine Gruppengeschichte (siehe Kapitel 8) sehr kompliziert macht.

7.1.2 Toggle2

Um spezielle Individuen auszuwählen, ist es oftmals wesentlich einfacher, die inverse Menge zu selektieren. Mit Hilfe einer bestimmten Methode, dem Toggle2 (siehe Kapitel 6), kann man genau so vorgehen, wie man es instinktiv machen würde. Zuerst kann man dabei die inverse Menge selektieren, um dann anschließend eine Inversion durchzuführen.

7.2 Selektionsarten in der Parallelen Koordinaten Darstellung

Die Parallele Koordinaten Darstellung stellt eine ganz besondere Art der Datenvisualisierung dar. Daraus ergeben sich auch spezielle, nur für eine solche Darstellung sinnvolle Selektionsarten. Einerseits kann man die aus einem Dotplot oder Boxplot bekannten Selektionen verwenden, andererseits bieten sich noch weitere Selektionen, speziell die der Linien, an.

Welche Selektionsarten nun genau in einer Parallelen Koordinaten Darstellung angewendet werden können und was diese genau bewirken, soll nun im Weiteren erläutert werden.

7.2.1 Punktselektion an der Achse

Eine Punktselektion direkt an einer Achse der Parallelen Koordinaten Darstellung bewirkt die selbe Auswahl von Punkten, wie eine Selektion in einem Dotplot oder Boxplot. Im Wesentlichen bedeutet dies eine Auswahl von Punkten, die auf der entsprechenden Variablenachse innerhalb eines bestimmten Intervalls liegen.

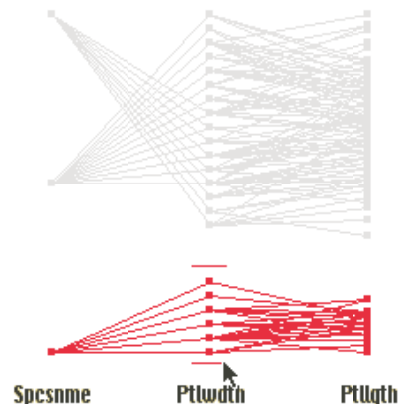


Abbildung 7.1: *Selektion einer Untergruppe*

Besonders geeignet erweist sich diese Art der Selektion zum Markieren von einzelnen Ausreißern. Will man diese später entfernen, kann man durch Toggle2 eine Untergruppe erhalten, die keine Ausreißer mehr enthält. Ebenso kann man mit Hilfe der Achsen Selektion auch einzelne Punktgruppierungen an einer Variablen auswählen und ebenfalls eine neue Gruppe erzeugen (siehe Kapitel 8).

7.2.2 Linienselektion mit Hilfe einer Linie (Pinch, Scherung)

Inselberg hat vorgeschlagen, für die Linienselektion eine einfache Linie zu verwenden (Inselberg personal communication). Bei einer derartigen Selektion werden alle Individuen selektiert, deren Linienzüge die angegebene Linie schneiden. Vor allem kann man dadurch Linienanhäufungen einfach und effektiv selektieren. Ebenfalls können so annähernd parallele Liniengruppen selektiert werden, auch wenn es störende, kreuzende Linien gibt.

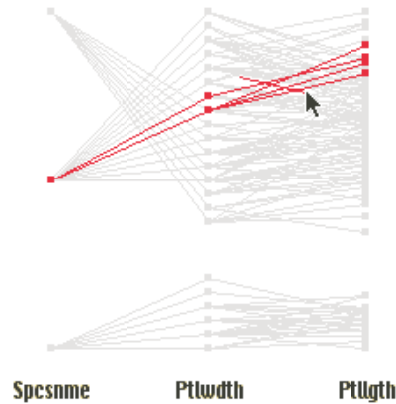


Abbildung 7.2: *Individuenselektion*

Durch die variable Länge und Steigung der Linie kann man ebenfalls sehr genau einzelne Linien selektieren und hat dadurch ein grundlegendes Werkzeug zur Individuenselektion in der Parallelen Koordinaten Darstellung geschaffen.

7.2.3 Linienselektion mit Hilfe einer Dragbox (doppelte Scherung)

Eine Selektion mit einer Dragbox entspricht der Selektion aller Individuen, die das Rechteck schneiden oder berühren. Auch wenn im Normalfall die Selektion mit einer Dragbox eine grundlegende Selektionsweise darstellt, bedeutet diese im Endeffekt nur eine zusammengelegte zweifache Linienselektion mit Hilfe der zwei Diagonalen des Rechteckes. Dies ist der Fall, da alle Linien, welche durch

eine Rechtecksselektion ausgewählt werden, entweder die eine oder die andere Diagonale dieses Rechtecks schneiden bzw. berühren. Aus diesem Grund ist die standardmäßige Voreinstellung der Linienselektion mit Hilfe nur einer Linie in CASSATT verwirklicht.

Es existieren aber dennoch sinnvolle Einsatzmöglichkeiten für diese Art der Selektion. Beispielsweise ist es dadurch möglich, bestimmte Linienknoten auszuwählen, um negative Korrelationen deutlich zu machen. Ebenfalls kann man viele parallele Linien, also eine parallele Korrelation, selektieren, indem man eine der anliegenden Achsen invertiert und den entstandenen Knoten auswählt (siehe Abb. 7.3).

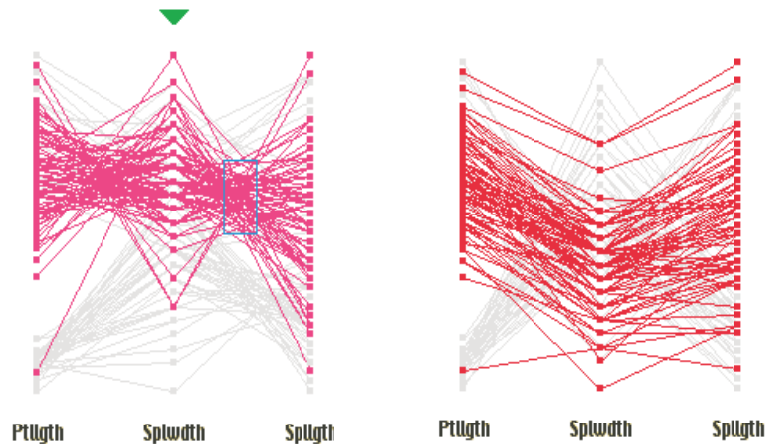


Abbildung 7.3: Rechtecksselektion des Knotens zwischen den zwei ersten Achsen mit nachfolgender Invertierung einer Variablenachse

7.2.4 Linienselektion mit Hilfe eines Winkelintervalls

Besonders intuitiv ist in der Parallelen Koordinaten Darstellung eine Selektion, mit der man annähernd parallele Linien auswählen kann. Dadurch wäre der relativ umständlichen Weg über die Invertierung einer Variablenachse nicht mehr notwendig.

Da die einzelnen Linienabschnitte immer an einer Variablen beginnen, muß man im Wesentlichen nur den Winkel bzw. das entsprechende Winkelintervall angeben und erhält damit ein neues Selektionswerkzeug – die Winkelselektion.

Ein solche Angabe der Winkel kann natürlich durch Textfelder, in die man die entsprechenden Winkelwerte tippen muß, verwirklicht werden. Für den Benutzer etwas angenehmer ist allerdings dennoch die Auswahl der Winkel per Maus in einer graphischen Darstellung, wie z.B. in einem Kreis (siehe Abb. 7.4).

Als eine weitere Erweiterung hierzu könnte man sich auch noch vorstellen, anstelle eines Kreises, einen Rosettenplot zu setzen, welcher im wesentlichen ei-

ner bestimmten Darstellung eines Histogramms über die einzelnen Winkel der Individuen entspricht und so die Verteilungen der Linienwinkel preisgeben kann.

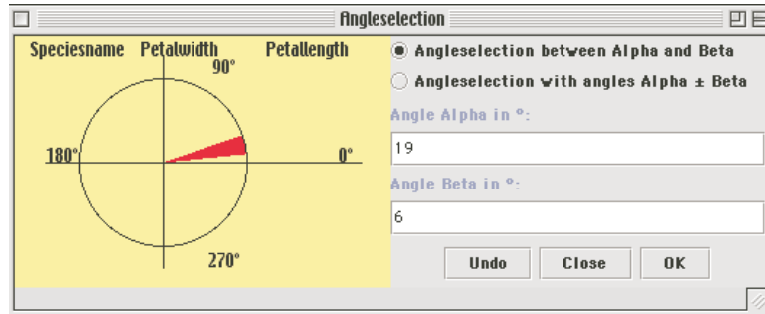


Abbildung 7.4: Dialogfenster zur Winkelselektion

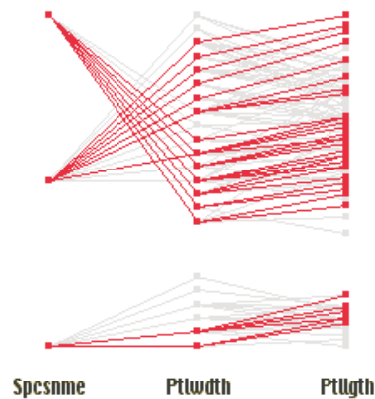


Abbildung 7.5: Ergebnis der Winkelselektion aus Abb. 7.4

Diese Selektion ist auch dann von großem Nutzen, wenn man Individuen finden will, die positive bzw. negative Zusammenhänge aufweisen. Wie schon desöfteren erwähnt, erkennt man positive Korrelationen durch das Auftreten von annähernd parallelen Linien, weshalb es einfach ist, die zugehörigen Linien mit Hilfe der Winkelselektion auszuwählen. Um negative Korrelationen aufzufinden, muß allerdings vor der Selektion eine Achse invertiert werden, um so annähernd parallele Linien zu erhalten.

Ein weiterer Vorteil der Winkelselektion ist die Möglichkeit, Ausreißer zu finden. Ausreißer besitzen entweder in der normalen Parallelen Koordinaten Darstellung oder in der Darstellung mit einer invertierten Achse Linienzüge, deren Steigungswinkel einen extremen Wert aufweist. Zum Auffinden von solchen Ausreißern kann man daher ebenfalls mit der Winkelselektion arbeiten, indem man in beiden Darstellungen Winkelintervalle wählt, die außerhalb der Hauptliniensteigungen liegen.

Eine etwas ungewöhnliche, aber sehr sinnvolle Anwendung der Parallelen Koordinaten ist der Vergleich von Residuen verschiedener Modelle. Bei der Analyse der verschiedenen Modelle, ist man daran interessiert, ähnliche Fälle zu finden. Damit ist gemeint, daß man versucht, einzelne Punkte zu sehen, die von den verschiedenen Modellen gleich behandelt werden. Aus diesem Grund kann man dabei sehr gut die Winkelselektion verwenden.

7.3 Translation der Selektionen der Parallelen Koordinaten Darstellung auf andere Darstellungen

In diesem Abschnitt soll nun näher auf die Bedeutungen der einzelnen Selektionsarten eingegangen werden und auf die Möglichkeiten, die diese Werkzeuge schaffen. Da der Mensch im Allgemeinen an die 3 dimensionale Welt gewöhnt ist, werde ich in diesem Kapitel die Selektionsarten in der Parallelen Koordinaten Darstellung auf zwei- bzw. dreidimensionale Darstellungen übertragen, damit man eine Vorstellung bekommt, was genau bestimmte Selektionen bewirken können.

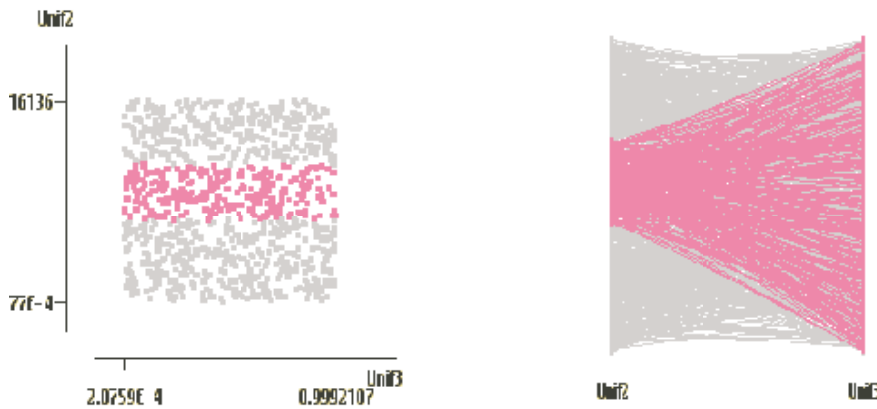
7.3.1 Eindimensionale Selektion

Grundsätzlich will man bei einer eindimensionalen Selektion ein Intervall auswählen, das heißt, daß man eine Menge

$$I = \{x \mid x_1 \leq x \leq x_2\}$$

erhält. Es ist einleuchtend, daß man diese Menge in eindimensionalen Darstellungen, wie einem Dotplot, selektieren kann. Bei einem Histogramm kann man prinzipiell auch derartige Selektionen durchführen, die Intervallgrenzen hängen jedoch von der zugehörigen Binbreite ab.

In der Parallelen Koordinaten Darstellung hat man ebenfalls die Möglichkeit, Individuen an nur einer Variablen zu selektieren. Dazu kann man die Punktselektion an der Achse verwenden. Im wesentlichen entspricht diese Selektion genau einer Selektion im Dotplot, so daß man diese Selektionen ineinander überführen kann. Mit Hilfe dieser Methode kann man schnell und einfach einzelne Punktgruppen an einer Variablen selektieren. Da diese Selektion aber eine Basisselektion darstellt, werde ich nicht näher auf die eindimensionale Selektion eingehen, sondern mich auf die höher-dimensionalen Selektionen konzentrieren.

