Institut für Visualisierung und Interaktive Systeme

Universität Stuttgart
Universitätsstraße 38
D–70569 Stuttgart

Masterarbeit Nr. 125

Interaktive, visuelle Fehleranalyse für die Chip- und Schaltkreisüberprüfung

Lan Jiang

Studiengang:	Softwaretechnik
Prüfer/in:	Prof. Dr. Thomas Ertl
Betreuer/in:	Dr. Steffen Koch, Dipl.-Phys. Qi Han, M. Sc. Markus John
Beginn am:	20. Oktober 2016
Beendet am:	21. April 2017
CR-Nummer:	D.2.2, H.5.2

Kurzfassung

Mit immer komplexeren modernen Computerchips und Schaltungen sind die Überprüfungen der Hardware in der Industrie anspruchsvoll und unerlässlich. Um die Qualität der Hardware zu gewährleisten werden regelmäßig große Mengen von Chipüberprüfungen durchgeführt. Dabei entstehen hochdimensionale Datensätze mit einer großen Menge von Testfällen, welche analysiert werden können, um Einsichten zur Verbesserung der Hardware zu gewinnen. Ein Ansatz zur Analyse dieser Daten existiert noch nicht.
In Rahmen dieser Masterarbeit wurde das `Chip Testing Error Detection System (CTEDS)` für die Analyse der hochdimensionalen Datensätze mit Testfällen umgesetzt. Mit dem System sollen die potenziellen Fehlerquellen der in den Testfällen aufgetauchten Fehler festgestellt werden. Das System ermöglicht einen analytischen Prozess der multivariaten Datensätzen sowie drei interaktive Ansichten für die Darstellung der Datenelemente.
Die Ansichten bieten drei Visualisierungstechniken jeweils für die Übersichtsdarstellung der Daten, die Korrelationsanalyse der Parameter und die Darstellung der hochdimensionalen Strukturen an. Dabei wurden geeignete Interaktionen entwickelt: Freie Selektion der Datenpunkte, Transformation der Datenelemente von einer Visualisierung in eine andere Visualisierung und dynamische Generierung der Korrelationen zwischen einem Parameter-Tupel. Die drei Ansichten sind mithilfe der Technik *Bruhsing-Linking* verknüpft. Die Kombination der interaktiven Visualisierungen ermöglicht eine effiziente visuelle Korrelationsanalyse bezüglich vorgegebener Fehler. Es ist dadurch möglich, potenzielle Fehlerquellen zu erkennen.
Das System dient als eine interaktive Darstellungsplattform sowohl für die Darstellung der Beziehungen zwischen Parametern als auch für die kausale Analyse der Fehler.

Abstract

With more complex modern computer chips and circuits, the testing of hardware in the industry is challenging and indispensable. In order to ensure the quality of the hardware, a large amount of chip testing is carried out in a short period of time. This results in high-dimensional data sets with a large amount of test cases, which can be analyzed to obtain insights for improving the hardware. An approach which helps users analyzing and making sense of this type of data is still missing.
Within the scope of this master thesis the `Chip Testing Error Detection System (CTEDS)` was implemented to analyze high-dimensional data sets with test cases. This system is designed to identify the potential sources of errors that are found in the test cases. It enables an interactive analytical process of the multivariate data sets, interactive views for the representation of different aspects of the data.
The views offer three visualization techniques for showing the overview of the data, conducting correlation analysis of the parameters and displaying high-dimensional structures. Furthermore, appropriate interactions are developed for the views: free selection of data points, transformation of data elements from one visualization into another visualization and on demand visualization of correlations between different parameters. Finally, all three views are connected by brushing and linking technique. The combination of the interactive visualizations allows an efficient visual correlation analysis regarding given errors. Potential sources of errors are thereby recognizable.
This system serves as an interactive presentation platform for the representation of relationship among the parameters as well as for the causal analysis of the errors.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Verzeichnis der Algorithmen

1 Einleitung

Die rasante Entwicklung der Informationstechnologie erzeugt riesige Datenmengen mit zahlreichen Attributen. Die hochdimensionalen Datensätze bieten enorme Möglichkeiten, Verhaltensmuster der Daten zu untersuchen und künftige Entwicklungen vorherzusagen. Umfassende Einsichten kommen häufig aus komplizierten Zusammenhängen unter Datenattributen. In zahlreichen Forschungsgebieten wird großer Wert auf die Untersuchung der Zusammenhänge gelegt. Dabei spielt die Analyse heterogener Daten eine entscheidende Rolle.

Neben der Wichtigkeit bei den wissenschaftlichen Forschungen gewinnt die multivariate Datenanalyse in der Industrie auch zunehmend an Bedeutung. Mit immer komplexeren modernen Computerchips und Schaltungen sind die Überprüfungen solcher Hardware in der Industrie anspruchsvoll und unerlässlich. Um die Qualität der Hardware zu gewährleisten werden regelmäßig Chipüberprüfungen durchgeführt. Dazu werden große Mengen von Chipüberprüfungen durchgeführt. Dabei entstehen hochdimensionale Datensätze mit vielen Testfällen.
Die Datensätze enthalten normalerweise eine große Menge von heterogenen Parametern, die bei jedem Testfall vorkommen. Um wesentliche Zusammenhänge unter den Parametern der Testfälle zu gewinnen ist ein gutes Verständnis der hochdimensionalen Daten erforderlich. Jedoch übersteigt hochdimensionaler Raum die menschliche Vorstellungskraft. Deswegen werden wirksame Werkzeuge benötigt, um das Verständnis zu verstärken. Eine übliche Möglichkeit dafür liegt in der Anwendung von Visualisierungstechniken, mit denen die Elemente der Daten graphisch dargestellt werden. Da es keine einfache Abbildung der mehrfachen Dimensionen auf einen zweidimensionalen Raum gibt, werden einerseits anspruchsvollere Visualisierungstechniken für multivariate Daten als standardmäßige Diagramme benötigt. Andererseits sind effiziente Interaktionstechniken notwendig, um die Daten zu manipulieren und versteckte Zusammenhänge aufzudecken.
Unter diesem Gesichtspunkt ist Visual Analytics in der Lage, Visualisierungen und Interaktionen zusammenzusetzen, um Erkenntnisse aus großen und komplexen Datensätzen zu gewinnen. Visual Analytics beschäftigt sich hauptsächlich mit der Kopplung von interaktiven visuellen Darstellungen mit zugrundeliegenden analytischen Prozessen. Dabei werden Techniken aus der Informationsvisualisierung und aus der Datenanalyse angewendet.

In einer Zusammenarbeit mit Advantest[1], dem weltweit größten Anbieter von automatischen Prüfgeräten für die Halbleiterindustrie, soll ein Visual-Analytics-Ansatz für die Analyse der hochdimensionalen Datensätzen mit Testfällen umgesetzt werden. Damit sollen die potenziellen Fehlerquellen der in den Testfällen aufgetauchten Fehler festgestellt werden. Eine Übersicht der erkannten Fehler unter Berücksichtigung ihrer Beziehung zum hochdimensionalen Parameterraum sollte ermöglicht werden.

[1] Advantest: https://www.advantest.com/home

In dieser Masterarbeit wird die Entwicklung eines Visual-Analytics-Systems beschrieben: Das `Chip Testing Error Detection System (CTEDS)`. Basierend auf der Java-Bibliothek für Informationsvisualisierung *prefuse* wird das System `CTEDS` als eine Java-Anwendung implementiert, das plattformübergreifend funktioniert und von jedem Rechner aus jederzeit durchgeführt werden kann. Die Eingangsdatensätze für das System können entweder große CSV-Datei oder TSV-Datei sein, die mehrere Zeilen für Testfälle und mehrere Spalten für Parameter enthalten.

Das System beschäftigt sich mit einem analytischen Prozess eines multivariaten Datensatzes, drei verschiedenen visuellen Darstellungen der Datenelementen und der Manipulation der Daten mithilfe von Interaktionstechniken. Bei dem analytischen Prozess des Datensatzes werden die Rohdaten nach Anforderungen bearbeitet, gefiltert und in Datenstrukturen abgespeichert. Mit den Datenstrukturen werden in dem System drei Visualisierungstechniken jeweils für Übersichtsdarstellung, Korrelationsanalyse und hochdimensionale Strukturen eingesetzt. Die Visualisierungen bilden ein Teil der direkten Schnittstelle zwischen Nutzer und Computer. Dabei handelt es sich auch um eine Explorationsvisualisierung auf dem Datensatz. Mittels geeigneter Interaktionen kann der Nutzer den Analyseprozess steuern und auf die Ergebnisse Einfluss ausüben. Zu den möglichen Interaktionstechniken zählen Filterung, Selektion, dynamische Anfragengenerierung und Clustering.

Da das System auf die Feststellung der Fehlerquellen abzielt, kommt die Korrelationsanalyse unter Parametern eine hohe Bedeutung zu. Eine Anforderung daran bezieht sich auf geeignete interaktive grafische Darstellungen der Parameter und ihrer Korrelationen. Die Visualisierungen sollen dem Nutzer die Möglichkeit geben, einen Überblick über die Daten und die Parameter schnell zu gewinnen und die zugrundeliegenden Beziehungen zu erkennen. Dabei werden die Korrelationen anhand von den Eigenschaften der Parameter berechnet.
Eine weitere Anforderung an das System ist die effiziente und interaktive Verknüpfung der erstellten visuellen Ansichten. Demgemäß zählt das System zu dem Verfahren Multiple-Coordinated-Views [Rob07], dass die Änderungen der visuellen Elemente in einer Ansicht die visuellen Elemente anderer Ansichten beeinflussen können.

Das System wird anschließend mit einem von Advantest angebotenen Datensatz ausgewertet. Dabei werden die Ergebnisse zur Umsetzung der Visualisierungen und Interaktionen erläutert, ob die Anforderungen an das System erfüllt werden. Zusätzlich werden die möglichen Einschränkungen und Schwachstellen des Systems verdeutlicht und diskutiert, die im Rahmen dieser Masterarbeit nicht beseitigt werden können.

Gliederung

Die Arbeit ist in folgender Weise gegliedert:

Kapitel 2 - Grundlagen: Hier werden die Grundlagen für diese Arbeit beschrieben. In diesem Kapitel werden Visual Analytics, Informationsvisualisierung und multivariate Datenanalyse grundsätzlich erläutert.

Kapitel 3 - Verwandte Arbeiten und Verwendetes Framework: In diesem Kapitel werden die verwandten Arbeiten vorgestellt, die sich mit der Thematik der interaktiven Visualisierung der

multivariaten Daten befassen. In einem anschließenden Abschnitt wird das verwendete Framework vorgestellt.

Kapitel 4 - Forschungsfragen und Systementwurf: Hier handelt es um den Entwurf des Systems. Zunächst werden eine Reihe von Forschungsfragen erläutert. Sie bilden die Anforderungen an das System und bilden damit die Grundlage für den Systementwurf. Anschließend wird der Systementwurf vorgestellt. Dabei werden der Arbeitsablauf des Systems und die Benutzeroberfläche des Systems erläutert.

Kapitel 5 - Implementierung und Visualisierung: Der Entwicklungsprozess des Systems und die interaktiven Visualisierungen werden in diesem Kapitel detailliert vorgestellt. Die Funktionalitäten des Systems werden hauptsächlich erläutert.

Kapitel 6 - Auswertung: Eine Auswertung des umgesetzten Systems ist in diesem Kapitel zu finden. Dabei wird der verwendete Testdatensatz zuerst vorgestellt. Die Ergebnisse des Systems werden durch einen Anwendungsfall im Anschluss vorgestellt.

Kapitel 7 - Zusammenfassung und Ausblick: Die vorliegende Arbeit wird hier zusammengefasst und ein Ausblick über künftige Arbeiten wird vorgestellt.

2 Grundlagen

In diesem Kapitel werden die Grundlagen für die Arbeit erläutert. Zuerst wird eine grundsätzliche Erklärung zu Visual Analytics (VA) in Abschnitt 2.1 vorgestellt. Anschließend wird das Konzept über Informationsvisualisierung einschließlich der Zusammenhänge mit Visual Analytics in Abschnitt 2.2 erklärt. Zuletzt erfolgt eine allgemeine Beschreibung der multivariaten Datenanalyse in Abschnitt 2.3. Dabei werden die Verfahren der Dimensionsreduktion generell erklärt.

2.1 Visual Analytics

Um wichtige Informationen aus extrem großen und komplexen Datensätzen zu gewinnen stehen zahlreiche Methoden zur Verfügung, dabei wird häufig Visual Analytics (VA) eingesetzt. Im Gegensatz zu reiner Informationsvisualisierung bietet VA dem Nutzer die Möglichkeit, eigene Einflüsse auf die automatisch generierten Ergebnisse zu haben. Keim et al. haben in [SKNP04] einen Ansatz beschrieben, bei dem die menschliche visuelle Wahrnehmung bei der Untersuchung und Analyse der großen Datensätze eine bedeutende Rolle spielt. Aus diesem Grund sind Interaktionsmethoden ein wichtiger Bestandteil von diesem Ansatz, womit man den Analyseprozess über die Daten und Informationen steuern kann. Heutzutage können viele notwendige Informationsanalysen aus großen Datenmengen mittels VA gewonnen und optimiert werden. Daher untersucht die Forschung um VA umfangreiche interdisziplinäre Aspekte von der Datenanalyse bis hinzu Informationsvisualisierung und Menschen-Computer-Interaktionen.

Visual Analytics besteht aus mehreren Komponenten, die sich gegenseitig beeinflussen und zusammenhängen können. Dadurch werden die wichtigen Ergebnisse und Kenntnisse aus den Daten generiert. Der Visual-Analytics-Prozess besteht aus vier Komponenten: Daten, Visualisierung, Modelle und Kenntnisse. Der Prozess ist in Abbildung 2.1 zu betrachten.
Wie bei allen Anwendungsfällen sollten die heterogenen Daten vor der visuellen oder automatischen Analyse integriert werden. Deswegen ist die Datenbearbeitung der erste Schritt des ganzen Prozesses für weitere sinnvolle Analysen. Die eingegangen Daten werden dementsprechend in geeignete Formen transformiert, damit verschiedene Darstellungen aus den Daten erzeugt werden können. In diesem Schritt werden die originalen Daten gegebenenfalls bereinigt, normalisiert oder gruppiert.
Nach der Transformation und Bearbeitung der Daten sind zwei Möglichkeiten für weitere Analysen verfügbar. Entweder werden die Daten durch visuelle Methoden gemappt oder durch automatische Methoden analysiert. Die Möglichkeiten führen zu unterschiedlichen Darstellungen der Daten. Jedoch stehen die Darstellungen nicht unabhängig voneinander. Den Schritt von den Daten zur Visualisierung bezeichnet man als "Mapping", wo die Parameter der Daten visuell repräsentiert werden. Eine Eigenschaft von der visuellen Methode liegt darin, dass die Anwender Interaktionen mit der transformierten Daten durchführen können, indem sie die Parameter der Daten oder Algorithmen modifizieren. Der

Data-Mining-Schritt der automatischen Methode generiert Modelle aus den Daten, die ebenfalls durch Interaktionen der Anwender verbessert werden können. Der Wechsel zwischen den visuellen und automatischen Methoden ist charakteristisch für den Visual-Analytics-Prozess. Somit führt er zur kontinuierlichen Verfeinerung der vorläufigen Ergebnisse. Im Visual-Analytics-Prozess kann man aus Visualisierung, automatischer Analyse sowie den vorangegangenen Interaktionen zwischen Visualisierungen, Modellen und den Anwendern wichtige Wissen gewinnen.

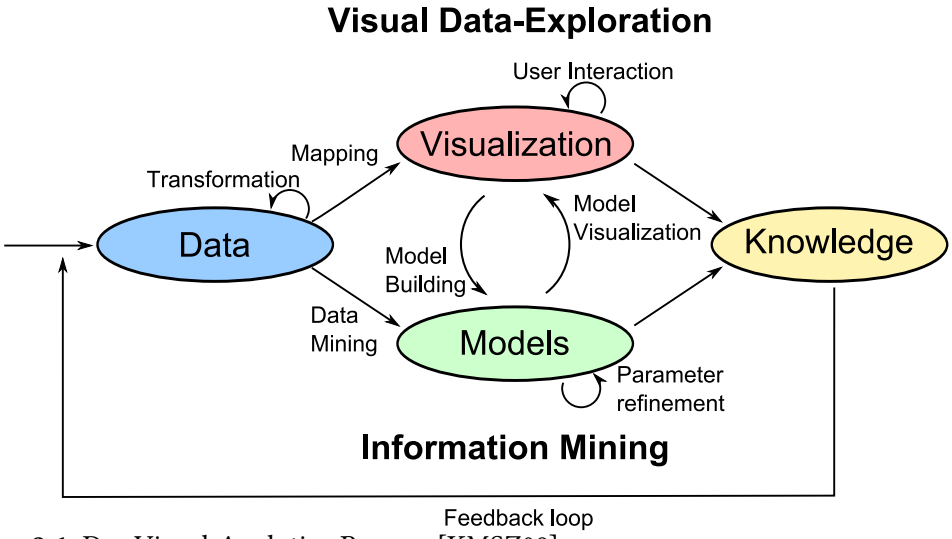


Abbildung 2.1: Der Visual-Analytics-Prozess [KMSZ09]

Mit dem schnellen Wachstum der digitalen Daten aus mehreren Quellen kommt die Weiterentwicklung der Datenanalyse durch Visual Analytics eine hohe Bedeutung zu. In [KMSZ06] haben Keim et al. über die technischen Herausforderungen des Forschungsgebietes diskutiert. Eine Lösung wird dabei zu den Herausforderungen erhoben, die einen effizienteren und präziseren Ansatz der Visual Analytics ermöglichen kann.

Nach [KMSZ06] kann Visual Analytics als ein integrierter Ansatz betrachtet werden. Damit werden Visualisierung, menschliche Faktoren und Datenanalyse kombiniert. Abbildung 2.2 stellt den detaillierten Umfang des Visual-Analytics. Von der Abbildung ist es auszugehen, dass sich Informationsanalytik, Geoanalytik und wissenschaftlicher Analytik in Visual Analytics integrieren. Die Skalierbarkeit ist eine zentrale Herausforderung des Visual Analytics, da sie die Fähigkeit der Verarbeitung großer Datensätze mittels des Rechenaufwands und angemessener Renderingverfahren bestimmt. Eine weitere Herausforderung in Visual Analytics ist die Interpretierbarkeit, welche als eine Fähigkeit bezeichnet wird, Daten zu erkennen und zu verstehen. Außerdem ist die Benutzerakzeptanz auch eine Herausforderung. Die Visual-Analytics-Anwendungen müssen zu den Anforderungen der zukünftigen Anwender passen, um mögliche Nutzungsbarrieren zu überwinden und das volle Potenzial des Visual Analytics zu erschließen.
Aufgrund von den oben erwähnten Herausforderungen umfasst daher der Visual-Analytics-Prozess die Anwendung von automatischen Analysemethoden vor und nach der interaktiven visuellen Darstellung. Deswegen haben Keim et al. in [KMSZ06] eine Lösung zu den Problemen vorgestellt, die

ein neues Visual-Analytics-Mantra definiert: *Analyse First - Show the Important - Zoom, Filter and Analyse Further - Details on Demand.* Im Unterschied zu dem Mantra von Shneiderman [Shn96] werden Daten unter dem neuen Mantra von Keim et al. zuerst analysiert und dann dargestellt. Ohne die Datenanalyse ist die Visualisierung der Rohdaten undurchführbar und kein Kenntnis kann daraus gewonnen werden.

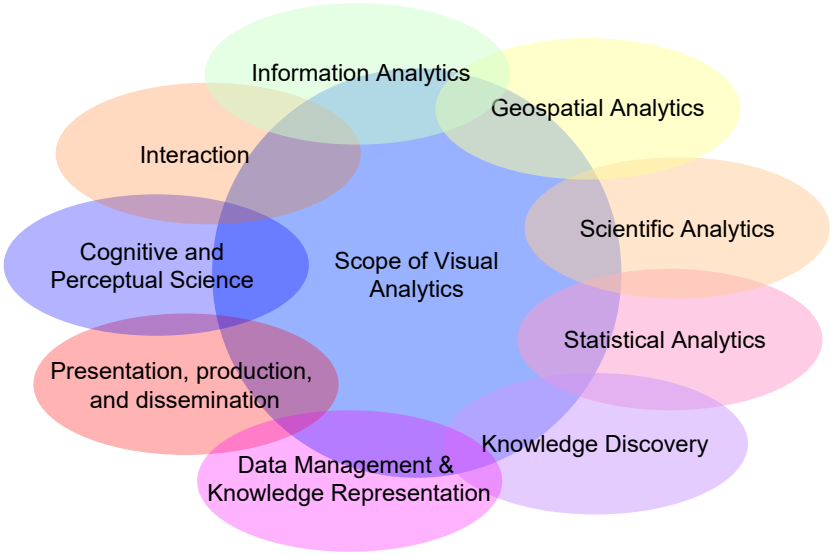


Abbildung 2.2: Visual-Analytics als ein stark interdisziplinäres Forschungsgebiet [KMSZ06]

2.2 Informationsvisualisierung

Informationsvisualisierung ist ein Forschungsgebiet, das sich mit der grafischen Repräsentation großer Mengen von Daten beschäftigt. Die grafischen Darstellungsmethoden sollen dazu beitragen, die Daten auszuwerten und aus ihnen neue Erkenntnisse zu gewinnen. Im Vergleich zu Visual Analytics befasst sich Informationsvisualisierung abstrakte Datenstrukturen wie Bäume oder Graphen. Visual Analytics beschäftigt sich besonders mit der Kopplung von interaktiven visuellen Darstellungen mit zugrundeliegenden analytischen Prozessen, sodass komplexe Aktivitäten effektiv durchgeführt werden können [SB03].

Informationsvisualisierung kann generell mittels eines Referenzmodells in Abbildung 2.3 beschrieben werden. Dabei wird der Visualisierungsprozess in vier Schritte zerlegt: Quelldaten, Datentabellen, visuelle Strukturen und Ansichten. Um die Quelldaten grafisch darzustellen werden mehrere Transformationen dazu verwendet. Die Rohdaten werden zuerst durch eine Datentransformation in Datentabellen transformiert. Dabei lassen sich die Daten filtern, vereinigen und extrahieren. Die Datentabellen werden in visuelle Strukturen umgewandelt, indem die Datenstrukturen in visuellen Variablen abzubilden sind. Das endgültige Rendering verwandelt die visuellen Strukturen in ein Bild in einer Ansicht. Dabei werden die visuellen Perspektiven wie Zooming, Panning modifiziert. Die gesamten Transformationen werden mit mehreren spezifischen Parametern durchgeführt. Somit werden die Interaktionen während des ganzen Prozesses so eingeführt, dass der Nutzer die jeweilige

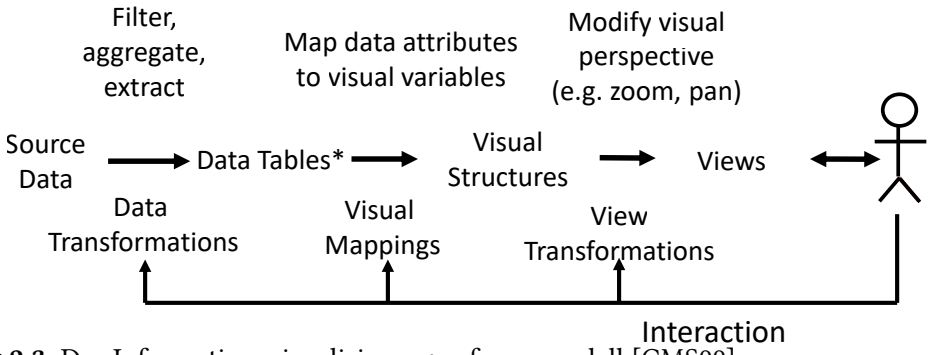


Abbildung 2.3: Das Informationsvisualisierungsreferenzmodell [CMS99]

Transformation ansteuern kann. Die Änderungen durch Interaktionen werden dem Nutzer sofort angezeigt.

Nach [MW14] gilt dieses Referenzmodell generell sowohl für die Netzwerk- als auch für die multivariate Visualisierung. Viele Interaktionstechniken wurden entwickelt, die speziell für die Eigenschaften der Datentypen sind. Card et al. hat in [CMS99] behauptet, dass Informationsvisualisierung Teil der direkten Schnittstelle zwischen Nutzer und Computer ist, mit der die menschlichen kognitiven Fähigkeiten verstärkt werden können. Durch die Darstellung einer großen Menge von Daten in einem kleinem Raum werden die Suchaufwände verringert. Mit visuellen Informationen, die durch ihre Zeitbeziehungen organisiert sind, werden potenzielle Muster besser erkannt. Die Zusammenhänge innerhalb der Daten werden durch Informationsvisualisierung deutlich dargestellt. Im Gegensatz zu statischen Diagrammen werden die Erforschung der Parameterwerte durch manipulierbare Visualisierungen ermöglicht.

2.3 Multivariate Dantenanalyse

Multivariate Daten sind Daten, deren Analyse auf mehrere unterschiedlichen Variablen basiert. Bei der Messung mehrerer Variablen auf einer komplexen experimentellen Einheit ist es oft notwendig, die Variablen gleichzeitig zu analysieren, anstatt sie zu isolieren und einzeln zu betrachten, weil die Variablen meistens miteinander korreliert sind. Multivariate Analysen ermöglichen es, die gemeinsame Leistung solcher Variablen zu erforschen und die Wirkung jeder Variable gegenüber den anderen zu bestimmen.

Wie alle anderen Messungen haben die Variablen der multivariaten Daten ebenfalls Skalenniveaus, welche die Variablen in unterschiedlichen Kategorien unterscheiden. Es kann häufig sein, dass die Variablen auf einer gleichen Weise gemessen werden, wie beispielsweise die Daten 1 und 2 in der Tabelle 2.1. Jedoch enthalten die meisten multivariaten Daten nicht nur Variablen mit gleichem Skalenniveau, sondern auch andere Skalenniveaus wie die letzten zwei Daten 3 und 4 in Tabelle 2.1. Aufgrund der Merkmalen der multivariaten Daten werden entsprechende Analysemethoden zu den multivariaten Daten angewendet. Zu dem Zweck der Analysemethoden hat Alvin Rencher [Ren12] beschrieben, dass die Vereinfachung der Daten die hauptsächliche Aufgabe der Methoden sein sollte,

damit die durch korrelierte Variablen bereitgestellten überlappenden Informationen der multivariaten Daten entwirrt werden und die grundlegende Struktur der Daten entdeckt wird.

Dateneinheiten	**Variablen**
1. Studenten	Mehrere Prüfungsnoten von einer Vorlesung
2. Studenten	Prüfungsnoten von verschiedenen Fächern
3. Mensch	Größe, Gewicht, Körperfettanteil, IQ
4. Hergestellte Artikel	Verschiedene Messungen zur Überprüfung der Einhaltung der Spezifikation

Tabelle 2.1: Beispiele für multivariate Daten [Ren12]

Die multivariate Datenanalyse kann hauptsächlich in zwei Bereiche unterteilt werden. Michel Mellinger [Mel87] hat erläutert, dass die multivariate Datenanalysemethoden in Faktorenanalyse und Klassifikationsverfahren gegliedert werden können.

Faktorenanalyse bezieht sich auf die Berechnung der neuen Variablen aus den originalen Variablen der Daten. Die neuen Variablen werden in dem Fall als Faktoren bezeichnet, welche die lineare Kombination der originalen Variablen sind. Daher handelt es sich bei der Faktorenanalyse um die Reduzierung der Redundanz zwischen den Variablen, indem eine kleinere Anzahl von Faktoren angewendet werden [Shi12]. Um die Faktoren zu berechnen, müssen die Daten zuerst in eine Matrix umgewandelt werden und deren Eigenwerte und Eigenvektoren zusammen die Faktoren definieren. Unterschiedliche Transformationen von Daten in Matrix verursachen unterschiedliche Faktorenanalysemethoden. Die Hauptkomponentenanalyse (PCA) [SWG87] hat eine Korrelationsmatrix der Variablen für die Berechnung der Faktoren [Mel87]. Mit homogenen Variablen, die gleiche Varianz haben, kann Korvarianzmatrix auch in PCA angewendet werden. Die Korrespondenzanalyse (CA) [Gre07] ist ein Verfahren von Faktorenanalyse, mit dem die Beziehungen der Variablen einer Kontingenztafel graphisch repräsentiert werden.

Mellinger [Mel87] beschreibt Klassifikationsverfahren ebenfalls als Clusteranalyse, im Gegensatz dazu haben Skiker [Shi12], Rencher [Ren12] und Timm [Tim02] behauptet, dass Clusteranalyse und Klassifikationsverfahren wesentliche Unterschiede haben. In Klassifikationsverfahren werden die Dateneinheiten einer bekannten Anzahl vordefinierter Gruppen oder Populationen zugewiesen, während in Clusteranalyse weder die Anzahl der Gruppen noch die Gruppen selbst im Voraus bekannt sind [Ren12]. Jedoch hat die Clusteranalyse das Ziel, ein Klassifikationsschema zu entwickeln, das die Dateneinheiten in k verschiedene Gruppen (Cluster) zu teilen [Tim02]. Zur Aufdeckung der Cluster der Daten muss eine Näherungsmessung definiert werden, die entweder als Grad der Entfernung oder als Assoziationsgrad definiert werden. Die Entscheidung der Näherungsmessung hängt von dem Skalenniveau der Variablen − Nominalskala, Ordinalskala, Intervallskala, Verhältnisskala − und den Variablentypen − kontinuierlich und kategorisch − ab. Die Näherungsmessung wird als eine $N \times N$ Matrix dargestellt, die Einträge der Matrix repräsentieren die Unähnlichkeiten D_{rs} oder die Ähnlichkeiten S_{rs} zwischen den r-ten und s-ten Dateneinheiten. Obwohl die Clusteranalyse die Ähnlichkeiten zwischen den Dateneinheiten numerisch darstellen kann, ist eine grafische Darstellung der Dateneinheiten in einer niedrigen Dimension durch Clusteranalyse jedoch nicht leicht durchzuführen. Zur Lösung der Probleme werden Methoden der Dimensionsreduktion angewendet.

Dimensionsreduktion

Dimensionsreduktion ist ein Prozess der Verringerung der Anzahl der Zufallsvariablen unter Berücksichtigung, indem man eine Menge von Hauptvariablen erhält. Für die multivariate Datenanalyse ist die Dimensionsreduktion ein wichtiges Verfahren um die grundlegenden Informationen und Strukturen der Daten zu erfahren, indem die hohen Dimensionen der Daten in niedrige Dimensionen projiziert werden. Mehrere Methoden und Verfahren in dem Bereich wurden entwickelt, beispielsweise zählen die Verfahren wie Hauptkomponentenanalyse (PCA) [SWG87], Unabhängigkeitsanalyse (ICA) [Com94] zu den klassischen Methoden der Dimensionsreduktion. Roweis et al. [RS00] hat ein anderes Verfahren mit dem Namen Locally-Linear-Embedding (LLE) vorgestellt, mit dem es unnötig wäre, paarweise Abstände zwischen weit voneinander getrennten Datenpunkten abzuschätzen. Im Gegensatz zu früheren Methoden behebt LLE globale nichtlineare Struktur von lokal linearen Zuständen.

Multidimensionale Skalierung

Multidimensionale Skalierung (MDS) [CC08] ist ein genereller Begriff für Techniken zur Konstruktion einer Abbildung für allgemein mehrdimensionale Daten in eine niedrige Dimension in Bezug auf die gegebene paarweise Näherungsmessung. Meistens wird MDS zur Dimensionsreduktion verwendet, um mehrdimensionale Daten in den euklidischen niedrig-dimensionalen Raum zu visualisieren [BCQF10]. Wie bei der Clusteranalyse berechnet MDS die Näherungsmessung ebenfalls als eine $N \times N$ Matrix mit N Dateneinheiten. Die Einträge der Matrix sind in dem Fall D_{rs}, welche die Unähnlichkeiten zwischen den Dateneinheiten darstellen. Die Matrix ist symmetrisch $D_{rs} = D_{sr}$, nicht-negativ $D_{rs} > 0$, die diagonalen Elemente der Matrix sind null $D_{ii} = 0$. Das Ziel der MDS-Techniken besteht darin, eine Konfiguration der gegebenen hoch-dimensionalen Daten in einen niedrigdimensionalen euklidischen Raum zu konstruieren. Der Abstand zwischen einem Paar von Punkten wird in der Konfiguration dem entsprechenden Unähnlichkeitswert so weit wie möglich angenähert. Die Ausgabe der MDS Algorithmen könnte als eine $N \times L$ Matrix dargestellt werden. Die Zeilen der Matrix sind die Datenpunkte x_i mit $i = 0 \cdots N$ in einem L-dimensionalen Raum. Für $L = 2$ hat jeder Datenpunkt eine $2D$-Koordinate, die grafisch in Streudiagramm repräsentiert werden kann.