Abbildung 7.6: *eindimensionale Selektionen*

7.3.2 Zweidimensionale Selektion

Intuitiv denkt man bei einer zweidimensionalen Selektion an die Auswahl von Punkten in einer zweidimensionalen Darstellung, wie einem Scatterplot. Eine Selektion im Scatterplot mit Hilfe einer Dragbox selektiert beispielsweise die Menge

$$I = \{(x, y) \mid x_1 \leq x \leq x_2 \wedge y_1 \leq y \leq y_2\}.$$

Diese Selektion läßt sich auf 2 eindimensionale Selektionen zurückführen, indem man Menge I in die Mengen

$$I_1 = \{(x, y) \mid x_1 \leq x \leq x_2\} \text{ und } I_2 = \{(x, y) \mid y_1 \leq y \leq y_2\}$$

aufteilt und die zugehörigen Selektionen mit Schnitt verbindet. Daher ist eine derartige Selektion auch in der Parallelen Koordinaten Darstellung nachzuvollziehen. Hier muß allerdings auch noch erwähnt werden, daß man theoretisch diese Selektion auch in der Parallelen Koordinaten Darstellung durchführen kann, der Zweck der oben genannten Selektion im Scatterplot aber in der Parallelen Koordinaten Darstellung nicht genau nachvollzogen werden kann. So will man in einem Scatterplot oft eine ganz bestimmte Punktgruppe mit der Dragbox selektieren, die man so einfach gar nicht in der Parallelen Koordinaten Darstellung erkennen kann.

Da jede Darstellung eigene spezifische Selektionsmöglichkeiten besitzt, die bestimmte Zwecke verfolgen, soll nicht versucht werden alle möglichen Selektionen auch in einer Parallelen Koordinaten Darstellung einzusetzen. Vielmehr soll eine sinnvolle Kombination der verschiedenen Selektionsmöglichkeiten beibehalten werden.

Zum besseren Verständnis der Selektionen in einer Parallelen Koordinaten Darstellung will ich in diesem Kapitel näher auf die Überführung von verschie-

denen Selektionen in einer Parallelen Koordinaten Darstellung auf Selektionen in anderen Darstellungen eingehen. Dabei wird hauptsächlich der zweidimensionale Scatterplot verwendet. Allerdings muß man festhalten, daß die weiteren Ausführungen darauf beruhen, daß jeweils die gleichen Skalierungen in den Graphiken verwendet werden. Anderenfalls muß man dies durch entsprechende Transformationen berichtigen.

Nun möchte ich näher auf die zweidimensionale Selektion in einer Parallelen Koordinaten Darstellung eingehen. Man kann eine solche Selektion erhalten, indem man zwischen zwei Achsen selektiert. Wie schon im vorherigen Kapitel erwähnt wurde, existieren in der Parallelen Koordinaten Darstellung verschiedenste Werkzeuge, um zwischen den Variablenachsen zu selektieren.

Pinch

Eine Basisselektion stellt hierbei die Linenselektion zwischen zwei Variablen mit Hilfe einer Strecke s dar. Diese Strecke s hat bestimmte Merkmale, mit deren Hilfe man eine Überführung in einen zweidimensionalen Scatterplot erklären kann.

Es ist klar, daß die zur Strecke s gehörenden Gerade g einen zweidimensionalen Punkt $A(x_A, y_A)$ darstellt. Dieser Punkt A stellt im folgenden eine Schlüsselrolle dar. Die Länge der Strecke s kann mit Hilfe der Streckenendpunkte $E_1(x', y')$ und $E_2(x'', y'')$ festgelegt werden. Diese Endpunkte in der Parallelen Koordinaten Darstellung entsprechen jeweils einer Geraden im Scatterplot (siehe Abb. 7.7).

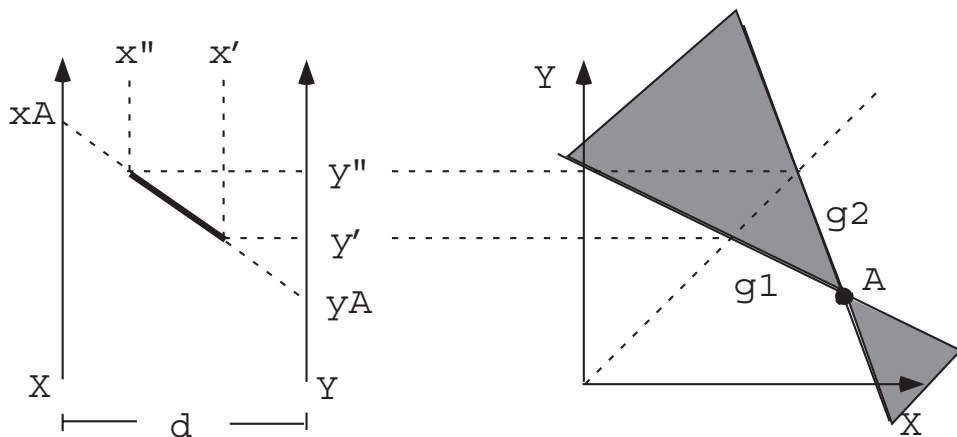


Abbildung 7.7: Geometrische Veranschaulichung der Linienselektion

Diese 2 so gefundenen Geraden haben ganz spezielle Eigenschaften. Einerseits enthalten beide Geraden den Punkt A , da die Streckenendpunkte E_1 und E_2 auf der Liniendarstellung des Punktes A , also auf g , liegen, andererseits schneiden diese Geraden die Winkelhalbierende des Scatterplots in den y -Werten der Streckenendpunkte, d.h. in y' und y'' . Aus diesen Überlegungen kann man die

Gleichungen der Geraden ableiten, wie ich in den unteren Ausführungen zeigen werde.

Nun will ich aber näher auf den genauen Selektionsbereich dieser Selektion eingehen. Nachdem man die 2 Geraden in einem Scatterplot eingetragen hat, kann man sich überlegen, daß diese 2 Geraden Randwerte der Selektion darstellen. Mit der zusätzlichen Information, daß auch die Punkte auf der Winkelhalbierenden, deren y -Werte zwischen y' und y'' liegen, zu dem Selektionsbereich gehören, kann man den Selektionsbereich als den Bereich ausmachen, der zwischen den 2 Geraden liegt. Damit ist auch der Bereich des Gegenwinkels gemeint (vgl. Abb. 7.7 und 7.8).

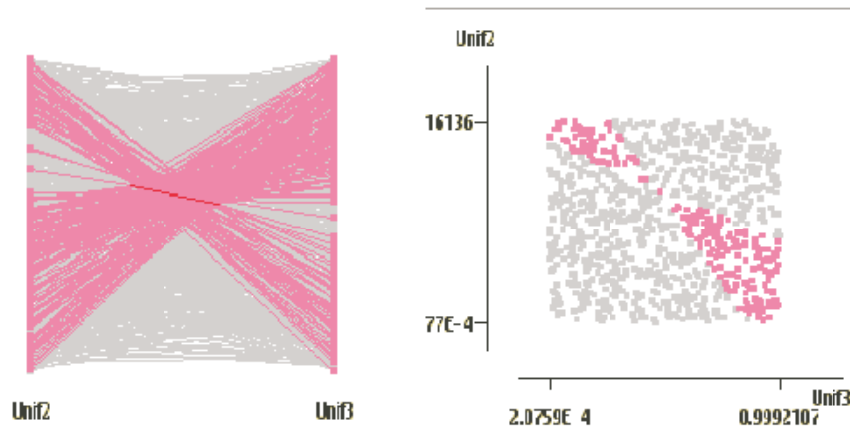


Abbildung 7.8: Selektion mit Hilfe einer Linie

Die Menge der Punkte, die zu diesem Selektionsbereich gehören, lassen sich demnach auch mathematisch berechnen, wie die folgenden Ausführungen zeigen.

Zur leichteren Berechnung gelte $y' < y''$. Durch Anwendung der Gleichung aus (3.1) erhält man für die Punkte E_1 und E_2 die Gleichungen $g_1(x)$ und $g_2(x)$.

$$g_1(x) = m'x + \frac{y'd}{x'}, \text{ mit } m' = 1 - \frac{d}{x'} \quad \text{falls } x' \neq 0$$

$$g_1 : x = y', \text{ mit } m' = \infty \quad \text{falls } x' = 0$$

$$g_2(x) = m''x + \frac{y''d}{x''}, \text{ mit } m'' = 1 - \frac{d}{x''} \quad \text{falls } x'' \neq 0$$

$$g_2 : x = y'', \text{ mit } m'' = \infty \quad \text{falls } x'' = 0$$

g_m sei eine Geradenschar durch den Punkt $A(x_A, y_A)$ mit Steigungen m .

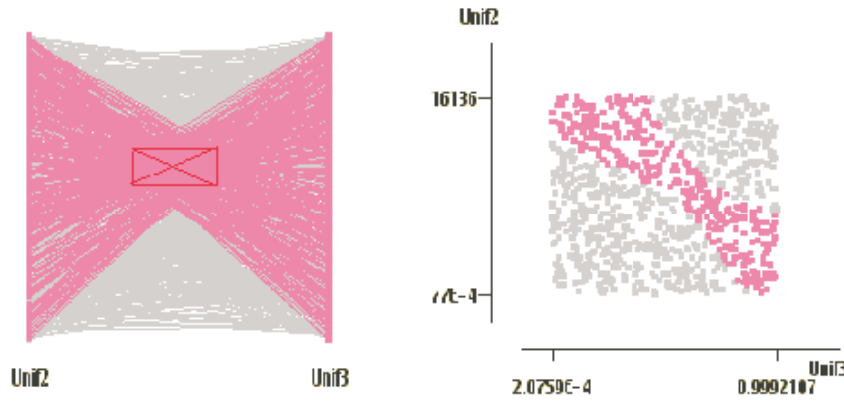


Abbildung 7.10: Selektion mit Hilfe einer Dragbox

$$\begin{aligned}
 g_1 : \quad x = y' , \text{ mit } m' = \infty & & \text{falls } x' = 0 \\
 g_2(x) = m''x + \frac{y''d}{x''} , \text{ mit } m'' = 1 - \frac{d}{x''} & & \text{falls } x'' \neq 0 \\
 g_2 : \quad x = y'' , \text{ mit } m'' = \infty & & \text{falls } x'' = 0 \\
 g_3(x) = m'''x + \frac{y'd}{x''} , \text{ mit } m''' = m'' & & \text{falls } x'' \neq 0 \\
 g_3 : \quad x = y' , \text{ mit } m'' = \infty & & \text{falls } x'' = 0 \\
 g_4(x) = m''''x + \frac{y''d}{x'} , \text{ mit } m'''' = m' & & \text{falls } x' \neq 0 \\
 g_4 : \quad x = y'' , \text{ mit } m' = \infty & & \text{falls } x' = 0
 \end{aligned}$$

Analog verwendet man hier 2 Geradenscharen g_{1m} und g_{2m} , wobei g_{1m} eine Geradenschar durch den Punkt $A(x_A, y_A)$ mit Steigungen m und g_{2m} eine Geradenschar durch den Punkt $B(x_B, y_B)$ mit Steigungen m darstellt.

$$g_{1m} = \{(x, y) \mid y = m(x - x_A) + y_A\}$$

$$g_{2m} = \{(x, y) \mid y = m(x - x_B) + y_B\}$$

Der Selektionsbereich I stellt eine Vereinigung von Selektionsbereichen dar wie sie beim Pinch aufgetreten sind. damit erhält man:

$I = I_1 \cup I_2$, wobei

$$I_1 = \begin{cases} \{(x, y) \mid (x, y) \in g_{1m}, m'' \leq m \leq m'\} & : \text{ falls } x_A < y_A \\ \{(x, y) \mid (x, y) \in g_{1m}, m' \leq m \leq m''\} & : \text{ falls } x_A > y_A \\ \{(x_A, y_A)\} & : \text{ falls } x_A = y_A \end{cases}$$

$$I_2 = \begin{cases} \{(x, y) \mid (x, y) \in g_{2m}, m' \leq m \leq m''\} & : \text{ falls } x_B < y_B \\ \{(x, y) \mid (x, y) \in g_{2m}, m'' \leq m \leq m'\} & : \text{ falls } x_B > y_B \\ \{(x_B, y_B)\} & : \text{ falls } x_B = y_B \end{cases}$$

Einen Extremfall stellt die Selektion mit Hilfe einer Dragbox dar, die von der linken bis zur rechten Variablen verläuft, d.h. $x'' = 0, x' = d, y' = y_A, y'' = y_B$ (siehe Abb. 7.11 und 7.12). Durch die Besonderheit dieser Werte erhält man bei der Übertragung auf den Scatterplot waagrechte bzw. senkrechte Geraden als Grenzen für den Selektionsbereich.