3 Verwandte Arbeiten und Verwendetes Framework

In diesem Kapitel werden die verwandten Arbeiten sowie die verwendeten Technologien zur Entwicklung des Zielsystems vorgestellt. Zunächst werden die Arbeiten zur Erläuterung der Thematik über multivariate Datenanalyse in Abschnitt 3.1 vorgestellt. Gefolgt ist eine Diskussion über die verwandten Arbeiten in Abschnitt 3.2. Anschließend wird in Abschnitt 3.3 das grundlegendes Framework vorgestellt, welches zur Realisierung des Zielsystems verwendet werden.

3.1 Zielverwandte Arbeiten

Für die Visualisierung der multivariaten Daten und der Korrelationsanalyse zwischen den Variablen wurden bereits verschiedene Herangehensweise entwickelt.
Stahnke et al. [SDMT16] entwickeln ein integriertes Verfahren, um die Visualisierungstechnik der Multidimensionalen Skalierung (MDS) [CC08] zu erweitern und angemessene Interaktionen mit der Visualisierung zu ermöglichen. Da multivariate Daten hochdimensional sind und präzise menschliche Erkennung der hochdimensionalen Daten unmöglich ist, entwickeln Wang et al. [WM16] ein Framework zur Beseitigung der von der schwachen menschlichen Erkennung verursachten Probleme. Das Framework serialisiert die Erforschung vom hochdimensionalen Raum auf eine kontinuierliche Weise entlang einer Reihe von verallgemeinerten 3D-Unterräumen. Diese Serialisierung ermöglicht es, die komplexen Interaktionen und Darstellungen vom hochdimensionalen Raum abzuschaffen und durch übliche Paradigmen zu ersetzen.
Die Korrelationsanalyse kann die komplexen Zusammenhänge aufdecken, die häufig bei den Variablen in multivariaten Daten existieren. Ein geeignetes System zur Visualisierung der Korrelationen in multivariaten Daten mit unterschiedlichen Variablentypen spielt eine wesentliche Rolle für eine effiziente Korrelationsanalyse. Zhang et al. [ZMZM15] entwickeln ein System, um die Korrelationsanalyse der numerischen und kategorischen Variablen mittels einer Correlation-Map zu visualisieren. Johansson et al. [JJ10] entwickeln ein System für die Korrelationsanalyse von kategorischen und gemischten Daten, in dem einen Quantifizierungsprozess aus [JJJ08] eingesetzt wird.

3.1.1 Probing Projections: Interaction Techniques for Interpreting Arrangements and Errors of Dimensionality Reductions

Stahnke et al. [SDMT16] entwickeln ein System *Probing Projections* (siehe Abbildung 3.1) mit dem Konzept *Probing*. Das System interpretiert die Bedeutung und Qualität der Datenvisualisierungen, die auf Dimensionsreduktion beruhen. Das System zielt darauf ab, die Abfrage der Projektion mit der

Interpretation der Daten zu integrieren und diese als zwei notwendigerweise verknüpfte Aktivitäten zu behandeln. Dieses System besteht aus mehreren interaktiven Visualisierungen. Dabei wird die Multidimensionale Skalierung (MDS) für die multivariate Datenanalyse einsetzt. Die resultierende Projektionsfehler werden ebenfalls untersucht.
Eine überlegene Funktionalität von dem System liegt daran, dass vor der Datenanalyse die Untersuchung der Projektion und der Daten durchgeführt werden. Die Untersuchung der Projektion trägt dazu bei, dass die Nutzer die Zuverlässigkeit der Positionierung in der Projektion und die Zusammenhänge zwischen der erzeugten Visualisierung und den originalen Daten überprüfen können. Zu dem Zweck werden die Gesamtfehler der Projektion, die Verteilung der Fehler über die Projektion, den Zusammenhang zwischen Unähnlichkeiten und Distanzen der Datenvariablen durch passende Interaktionen visualisiert. Die Untersuchung der Daten kann die Ursachen der Ähnlichkeiten und Unähnlichkeiten der Variablen in der Projektion ermitteln.
Der Entwurf des Systems besteht aus einer Komponente für MDS-Projektion und einer Komponente für Kontrolle der Visualisierung. In der MDS-Projektion sind die Dateneinheiten als Punkte dargestellt. Die Punkte können sich gruppieren, indem man die Punkte als Cluster selektiert. Die selektierten Punkten lassen sich miteinander vergleichen. Da die originalen Daten mehrdimensional sind, werden die Werte der Dimensionen für jeden Punkt oder jedes Cluster als Heatmap [Fri09] visualisiert werden. Um die Untersuchung der Projektionsfehler zu unterstützen, bietet das System die Funktionalitäten wie Darstellung des Fehlerkreis um die Punkte, Distanzkorrektur und ein Dendrogram, das die Punkte hinsichtlich ihrer Position in der Clustering-Hierarchie visualisiert.

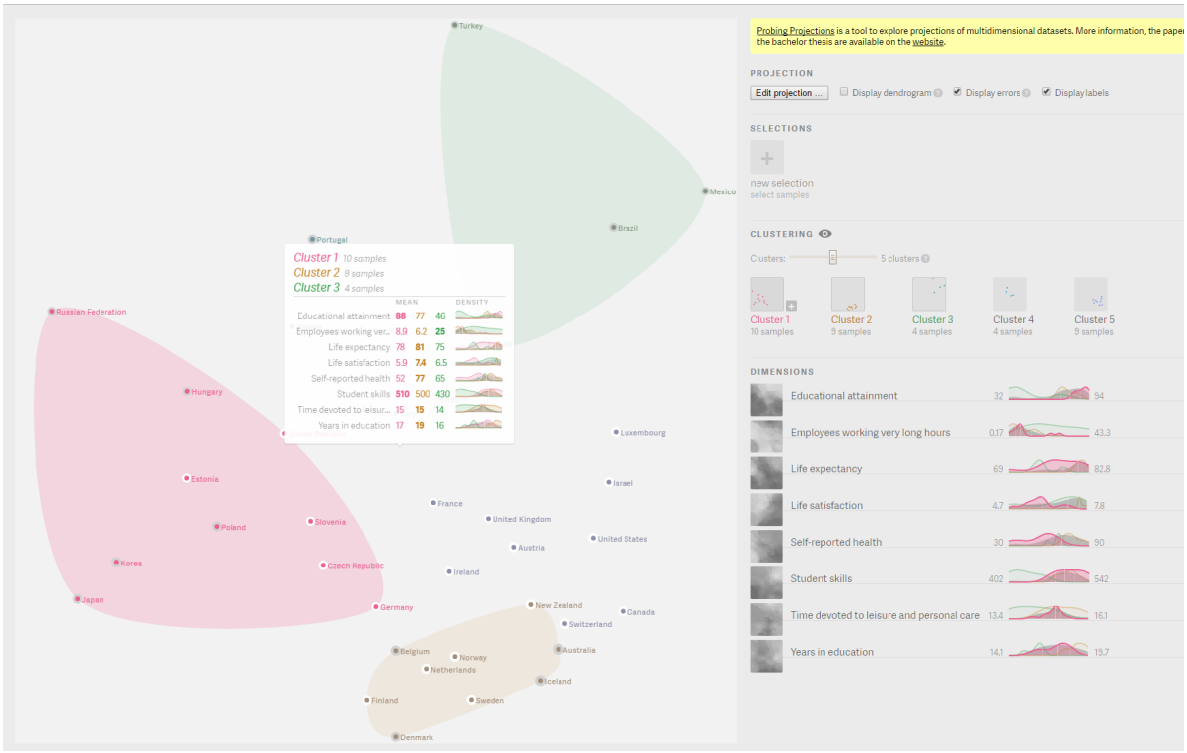


Abbildung 3.1: Übersicht des Systems *Probing Projections* [SDMT16]

3.1.2 The Subspace Voyager: Exploring High-Dimensional Data along a Continuum of Salient 3D Subspaces

Wang et al. [WM16] entwickeln ein Framework *Subspace Voyager* zur Beseitigung der Probleme beim hochdimensionalen Raum, dass durch menschliche Erkennung die hochdimensionalen Strukturen schwer erkannt werden. Das Framework serialisiert die Erforschung vom hochdimensionalen Raum auf eine kontinuierliche Weise entlang einer Reihe von verallgemeinerten 3D-Unterräumen. Diese Serialisierung ermöglicht es, die komplexen Interaktionen und Darstellungen vom hochdimensionalen Raum abzuschaffen und durch die normalen Paradigmen zu ersetzen. Die Paradigmen sind in dem System Trackball, Maps, Wold-Clouds und Gallery-Views.
Das System enthält drei wichtige Komponenten, die jeweils als *Subspace Explorer* (SE), *Subspace Trail Map* (STM) und *View Gallery* (VG) benannt werden. Beim SE (siehe Abbildung 3.2) wird ein Trackball-Interface neben der Visualisierung des Scatterplots eingesetzt, damit die aktuellen Richtungen der projizierten Dimensionsachsenvektoren visualisiert werden. Die Dimensionsachsenvektoren liegen als Kennzeichnungen außerhalb der kreisförmigen Grenze. STM stellt eine Word-Cloud von Attributen der Daten dar, wobei die Größe der Wörter der Relevanz ihrer zugehörigen Dimensionen im aktuellen Unterraum entspricht. Beim STM werden die untersuchten 3D-Unterräume entweder als Punkte oder als Dreiecke visualisiert. Durch die Brush-Linking-Technik werden die Interaktionen zwischen den drei Komponenten ermöglicht, wobei VG die 3D-Unterräume darstellt, die in SE oder STM untersucht werden.

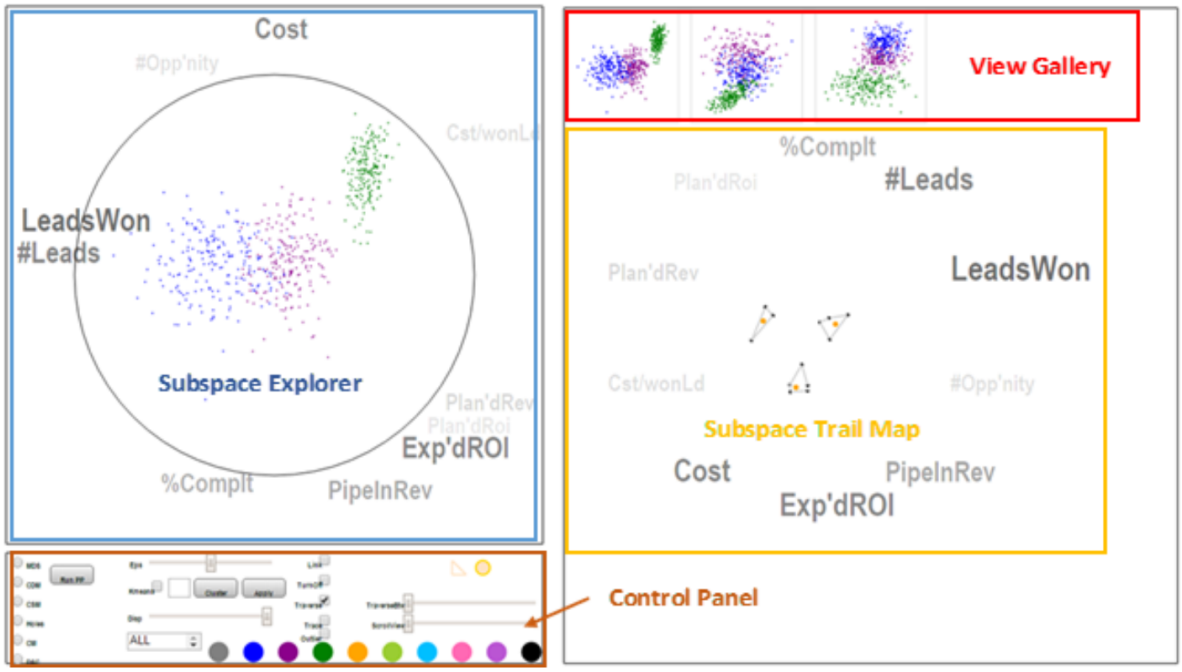


Abbildung 3.2: Übersicht des Systems *Subspace Voyager* [WM16]

3.1.3 Visual Correlation Analysis of Numerical and Categorical Data on the Correlation Map

Zhang et al. [ZMZM15] entwickeln ein System um die Korrelationen der numerischen und kategorischen Variablen der multivariaten Daten mithilfe einer Correlation-Map interaktiv zu analysieren und zu visualisieren. Dieses System hat zwei Visualisierungsmöglichkeiten für die Korrelationsanalyse (siehe Abbildung 3.3): ein Netzwerk-Graph und eine Darstellung der parallele Koordinaten. Neue Konzepte werden in dem System eingeführt: anstatt nur Korrelationen zwischen kategorischen oder numerischen Variablen zu analysieren, bietet das System einen Mechanismus an, sowohl kategorische Variablen als auch numerische Variablen zusammen zu bearbeiten. Ein mehrstufiger semantischer Zoom-Ansatz des Systems ermöglicht gute Skalierbarkeit für multivariate Daten. In dem System sind interaktive Techniken entwickelt, welche die Auswirkung der Werte der Korrelationen untersuchen. Korrelierten Variablen bilden Unterräume, in denen die Daten in Scatterplot projiziert werden und somit die Beziehungen zwischen den Daten visualisiert werden.
Für die Korrelationsanalyse zwischen kategorischen und numerischen Variablen entwickeln Zhang et al. [ZMZM15] einen Algorithmus, indem man die kategorischen Variablen eines Datensatzes in numerische Variablen umwandelt. Angenommen hat ein Datensatz Ω eine kategorische Variable v_c und eine numerische Variable v_n mit N Datenpunkten und M Ebenen in v_c. Sei M_i die gesamte Anzahl von Datenpunkten, die zu der kategorischen Ebene $v_c(i)$ gehören. Sei $v_n^i(j)$ der j-te numerische Datenpunkt, der zu der kategorischen Ebene $v_c(i)$ gehört. Der Algorithmus zielt darauf ab, jede kategorische Ebene $v_c(i)$ von v_c in numerische Variablen $v_c^{'}(i)$ zu transformiert, sodass die Korrelation zwischen den Variablen maximiert ist. Wenn alle Variablen numerisch sind, werden die Korrelationen anhand des Korrelationskoeffizient r von Pearson berechnet. Daher liegen die Werte der Korrelationen im Intervall $[-1, 1]$.
Um die Visualisierung der Daten mit mehreren Variablen in einer kleinen Bildschirmfläche zu optimieren, unterstützt das System die Funktion *Multi-Scale Zooming*, die anhand der Korrelationsstärke die Knoten in der Correlation-Map fusionieren kann, um die Anzahl der Knoten und Kanten zu reduzieren. Die Fusionierung von zwei Knoten hängt davon ab, ob die Knoten untereinander positiv korreliert sind. Mit einer positiven Korrelation haben die Knoten die gleiche Korrelation zu dem anderen Knoten nach der Fusionierung.
Da bei den Parallelen Koordinaten die Rohdaten dargestellt werden und die Achsen horizontale Ordnung haben, ist es schwer die Korrelationen zwischen mehr als drei Variablen zu erkennen. Dieses System bietet daher *Subspace-Scatterplot* an, das die Daten in Unterräumen in Scatterplot projiziert. Zur Generierung der Unterräume in der Correlation-Map wird Delaunay-Triangulierung [Sei95] angewendet. Dabei werden die Kanten anhand der Korrelationsstärke aufsteigend sortiert. Innerhalb jedes Unterraums wird ein Scatterplot generiert.

3.1.4 Visual Analysis of Mixed Data Sets Using Interactive Quantification

Johansson et al. [JJ10] entwickeln ein interaktives System *MiDAVisT* (Mixed Data Analysis Visualization Tool) für die Analyse von kategorischen und gemischten Datensätzen. Das System bietet eine algorithmische und von Nutzern kontrollierte Quantifizierung von kategorischen Variablen. Die Quantifizierung ermöglicht eine Analyse sowohl mit algorithmischen Methoden als auch mit visuellen Darstellungen, die für rein numerische Datensätze entwickelt werden. *MiDAVisT* bietet

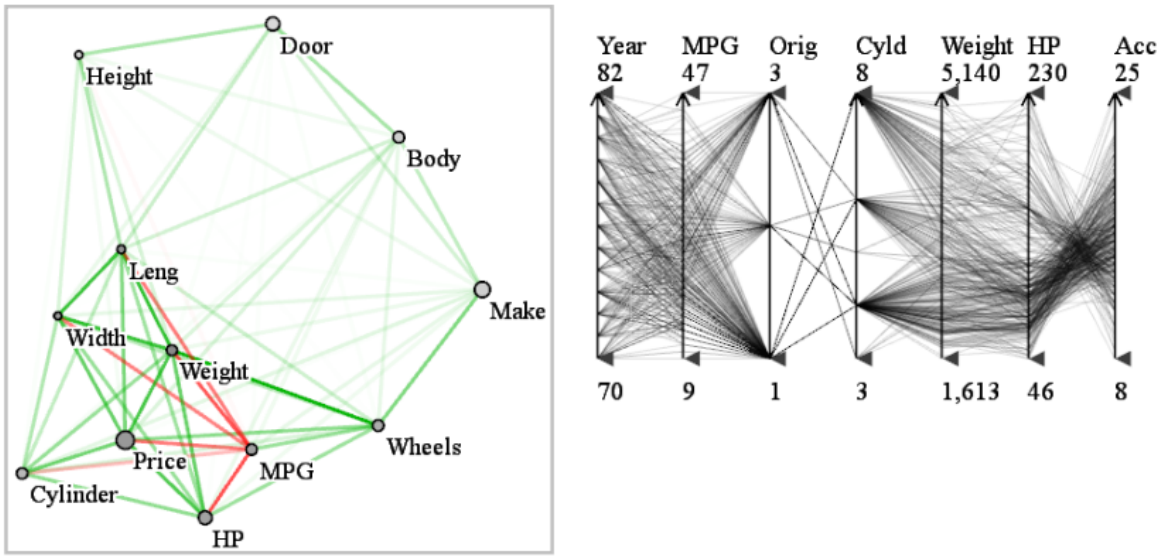


Abbildung 3.3: Zwei koordinierte Displays für Dual-Space-Analytics des Systems: (a) Correlation-Map stellt die Korrelationen als Kanten und Variablen als Knoten dar, (b) Die Darstellung der Parallelen Koordinaten repräsentiert die gleiche Korrelationen als Polyliniengruppe und Variablen als vertikale Achsen. [ZMZM15]

ebenfalls eine Umgebung mit Multiple-Coordinated-Views (MCV), wo die visuellen Darstellungen und die algorithmischen Analysemethoden, die speziell für numerische Daten entwickelt sind, für die Analyse der gemischten Daten zur Verfügung stehen.
Der interaktive Quantifizierungsprozess von *MiDAVisT* basiert auf der Vorversion des Prozesses in [JJJ08]. In diesem erweiterten Quantifizierungsprozess handelt es sich um die Quantifizierung von kategorischen Daten und die Identifikation der Ähnlichkeiten und Zusammenhänge zwischen Kategorien. Dass die Zusammenhänge innerhalb der Daten durch sinnvolle numerische Repräsentationen ersetzt werden, ist für die Quantifizierung eine hohe Bedeutung. Die Identifikation der Ähnlichkeiten zwischen Kategorien ist mit der Methode Correspondence-Analysis (CA) [Gre07] möglich. CA wird auf Häufigkeitstabellen durchgeführt, wobei jede Zelle die Häufigkeit einer Kombination von zwei Kategorien darstellt und somit eine Quantifizierung auf Grund der Zusammenhänge innerhalb von den kategorischen Variablen ermöglicht wird. Wenn es um einen Datensatz mit kategorischen und numerischen Variablen handelt, müssen die numerischen Variablen vor der Durchführung von CA in die Häufigkeitstabelle integriert werden. Der Prozess der Kategorisierung muss so ausgeführt werden, dass die bestehenden numerischen Zusammenhänge gewahrt bleiben. Auf diese Weise basiert die Quantifizierung auf den Zusammenhängen zwischen kategorischen und numerischen Variablen. In *MiDAVisT* ist die Kategorisierung der numerischen Variablen sowohl durch die Interaktionen der Nutzer als auch durch die Durchführung des K-Means-Clusering-Algorithmus [Mac02] möglich.
Zusätzlich zur Quantifizierung bietet *MiDAVisT* auch eine interaktive Umgebung für die visuelle Analyse von kategorischen und gemischten Datensätzen mit MCV. Drei häufig angewendeten visuellen Darstellungen für numerische Daten stehen in *MiDAVisT* zur Verfügung. In Abbildung 3.4 ist es zu sehen, dass die MCV sich aus einer Scatterplot-Matrix, einem Table-Lens [RC94] und den

Parallelen Koordinaten zusammensetzen. Die Darstellungen sind so koordiniert, dass jede Auswahl oder Markierung Von Punkten in einer Darstellung sofort in den anderen reflektiert wird.

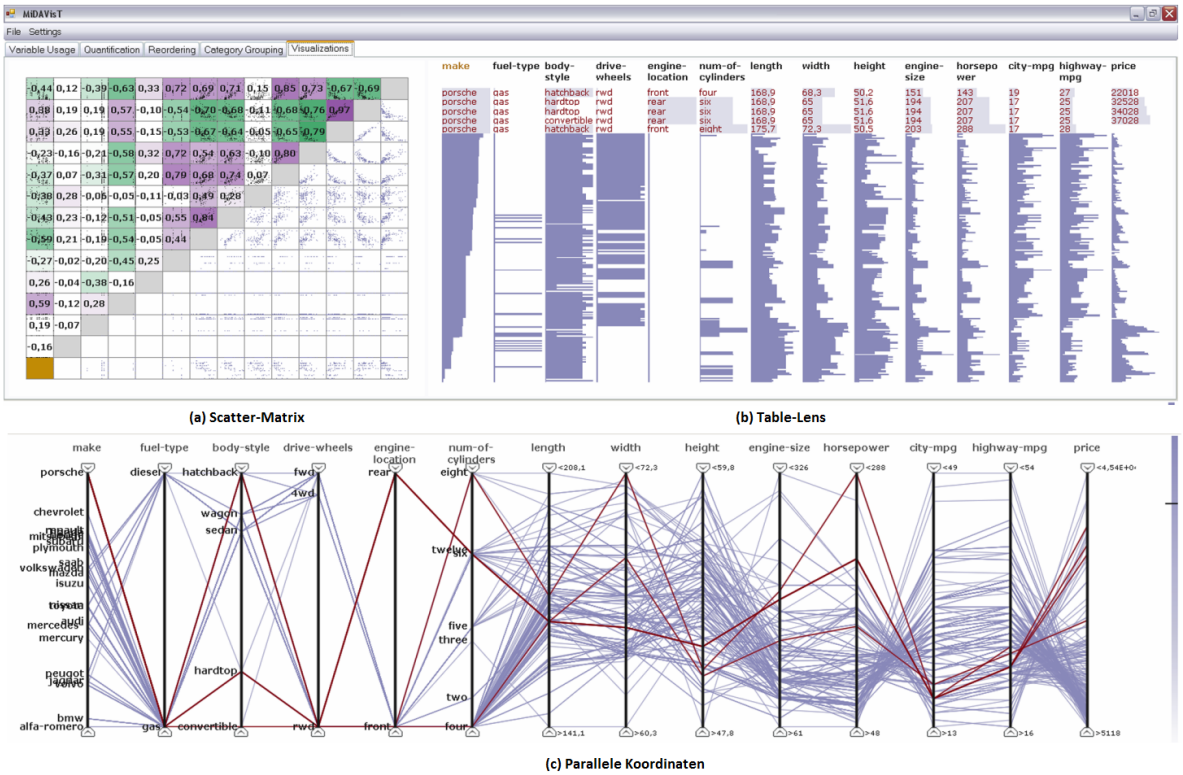


Abbildung 3.4: *MiDAVisT* mit Multiple-Coordinated-Views: (a) Scatter-Matrix zeigt die Korrelationen der Variablenpaare mit grüner und violetter Farbe an. (b) Table-Lens. (c) Parallele Koordinaten. In (b) und (c) sind alle Datenpunkte *Porsche* selektiert und mit roter Farbe markiert.

3.2 Diskussion über die verwandten Arbeiten

In Abschnitt 3.1 wurden bestehende Systeme vorgestellt, welche die bestehenden Systeme eine effiziente visuelle Analyse für hochdimensionale Daten ermöglichen. Größtenteils sind die Daten für die vorgestellten Systeme synthetische multivariaten Daten. Die Anzahl von Dimensionen solcher Daten ist normalerweise kleiner. Außerdem werden die multivariaten Daten in den vorhandenen Systemen als Ganzes betrachtet und analysiert. Dies liefert einen guten Überblick über die Gesamtzusammenhänge. Jedoch helfen die Ansätze nicht bei der Analyse einzelner Variablen mit spezifischen Anforderungen.
Weitere Einschränkungen bei den bestehenden Systemen liegen darin, dass die Systeme nur allgemeine Probleme bei der multivariaten Datenanalyse berücksichtigen. Beispielsweise zielt das System von Stahnke et al. [SDMT16] auf die Optimierung der Dimensionsreduktion auf einem 2D-Projektionsraum ab. In dem System wird der Einfluss der Datendimensionen auf der Projektion visualisiert, ohne die Zusammenhänge zwischen den Dimensionen anzuzeigen. Ein anderes Beispiel ist das System von

Wang et al. [WM16], das für ein besseres Verständnis der hochdimensionalen Daten entwickelt wurde. In dem System wird der originale Datenraum in mehrere 3D-Unterräume geteilt. Die Eigenschaften der einzelnen Variablen (Dimensionen) von den Daten werden analysiert, ohne die Zusammenhänge zwischen den Variablen zu betrachten. Die Systeme von Zhang et al. [ZMZM15] und Johansson et al. [JJ10] betrachten die multivariaten Daten als ein Ganzes. Dabei werden die Korrelationen zwischen den Variablen paarweise analysiert. Jedoch können starke Korrelationen zwischen zwei Mengen von Variablen entstehen.
Die Entstehung der oben genannten Einschränkungen der vorhandenen Systemen hängen teilweise von den Anforderungen der Systeme ab. Die meisten Systeme werden für wissenschaftliche Forschungen entwickelt und deswegen sind keine spezifischen Funktionalitäten verlangt. Im Vergleich dazu sind die Anforderungen aus der Industrie anders. In der Hardwareindustrie gibt es häufig große multivariaten Daten aus Chipüberprüfungen. Die Daten enthalten heterogene Parameter in großer Anzahl. Die Parameter werden als Eingabe für moderne integrierte Schaltungen konfiguriert. Es ist unmöglich, eine vollständige Abdeckung der Testfälle der Chipüberprüfung zu erreichen. Somit wird eine Teilmenge von Samples aus dem Parameterraum zum Testen verwendet. Es ist wichtig, die möglichen Parameter zu entdecken, welche fehlerhafte Testfälle verursachen. Um dieses Ziel zu erreichen wird im Rahmen dieser Masterarbeit ein Visual-Analytics-System entwickelt, mit dem die Experten bei der Entdeckung der potenziellen Fehlerquellen unterstützt werden können.

3.3 Verwendetes Framework

Zur Verwirklichung des Zwecks der Arbeit wird das Visualisierungswerkzeug *Prefuse*[1] [HCL05] verwendet, das als Open-Source im Internet zur Verfügung steht. Das Framework ist die Basis für die Implementierung des Zielsystems im Rahmen der Masterarbeit. Folgend wird das Framework und seine Funktionalitäten vorgestellt.

Prefuse ist ein Java-basiertes Toolkit für die Anwendungen der interaktiven Informationsvisualisierungen, das von Heer et al. [HCL05] entwickelt wurde. Es unterstützt eine Vielzahl von Funktionen für Datenmodellierung, Visualisierung und Interaktion. Hauptsächlich bietet *Prefuse* optimierte Datenstrukturen für Tabellen, Graphen und Bäume, zahlreiche Kodierverfahren für Layout und Visualisierung. Animationen, dynamische Abfragen, integrierte Suche und Datenbankanbindung werden ebenfalls in *Prefuse* unterstützt. Die oben genannten Funktionalitäten sind die Kernelemente von *Prefuse*, die für die Implementierung interaktiver Anwendungen erforderlich sind.
Da *Prefuse* ein Java-basiertes Software-Framework ist, kann man dieses Framework in beliebigen Java-Programmen einfügen und es entsprechend erweitern. Dank der Veröffentlichung von *Prefuse* auf einer öffentlichen Plattform ist der Quellcode für jeden verfügbar, das ist sehr hilfreich für die eigene Implementierung der Anwendungen.

Der Visualisierungsprozess von *Prefuse* besteht aus vier Schritten, die in Abbildung 3.5 anschaulich dargestellt sind. Der erste Schritt des Prozesses sind Abstrakte Daten (Abstract Data) für Visualisierung. Durch bestehende Schnittstellen und Implementierungen in *Prefuse* können die abstrakten Daten in verschiedene Datenstrukturen umgewandelt werden, die für unstrukturierte Daten wie

[1] Prefuse: http://prefuse.org/

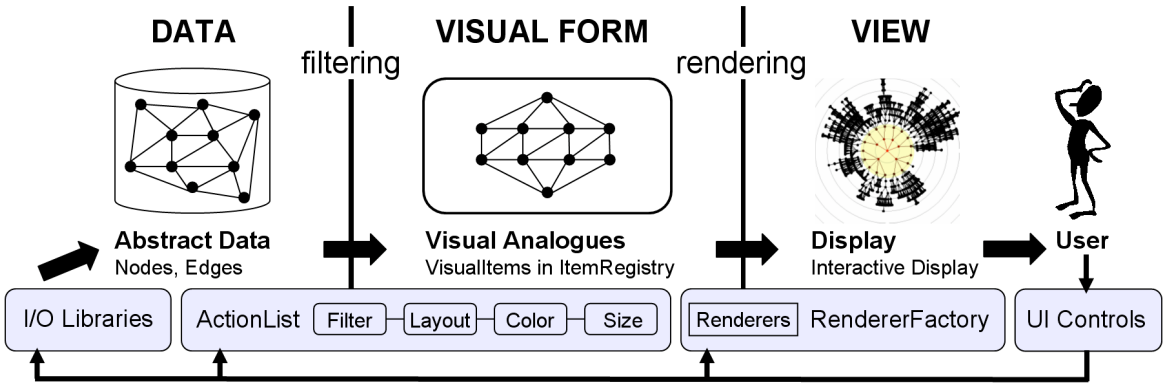


Abbildung 3.5: Die Funktionspipeline des Visualisierungswerkzeugs *Prefuse* [HCL05]

Daten für Scatterplot oder strukturierte Daten wie Daten für Tabellen, Graphen und Bäume sind. Wenn die Daten in Graphen oder Bäume umgewandelt sind, werden die Daten entsprechend in Datentypen wie `Node` oder `Edge` gespeichert.
Nach der Bearbeitung der abstrakten Daten werden die Daten in geeigneten visuellen Repräsentationen (Visual Analogues) nachgebildet. Dieser Schritt wird in der Regel als `Filterung` genannt, weil die abstrakten Daten zu visualisierbaren Inhalten reduziert werden. Dies ermöglicht mehrere Visualisierungen eines gemeinsam genutzten Datensatzes durch Verwendung von separaten Filtern und verschiedene Ansichten einer bestimmten Visualisierung durch Wiederverwendung der gleichen gefilterten Elemente. Wenn die Daten in visualisierbaren Repräsentationen nachgebildet sind, werden den Daten einige visuelle Eigenschaften wie Position, Farbe, Größe und Schriftart zugeordnet. Um die Zuordnung der visuellen Eigenschaften zu ermöglichen, wird ein Mechanismus in *Prefuse* eingesetzt, der als `Actions` benannt ist. Mit `Actions` sind Funktionen wie Filterung, Zuordnung der Layout und Farbe möglich. Sind die visuelle Repräsentationen der Daten bereitgestellt, kümmert sich der dritte Schritt `Rendering` und `Display` um das tatsächliche Erscheinen der Daten auf dem Bildschirm. Am Ende des Prozesses werden die Interaktionen von Nutzern durch eine Schnittstelle `ControlListener` in `Display` unterstützt, indem die Maus- und Tastaturereignisse auf visuelle Objekten berücksichtigt werden.
Anwendungen, die mit *Prefuse* implementiert wurden, zeigen die Flexibilität und Leistungsfähigkeit der Architektur des Toolkit. *Prefuse* hat daher zu der Systematisierung der Forschung der Informationsvisualisierung beigetragen und mehr Interaktivitäten in die Bereiche der Datenanalyse und Datenerforschung eingebracht.