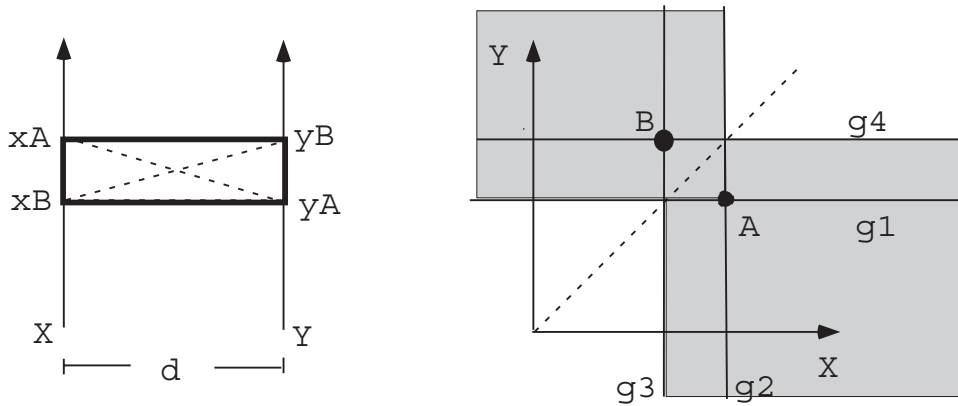


Abbildung 7.11: Geometrische Veranschaulichung der Selektion mit Hilfe einer Dragbox, wobei $x'' = 0, x' = d, y' = y_A, y'' = y_B$

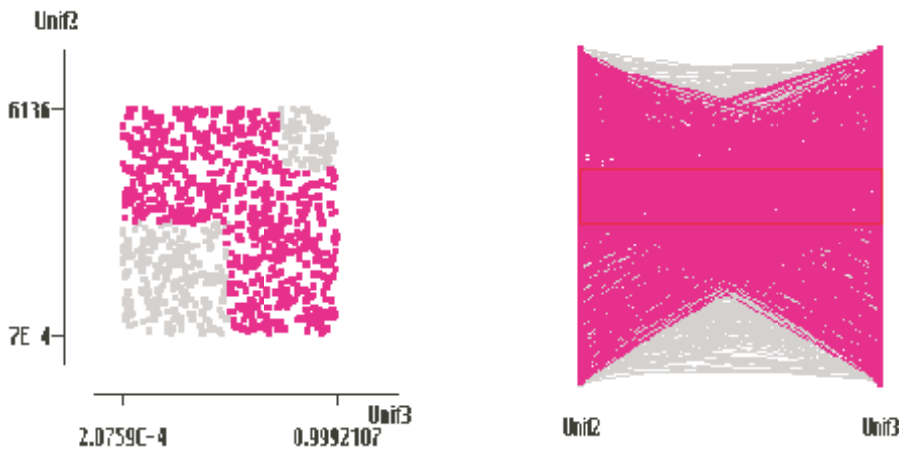


Abbildung 7.12: Spezialfall

Der Selektionsbereich I setzt sich im Wesentlichen genau wie bei einer normalen Dragboxselektion zusammen, nur mit der Vereinfachung, daß die Geraden-

gleichungen einfacher angegeben werden können.

Es gelte $y_A < y_B$.

$g_1 = y_A$, mit $m' = 0$

$g_2 = y_B$, mit $m'' = \infty$

$g_3 = y_A$, mit $m''' = m'' = \infty$

$g_4 = y_B$, mit $m'''' = m' = 0$.

Winkelselektion

Eine weitere zweidimensionale Selektionsart stellt die vorher beschriebene Winkelselektion dar. Bei dieser Selektion werden nur Individuen selektiert, deren Liniendarstellungen in der Parallelen Koordinaten Darstellungen Steigungen besitzen, die in einem vorgegebenen Steigungsintervall liegen. Die Intervallgrenzen der Steigung werden über die zugehörigen Winkel festgelegt.

Um diesen Selektionsbereich näher spezifizieren zu können, werde ich zuerst erklären, wo die Individuen liegen, deren Liniendarstellungen in der Parallelen Koordinaten Darstellung parallel sind. Da parallele Linien denselben Steigungswinkel besitzen, kann man hier den zugehörigen Winkel α annehmen. Man kann allgemein annehmen, daß $\alpha \in (-90, 90)$, da alle anderen Winkel durch die Transformation $\alpha = 180 + \alpha$ erhalten werden können.

Damit ergibt sich als Steigung $m = \tan \alpha = \frac{p}{d}$ (vgl. Abb. 7.13). Alle Individuen, die diese Steigung in der Liniendarstellung besitzen, haben somit die Koordinatenwerte $(x, x + p)$. Daraus kann man ersehen, daß diese Individuen auf einer um p verschobenen Parallelen zur Winkelhalbierenden liegen, also auf $g_1(x) = x + p$, mit $p = d \tan \alpha$.

Da aber ein ganzes Winkelintervall ausgewählt werden soll, wird noch einen zweiter Winkel β angegeben, der den Steigungswinkel der zweiten Geraden angibt. Analog ergibt sich somit im Scatterplot die Gerade $g_2(x) = x + q$, mit $q = d \tan \beta$ (siehe Abb. 7.14).

Durch Überlegung erkennt man, daß sich der Bereich als Selektionsbereich ausmachen läßt, der zwischen den 2 Geraden g_1 und g_2 liegt. Da beide Geraden jeweils Steigung $m = 1$ besitzen, entspricht der genannte Bereich einem Band, das parallel zur Winkelhalbierenden des 1. Quadranten verläuft (siehe Abb. 7.14 und 7.15).

Auch hier kann man den Selektionsbereich I mathematisch angeben:

Es gelte hier allerdings $\alpha \geq \beta$

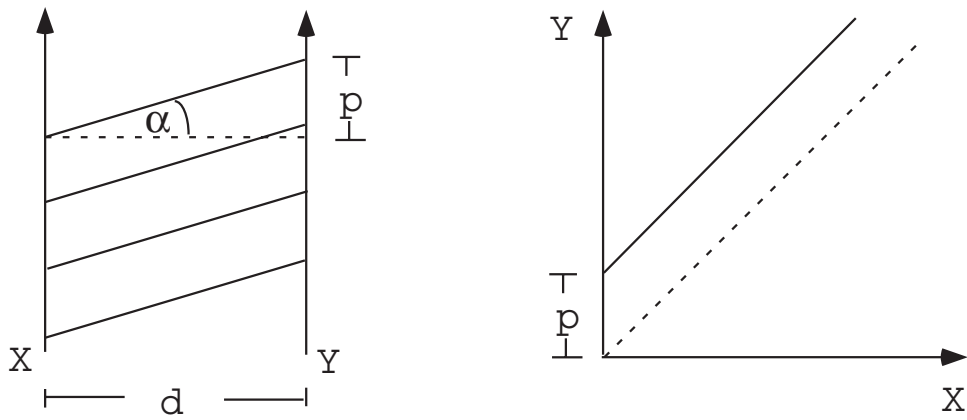


Abbildung 7.13: Übertragung eines Winkels von der Parallelen Koordinaten Darstellung auf den Scatterplot

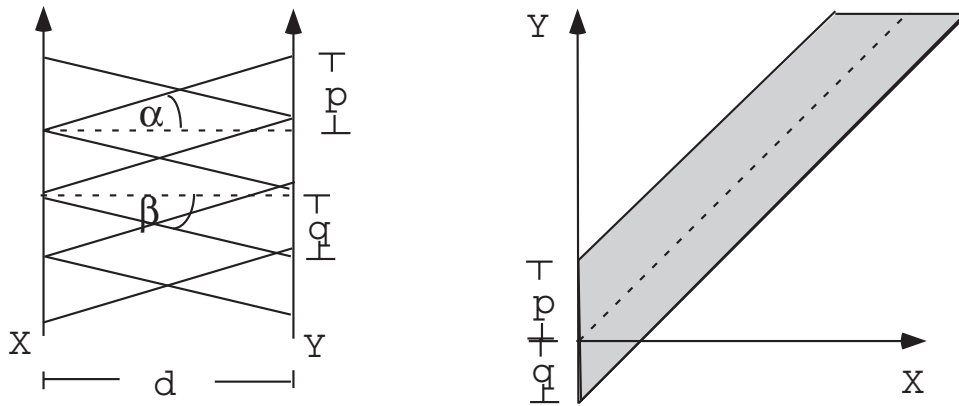


Abbildung 7.14: Geometrische Veranschaulichung der Winkelselektion

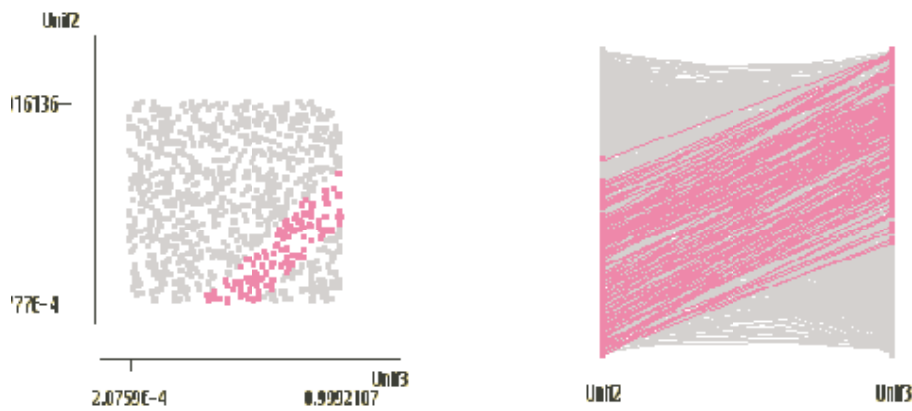


Abbildung 7.15: Selektion mit Hilfe einer Winkelintervallangabe

und $\alpha, \beta \in (-90, 90)$.

$$I = \{(x, y) \mid y = x + t, q \leq t \leq p\}$$

mit $p = d \tan \alpha$ und $q = d \tan \beta$.

7.3.3 Drei- und mehrdimensionale Selektion

Anfangs will ich hierzu erwähnen, daß es möglich ist, über mehrere Dimensionen zu selektieren, indem man schrittweise einzelne Selektionen vereinigend verketten. So kann man beispielsweise eine dreidimensionale Selektion erreichen, indem man eine zweidimensionale Selektion, z.B. im Scatterplot, und danach eine eindimensionale Selektion durchführt. Dadurch kann man schrittweise die Dimension erhöhen. Durch die verschiedenen Selektionsmodi kann man schon bei 3 Dimensionen verschiedenartigste Mengen erhalten. Beispielsweise erhält man bei 3 aufeinanderfolgenden Punktselektionen im Schnittmodus als Selektionsbereich die Menge

$$I = \{(x, y, z) \mid x_1 \leq x \leq x_2 \wedge y_1 \leq y \leq y_2 \wedge z_1 \leq z \leq z_2\}.$$

Diese Menge stellt im dreidimensionalen Raum einen Quader dar. Führt man genau diese Selektion im Vereinigungsmodus durch, so erhält man eine wesentlich größere Menge, die nur noch schwer vorstellbar ist. Es handelt sich hier um orthogonale "Wände" im Raum, die vereinigt werden. Solche Mengen kann man dennoch relativ einfach mathematisch angeben:

$$I = \{(x, y, z) \mid x_1 \leq x \leq x_2\} \cup \{(x, y, z) \mid y_1 \leq y \leq y_2\} \\ \cup \{(x, y, z) \mid z_1 \leq z \leq z_2\}$$

Schon hier merkt man, welche Möglichkeiten sich durch Verkettung mit verschiedenen Selektionsmodi ergeben. Verwendet man nun neben den einfachen Selektionen aus den Standarddarstellungen auch kompliziertere zweidimensionale Selektionen aus der Parallelen Koordinaten Darstellung, so erhält man noch kompliziertere Mengen. Es ist schnell einzusehen, daß mehrdimensionale Selektionen nicht mehr einfach dargestellt werden können, da diese Mengen an die Grenzen des Vorstellungsvermögens gehen. Selbst in dreidimensionalen Scatterplots lassen solche Mengen nicht mehr erkennen und erst recht nicht selektieren. Es ist daher zu überlegen, ob man überhaupt höherdimensionale Selektionswerkzeuge einführen soll, da man zu schnell den Überblick über das was man tut, verlieren kann.

Aus diesem Grund gibt es in CASSATT neben den Selektionssequenzen nur

ein höherdimensionales Selektionswerkzeug. Mit Hilfe einer Dragbox kann man über mehrere Variablen hinweg selektieren. Dabei ergibt sich eine vereinigende Verkettung von Dragboxselektionen. Bei der Übertragung auf traditionelle Darstellungen kommt aber, wie schon bei dreidimensionalen Selektionen, an die Grenze des Darstellbaren.

Solche Selektionen können genutzt werden, um beispielsweise alle Individuen zu selektieren, die bei bestimmten Variablen mindestens einmal einen bestimmten Minimalwert überschreiten.

7.3.4 Selektion bei invertierten Achsen

An dieser Stelle will ich noch auf einen Sonderfall bei den Selektionen eingehen. Da man in der Parallelen Koordinaten Darstellung die Möglichkeit hat, einzelne Achsen zu invertieren, ergeben sich noch weitere Selektionmöglichkeiten. Diese Möglichkeiten wurden auch schon am Anfang dieses Kapitels beschrieben.

Die Achsen werden beim Invertieren genau am Mittelpunkt gespiegelt, weshalb man die zugehörige Transformation angeben kann.

$$T : T(x) = (\maxValue(X) + \minValue(X)) - x,$$

wobei $\maxValue(X)$ den maximalen Skalenwert und $\minValue(X)$ den minimalen Skalenwert der entsprechenden Variablen angeben.

Diese Transformation verändern die endgültigen Selektionsbereiche, die man aus den vorherigen Abschnitten kennengelernt hat. Bei den zweidimensionalen Selektionen kann man dies im Scatterplot nachvollziehen, indem man dort ebenfalls eine Achse an der entsprechenden Stelle spiegelt und nach der Selektion diese wieder in ihren ursprünglichen Zustand bringt.

So kann man beispielsweise bei der Winkelselektion durch Invertierung einer Achse auch zur Winkelhalbierenden des II Quadranten parallele Bereiche auswählen (siehe Abb. 7.16).

Die neuen Selektionsbereiche entsprechen im Scatterplot also genau den Bereichen, die man von den Selektionsarten her kennt, mit der Ausnahme, daß man dort nach der Selektion die Variablenachse der entsprechenden Variablen ebenfalls invertieren muß. Diese Erkenntnis läßt sich natürlich auch auf höhere Dimensionen anwenden, was aber wiederum nur schlecht dargestellt werden kann.