4 Forschungsfragen und Systementwurf

In diesem Kapitel werden die Konzepte des Systems vorgestellt. Zuerst werden die vordefinierten Forschungsfragen zum Entwurf des Systems in Abschnitt 4.1 erläutert. Anschließend wird der Systementwurf in Abschnitt 4.2 diskutiert. Dabei wird der grundlegende Arbeitsablauf des Systems erläutert. Eine Übersicht über die Benutzeroberfläche des Systems wird gefolgt vorgestellt.

4.1 Forschungsfragen

Auf Grundlage von Diskussionen mit Experten wurden eine Reihe von Forschungsfragen aufgestellt. Sie bilden die Anforderungen an das System und bilden damit die Grundlage für den Systementwurf. Die Forschungsfragen sind in drei Bereiche aufgeteilt: Multivariate Visualisierung, Fehler und Parametereinstellung sowie Koordination. Diese werden im Folgenden näher erläutert.

Multivariate Visualisierung

Da das System multivariate Datensätze verwendet, ist eine Übersicht der Daten in dem System wichtig. Die Forschungsfragen hier beziehen sich auf die Umsetzung der Übersicht in dem System.

1. *Ist eine Übersicht über die Daten möglich?*
Die Datensätze sind hochdimensional. Sie enthalten meistens eine große Menge von Datenpunkten. Es ist wichtig, dem Nutzer eine Übersicht über die Datenpunkte auf eine geeignete Weise zu ermöglichen.

2. *Wie wird die Übersicht generiert und visualisert?*
Mehrere Möglichkeiten zur Erzeugung einer Übersicht sind vorhanden. Die Auswahl der Visualisierungstechniken beziehungsweise das Hervorheben der wichtigen Informationen der Daten spielt eine entscheidende Rolle für ein funktionsfähiges System.

3. *Wie wird die Visualisierung der Übersicht durch Interaktionstechniken erweitert?*
Mit einer statischen Visualisierung der Übersicht ist die Datenanalyse nicht flexibel. Geeignete Interaktionen werden neben der Visualisierung benötigt. Damit kann der Nutzer die versteckten Informationen in den Daten erkennen.

Fehler und Parametereinstellung

Das Ziel des Systems ist die Feststellung der Fehlerquellen. Die Aufdeckung der Korrelationen zwischen den Fehlern und Parametern sind wichtig:

1. *Wie werden die Fehler und die entsprechenden Parameter erkannt?*
Die Fehler und die entsprechenden Parameter sollen so eingestellt werden, dass der Nutzer die Zusammenhänge zwischen den Fehlern und den Parametern leicht entdecken kann und somit weitere Analysen durchführen kann.

2. *Wie wird die Navigation durch potenzielle interessante Parameter-Tupel ermöglicht?*
Die Fehler können nicht nur von einem einzelnen Parameter verursacht werden, sondern auch von möglichen Parameter-Tupeln. Die Navigation durch die Parameter-Tupel erlaubt dem Nutzer, interessante Parameter zu untersuchen. Auf die Navigationsmöglichkeiten soll in dem System deutlich hingewiesen werden.

3. *Wie kann der Nutzer die Visualisierung für ein gegebenes Parameter-Tupel ändern?*
Dem Nutzer soll ermöglicht werden, die Visualisierungen des wichtigen Parameter-Tupels interaktiv zu ändern.

Koordination

Für ein Visual-Analytics-System sind Interaktionen ein wichtiger Bestandteil. Die Visualisierungen werden dadurch verbunden. Der Nutzer kann damit mehr Wissen über die Daten entdecken.

1. *Wie werden die Visualisierungen miteinander gekoppelt?*
Die Visualisierung für die Übersicht soll in Zusammenhang mit der Visualisierung für Parametereinstellung gebracht werden.

2. *Wie gut wirken sich die Visualisierungen und die Interaktionen auf die Datenanalyse aus?*
Die Auswirkungen der Visualisierungen und Interaktionen auf die Analyse und Untersuchung der Fehlerquelle sind beachtenswert.

4.2 Systementwurf

Mithilfe von verschiedenen interaktiven Visualisierungen soll dem Nutzer ermöglicht werden, Zusammenhänge zwischen Fehlern und Parametern zu erkennen, wie in Abschnitt 4.1 beschrieben.
Dazu bietet das System dem Nutzer drei verknüpfte Visualisierungen. In der Ansicht der Multidimensionalen Skalierung (MDS) kann der Nutzer einen Überblick über die Verteilung der Daten bekommen. In der Ansicht des Node-Link-Graphen (NLG) kann der Nutzer die Korrelationen zwischen Fehlern und Parametern erkennen. Die Ansicht der Parallelen Koordinaten (PK) hilft dem Nutzer die hochdimensionalen Strukturen nachzuvollziehen.
Das Zusammenspiel der Ansichten hilft dem Nutzer schrittweise Einsichten in die Daten zu gewinnen. Im folgenden Abschnitt wird dieser Arbeitsablauf näher beschrieben.

4.2.1 Arbeitsablauf

Der Arbeitsablauf beginnt mit einem Überblick über die Verteilung der Daten in der MDS-Ansicht. Durch eine interaktive Selektion in der MDS-Ansicht kann der Nutzer ein Cluster der Datenpunkte generieren. Durch Klicken auf das selektierte Cluster kann der Nutzer zu der NLG-Ansicht geführt werden. Dabei wird ein Node-Link-Graph für die Korrelationen zwischen den Fehlern und Parametern dargestellt. Der Nutzer kann die Korrelationen innerhalb einer Parametergruppe mittels eines Kreisgraphen anschauen. Mittels eines Bipartiten Graphen kann der Nutzer die Korrelationen zwischen den Parametergruppen anschauen. Durch Klicken auf einen Parameter in dem Bipartiten Graph kann der Nutzer die Korrelationen bezüglich des Parameters generieren. Der Nutzer kann die Korrelationsordnung beliebig ändern, um stark korrelierte Parameter-Tupel zu finden. Um genaue Zusammenhänge der Parameter-Tupel zu bekommen kann der Nutzer zu der PK-Ansicht wechseln. Dabei werden die zweidimensionalen Strukturen zwischen den Parametern durch die parallelen Achsen angezeigt. Der Nutzer kann die Parameter-Tupel durch Interaktionen genauer analysieren. Somit können potenzielle Fehlerquellen anhand von den dargestellten Parameter-Tupeln festgestellt werden.

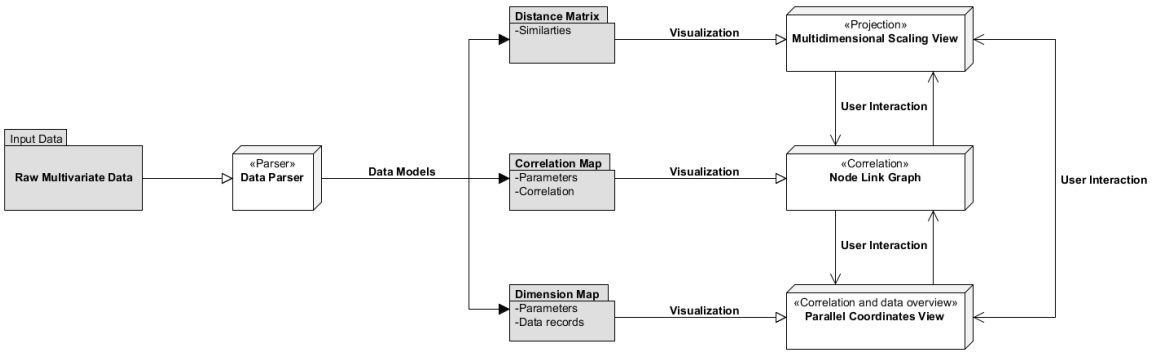


Abbildung 4.1: Architekturübersicht des CTEDS

Eine Übersicht über die Softwarearchitektur zur Umsetzung des Arbeitsablaufs ist in Abbildung 4.1 zu sehen. In Kapitel 5 wird die konkrete Implementierung detailliert beschrieben. Die Architektur besteht aus mehreren Komponenten. Die zu bearbeitenden und zu analysierenden Rohdaten sind hochdimensionale Daten. Die Datenbearbeitung findet bei dem ersten Schritt in einer Komponente Data Parser statt. Die Daten werden eingelesen und in computergestützte Datenstrukturen umgewandelt. Sind die Parameter der Rohdaten in unterschiedliche Kategorien geteilt, behalten die bearbeiteten Daten ebenfalls die entsprechenden Kategorien.
Nach der Datenbearbeitung werden die geeigneten Datenmodelle für verwendete Visualisierungen generiert. In CTEDS werden die drei Datenmodelle generiert. Das erste Datenmodell ist eine Distanzmatrix für die Dimensionsreduktion. Das zweite Datenmodell ist eine Korrelationen-Map für die Korrelationsanalyse. Das letzte Datenmodell ist eine Dimensionen-Map für die Übersicht der hochdimensionalen Strukturen.
Der weitere Schritt des Systemprozesses bezieht sich auf die Umsetzung der Visualisierungen für die bearbeiteten Daten. Um gute Ergebnisse zu bekommen, müssen die Datenmodelle richtige Elemente enthalten. Zur Generierung der Distanzmatrix werden die Un-/Ähnlichkeiten zwischen den

Datenpunkten berechnet. Dabei werden Distanz- und Ähnlichkeitsmaße für kategorische und numerische Parameter angewendet. Zur Berechnung der Korrelationen zwischen Parametern wird ein Algorithmus für bedingte Entropie eingesetzt. Für die Visualisierung der Übersicht über Dimensionen ist ein Filteralgorithmus an den Parametern angewendet, um die weniger informative Parameter zu entfernen. Somit wird die Anzahl der zu betrachtenden Dimensionen reduziert.
In dem letzten Schritt des Prozesses vom CTEDS sind drei verknüpfte Komponenten (siehe Abbildung 4.1): Multidimensional Scaling View, Node Link Graph und Parallel Coordinates View. Die Komponenten sind drei Ansichten, in denen die Visualisierungen und Interaktionen umgesetzt werden. Die verkoppelten Ansichten können dem Nutzer erlauben, eine interaktive Datenanalyse durchzuführen.

4.2.2 Benutzeroberfläche

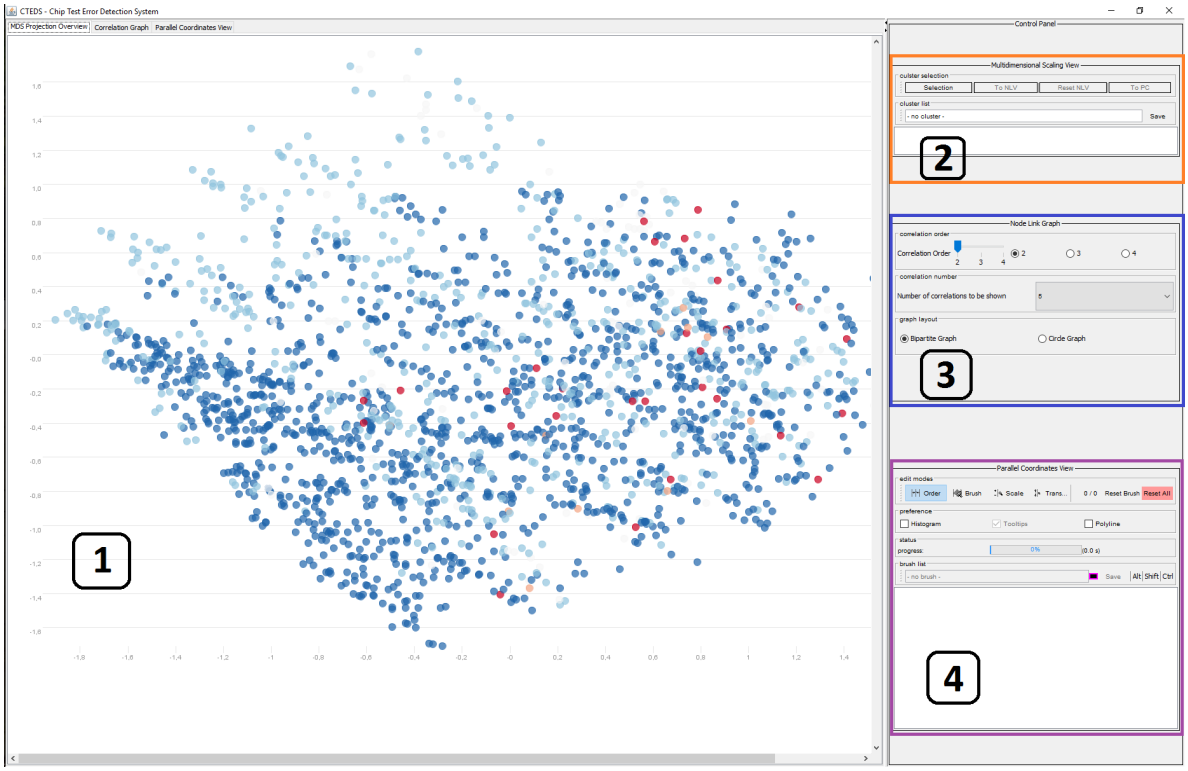
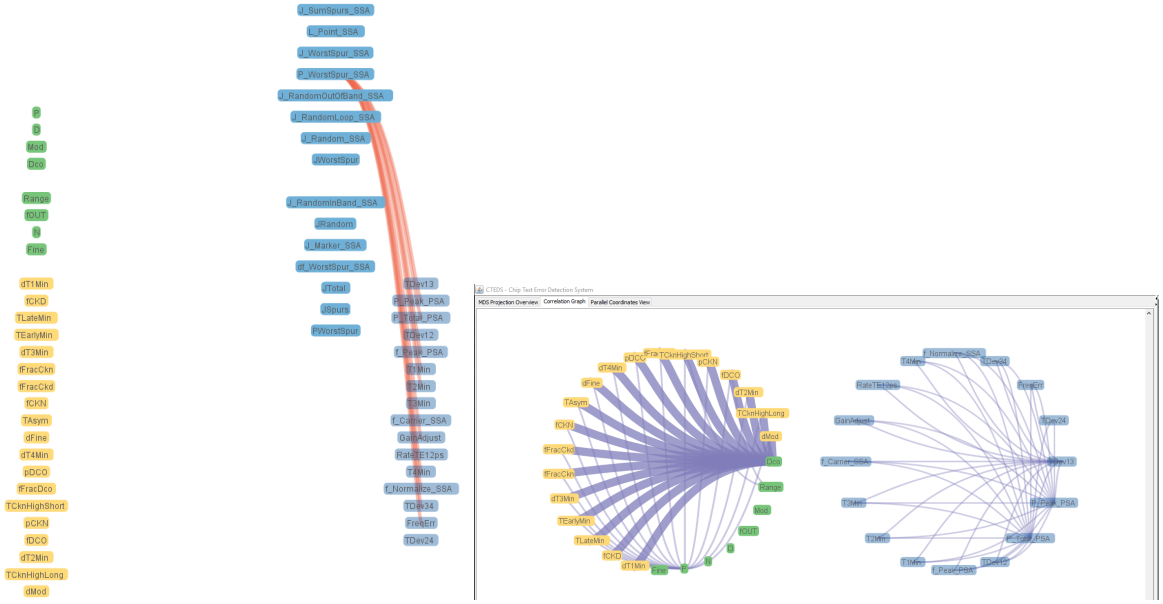


Abbildung 4.2: Die Benutzeroberfläche von CTEDS mit zwei Funktionsbereichen: Visualisierungsbereich (links: Nummer 1) und Steuerungsbereich (rechts: Nummer 2 bis 4).

In diesem Abschnitt wird eine Übersicht über die Benutzeroberfläche vom CTEDS erläutert. Abbildung 4.2 zeigt die Benutzeroberfläche des Systems. Die Benutzeroberfläche besteht aus zwei Funktionsbereichen: ein Bereich für die Visualisierungen und ein Bereich für die Steuerungen.
Der Bereich für die Visualisierungen (siehe Abbildung 4.2 (1)) ist ein Registerfeld (Tab-Panel) mit drei Tabs, die jeweils eine Visualisierung präsentieren. In der Abbildung ist die Ansicht für Multidimensionale Skalierung zu sehen.
Der rechte Bereich (siehe Abbildung 4.2 (2−4)) ist für die Steuerung. Dort werden die Interaktionen

innerhalb von einer Ansicht oder zwischen den Ansichten ermöglicht. Die Unterbereiche sind mit 2 bis 4 nummeriert. Der Bereich 2 ist für die Interaktion zwischen der Ansicht der Multidimensionale Skalierung und anderen Ansichten. Der Bereich 3 steuert die Interaktionen innerhalb der Ansicht des Node-Link-Graphen. Der Bereich 4 ermöglicht die Interaktionen innerhalb der Ansicht der Parallelen Koordinaten.

Im Bereich 1 von der Abbildung 4.2 können neben der MDS-Ansicht auch die NLG-Ansicht und die PK-Ansicht dargestellt werden, welche durch Auswählen der Registerkarten gewechselt werden.
Die MDS-Ansicht ermöglicht eine übersichtliche Visualisierung der hochdimensionalen Daten auf einem reduzierten Dimensionenraum (siehe Abbildung 4.2 (1)). Die Testfälle in den Daten werden so räumlich angeordnet, dass die Abstände (Distanzen) zwischen den Datenpunkten (Testfällen) im Raum möglichst exakt den erhobenen Un-/Ähnlichkeiten entsprechen. Durch Selektion in der MDS-Ansicht kann ein Cluster gebildet werden. Durch die Interaktionen im Bereich 2 kann das Cluster bearbeitet werden.



(a) (b)

Abbildung 4.3: Die Graphenlayouts in der NLG-Ansicht

Die NLG-Ansicht bezieht sich auf die Visualisierung der Korrelationen zwischen den Datenparametern. Die Parameter werden als Knoten und die Korrelationen zwischen den Parametern werden als Kanten des Graphen dargestellt. Um die Unordnung der Kanten zu vermeiden, werden zwei Graphenlayouts eingeführt (siehe Abbildung 4.3(a) und Abbildung 4.3(b)). Die Darstellung der Korrelationen werden durch die Interaktionen im Bereich 3 in Abbildung 4.2 ermöglicht.

Die PK-Ansicht ermöglicht eine effiziente Methode zur Visualisierung von hochdimensionalen Strukturen und multivariaten Daten (siehe Abbildung 4.4). Die Vorteile der Visualisierungen von MDS und NLG lassen sich in PK kombinieren. Die gesamten Testfälle werden als Linienzüge visualisiert, die eine Übersicht über die Verteilung der Datenpunkte ermöglichen. Die Datenparameter werden als

parallele vertikale Achsen veranschaulicht. Die Korrelationen zwischen zwei Parametern werden durch die Muster der Linienzüge dargestellt. Durch die Interaktionen im Bereich 4 in in Abbildung 4.2 können die Achsen beliebig angeordnet und Testfälle ausgewählt werden.

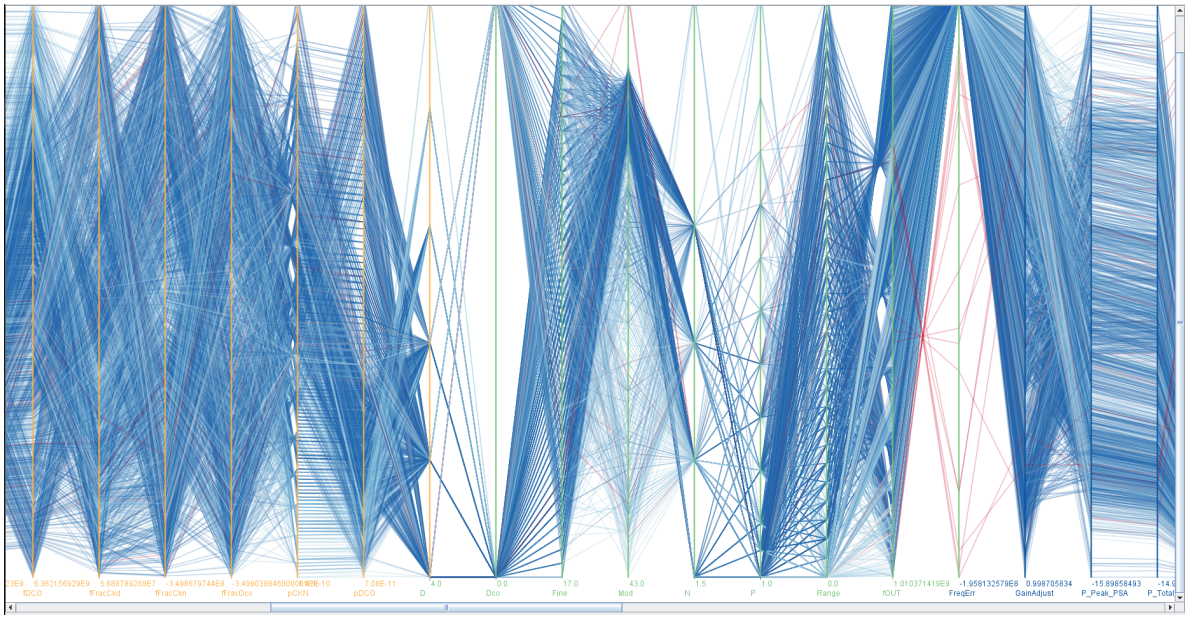


Abbildung 4.4: Die PK-Ansicht

5 Implementierung und Visualisierung

In diesem Kapitel wird die Implementierung des `Chip Testing Error Detection System (CTEDS)` vorgestellt. Wie in Kapitel 4 beschrieben ist `CTEDS` ein Java-basiertes Visual-Analytics-System für die visuelle und interaktive Analyse der hochdimensionalen Daten. `CTEDS` enthält drei gekoppelte Ansichten für Visualisierungen und Interaktionen. Zuerst wird die Datenbearbeitung in Abschnitt 5.1 vorgestellt. Dabei werden die Rohdaten für spezifische Datenmodelle vorgearbeitet. Gefolgt ist die Beschreibung der Ansicht für Multidimensionale Skalierung (MDS) in Abschnitt 5.2, mit der die Projektion der hochdimensionalen Daten auf einem 2D-Raum visualisiert wird und dem Nutzer eine Übersicht der Daten ermöglicht wird. Anschließend wird die Ansicht für Node-Link-Graph (NLG) in Abschnitt 5.3 vorgestellt. Damit kann der Nutzer die Korrelationen zwischen Parametern analysieren und potenzielle Fehlerquellen untersuchen. Die Ansicht für Parallele Koordinaten (PK) wird in Abschnitt 5.4 erläutert. Dabei kann der Nutzer eine ausführliche Übersicht über die hochdimensionalen Strukturen bekommen. Die korrelierten Parameter können in der PK-Ansicht ebenfalls genauer angezeigt werden. Zuletzt werden in Abschnitt 5.5 die Interaktionsmöglichkeiten zwischen den Ansichten des `CTEDS` beschrieben.

5.1 Datenbearbeitung

Um die Daten mit `CTEDS` zu visualisieren und interaktiv zu analysieren, müssen sie zuerst vorverarbeitet werden. Die gültigen Rohdaten für `CTEDS` sind entweder in CSV-Daten oder in TSV-Daten mit $N + 1$ Zeilen $R = \{r_0, r_1, \cdots, r_n\}$ und M Spalten $C = \{c_1, \ldots, c_m\}$ gespeichert. Die Zeilen $r_1, \ldots, r_n$ sind die Testfälle einer Chipüberprüfung. Die Zeile r_0 definiert die Namen der Parameter eines Testfalls. Die Spalten $c_1, \ldots, c_m$ enthalten die gesamten Parametereinträge der Chipüberprüfung. Der ganze Datenbearbeitungsprozess findet in der Komponente Data Parser des Systemprozesses (siehe in Abbildung 4.1) statt.

Im Data Parser werden die Rohdaten eingelesen, geparst, normalisiert und in Datenmodelle umgewandelt. In Abbildung 5.1 ist die Pipeline des Data Parser dargestellt.
Nach dem Einlesen werden die Rohdaten so analysiert, dass die Parameternamen und ihre Kategorien mit den Einträgen der Parameter in ein Datenmodell umgewandelt werden. Aus dem erzeugten Datenmodell werden zwei weitere spezifische Datenmodelle generiert, die HashMap-Datenstrukturen für die Korrelationsanalyse und die Dimensionsanalyse. Zur Kostenersparnis der Rechenleistung werden die Korrelationen zwischen den Parametern anhand der Korrelation-HashMap vorberechnet, die für die Visualisierung der Korrelationen im Node-Link Graph (siehe Abbildung 4.1) erforderlich sind. Die Dimension-HashMap ist grundlegend für die Visualisierung der Dimensionen im Parallel Coordinates View (siehe Abbildung 4.1).

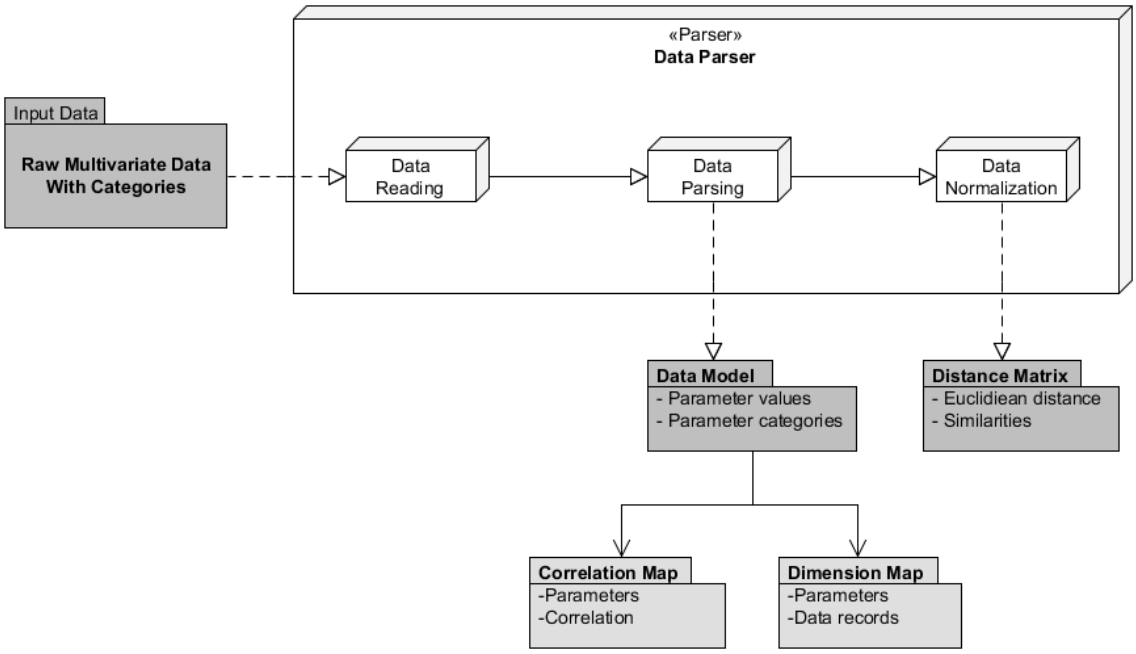


Abbildung 5.1: Die Pipeline des DataParser

Wenn in den Rohdaten numerische und kategorische Parameter vorhanden sind, werden die numerischen Parametereinträge nach dem Parsing normalisiert. Die Einträge der numerischen Parameter haben nach der Normalisierung die gleiche Skalierung. Die Gleichung 5.1 stellt die Normalisierung dar. Dabei ist c_i der i-te Eintrag des Parameters c. c_{max} und c_{min} sind das Maximum und das Minimum des Parameters c. c^i_{norm} definiert den normalisierten Wert von c_i. Haben die Einträge des Parameters den gleichen Wert, sind die normalisierten Werte 0. So liegen die Parametereinträge nach der Normalisierung in dem Intervall $[0, 1]$.

(5.1)

$$c^i_{norm} = \begin{cases} \frac{c_i - c_{min}}{c_{max} - c_{min}} & \text{if } c_{max} \neq c_{min} \\ 0 & \text{else} \end{cases}$$

Aus den normalisierten Parametern entsteht eine $N \times N$ Distanzmatrix aus den bearbeiteten Daten. Mit der Distanzmatrix wird die Multidimensionale Skalierung berechnet, die in Abschnitt 5.2.2 beschrieben wird. Für numerische und kategorische Parameter stehen zwei Distanzfunktionen zur Verfügung. Der Euklidischer Abstand (siehe Gleichung 5.2) berechnet die Distanz zwischen zwei Objekten im Raum, die numerische Parameter enthalten. Aus den Testfällen in den Daten $R = \{r_0, r_1, \ldots, r_n\}$ bekommt man $\binom{n}{2}$ Unähnlichkeiten unter den Testfällen.

(5.2)

$$d(r_i, r_j) = \| r_i - r_j \|_2 = \sqrt{\sum_{k=1}^{n} ((r_i)_k - (r_j)_k)^2}$$

Testfall i	wahr	wahr	falsch	falsch
Testfall j	wahr	falsch	wahr	falsch
Ähnlichkeit$_{ijk}$	1	0	0	0
Gewicht$_{ijk}$	1	1	1	0

Tabelle 5.1: Gower-Ähnlichkeit-Koeffizienten von Testfällen i und j mit einem binären Parameter k.

Zur Berechnung der Ähnlichkeiten von den Objekten mit kategorischen Parametern werden die Gower-Ähnlichkeit-Koeffizienten [Hov16] angewendet. Für binäre Parameter definiert Gower die Komponente der Ähnlichkeit und das Gewicht, die in Tabelle 5.1 zu sehen sind. In der Tabelle sind die Ähnlichkeit und das Gewicht von zwei Testfällen i und j mit dem binären Parameter k dargestellt. Wenn i und j beide **wahr** sind, ist die Ähnlichkeit 1, sonst 0. Analog ist für nominale Parameter die Ähnlichkeit 1 wenn i und j den gleichen Wert haben, sonst ist die Ähnlichkeit 0.
Nach der Datenbearbeitung werden die generierten Datenmodelle für die Visualisierungen verwendet. In den folgenden Abschnitten werden die drei Visualisierungen und die Interaktionen beschrieben.

5.2 Dimensionsreduktion

Entsprechend der in Abschnitt 4.1 definierten Forschungsfragen zur multivariaten Visualisierung soll dem Nutzer zuerst eine Übersicht der Daten zur Verfügung stehen. Die Ansicht für die Multidimensionale Skalierung (MDS) stellt eine übersichtliche Visualisierung in `CTEDS` dar. Dabei bezieht sich die Visualisierung auf die Dimensionsreduktion der multivariaten Daten auf einem 2D-Projektionsraum.

5.2.1 Kategorien der Datendimensionen

Die zu analysierenden Daten des Systems sind multivariat und hochdimensional. Bei der Chipüberprüfung kommen folgende Parametertypen häufig in Testfällen vor. Um die Beschreibung der Parametertypen zu vereinfachen werden die Buchstaben in Klammern als Abkürzung zur Repräsentation der Parametertypen verwendet.

- Eingangsparameter (**E**) sind die Eingabeparameter eines Chiptestfalls, die eine Teilmenge von dem gesamten Parameterraum für Chipüberprüfung sind.
- Berechnete Parameter (**B**) sind die aus den Eingangsparametern berechneten Parameter.
- Gemessene Debug-Informationen (**D**) sind die Messungen der Debug-Informationen eines Chiptestfalls.
- Zielparameter (**Z**) sind ein Teil von den gemessenen Informationen. Die Parameter kennzeichnen fehlerhafte und fehlerfreie Testfälle. Die Ursachen der Zielparameter zu finden ist das Ziel vom `CTEDS`.