7.4 Selektionssequenzen

Es ist einzusehen, daß man mit einfachen Selektionen gewöhnlich keine komplizierte Untergruppen selektieren kann. Dazu sind verschiedene Selektionen notwendig, die hintereinander ausgeführt werden und mit bestimmten Mengenoperationen verknüpft werden. Um dies zu ermöglichen existieren verschiedene Se-

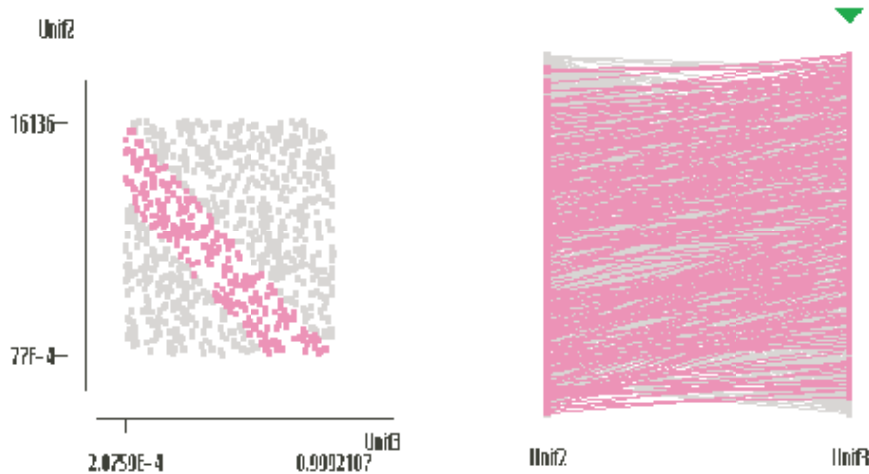


Abbildung 7.16: *Selektion mit Hilfe einer Winkelintervallangabe mit zusätzlicher Transformation*

lektionsmodi. Dabei kann man grundsätzlich 5 Modi unterscheiden (siehe Abb. 7.17).

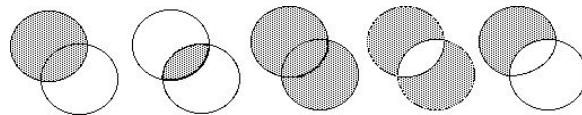


Abbildung 7.17: *Die 5 Selektionsmodi: Ersetzen, Schnitt, Vereinigung, Exklusion und Differenz, wobei die unteren Bereiche die ursprünglichen Selektionsmengen darstellen und die oberen Bereiche jeweils die neue Selektion*

Im Toolfenster (siehe Abb. 6.2) am oberen Rand des Bildschirms kann man durch Klicken den gewünschten Modus einstellen. Dennoch sollte beachtet werden, daß die Differenz nicht implementiert ist, da man im Bereich der Datenanalyse diesen Modus in der Regel nicht benötigt. Hinzu kommt die Tatsache, daß man ebenfalls mit Selektionsverknüpfungen anderer Selektionsmodi und Werkzeugen, wie dem Toggle2, dasselbe Ergebnis erreichen kann.

Standardmäßig ist das **Ersetzen** (Replace-Modus) eingestellt, was im Wesentlichen nur einer Aneinanderreihung einfacher Selektionen entspricht, ohne die vorherigen Selektion zu beachten. Es werden also nur die neu ausgewählten Punkte selektiert und die vorher selektierten Punkte nicht beachtet.

Der **Schnitt** (Intersection-Modus) bedeutet, daß man die Menge der schon selektierten Punkte mit der Menge der neu selektierten Punkte schneidet. Die Ergebnismenge stellt die Menge der Punkte dar, die in den Darstellungen nun markiert werden.

Analog werden bei der **Vereinigung** (Union-Modus) die zwei Selektionsmen-

gen vereinigt.

Ähnlich verhält es sich mit der **Exklusion**, die man ebenfalls auch im Replace-Modus durch Drücken der “Shift”- Taste während einer Selektion erhalten kann. Dabei werden die Punkte, die schon selektiert waren, mit den neuselektierten Punkten verglichen. Die Punkte, die in beiden Mengen vorhanden sind, werden dann deselektiert. Dies entspricht dem \wedge - Operator aus der Informatik bzw. der symmetrischen Differenz in der Mengenalgebra.

Weiterhin ist zu erwähnen, daß es zusätzlich noch die Möglichkeit der kompletten **Deselektion** gibt, indem man einfach mit der Maus auf eine beliebige Darstellung klickt. Diese Methode nennt man “clear”.

Aufgrund der Selektionsart und des Selektionsmodus ergibt sich automatisch bei der Ausführung mehrerer Selektionen eine Aneinanderreihung von Mengenoperationen, also eine Selektionssequenz. Den relevanten Anfang einer solchen Sequenz bildet entweder die erste Selektion nach einem “clear” oder eine Selektion im Replace-Modus, da bei beiden Selektionen die Vergangenheit nicht beachtet wird.

Zusätzlich ist zu bemerken, daß es relativ häufig passiert, daß man eine im Nachhinein ungewollte Selektion oder Deselektion ausführt. Gegen dieses Mißgeschick ist in CASSATT eine Sicherung, das **Undo**, eingebaut. Dies heißt, daß man die letzte Aktion rückgängig machen kann. Einerseits gibt es dafür einen Menüpunkt, und andererseits kann man auch mit einem Shortcut arbeiten.

Weiterhin ist das **Toggle2** ein Bestandteil einer Selektionssequenz, da es ebenfalls eine spezielle Selektion darstellt. Allerdings hat diese Art der Selektion nichts mit dem eingestellten Modus zu tun, da eine Inversion keine Operation zwischen zwei Mengen darstellt.

Meistens erhält man nach einer bestimmten Anzahl von Selektionen eine Untergruppe, die man sich näher betrachten will oder auch mit anderen Untergruppen vergleichen will. Aus diesem Grund möchte ich nun näher auf die Möglichkeiten der Gruppenerstellung und Gruppenbearbeitung in CASSATT eingehen. Da dieses Thema einen sehr großen Bereich darstellt, werde ich diesen Punkt in einem eigenen Kapitel ausführen.

Kapitel 8

Gruppen in CASSATT

8.1 Gruppenerstellung

Eine besonders interessante Menge von Individuen kann man in CASSATT zu einer Gruppe zusammenfassen, indem man einfach im Toolfenster (siehe Abb.6.2) auf den Knopf "Make Group" drückt, während die gewünschten Individuen selektiert sind.

Als Folge davon erscheint ein Dialogfenster, in welchem man eine Farbe und einen Namen für diese Gruppe angeben kann. Akzeptiert man die Eingaben, erscheint im Hauptfenster ein neuer Listeneintrag in der Gruppenliste mit dem Namen und der Anzahl der Gruppenmitglieder. Zusätzlich ist es auch noch möglich, über das Menü im Hauptfenster die Gruppeneigenschaften später nochmals zu bearbeiten (siehe Abb. 8.1).

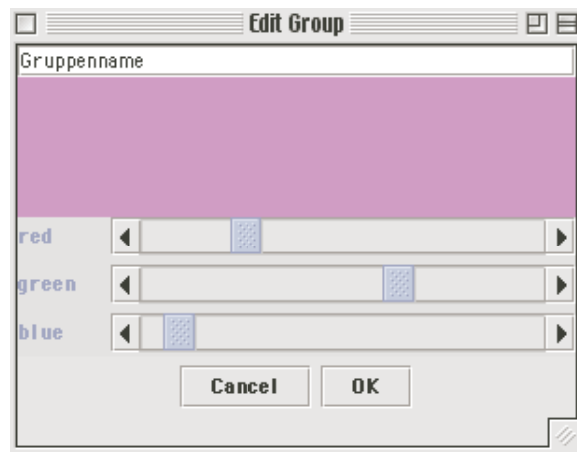


Abbildung 8.1: Dialogfenster zur Gruppenerzeugung und Gruppenbearbeitung

8.2 Gruppenselektion

An dieser Stelle möchte ich noch einen Nachtrag zur Selektion liefern. Aufgrund der Möglichkeit einzelne Gruppen zu erstellen, ergibt sich eine neue Selektionsmöglichkeit. Oft will man beispielsweise schon vorhandene Gruppen selektieren.

Ein sehr umständlicher Weg wäre daher, die spezielle Gruppe erneut auszuwählen. Es bietet sich daher an, die Möglichkeit zu schaffen, direkt eine Gruppe zu selektieren. So kann in CASSATT in der Liste der vorhandenen Gruppen die gewünschte Gruppe mit der Maus selektiert werden.

Es ist aber wichtig, daß aus Konsistenzgründen unbedingt der eingestellte Selektionsmodus beachtet werden muß.

8.3 Gruppen in Parallelen Koordinaten Darstellungen

Eine wichtige Eigenschaft einer Gruppe für die Darstellung ist deren Farbe. Grundsätzlich existieren 2 verschiedene Gruppendarstellungen der Parallelen Koordinaten: Die Darstellung von nur einer Gruppe und die gleichzeitige Darstellung mehrerer Gruppen, wobei die erste Möglichkeit lediglich ein Spezialfall der letzteren Darstellung ist (siehe Abb. 8.2 und 8.3). Die jeweilige Skala richtet sich bei den verschiedenen Darstellungen jeweils nach den Beobachtungen, die zu den ausgewählten Gruppen gehören.

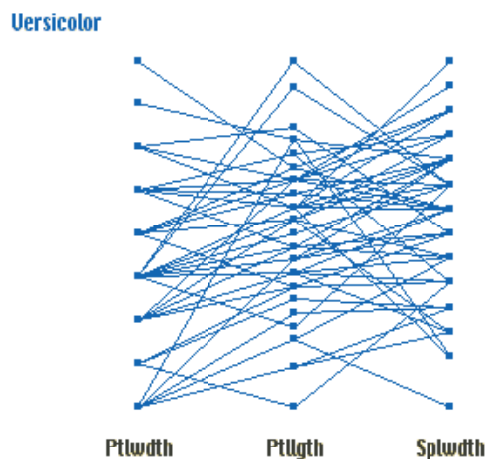
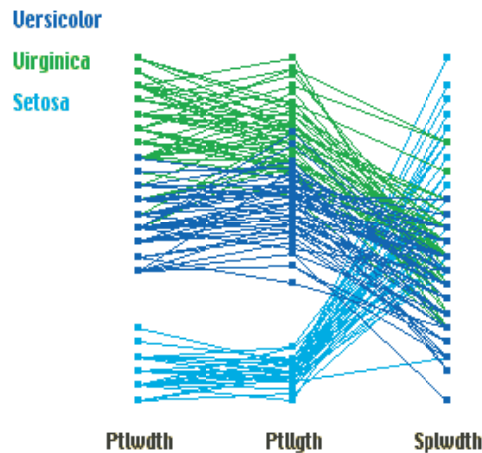


Abbildung 8.2: *Darstellung mit einer Gruppe*

Die entsprechenden Darstellungen kann man bewerkstelligen, indem man nur eine Gruppe bzw. mehrere Gruppen im Hauptfenster auswählt und die Parallele Gruppendarstellung wählt. Man erhält daraufhin einer Parallelen Koordinaten

Abbildung 8.3: *Darstellung mit drei Gruppen*

Darstellung, die allerdings nur die Individuen dieser Gruppe bzw. dieser Gruppen enthält, jedoch alle bisher genannten Funktionen und Eigenschaften einer Parallelen Koordinaten Darstellung beibehält.

Bei der Selektion muß nun beachtet werden, daß man nur noch Punkte selektieren kann, die man auch tatsächlich in der Darstellung sehen kann. Diese Darstellung kann von einer normalen Parallelen Koordinaten Darstellung durch die zugehörigen Gruppennamen in der linken oberen Ecke der Darstellung unterschieden werden.

Um die einzelnen Gruppenmitglieder in einer Darstellung unterscheiden zu können, werden die Gruppen nacheinander, jeweils in ihrer entsprechenden Farbe gezeichnet. Die Gruppe, die als erstes gezeichnet wurde steht in der erwähnten Liste ganz unten usw. Damit ergibt sich jedoch ein schon beim Toggle1 beschriebenes Problem des Overplottings. Die Punkte der verschiedenen Gruppen können sich gegenseitig überdecken, so daß es im Normalfall passiert, daß man nur die als letztes gezeichnete Gruppe komplett erkennen kann.

Daher hat CASSATT die Fähigkeit, die Zeichnungsreihenfolge der einzelnen Gruppen zu verändern (siehe Abb. 8.4). Wie beim Vertauschen der Variablen, kann man in einer Parallelen Gruppen Darstellung per “Drag & Drop” einen Gruppennamen packen und an eine neue Position bringen. Zieht man einen Gruppennamen direkt über einen anderen, werden deren ursprünglichen Positionen vertauscht. Ansonsten wird die Gruppe zwischen die gewünschten Gruppen eingeordnet.

Beim Zeichnen tritt ein anderes, äußerst wichtiges und hilfreiches Phänomen auf. Alle Gruppen, bis auf die zuletzt gezeichnete Gruppe, erhalten eine hellere Farbe, so daß man den Kontrast zu der letzten, also obersten Gruppe besonders deutlich erkennen kann. Auch beim Vertauschen wird dies jedesmal beachtet. Diese Farbabschwächung der hinteren Individuen wird als Ghost Plot bezeichnet,

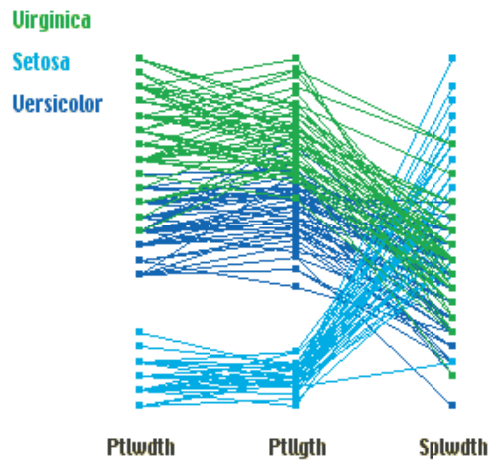


Abbildung 8.4: Darstellung mit drei Gruppen mit geänderter Reihenfolge

da diese Individuen wie Geister im Hintergrund stehen.