Mit den vier Hauptkategorien können die Parameter sich zu zwei großen Gruppen zusammensetzen. Der Parametertyp **B** stammt grundsätzlich von dem Parametertyp **E**, die beiden Typen gehören zu der Parametergruppe Eingabeparameter. Analog stammen die Parametertypen **D** und **Z** gleichzeitig

von den Messungen, daher gehören sie zu der Parametergruppe Ausgabeparameter.
Die Anzahl der Parameter kann zwischen Hunderten und Millionen liegen. Mit einer riesigen Parameteranzahl verliert man schnell den Überblick über die Eigenschaften und Zusammenhänge der Parameter. Das kann zur schwierigen Analyse der Daten führen.
Aufgrund von der großen Anzahl unterschiedlicher Parameter müssen die Datenparameter ihren entsprechenden Kategorien (**E, B, D, Z**) vor der Entwicklung des Systems zugeordnet werden. Die Datenparameter werden auch als Datendimensionen bezeichnet. Die Zuordnung der Kategorien ist eine grundlegende Voraussetzung für effiziente Visualisierungen der Daten und ihre Dimensionen.

5.2.2 MDS-Koordinaten

Vor der Visualisierung mithilfe der Technik MDS müssen die Koordinaten der Zieldimension für die Objekte der hochdimensionalen Daten bestimmt werden. Zur Berechnung der entsprechenden MDS-Koordinaten wird der Algorithmus 5.1 von [Pic09] für klassische Skalierung eingesetzt.

Algorithmus 5.1 Klassische Skalierung Algorithmus [Pic09]

function CLASSICALSCALING
Input: dissimilarity matrix $D \in \mathbb{R}^{n \times n}$, dimensionality $d \in \mathbb{N}$
Output: coordinate vectors $x_1, \ldots, x_d \in \mathbb{R}^n$
$B \leftarrow -\frac{1}{2} J D^2 J$
$(\lambda_1, \mu_1), \ldots, (\lambda_d, \mu_d) \leftarrow$ decompose B
for $i = 1, \ldots, d$ **do**
$x_i \leftarrow \sqrt{(\max \lambda_i, 0)} \cdot \mu_i$
end for
end function

(5.3)

$$B = UAU^T = \sum_{i=1}^{n} \lambda_i \mu_i \mu_i^T \quad , A = diag(\lambda_1, \ldots, \lambda_n), U = [\mu_1, \ldots, \mu_n]$$

$$X = [X^1, \ldots, X^d] = [\sqrt{(\lambda_1)} \cdot \mu_1, \ldots, \sqrt{(\lambda_d)} \cdot \mu_d]$$

Es ist von dem Algorithmus 5.1 auszugehen, dass der grundlegender Ausgangspunkt in klassischer Skalierung die Distanzmatrix $D = (d_{ij}) \in \mathbb{R}^{n \times n}$ ist. Die Distanzmatrix wird im Data Parser generiert. Die gewünschte Ausgabe des Algorithmus sind die Koordinaten $x_1, \ldots, x_n$ in d-Dimension. Der euklidische Abstand zwischen den Koordinaten entspricht in dem Algorithmus der Distanz d_{ij} in der Distanzmatrix: $\| x_i - x_j \| = d_{ij}$.
Die Ausgangsmatrix D wird bereits in Abschnitt 5.1 vorgestellt, D^2 repräsentiert D mit allen Einträgen in Quadrat. $J = I - \frac{1}{n} 1_n 1_n^T$ repräsentiert eine Zentriermatrix mit $1_n = [1, \ldots, 1]^T \in \mathbb{R}^n$ und der Einheitsmatrix $I = diag(1_n)$. Die Koordinatenmatrix $X \in \mathbb{R}^{n \times d}$ ist die Ausgabe des Algorithmus und somit $XX^T = B$. λ_i, μ_i sind die Eigenwerte und Eigenvektoren mit einheitlicher Länge von B. Es ist möglich, B wie die Gleichung 5.3 zu zerlegen. Entsprechend ist die Koordinatenmatrix X zu berechnen. Mit der Koordinatenmatrix X wird die Visualisierung mithilfe der MDS ermöglicht.

5.2.3 Visualisierung der MDS

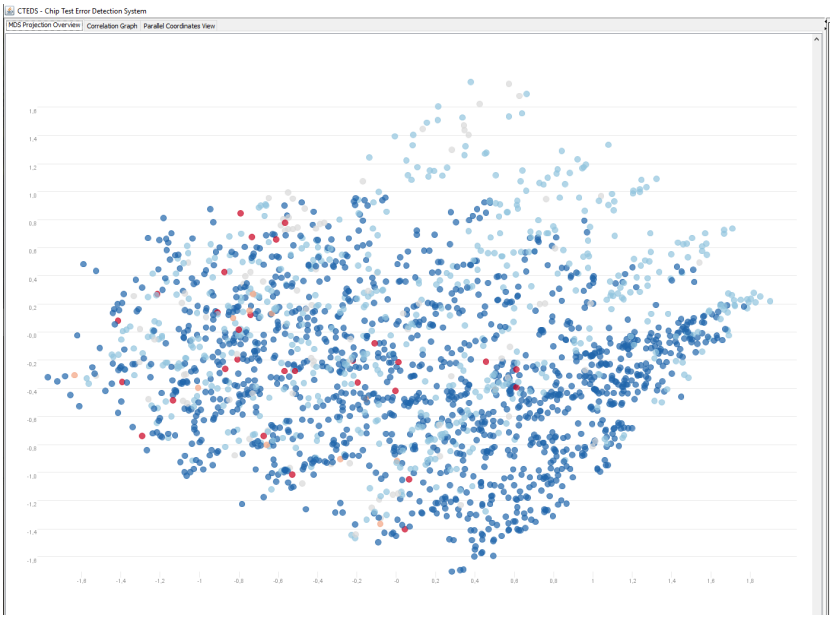


Abbildung 5.2: Ansicht der Multidimensionalen Skalierung mit fünf Farbkaregorien zur Unterscheidung der Testfälle

Die MDS-Ansicht visualisiert die von der Distanzmatrix generierten MDS-Koordinatenmatrix im Projektionsraum. In CTEDS ist der Projektionsraum zweidimensional, die Koordinatenmatrix ist ebenfalls zweidimensional definiert. Da die MDS-Ansicht als eine Übersichtsvisualisierung für die multivariaten Daten dient, wird die generierte Koordinatenmatrix als ein 2D-Streudiagramm (Scatterplot) visualisiert.

Die Abbildung 5.2 zeigt die MDS-Ansicht in der Form eines Streudiagramms. Die Achsen des Streudiagramms werden von der 2D-Koordinatenmatrix bestimmt. Jedem Testfall wird nach der Berechnung der MDS-Koordinaten in Abschnitt 5.2.2 ein Koordinatenpaar zugewiesen. Die Koordinatenpaare sind in ein kartesisches Koordinatensystem eingetragen. Dementsprechend definieren die Koordinatenpaare die Werte und den Umfang der X- und Y-Achsen des Streudiagramms. Die Testfälle der originalen Daten werden in der Ansicht als Punkte dargestellt. Die Punkte streuen in dem Streudiagramm anhand ihrer Koordinatenwerte. Die Distanz zwischen den abgebildeten Punkten i und j entspricht dem Wert d_{ij} in der vorberechneten Distanzmatrix.

Da CTEDS auf die Feststellung der Fehlerquellen abzielt, ist es wichtig, die Eigenschaften der Testfälle in der Visualisierung widerzuspiegeln. Ob die Testfälle fehlerhaft oder fehlerfrei sind, wird durch die Werte der Zielparameter (**Z**) bestimmt.
Mit den projizierten Punkten im Streudiagramm ist es möglich, mit verschiedenen Farben eines Farbschemas die Testfälle zu unterscheiden. Der Wert des **Z**-Parameters jedes Testfalls bestimmt die Farbe des entsprechenden Punkts im Streudiagramm. Wie in Abbildung 5.2 dargestellt haben die abgebildeten Punkte fünf unterschiedlichen Farben. Die Auswahl der Farben für die Testfälle ist in Abbildung 5.3 dargestellt. Die Farben sind divergierend ausgewählt. Von links nach rechts repräsentieren die Farben in der Farbpalette die Eigenschaften der Testfälle von fehlerhaft bis fehlerfrei.

Die rote Farbe auf der linken Seite und die dunkelblaue Farbe auf der rechten Seite vertreten jeweils die fehlerhaften und die fehlerfreien Testfälle. Die hellrote und hellblaue Farben repräsentieren jeweils die weniger fehlerhaften Testfälle und die weniger fehlerfreien Testfälle. In der Mitte der Farbpalette vertritt die graue Farbe die neutralen Testfälle.

Die Visualisierung erlaubt es dem Nutzer, die Datenpunkte anhand ihrer Ähnlichkeiten in Clustern erkennen. Der Nutzer kann die Cluster durch Interaktionen visuell ändern, um tiefe Informationen von den Clustern zu gewinnen. Folgend werden die Interaktionsmöglichkeiten in der MDS-Ansicht vorgestellt, die dem Nutzer weitere interaktive Analyse ermöglichen.

Abbildung 5.3: Die Auswahl der Farben für die MDS-Punkte

5.2.4 Interaktionsmöglichkeiten

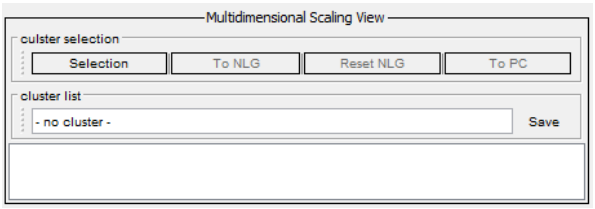


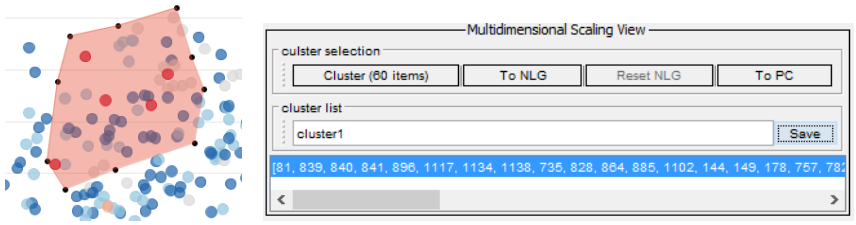
Abbildung 5.4: Der Steuerungsbereich der MDS-Ansicht in `CTEDS`

Der Steuerungsbereich für die MDS-Ansicht ist ein Unterbereich in `CTEDS`. In Abbildung 5.4 ist der Bereich für die Interaktionen der MDS-Ansicht zu sehen. Zwei Funktionen sind möglich: Selektion der Cluster im cluster selection und Speicherung der Cluster in einer Liste cluster list. Folgend wird zuerst die Lasso-Selektion Funktion vorgestellt. Anschließend wird die Funktion für Speicherung der Cluster beschrieben.

Lasso-Selektion

Interaktive Untersuchung der Untermenge einer MDS-Projektion zählt zu einem der wichtigsten Konzepte vom `CTEDS`. Neben der Übersichtsvisualisierung ist es erforderlich, den Nutzer durch Interaktionen bei der Analyse der Datenpunkte zu unterstützen. Interaktive Selektion der Punkte in dem Projektionsraum ist eine Möglichkeit für weitere Analyse der interessanten Punkte. Die Punkte können ein Cluster bilden und in anderen visuellen Ansichten dargestellt werden. Die Methode wird als *Brush-Linking* bezeichnet.

Übliche *Brushing*-Methoden ermöglichen die Selektion der Datenpunkte mit einer Maus. Die selektierte Teilmenge wird dabei hervorhoben. Durch konventionelle Hervorhebung der Teilmenge verlieren die nicht-selektierten Datenpunkte ihre Farben, was zum Verlust der Details der Datenpunkte führen kann. Um die Eigenschaften der Testfälle zu behalten werden die selektierten Punkte in der MDS-Projektion nicht hervorhoben. In Abbildung 5.5(a) ist die *Brushing*-Funktion vom `CTEDS` zu sehen. Die zu selektierende Teilmenge der Datenpunkte lässt sich von einem halb durchsichtigen orangen



(a) (b)

Abbildung 5.5: (a) Lasso-Selektion: Selektierte Datenpunkte in der MDS-Projektion sind von einem orangen Polygon markiert. (b) Ein Cluster ist durch die Lasso-Selektion der Punkte generiert. Das Cluster hat 60 Punkte und wird als cluster1 in der Cluster-Liste gespeichert.

Polygon bedecken. Die Datenpunkte gehören zu der selektierten Teilmenge und bilden ein Cluster, wenn ihre Mittelpunkte innerhalb des Polygons liegen.

Die Form einer konventionellen *Brushing*-Selektion ist meistens ein Rechteck mit dem Maus-Klick-Drag. Allerdings ist ein rechteckiger Selektionsbereich nicht flexibel genug für eine freie Auswahl in einem Streudiagramm. Um die freie Auswahl zu ermöglichen ist der Selektionsbereich in CTEDS ein freies Polygon in beliebiger Form. Die Abbildungen in Abbildung 5.6 stellen die möglichen Polygone für eine freie Lasso-Selektion dar. Alle Polygone bestehen aus mehreren Punkten, mit denen die Orientierung der Kanten der Polygone festgelegt wird. Alle funktionsfähigen Polygone gehen von einem Ausgangspolygon aus, das in Abbildung 5.6(a) zu sehen ist. Durch einen Maus-Klick mit der linken Maustaste wird ein Punkt in der MDS-Projektion erzeugt wird. Das Ausgangspolygon besteht aus zwei Punkten P_1 und P_2, die eine Linie $\mathbf{L}_{P_1P_2}$ bilden. Durch einen weiteren Maus-Klick außerhalb der Linie $\mathbf{L}_{P_1P_2}$ entsteht ein dritter Punkt P_3, der das Ausgangspolygon erweitert. Und somit ist ein neues konvexes Polygon für die Lasso-Selektion möglich. Ein Beispiel für ein Polygon $\mathbf{P}_{P_1P_2P_3}$ mit drei Punkten ist in Abbildung 5.6(b) zu sehen. Analog sind konvexe Polygonen mit N Punkten $\{P_1 \dots P_n\}$ für eine größere Selektion möglich, solange der neu hinzugefügte Punkt P_n stets außerhalb des bereits erzeugten Polygons $\mathbf{P}_{P_1 \dots P_{n-1}}$ liegt. Die Polygone in Abbildung 5.6(c) und Abbildung 5.6(d) zeigen beispielsweise die konvexen Polygone mit jeweils vier und fünf Punkten.
Das Selektionspolygon kann nicht nur konvex sondern auch konkav sein. Ein Beispiel für ein konkaves Polygon ist in Abbildung 5.6(e) zu zeigen. Acht Punkte $P_1, \dots, P_8$ bilden ein konkaves Polygon. Der letzte Punkt P_8 liegt innerhalb des Polygon $\mathbf{P}_{P_1 \dots P_7}$. Konkave Polygone sind hilfreich für spezielle Selektionen der Datenpunkte.

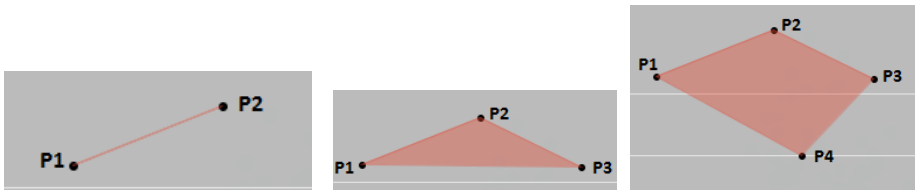
Cluster

Durch die Lasso-Selektion in Abbildung 5.5(a) wird ein Cluster generiert. Von der Abbildung 5.4 ist es auszugehen, dass die Tasten To NLG und To PC im cluster selection Bereich ursprünglich deaktiviert sind. Die Cluster-Liste ist auch leer. Mit einer Lasso-Selektion wird die Anzahl des aktuellen Clusters in dem cluster selection Bereich gezeigt. Wie in Abbildung 5.5(b) dargestellt ermöglicht die Selektion ein Cluster mit 60 Punkten. Außerdem wird das Cluster als cluster1 benannt und in der Cluster-Liste gespeichert. Das gespeicherte Cluster in dem cluster list Bereich zeigt den Index der selektierten Punkte. Klickt man auf das Cluster, werden die Buttons To NLG und To PC im cluster selection

Bereich aktiviert. Die Buttons ermöglichen Interaktionen zwischen der MDS-Ansicht und anderen Ansichten.
Der Nutzer kann auf den Button To NLG klicken, wenn ein Cluster in der Cluster-Liste ausgewählt wird. Dadurch wird das Cluster in einen Bipartiten Graph transformiert, der in der Ansicht des Node-Link-Graphen dargestellt wird. Folgend wird die NLG-Ansicht vorgestellt. Dabei werden zwei Graphenlayouts für die Korrelationen visualisiert. Einige Interaktionen werden auch für die Korrelationsanalyse eingeführt.