Noch eine weitere besondere Eigenschaft verbirgt sich in einem solchen Gruppenfenster. Man erhält ein zusätzliches Abfrageobjekt, die Gruppe. Bei einer derartigen Abfrage erscheint, im Gegensatz zu den temporären Anzeigen der bisherigen Abfragen, ein neues Fenster mit der Information über die bestimmte Gruppe. Wie die erwähnte Gruppeninformation aufgebaut ist, werde ich im nächsten Abschnitt näher beschreiben.

8.4 Gruppeninformation und Selektionsgeschichte

Da eine Gruppe einer bestimmten Menge von bereits selektierten Individuen entspricht, ist es oft wünschenswert im Nachhinein zu erfahren, wie man auf diese spezielle Auswahl gekommen ist. Wichtig ist es also, die dazugehörige Selektionssequenz nachvollziehen zu können. Dazu müssen aber die wichtigsten Informationen aus der zugehörigen Selektionssequenz herausgefiltert werden.

Gleichzeitig ist es sinnvoll dabei redundante Selektionen auszusortieren. Vorerst ist es wichtig, sich zu überlegen, welche besonderen Merkmale eine Selektion ausmachen. Die wichtigsten Merkmale der Selektionen kann man in 5 Punkten beschreiben:

- Darstellungsart, in der selektiert wurde
- Art der Selektion
- betroffene Variablen
- Werte zur Selektion / Skizze

- Selektionsmodus

Diese Merkmale werden dann in CASSATT in einer Matrix eingetragen. Wichtig ist es auf jeden Fall die Darstellungsart, wie z.B. Scatterplot oder Dotplot, zu kennen, in der diese Selektion ausgeführt wurde. Bei einer Parallelen Koordinaten Darstellung sollte noch beachtet werden, ob diese gleichskaliert, standardisiert oder transformiert war. Zusätzlich ist es in einer Parallelen Gruppen Darstellung notwendig, die zugehörigen Gruppen zu kennen.

In CASSATT wird demnach in eine erste Spalte der Informationsmatrix die spezielle Darstellungsart eingegeben. Bei einer Parallelen Koordinaten Darstellung und bei einer Parallelen Gruppen Darstellung wird zusätzlich ein Stern hinter dem Namen angetragen, falls die Darstellung nicht standardisiert war. Bei einer Parallelen Gruppen Darstellung kommen dann noch zusätzlich, wie erwähnt, die Namen der Gruppen hinzu.

Als weiteres muß man über die durchgeführte Selektionsart Bescheid wissen. Es gibt vor allem in einer Parallelen Koordinaten Darstellung verschiedene Arten der Selektion, wie z.B. Winkelselektion, Linienselektion oder Achsensелеktion (siehe Kapitel 7). Da diese Information relativ wichtig ist, wird hierfür ein entsprechender Eintrag in der Informationsmatrix getätigt.

Einen weiteren relevanten Punkt stellen die Variablennamen dar. Daher werden je nach Art der Selektion bzw. Darstellungsart ein oder zwei Variablen angegeben. Bei einer Selektion im Scatterplot wird man beispielsweise zwei Variablen benötigen, bei einer Punktselektion in einer Parallelen Koordinaten Darstellung nur eine Variable. Wichtig ist es auch anzugeben, ob eine verwendete Variablenachse invertiert war. In CASSATT werden daher in der Informationsmatrix nicht nur die betroffenen Variablen, sondern ebenfalls deren Orientierung angezeigt. Ist eine Variable invertiert, so wird dies, wie schon bei der Skalierung, durch einen Stern hinter dem Variablennamen gekennzeichnet.

Um nun zu den wichtigsten Werten der Selektion zu kommen, sollte man sich noch kurz überlegen, welche Art der Darstellung am einleuchtendsten ist. Bei der Winkelselektion ist es eigentlich ausreichend, ein Bild der zwei dazugehörigen Winkel zu zeichnen. Ebenso sollte es ausreichend sein, einen kleinen Dotplot bzw. einen kleinen Scatterplot der neuen Punktselektionen zu erhalten. Bei der Linienselektion in einer Parallelen Koordinaten Darstellung fällt auf, daß es sehr schwierig ist, wichtige Werte anzugeben.

Da eine solche Selektionsgeschichte nur Gedankenhiebe enthalten soll, reicht es aus, eine kleine Parallele Koordinaten Darstellung der betroffenen Variablen mit den selektierten Individuen zu zeichnen. In der 4. Spalte erhält man aus diesem Grund eine kleine Graphik. Diese soll dem Betrachter eine Erinnerungshilfe sein. Um nun diese Skizze der Selektion genauer zu betrachten, besteht die Möglichkeit, wie bei der Abfrage über das Betätigen der Maus in Kombination mit der "Alt"-Taste nochmals Information zu erhalten. Es erscheint daraufhin eine vergrößerte und genauere Darstellung der in der Matrix vorhandenen Graphik. Als weitere

Abfrageebene wäre hier beispielsweise die Angabe der genauen Selektionswerte möglich.

Als letzter, aber besonders wichtiger Faktor, wird in der 5. Spalte der Matrix der zum Zeitpunkt der Selektion eingestellte Selektionsmodus angegeben. Es ist einleuchtend, daß der jeweilige Modus erheblichen Einfluß auf die endgültige Selektion hat. Da dieser Modus der Mengenoperation entspricht, die beim Verknüpfen der Selektion ausgeführt wird.





Scatterplot	Dots	Petalwidth Petallength		REPLACE
Grouplot Versicolor Virginica	Angle	Petalwidth Petallength *		UNION
Grouplot Versicolor Virginica	Lines	Petalwidth Petallength *	Gr1 Gr2 	SCHNITT
Grouplot Versicolor Virginica	Dots	Sepalwidth		UNION

Abbildung 8.5: *Beispiel einer Informationsmatrix*

Toggle2, Undo, Clear

An dieser Stelle muß unbedingt erwähnt werden, daß zu einer Selektion noch andere Methoden gehören, die in der Matrix aufgeführt werden müssen. Dies sind Methoden, die sich direkt auf die selektierten Individuen auswirken. Hier kann noch erwähnt werden, daß es sich bei diesen Methoden um klar definierte Selektionsvorschriften handelt, weshalb der eingestellte Selektionsmodus nicht beachtet werden muß.

In Kapitel 6 wurde das Toggle2, also die Selektionsinvertierung, schon soweit erklärt, daß diese Methode zu den Selektionen gezählt werden kann.

Ebenso beeinflußt das Undo die Selektion insofern, als daß es die letzte Selektion rückgängig macht. Dies bedeutet, daß auf die letzte Selektion keine Rücksicht mehr genommen werden muß. Aus diesem Grund kann man an dieser Stelle ansetzen, die Selektionsliste zu kürzen.

Beim Clear wird automatisch die komplette Selektion gelöscht und der Modus wieder auf “Replace” gestellt. Da dies einer kompletten Deselektion aller Beobachtungen entspricht, startet also eine neue Selektionssequenz. Allerdings muß man hier noch einen wichtigen Punkt beachten: Ein solches “clear” ist nicht immer der neue Startpunkt einer Selektionssequenz, da ein Undo diese Deselektion wieder rückgängig machen kann.

Die drei genannten Methoden spielen eine besondere Rolle in Bezug auf die Selektionssequenzen, weshalb auch diese Informationen in die Selektionsgeschich-

te mit eingehen müssen. Die Methoden clear und Undo müssen zwar nicht in der Informationsmatrix aufgelistet werden, spielen jedoch beim Aufstellen der endgültigen Selektionssequenz eine wichtige Rolle.

Im Wesentlichen kann man mit einer solchen Matrix jede Selektion gedanklich noch einmal nachvollziehen. Dennoch sollte überflüssige Information aus dieser Matrix eliminiert werden, was im nächsten Abschnitt näher erläutert werden soll.

Optimierungen der Selektionssequenz

Man kann sich überlegen, daß während einer Selektionssequenz oft unnütze oder redundante Selektionen vorgenommen werden. Hier spreche ich vor allem von verschiedenen Selektionen im "Replace" - Modus oder Methoden, wie Clear und Undo.

Daher können bei jeder Gruppenerstellung verschiedene Optimierungen der Selektionssequenz vorgenommen werden. Wie schon vorher beim Clear erwähnt, spielt das Undo eine übergeordnete Rolle. Daher kann man auf jeden Fall bei einem ersten Durchgang die Selektionen aus der Liste streichen, die direkt vor einem Undo stehen. Ebenfalls können jeweils die zugehörigen Einträge mit Undo gestrichen werden.

Da eine Selektionssequenz erst mit einer Selektion beginnt, die direkt auf ein Clear erfolgt bzw. deren Modus der Replace-Modus ist, können insbesondere alle Selektionen, die vor der letzten derartigen Selektion stattfanden, gestrichen werden. Es ist daher sinnvoll, diese Optimierungen von hinten an auszuführen.

Wichtig bei diesen Optimierungsschritten ist es allerdings, die genannte Reihenfolge zu beachten. Dabei ist gemeint, daß zuerst die vorhandenen Undos abgearbeitet werden und erst danach den Beginn der Sequenz gesucht wird. So wird die Wahl eines falschen Anfangspunktes vermieden.

Ein kleines Beispiel soll dies verdeutlichen:

Hat man eine Selektionssequenz von drei aufeinanderfolgenden Aktionen, wie beispielsweise folgende Sequenz:

Selektion1 — Selektion2(Replace) — Undo

Eine korrekte Optimierung ergäbe nur noch

Selektion1.

Würde man jedoch zuerst den Optimierungsschritt 2, also die Suche nach dem Sequenzanfang, ausführen, so erhielte fälschlicherweise als Zwischenergebnis Folgendes:

Selektion2 — Undo

Das Ausführen des noch fehlenden Schrittes, der Bearbeitung der Undos, ergäbe ein vollkommen falsches Ergebnis, nämlich, daß keine Selektion mehr vorliegt.

Kapitel 9

Beispiel: Zehnkampfdatensatz

Eine besonders für multivariate Datensätze geeignete Darstellungsform stellen die Parallelen Koordinaten (Inselberg, 1985 und Wegman, 1990) dar. Am Beispiel des Zehnkampfdatensatzes werde ich verschiedene Techniken zur Erkennung verschiedener Datenstrukturen vorstellen. Dabei werde ich auch auf unterschiedliche Möglichkeiten der speziell für diesen Zweck entwickelte Software CASSATT eingehen.

9.1 Datensatz

Die verwendeten Daten beziehen sich auf den Zehnkampfwettbewerb der Herren bei den Olympischen Spielen 1988. Ein solcher Wettbewerb erstreckt sich über zwei Tage, wobei an jedem Tag 5 Disziplinen in vorgeschriebener Reihenfolge durchgeführt werden müssen.

100m Lauf	(100m)
Weitsprung	(LongJump)
Kugelstoßen	(ShotPut)
Hochsprung	(HighJump)
400m Lauf	(400m)
110m Hürden	(110mHurdles)
Diskus	(Discus)
Stabhochsprung	(PoleVault)
Sperwurf	(Javelin)
1500m Lauf	(1500m)

Um die Plazierung der einzelnen Athleten zu bestimmen, werden die Ergebnisse der einzelnen Disziplinen in Punkte umgerechnet und so eine Gesamtpunktzahl errechnet. Die Umrechnung der einzelnen Disziplinen erfolgt über die aktuellen Tabellen aus den Internationalen Leichtathletik Mehrkampf Bestimmungen (IAAF).

Der originale Zehnkampfdatensatz besteht aus den tatsächlichen Leistungsnachweisen der 34 Teilnehmer und befindet sich als Textdatei auf der beiliegenden CD-ROM. Um die direkten Vergleiche zwischen den einzelnen Disziplinen herstellen zu können, müssen die Daten in Punkte umgerechnet werden. Aus diesem Grund wird in diesem Kapitel noch ein weiterer Datensatz für die Analyse verwendet, der die Punktwertungen der Athleten enthält. Dieser Datensatz befindet sich ebenfalls auf der CD-ROM.

Allerdings muß man bemerken, daß die originalen Punktetabellen des Jahres 1988 nicht mehr zu beschaffen waren, so daß für diese Transformation die aktuellen Tabellen verwendet wurden. Dadurch haben sich zwar geringe Änderungen ergeben, die jedoch für das Endergebnis dieser Analyse keine Rolle spielen.

9.2 Methoden

Das Ziel dieser graphischen Datenanalyse ist das Erkennen und Interpretieren von Strukturen und Zusammenhängen. Da es sich hier um einen mehrdimensionalen Datensatz handelt, bietet es sich an mit Parallelen Koordinaten Darstellungen zu arbeiten. Es ist hier von großem Vorteil, daß man mit dieser Darstellung mehrdimensionale Ausreißer, lineare Zusammenhänge oder Ähnlichkeiten erkennen kann (siehe Kapitel 3).

Da es sich um eine graphische Datenanalyse handelt, ist es empfehlenswert mit einer interaktiven Software zu arbeiten. Die folgenden Ausführungen wurden mit Hilfe der interaktiven Software CASSATT gemacht, die insbesondere für diese Art der Datenanalyse entwickelt wurde.

9.3 Analyse

Zu allererst sollte man sich einen Gesamtüberblick über alle Variablen verschaffen. Dazu ist es ratsam eine Parallelen Koordinaten Darstellung mit allen Variablen anzusehen. Bei den Daten mit den Leistungsnachweisen ist allerdings zu beachten, daß unterschiedliche Skalierungen vorliegen. Beispielsweise verhält sich eine Zeitwertung im Normalfall invers zu einer Distanz- oder Höhenmessung. Aus diesem Grund ist es empfehlenswert, die Variablenachsen, die Zeitmessungen angeben, zu invertieren, also “umzudrehen”. Dadurch erhält man eine Darstellung, in der die besseren Leistungen im oberen Bereich zu finden sind.

Blendet man nun zusätzlich die zugehörigen Boxplots auf die Variablenachsen, so kann man feststellen, daß es einen mehrdimensionalen Ausreißer gibt (siehe Abb. 9.1). Bei diesem Athleten handelt es sich um den mit Abstand schlechtesten Teilnehmer dieses Wettkampfes. Aufgrund der Tatsache, daß es sich um einen Ausreißer in fast allen Disziplinen handelt, sollte eine neue Gruppe gebildet werden, die nur die besten 33 Athleten umfaßt.

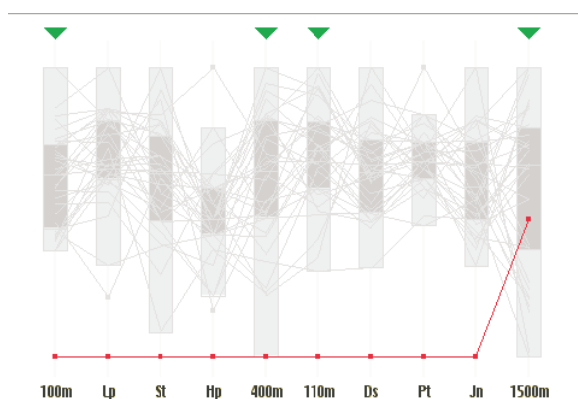


Abbildung 9.1: *Darstellung der Leistungsnachweise - mehrdimensionaler Ausreißer*

Stellt man nun die Athleten dieser Gruppe in einer Parallelen Koordinaten Darstellung dar, so entdeckt man im unteren Wertebereich 2 zweidimensionale Ausreißer (siehe Abb. 9.2). Da diese keine krassen Ausreißer des Gesamtdatensatzes darstellen (siehe Abb. 9.1), werde ich im weiteren Verlauf der Analyse die komplette Gruppe untersuchen.

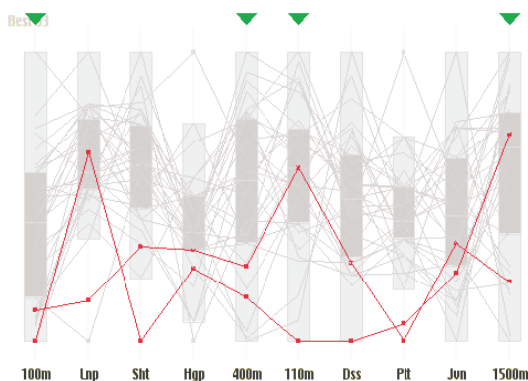


Abbildung 9.2: *Gruppendarstellung - 2 zweidimensionale Ausreißer*

Weiterhin können zwei eindimensionale Ausreißer entdeckt werden, die herausragende Leistungen in jeweils einer Disziplin erbracht haben (siehe Abb. 9.3). Man kann feststellen, daß es sich bei dem einen Ausreißer in der Variable Hochsprung um den endgültigen Sieger handelt, obwohl dieser in den anderen Disziplinen nur im vorderen Mittelfeld zu finden ist. Es ist zu vermuten, daß der Punktevorsprung im Hochsprung ausreichte, um Gesamtsieger zu werden. Ob diese Theorie stimmt, kann man sehr gut in dem Datensatz erkennen, der die einzelnen Punktwertungen enthält. Darauf werde ich später noch genauer eingehen.

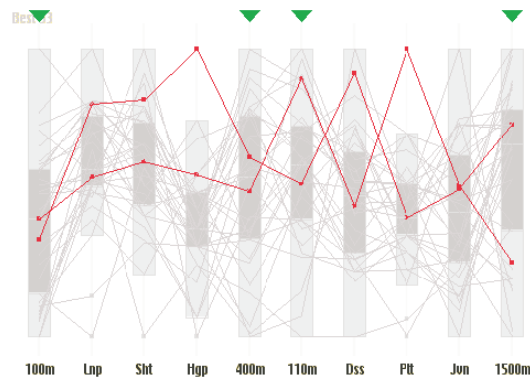


Abbildung 9.3: Auftreten von 2 positiven Ausreißern in den Disziplinen Hochsprung und Stabhochsprung

Der zweite Ausreißer in der Variable Stabhochsprung kommt trotz dieses Vorsprungs nur auf den 6. Platz. Aufgrund des Linienzuges dieses Sportlers im Vergleich zum Sieger, kann man erkennen, daß die Ursachen in anderen Disziplinen, wie 1500 m Lauf, Weitsprung (Lnp) oder auch Diskus, zu suchen sind.

Um jedoch genauere Schlußfolgerungen ziehen zu können, müssen die Daten allerdings vergleichbar sein. Dies erreicht man, indem man die zugehörigen Punktwertungen betrachtet. Von Vorteil ist zusätzlich, daß durch diese Transformation des Datensatzes auch die Inversion der einzelnen Achsen wegfällt.

Wie schon bei den Originaldaten werden wir eine Gruppe mit den 33 besten Athleten erzeugen und uns diese in einer Parallele Koordinaten Darstellung betrachten. Der Vorteil der Punktwertung ist, daß man die Daten gleichskaliert betrachten kann. So kann man feststellen, daß der schon vorher erwähnte Ausreißer im Hochsprung, also der Sieger, einen geringeren Punktevorsprung (mit 1061 Punkten nur 117 Punkte Differenz zum Nächstbesten) in dieser Disziplin hat, als der Ausreißer im Stabhochsprung (mit 1132 Punkten 160 Punkte Differenz zum Nächstbesten) (siehe Abb. 9.4).

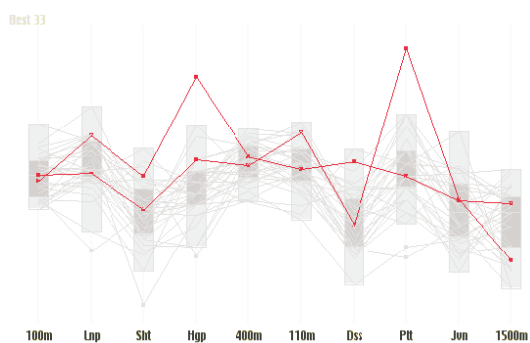


Abbildung 9.4: Gleichskalierte Darstellung – positive Ausreißer

Zusätzlich kann man bei der vorliegenden Gleichskalierung gut erkennen, daß der Sieger allgemein höhere Punktwertungen erhält und keine Ausreißer nach unten aufweist (siehe Abb. 9.4).

Die 10 Disziplinen werden bei jedem Wettkampf in einer festgelegten Reihenfolge durchgeführt. Dadurch ist es möglich, den zeitlichen Verlauf zu untersuchen.

Zwei Fragestellungen können an dieser Stelle untersucht werden:

- Hat die Zeit allgemein einen Einfluß auf die Sportler?
- Gibt es einen Sportler, der einen zeitlichen Einbruch oder Verletzungen hatte?

Hier lohnt es sich wiederum, mit den Punktbewertungen zu arbeiten, da die originalen Leistungsbewertungen nur den Vergleich zwischen den Athleten aufzeigen kann. Um Vergleiche zwischen den Sportarten anzustellen, werden die einzelnen Boxplots eingebildet (siehe Abb. 9.5).

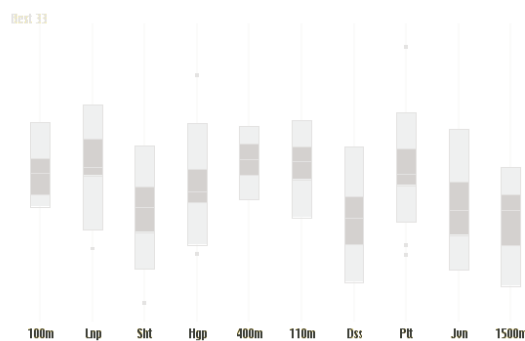


Abbildung 9.5: *Untersuchung des zeitlichen Verlaufs aller Sportarten*

Hier kann man keine eindeutigen Aussagen über den zeitlichen Verlauf machen. Weder verschlechtern sich die Punktwertungen von einem auf den nächsten Tag, noch kann man ein derartiges Phänomen innerhalb eines Tages erkennen (siehe Abb. 9.5). Die Punktwertungen der einzelnen Sportarten unterscheiden sich zwar untereinander, aber dies liegt wahrscheinlich an den Sportarten selber.

Um nun die zweite Fragestellung zu überprüfen, wird eine Parallele Koordinaten Darstellung in der Liniendarstellung erzeugt. Dabei kann man nun interaktiv verschiedene Sportler selektieren, die in einzelnen Disziplinen, die vor allem gegen Ende des Wettkampfes stattfanden, schlecht abgeschnitten haben (siehe Abb. 9.6 und 9.7).

Bei dem vorliegenden Datensatz kann man zwar verschiedene Sportler ausmachen, die sich im Vergleich zu den anderen Sportlern verschlechtern. Diese Tatsache kann aber nicht in Zusammenhang mit dem zeitlichen Verlauf gebracht werden, da solche Verschlechterungen nur in einzelnen Disziplinen auftreten, ohne daß die nachfolgenden Disziplinen mit beeinflußt werden. Auch gibt es keine Athleten, die ab einem bestimmten Zeitpunkt nur noch schlechte Leistungen zeigen, was auf Verletzungen oder ähnliches hinweisen könnte.

Wie man bei diesen Untersuchungen gemerkt hat, scheint es so zu sein, daß

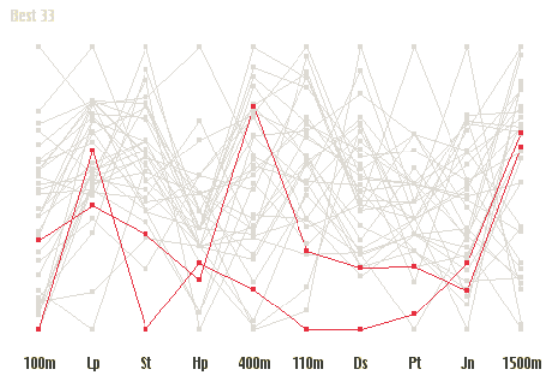


Abbildung 9.6: Untersuchung des zeitlichen Einflusses auf einzelne Sportler – 1

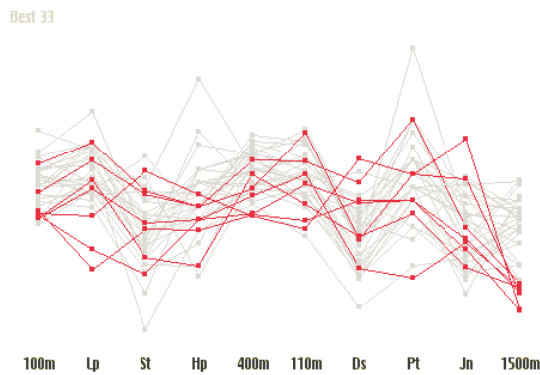


Abbildung 9.7: Untersuchung des zeitlichen Einflusses auf einzelne Sportler – 2

die einzelnen Disziplinen unterschiedliche Bewertungen aufweisen. Daher wäre es interessant den Einfluß der einzelnen Disziplin auf das Gesamtergebnis zu untersuchen. Hierzu lohnt es sich die Disziplinen nach dem Median zu sortieren (siehe Abb. 9.8).

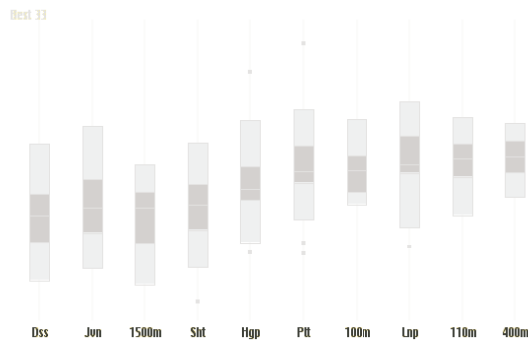


Abbildung 9.8: Anordnung der Variablen nach dem Median

Aus dieser Darstellung läßt sich ersehen, welche Disziplinen höhere Punktwertungen als andere Disziplinen ergeben. Man erkennt, daß die Punktwertungen in allen Wurfdisziplinen und im 1500 m Lauf eher gering ausfallen. Dagegen erhalten aber die Athleten bei den Kurz- und Mittelstrecken der Laufdisziplinen und beim Weitsprung relativ hohe Bewertungen (siehe Abb. 9.8).

Hier muß allerdings noch die Variabilität mit berücksichtigt werden. Besitzt eine Disziplin eine geringe Variabilität, so kann man davon ausgehen, daß diese Disziplin keinen großen Einfluß auf die Gesamtpunktzahl haben kann. Daher kann man die Variablen nach der Standardabweichung anordnen und untersuchen, welche Disziplinen dabei die größte bzw. kleinste Variabilität aufweisen (siehe Abb. 9.9).

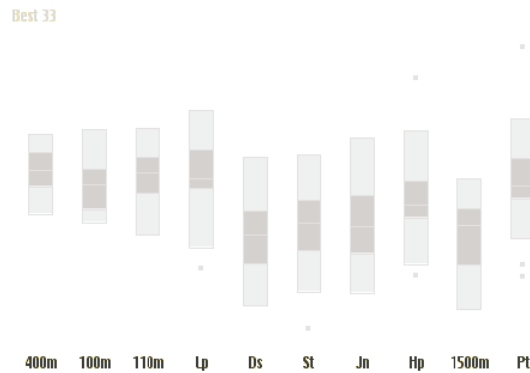


Abbildung 9.9: Anordnung der Variablen nach der Standardabweichung

Beispiele für eine sehr geringe Variabilität sind die Kurz- und Mittelstrecken der Laufdisziplin. Wie aus der Anordnung nach Punktwertungen (siehe Abb. 9.8) zu erkennen war, stellen diese Disziplinen auch Disziplinen dar, deren Punktwertungen relativ hoch sind. Das bedeutet demnach, daß in diesen Disziplinen zwar hohe Punktzahlen erreicht werden, durch die geringe Variabilität allerdings jeder Athlet relativ viele Punkte erhält. Dies führt auch dazu, daß in solchen Disziplinen keine bedeutenden Punktvorsprünge gemacht werden können.

Anders verhält es sich mit den Variablen Stabhochsprung, 1500m und Hochsprung. Bei diesen Sportarten werden relativ hohe Punktvorsprünge erzielt. Jedoch kann man dazu erwähnen, daß sich der Einfluß des 1500 m Laufes auf die Gesamtpunktzahl durch die geringen Punktwertungen etwas verringert (siehe Abb. 9.8).

Insgesamt muß man aber bemerken, daß die Variabilität mehr Einfluß auf die Gesamtpunktzahl hat, als der Mittelwert der Disziplinen. Zusammenfassend kann man also sagen, daß die Variablen Stabhochsprung und Hochsprung einen großen Einfluß, die Wurfdisziplinen dagegen eher einen geringen Einfluß ausüben.

Überprüfen kann man diese Ergebnis, indem man einen kleinen Teil der besonders guten Athleten selektiert und deren Punktverteilungen in den einzelnen

Disziplinen mit denen aller Sportler vergleicht. Man stellt fest, daß diese Athleten tatsächlich in den erwähnten Disziplinen im Vergleich zur Gesamtheit besser abgeschnitten haben (siehe Abb. 9.10).

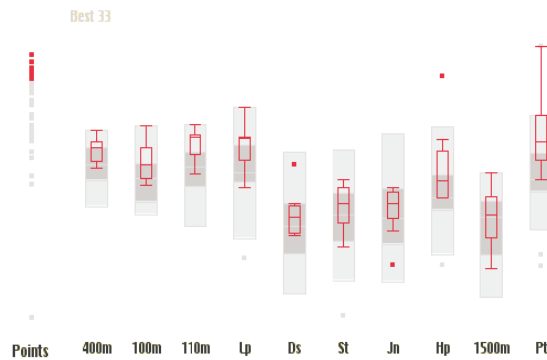


Abbildung 9.10: Markierung der besten Sportler zur Überprüfung einer Hypothese

Neben den unterschiedlichen Einflüssen der Disziplinen auf die Gesamtpunktzahl ist es ebenfalls interessant, Zusammenhänge zwischen einzelnen Sportarten festzustellen. Rein intuitiv vermutet man einen Zusammenhang jeweils zwischen den Lauf-, Sprung-, und Wurfdisziplinen. Aus diesem Grund kann man die einzelnen Disziplinen in der genannten Reihenfolge anordnen. Hierbei will ich auch darauf eingehen, warum verschiedene Skalierungen betrachtet werden sollten. Sieht man sich die so geordneten Sportarten in einer nach Minimum und Maximum skalierten Darstellung an, so erkennt man einen Athleten, der in den Wurfdisziplinen herausragende Leistungen erbracht hat (siehe Abb. 9.11). Bei diesem Sportler fällt auf, daß er in den Laufsportarten sehr schlecht im Vergleich zu den anderen Athleten abschneidet, bei den Laufsportarten im Mittelfeld liegt und bei den Wurfdisziplinen der Spitzengruppe angehört. In der gleichskalierten Darstellung hätte man diesen Sportler nicht als Sonderfall ausmachen können (siehe Abb. 9.12).

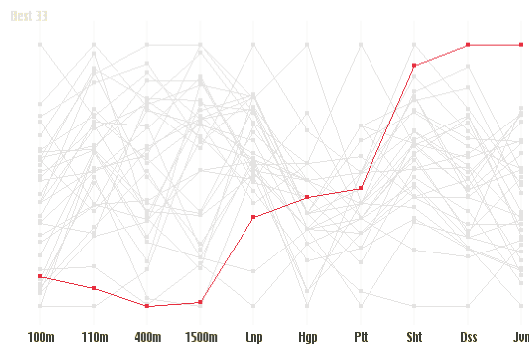


Abbildung 9.11: Besonderer Linienzug eines Athleten

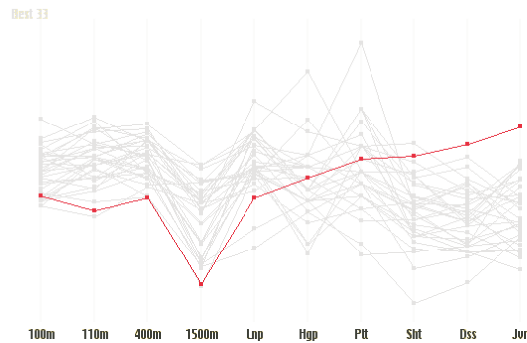


Abbildung 9.12: Gleichskalierte Darstellung von Abb. 9.11

Nun ist es ebenfalls interessant zu erfahren, ob Zusammenhänge zwischen den Sportarten bestehen. Zwischen den Laufdisziplinen 100 m, 110 m und 400 m kann man relativ gut ähnliche Strukturen der Linienzüge erkennen, was auf positive Zusammenhänge schließen läßt (siehe Abb. 9.13). Allerdings erkennt man, daß es zwischen dem 1500 m Lauf und dem 100 m Lauf sowohl parallele Linien als auch Knoten gibt. Dies bedeutet, daß diese Variablen unkorreliert sind (siehe Abb. 9.13).

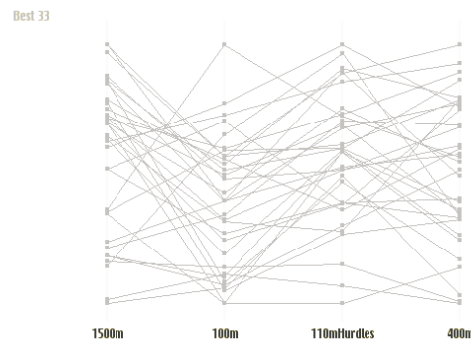


Abbildung 9.13: Untersuchung linearer Zusammenhänge zwischen den Laufdisziplinen

Ebenso kann man zwischen den Sprungdisziplinen (siehe Abb. 9.14) und den Wurfdisziplinen (siehe Abb. 9.15) lineare Zusammenhänge erkennen. Diese Korrelationen sind höher als die bei den Laufsportarten erkannte Korrelation, da weniger Linien vom Haupttrend abweichen.

Zum Schluß kann noch auf ähnliche Fälle eingegangen werden. Damit sind Sportler gemeint, die in verschiedenen Disziplinen vergleichbare Leistungen erbringen. Durch die Möglichkeit der aneinander gereihten Winkelselektionen zwischen Variablen, kann man einige ähnliche Sporttypen finden (siehe Abb. 9.16).

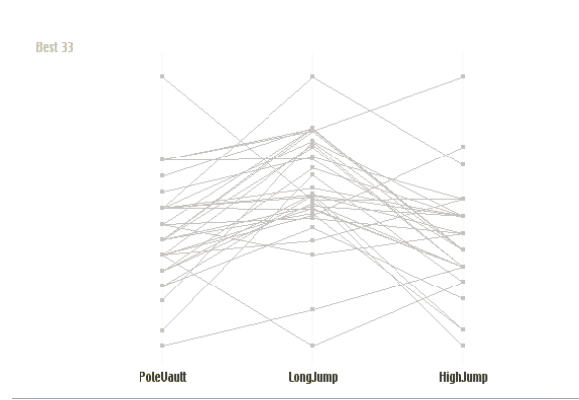


Abbildung 9.14: dreifache Korrelation zwischen den Sprungdisziplinen

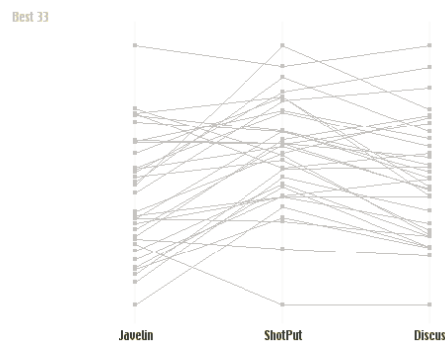


Abbildung 9.15: dreifache Korrelation bei den Wurfdisziplinen



Abbildung 9.16: Erkennen ähnlicher Sportler

9.4 Ausblick

Um noch mehr über die einzelnen Athleten zu erfahren, speziell welche Sportler ähnliche Stärken in einzelnen Sportarten aufweisen, kann man diese Datensatzmatrix transponieren. Die neuen Variablen sind die einzelnen Sportler, mit den entsprechenden Punkteinträgen der 10 Disziplinen in den Zeilen. Erstellt man

nun eine parallele Koordinaten Darstellung, so erhält man einen völlig neuen Blickwinkel auf diese Daten (siehe Abb. 9.17).

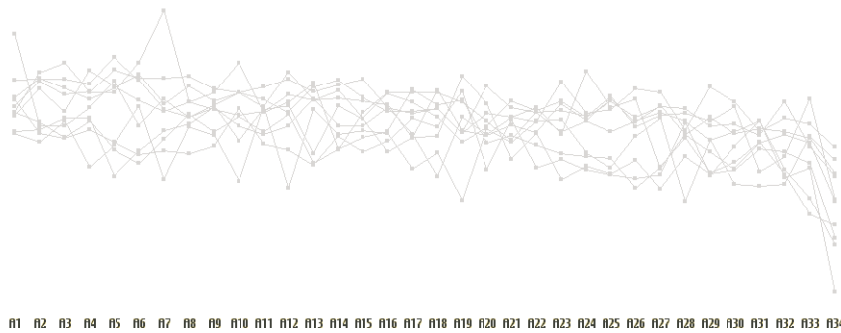


Abbildung 9.17: *Parallele Koordinaten Darstellung des transponierten Datensatzes*

Man kann nun die einzelnen Sportler in Bezug auf die Disziplinen vergleichen, da jeder Linienzug einer Disziplin entspricht. Da die Sportler in der Reihenfolge ihrer Platzierungen geordnet sind, erkennt man gut, wo welcher Athlet Punkte gesammelt hat. In dieser Darstellung fällt wiederum der Athlet auf Platz 34 als schlechter Ausreißer auf.

Als weiteren Punkt kann man in diesem Datensatz verschiedene Sportler finden, deren Leistungen in einigen Sportarten ähnlich sind. damit sind Athleten gemeint, deren stärkere und schwächere Disziplinen aus den selben Bereichen stammen. Dies ist, wie schon vorher erwähnt, mit Hilfe aneinander gereihter Selektionen und Variablenumordnungen möglich (siehe Abb. 9.18)

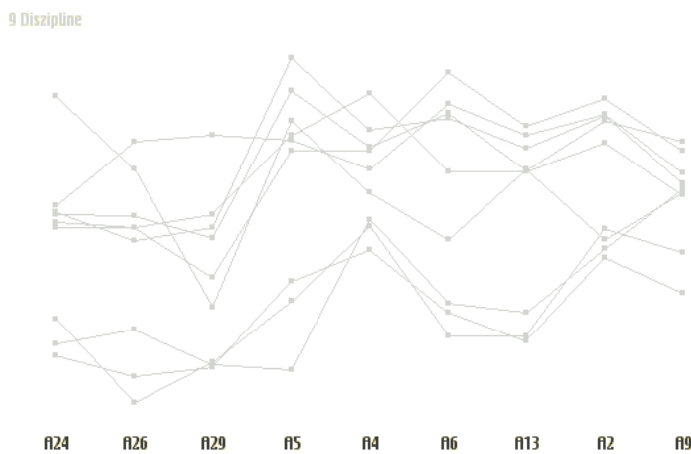


Abbildung 9.18: *Ähnliche Sportler im transponierten Datensatz*

9.5 Zusammenfassung

Als erstes Ergebnis kann man festhalten, daß es verschiedene Ausreißer gibt. So gibt es einen Sportler, der in 8 Disziplinen besonders schlechte Leistungen erbracht hat. Im Gegensatz dazu gibt es noch 2 Sportler, die jeweils besonders gute Leistungen in einer Disziplin gezeigt haben. Beide Athleten weisen sonst keine herausragenden Leistungen auf. Dennoch erweist sich der eine als endgültiger Sieger, was eher auf konstantere Leistungen während des gesamten Wettkampferlaufs zurückzuführen ist.

Bei der Untersuchung nach zeitlichen Zusammenhängen muß man feststellen, daß es dabei keine Besonderheiten gibt. Weder findet man einzelne Sportler, die sich zeitlich verbesserten oder verschlechterten, noch kann man einen allgemeinen zeitlichen Trend finden.

Ein verblüffendes Ergebnis liefert die Untersuchung der Punktwertungen. Dabei gibt es zwischen den Disziplinen sowohl starke Unterschiede bei der Punktzahl als auch bei der Variabilität. Man kann feststellen, daß die Disziplinen unterschiedliche Einflüsse auf die Gesamtpunktzahl haben. Die Einflüsse der einzelnen Disziplinen beruhen hauptsächlich auf der Kombination aus den durchschnittlichen Punkthöhen und den Variabilitäten der Punktbewertungen. Dabei erweisen sich die Disziplinen Stabhochsprung und Hochsprung als Disziplinen mit hohem Einfluß und die Wurfdisziplinen als Disziplinen mit geringerem Einfluß.

Bei der Untersuchung von Korrelationen zwischen einzelnen Sportarten kann man festhalten, daß es zwischen den Wurf- und Sprungdisziplinen relativ starke Zusammenhänge gibt. Bei den Laufdisziplinen muß man dagegen nochmals zwischen Kurz- bzw. Mitteldistanz und der Langstrecke unterscheiden, da alle Laufdisziplinen bis auf den 1500 m Lauf miteinander korreliert sind.

Als letzten Punkt kann man bemerken, daß man verschiedene Sportler finden kann, die in den einzelnen Sportarten ähnliche Leistungen erbringen, was die These unterstützt, daß es verschiedene Sportlertypen gibt.

An dieser Stelle möchte ich aber noch darauf hinweisen, daß die gewonnenen Ergebnisse noch nicht verallgemeinert werden sollten. Es sollten vorher noch die Zehnkampfergebnisse anderer Jahre untersucht werden, um eine größere Menge an Daten zu erhalten.

Kapitel 10

Zusammenfassung und Ausblick

I doubt if you know the effort to paint! The concentrations it requires to compose your picture, the difficulty of posing the models, of choosing the color scheme, of expressing the sentiment and telling your story! The trying and trying again and again, and oh, the failures, when you have to begin all over again! The long months spent in effort upon effort, making sketch after sketch After a time, you get keyed up and it “goes”, you paint quickly and do more in a few weeks than in the preceding weary months. When I am ‘en train’ nothing can stop me and it seems easy to paint, but I know very well it is the result of my previous efforts.

Louisine Havemeyer reports what Mary Cassatt said about painting. (zitiert in ...)

10.1 Zusammenfassung

Die vorliegende Arbeit gibt eine Zusammenfassung der Theorie über Parallele Koordinaten. Weiterhin werden Zusammenhänge zu anderen Graphiken erklärt und ausführlich beschrieben. Es wird auf einfache Datenstrukturen eingegangen und eine Übertragung der Ergebnisse auf mehrere Dimensionen geschaffen.

Der zweite Teil geht eher auf die Entwicklung der Software CASSATT ein. Dabei werden vorerst die Anforderungen an die Software näher beschrieben. Speziell wird das Thema Interaktivität angesprochen. Darauf folgt ein Überblick über die wichtigsten interaktiven Methoden, die die Software in Parallelen Koordinaten verwirklichen soll. Anhand von Beispielen wird aufgezeigt, welches Ziel diese Methoden verfolgen. Ebenfalls wird auch auf erweiterte Fähigkeiten eingegangen, die in Parallelen Koordinaten Darstellungen möglich sind. Ein großes Thema stellt dabei die Selektion dar. Es werden spezielle Möglichkeiten zur Selektion in Bezug auf Parallele Koordinaten dargestellt und ausführlich erläutert. Man bekommt einen Überblick über die Auswirkungen, die die jeweiligen Selektionsarten mit

sich bringen. Beim Übergang auf Selektionssequenzen kommt man schnell auf das Thema Gruppen. Dabei wird darauf eingegangen, wie Gruppen erstellt werden können und welche Möglichkeiten sich in den Parallelen Koordinaten Darstellungen damit ergeben. Bei der Darlegung der Gruppeneigenschaften wird das Thema der Selektionsgeschichte erörtert und erklärt, wie dies in CASSATT ausgeführt wird.

Im letzten Teil wird eine Datenanalyse mit Hilfe der Parallelen Koordinaten durchgeführt. Dabei bekommt man einen Überblick, wie man eine multivariate Datenanalyse ausführen und die Ergebnisse interpretieren kann.

10.2 Ausblick

Obwohl die Software CASSATT schon sehr viele Methoden implementiert hat, kann man sich noch weitere Funktionen überlegen, die noch nicht verwirklicht sind.

Es handelt sich einerseits um relativ einfache graphische Dinge, wie beispielsweise der Rotationplot bei der Winkelselektion. Weiterhin wurden auch schon Skalierungsmethoden, wie Zoom, erwähnt. Dabei wäre es auch noch interessant, eine Möglichkeit bereit zu stellen, wodurch die Skalierung einzelner Variablenachsen ermöglicht wird.

In Bezug auf die Sortierungen wäre es eventuell noch interessant, möglichst gute Anordnungen, eventuell mit Hilfe der einzelnen Korrelationen, zu finden. Auch kann man sich eine automatische Achsensortierung gut vorstellen, bei der der Benutzer wie bei einem Film die verschiedenen Anordnungen gezeigt bekommt.

Ein ganz anderer Punkt ist die Verarbeitung von kategoriellen Variablen in den Parallelen Koordinaten. Hier kann man sich gut vorstellen, daß die Größe der Punkte in der Parallelen Koordinaten Darstellung direkt proportional zur Individuenanzahl der entsprechenden Kategorie ist. Man erhält also einen Dotplot, dessen Punkte verschiedene Größen besitzen. Da bei Kategorien die Reihenfolge keine Rolle spielt, sollte die Möglichkeit geschaffen werden, die zugehörigen Punkte wie auch die Variablenachsen vertauschen zu können.

Als eine andere Erweiterungsart bieten sich analytische Methoden, wie Regression, an. Wie diese genau verwirklicht und dargestellt werden kann, muß man sich allerdings noch genauer überlegen.

Zum Schluß will ich noch erwähnen, daß nur die Möglichkeit besteht, vollständige Datensätze aus Dateien einzulesen. Dies wirft zwei Punkte auf, die man noch verbessern könnte. Einerseits sollte man eventuell einen Datenbankanschluß schaffen, wodurch es möglich wäre, auch ganze Datenbanken einzulesen. Wichtiger ist vielleicht noch die Tatsache, daß auch unvollständige Datensätze eingelesen werden können, da vollständige Datensätze eher eine Seltenheit darstellen. Die Darstellung von derartigen Daten in Parallelen Koordinaten Darstellung ist im

Wesentlichen nicht komplizierter als in einfachen Dotplots.

Mit der Implementierung derartiger Methoden könnte CASSATT zu einer konkurrenzfähigen Software heranwachsen, da die vollständige Bearbeitung von realen Datensätzen ermöglicht und unterstützt wird.

Anhang A

Shortcuts und Befehle in CASSATT

A.1 Datensätze

Da es sich bei CASSATT um eine Forschungssoftware handelt, muß man die Datensätze darauf prüfen, daß diese keine fehlenden Werte oder Leerzeichen in einzelnen Wörtern enthalten. Weist ein Datensatz einen Fehler auf, so erscheint beim Einlesen eine Warnung und der Einlesevorgang wird abgebrochen.

A.2 Shortcuts

DARSTELLUNG VERÄNDERN:

Linien hinzufügen:	Shift & L
Linien entfernen:	Shift & C
Boxen einblenden:	Shift & B
Punkte einblenden:	Shift & D
Linien und Boxen einblenden:	Shift & A
Konfidenzintervall hinzufügen:	Shift & K
Konfidenzintervall entfernen:	k
Variablenabstand vergrößern:	Shift & +
Variablenabstand verkleinern:	Shift & -
Punkte vergrößern:	+
Punkte verkleinern:	-

SELEKTION:

Toggle (Version 1):	Shift & T
Letzte Selektion rückgängig:	Shift & Z
Punktselektion - 1:	An der Achse Punkte mit der Maus auswählen - Curser beachten

Punktselektion - exklusiv:	Shift & wie Punktselektion - 1 im Replace - Modus
Linienselektion (Pinch)- 1:	zwischen den Achsen selektieren
Linienselektion (Pinch)- exklusiv:	Shift & zwischen den Achsen selektieren
Linienselektion (Dragbox)- 1:	r & zwischen den Achsen selektieren
Linienselektion (Dragbox)- exklusiv:	Shift & r & zwischen den Achsen selektieren
Winkelselektion - 1:	Mausklick links unter dem Variablennamens - Cursor beachten
Winkelselektion - exklusiv:	Shift & wie Winkelselektion - 1 im Replace - Modus

SONSTIGE INTERAKTIVEN METHODEN:

Toggle (Version 2):	Shift & S
Invertierung einer Achse:	Mausklick rechts unter dem Variablennamens - Cursor beachten
Abfrage:	Alt & Mausclick
Menüleiste entfernen:	m
Menüleiste hinzufügen:	Shift & M

A.3 Menübefehle

MENÜLEISTE IM HAUPTFENSTER

File:	Datensatz öffnen Datensatz schließen Programm beenden
Plots:	Dotplots erstellen Boxplots erstellen Scatterplots erstellen Parallele Koordinaten Darstellungen erstellen Parallele Gruppen Darstellungen erstellen
Selection & Group:	Gruppeneigenschaft bearbeiten Gruppeninformation abfragen Letzte Selektion rückgängig machen

MENÜLEISTE IM PARALLELEN KOORDINATEN DARSTELLUNGEN

Edit:	Farbe der nicht-selektierten Individuen abändern
-------	--

	Selektionsfarbe abändern Nicht-selektierte Fälle verstecken Selektierte Fälle verstecken Alle Fälle anzeigen
Selection:	Replace - Modus einstellen Exklusions - Modus einstellen Vereinigungs - Modus einstellen Schnitt - Modus einstellen Letzte Selektion rückgängig machen
Variables:	Achsen nach verschiedenen Kriterien sortieren Achsen nach verschiedenen Kriterien der selektierten Fälle sortieren Reihenfolge umdrehen
Scale:	Variablen nach Minimum und Maximum der jeweiligen Achse skalieren Variablen gleichskalieren Variablen transformieren, so daß Mittelwert auf einer Höhe liegt
Help:	Liste der Shortcuts anzeigen lassen

Literaturverzeichnis

- Andrews D.: *Plots of high-dimensional data*. Biometrics, 28, pp. 125-136, 1972.
- Avidan T. & Avidan S.: *ParallAX - A data mining tool based on parallel coordinates*. Computational Statistics, 14, pp. 79-89, 1999.
- Baecker R.M., Grudin J., Buxton W.A.S. & Greenberg S.: *Readings in Human-Computer Interaction: Toward the Year 2000*. Morgan Kaufmann Publishers, Inc., San Francisco, CA, 1995.
- Bassett E.W.: *Ibm's ibm fix*. Industrial Computing, 14(41), pp. 23-25, 1995.
- Geßler J.R.: *Statistische Graphik*. Birkhäuser Verlag, Basel, 1993.
- Inselberg A.: *Don't panic ... just do it parallel*. Computational Statistics, 14, pp. 53-77, 1999.
- Inselberg A.: *The Plane with Parallel Coordinates*. The Visual Computer, pp. 69-91, 1985.
- Inselberg A. & Dimsdale B.: *Parallel Coordinates: A Tool For Visualizing Multidimensional Geometry*. Proceedings of the First IEEE Conference on Visualization, pp. 361-378, 1990.
- Inselberg A. & Avidan T.: *The Automated Multidimensional Detective*.
- Klinke S.: *Data Structures for Computational Statistics*. Physica-Verlag Heidelberg, 1997.
- Larman C.: *Applying UML And Patterns – An Introduction To Object-Oriented Analysis And Design*. Prentice-Hall, PTR, NJ, 1998.
- Miller J.J. & Wegman E. J.: *Construction of line densities for parallel coordinate plots*. Computing and Graphics in Statistics, Springer-Verlag, pp. 107-123, 1991.
- Partsch H.A.: *Specification and Transformation of Programs*. Springer Verlag, Berlin Heidelberg, 1990.

- Schnell R.: *Graphisch gestützte Datenanalyse*. Oldenbourg Verlag, München, 1994.
- Theus M.: *Theorie und Anwendung Interaktiver Statistischer Graphik*. Augsburger mathematisch-naturwissenschaftliche Schriften, 14, Wißner Verlag, Augsburg, 1996.
- Unwin A.R.: *Requirements for interactive graphics software for exploratory data analysis*. Computational Statistics, 14, pp. 7-22, 1999.
- Wegman E. J.: *Hyperdimensional Data Analysis Using Parallel Coordinates*. Journal of the American Statistical Association, Vol. 411(85), pp. 664-675, 1990.
- Wegman E. J.: *The grand tour in k-dimensions*. Computing Science and Statistics: Proceedings of the 22nd Symposium on the Interface, Interface Foundation of North America, pp.127-136, 1991.
- Wegman E. J. & Luo Q.: *High Dimensional Clustering Using Parallel Coordinates and Grand Tour*. Computing Science and Statistics, 28, 352-360, 1997.
- Wilhelm A.F.X.: *Interactive Statistical Graphics: The Paradigm of Linked Views*. Habilitationsschrift, Mathematisch-Naturwissenschaftliche Fakultät der Universität Augsburg, 1999.
- Wilhelm A.F.X. & Wegman E. J. & Symanzik J.: *Visual clustering and classification: The Oronsay particle size data set revisited*. Computational Statistics, 14, pp. 109-146, 1999.
- Wilkinson L.: *The Grammar Of Graphics*. Springer Verlag, New York, Inc, 1999.
- Wills G.: *Nicheworks – interactive visualisation of very large graphs*. Journal of Computational and Graphical Statistics. forthcoming, 1999b.
- Wills G.: *Selection: 524,288 Ways to Say “This is Interesting”*. Proceedings of Visualisation ‘96’, 1996.
- Internetverweise:
- INSTITUT FR DEUTSCHE GEBÄRDENSPRACHE UND KOMMUNIKATION GEHÖRLOSER *Psychologie - Fachgebärdenlexikon* <http://www.signlang.uni-hamburg.de/Projekte/PsychLex.html>.
- Schmidtner R.: *Osinet*. <http://paedglo.psychol.uni-giessen.de/osinet/paedagog/instrukt/cul/ADAPINT2.HTM>.

Schwenke J.R. & Fergen B.J.: *Graphical Techniques for Displaying Multivariate Data Using SAS/GRAPH Software*
<http://www.sas.com/service/library/periodicals/obs/obswww22/>.

Winkler S.: *CASSATT* <http://www1.math.uni-augsburg.de/Cassatt/>.