5.3 Korrelationsanalyse

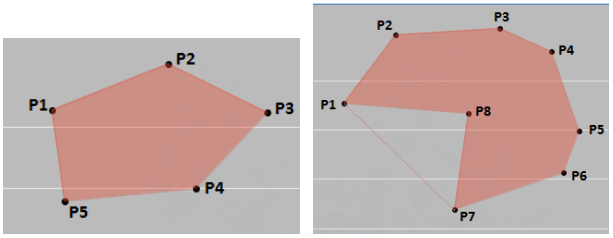
Nach der Übersichtsvisualisierung in der MDS-Ansicht kann der Nutzer die Korrelationen zwischen den Parametern in einem Node-Link-Graph analysieren. Die Ansicht für die Korrelationsanalyse (NLG-Ansicht) bezieht sich hauptsächlich auf die Visualisierung der Korrelationen zwischen den Parametern. Anders als die Datenpunkte in der MDS-Ansicht in Abschnitt 5.2 sind die visuellen Komponenten des Node-Link-Graphen die Parameter der multivariaten Daten. Die Parameter werden als Knoten des Graphen visualisiert. Die Korrelationen zwischen den Parametern werden als Kanten des Graphen visualisiert. In der Ansicht kann der Nutzer die Fehler und die entsprechenden Parameter mithilfe der Graphen erkennen. Die Navigation durch interessante Parameter-Tupel wird mithilfe der Interaktionen ermöglicht. Somit können die potenziellen Fehlerquellen festgestellt werden.



(a) Das Ausgangspolygon mit Punkten P1 und P2

(b) Ein konvexes Polygon mit drei Punkten

(c) Ein konvexes Polygon mit vier Punkten



(d) Ein konvexes Polygon mit fünf Punkten

(e) Ein konkaves Polygon mit Punkten P1 bis P8. Der Punkt P8 ermöglicht eine konkave Form des Polygons.

Abbildung 5.6: Lasso-Selektion-Polygone mit aufsteigender Anzahl der Polygonspunkte: (a) bis (d) sind konvexe Polygone. (e) zeigt ein konkaves Polygon.

5.3.1 Korrelationenberechnung

Die Berechnung der Korrelationen ist ein wichtiger Bestandteil der Korrelationsanalyse. In der Pipeline des Data Parser in Abbildung 5.1 befindet sich die Berechnung der Korrelationen nach der Phase des Parsing. Durch das Parsing der Rohdaten ist ein Datenmodell für die Parameterwerte und ihre Kategorien zu generieren, womit die Korrelationen zwischen den Parametern berechnet werden. Die Ergebnisse werden in einer Korrelationen-HashMap gespeichert. Anhand der Kategorien der Parameter können die Korrelationen so vorberechnet, dass die **Z**-Parameter mit den Parametern anderer Kategorien zusammen analysiert werden.
Der Grad der Zusammenhänge zwischen den **Z**-Parametern und anderen Parametern spielt eine wichtige Rolle. Solche Zusammenhänge werden in CTEDS nicht als statistische Korrelationen betrachtet, sondern als die Informationen über die **Z**-Parameter basierend auf die gegebenen Informationen über andere Parameter. Unter dieser Berücksichtigung wird die bedingte Entropie über die **Z**-Parameter als die Korrelationen berechnet.

Die bedingte Entropie $H(X|Y)$ einer Variable X wird anhand der bekannten Information über die Variable Y bestimmt. Laut der Definition der Entropie ist die bedingte Entropie $H(X|Y)$ folgend in Gleichung (5.4) definiert.

(5.4)

$$H(X|Y) = \sum_{y\in\mathcal{Y}} p(y)H(X|Y=y) = -\sum_{y\in\mathcal{Y}} p(y) \sum_{x\in\mathcal{X}} p(x|y)\log p(x|y) = \sum_{y\in\mathcal{Y},x\in\mathcal{X}} p(x,y)\log\frac{p(y)}{p(x,y)}$$

(5.5) $H(X|Y) = H(X,Y) - H(Y)$

Eine vereinfachte Rechnung abgleitet aus der Gleichung (5.4) zeigt die bedingte Entropie der Variable X in Gleichung (5.5). Die Unsicherheit von X gegeben Y ist gleich der Unsicherheit von X und Y abzüglich der Unsicherheit von Y. Dabei ist die Joint-Entropie $H(X,Y) = -\sum_x \sum_y P(x,y)\log_2[P(x,y)]$ so definiert. Analog ist eine Joint-Entropie für N Variablen $H(X_1,...,X_n) = -\sum_{x_1} ... \sum_{x_n} P(x_1,...,x_n)\log_2[P(x_1,...,x_n)]$ zu berechnen.
Um die bedingten Entropie der **Z**-Parameter $H(Z|X_1 \dots X_n)$ zu berechnen müssen zuerst die Joint-Entropie $H(Z,X_1,\dots,X_n)$ und die bekannten Entropie der N Variablen $H(X_1,\dots,X_n)$ bestimmt werden. Daher ist $H(Z|X_1 \dots X_n) = H(Z,X_1,\dots,X_n) - H(X_1,\dots,X_n)$. Um die Joint-Entropie zu ermöglichen müssen die Joint-Wahrscheinlichkeit berechnet werden.
Um eine gute Joint-Wahrscheinlichkeit der Parameter zu bekommen müssen die zu analysierenden numerischen Parameter auf eine gleiche Weise normalisiert werden. Jeder numerische Parameter von multivariaten Daten kann als ein Vektor V mit n Elementen betrachtet werden. Nach der Normalisierung wird V in einen neuen Vektor N der gleichen Länge mit ganzzahligen Elementen umgewandelt. Um eine relativ uniforme Verteilung der Elemente in V zu schaffen wird V in einige Klassen geteilt. Die Anzahl der Klassen wird durch die Anzahl der Parameter einer Joint-Wahrscheinlichkeit bestimmt. Beispielsweise hat eine Joint-Wahrscheinlichkeit $P(X_0,\dots,X_{m-1})$ m Parameter, die Anzahl der Klassen für jedes Parameter in P ist $|\text{Klassen}| = \sqrt[m+1]{n}$ bezüglich der Anzahl der Parameter m und der Länge des Parameters n.
Mit Joint-Wahrscheinlichkeiten für beliebige Anzahl der Parameter wird die Joint-Entropie der Parameter mit entsprechender Anzahl berechnet. Dadurch ist Berechnung der bedingten Entropie eines bestimmten Parameters möglich.

5.3.2 Node-Link-Graph

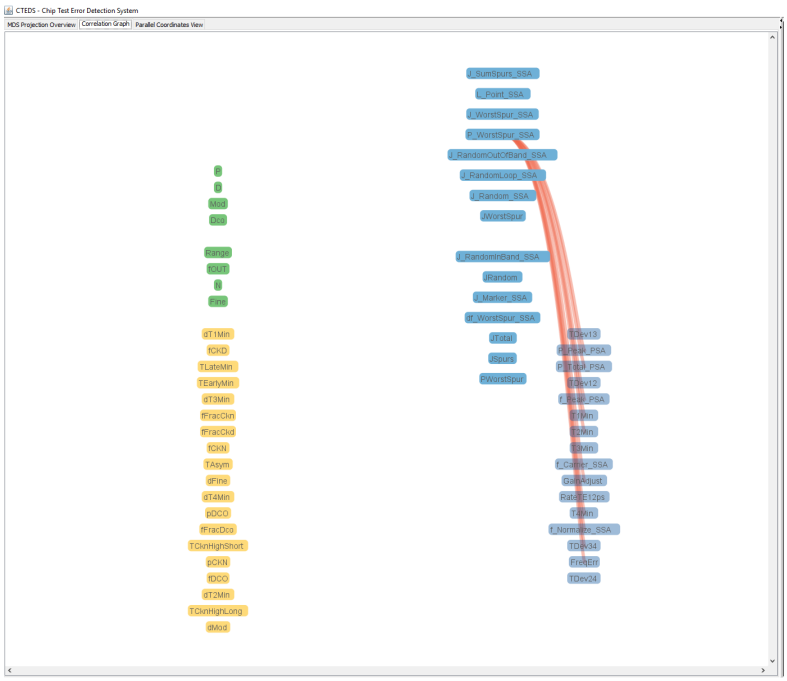


Abbildung 5.7: Übersicht des Node-Link-Graphen als Bipartiter Graph

In der NLG-Ansicht werden die Parameter der multivariaten Daten anhand ihrer Kategorien als Knoten des Graphen visualisiert. Die Korrelationen zwischen den Parametern werden als Kanten des Graphen visualisiert.
In Abbildung 5.7 ist die Übersicht des Node-Link-Graphen angezeigt. Die Parameter werden aufgrund ihrer Kategorien in unterschiedlichen Farben visualisiert, sie bilden einen Bipartiten Graph [Har69]. Ein Bipartiter Graph $G = (V, E)$ besteht aus zwei disjunkten Teilmengen von Knoten V, wobei keine Kanten zwischen den Knoten innerhalb beider Teilmengen verlaufen.
Die Parameter werden anhand ihrer Kategorien geteilt. Die Kategorien sind bereits in Abschnitt 5.2.1 definiert. Die Eingabeparameter mit Parametertypen **E** und **B** bilden sich eine vertikale Linie auf der linken Seite des Graphen. Gegenüber der Eingabeparameter sind die Ausgabeparameter mit den Parametertypen **Z** und **D** auch in einer vertikalen Linie auf der rechten Seite des Graphen dargestellt. Um die Kategorien der Parameter genau zu unterscheiden haben die Parameter jeder Kategorie eine Farbe. In Abbildung 5.7 sind die grünen Knoten die **E**-Parameter, die gelben Knoten sind die **B**-Parameter, die dunkelblauen Knoten sind die **Z**-Parameter und die hellblauen Knoten sind die **D**-Parameter. Andere Farbzuweisungen sind auch möglich für die Darstellung der Parameter. Es ist wichtig, die Farben konsequent in `CTEDS` zu behalten, damit die Parameter in anderen Ansichten visuelle Assoziationen haben.
In der NLG-Ansicht sind die zehn besten Korrelationen zwischen den **Z**- und **D**-Parametern in orange Kanten dargestellt. Anhand der Werte der Korrelationen haben die Kanten unterschiedliche Kantenbreite und Farbsättigung. Mit höheren Korrelationen sind die Kanten breiter und mehr gesättigt. Mit niedrigeren Korrelationen sind die Kanten dünner und weniger gesättigt.

5.3.3 Interaktionsmöglichkeiten

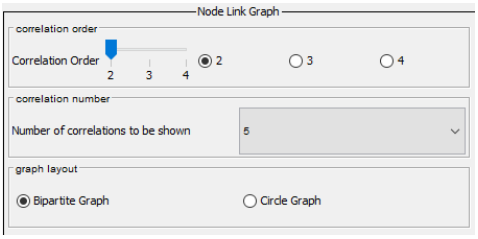


Abbildung 5.8: Steuerungsbereich für die NLG-Ansicht

Die Interaktionen bei der Ansicht des Node-Link-Graphen (NLG) werden in dem Steuerungsbereich ermöglicht. In Abbildung 5.8 ist der Bereich für die Interaktionen der NLG-Ansicht zu sehen. Zwei Interaktionsfunktionen sind verfügbar: Die Darstellung der Korrelationen in verschiedenen Ordnungen ist durch die Funktionen in Bereichen correlation order und correlation number möglich. Die Umschaltung des Graphenlayouts für Node-Link-Graph ist im Bereich graph layout möglich.

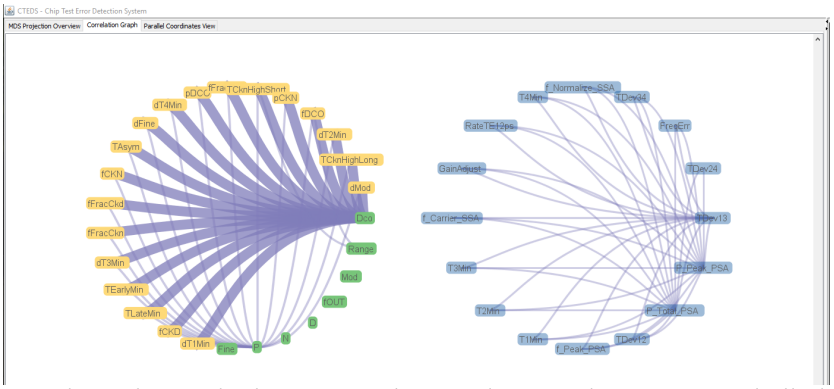


Abbildung 5.9: Node-Link-Graph als Kreisgraph zeigt die Korrelationen innerhalb der Eingabeparameter und innerhalb der Ausgabeparameter.

Graphenlayout

Das vorgegebene Graphenlayout von NLG ist der Bipartite Graph in Abbildung 5.7. Der Bipartite Graph visualisiert hauptsächlich die Kategorien der Parameter. Im Bereich graph layout des Steuerungsbereichs für die NLG-Ansicht sind zwei Radiobuttons für die Umschaltung des Graphenlayouts verfügbar. Wenn der Radiobutton circle graph gewählt wird, ändert sich der Node-Link-Graph in einen Kreisgraph in Abbildung 5.9. Der Kreisgraph besteht aus zwei Unterkreisgraphen. Auf der linken Seite ist der Unterkreisgraph der Eingabeparameter mit **E**- und **B**-Parametern. Die internen Korrelationen von den Eingabeparametern sind als Kanten des Unterkreisgraphen in hellviolett dargestellt. Auf der rechten Seite ist der Unterkreisgraph der **D**-Parameter. Die internen Korrelationen von den **D**-Parametern sind ebenfalls als Kanten des Unterkreisgraphen in hellviolett dargestellt. Anhand der Werte der Korrelationen haben alle Kanten unterschiedliche Kantenbreite und Farbsättigung.

Korrelationen

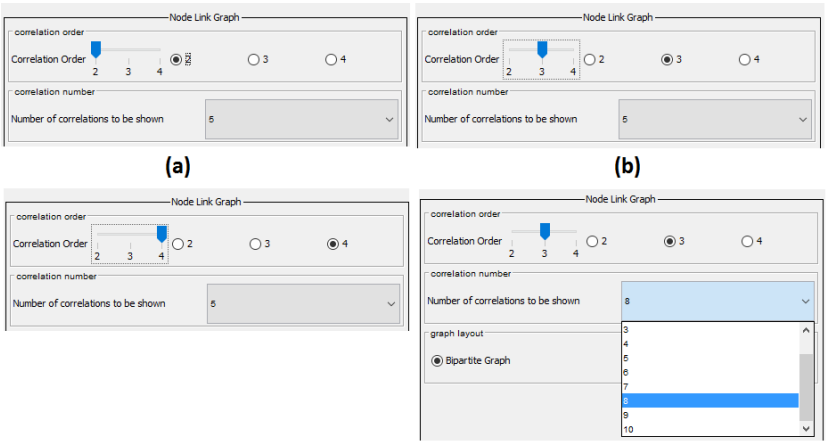


Abbildung 5.10: Systemsteuerung für die Korrelationen des Node-Link-Graphen. Die 5 besten Korrelationen werden gezeigt ((a) bis (c)): (a) Schieberegler und Radio-Button mit dem Wert 2 aktiviert die 2. Korrelationsordnung; (b) Schieberegler und Radio-Button mit dem Wert 3 aktiviert die 3. Korrelationsordnung; (c) Schieberegler und Radio-Button mit dem Wert 4 aktiviert die 4. Korrelationsordnung. (d) Die änderbare Anzahl der zu zeigenden Korrelationen.

Um die Fehlerquellen der **Z**-Parameter zu finden, ist es wichtig, die Korrelationen zwischen den **Z**-Parametern und den anderen Parametern zu analysieren. CTEDS bietet in der NLG-Ansicht Interaktionen für die Darstellung der Korrelationen in verschiedenen Ordnungen an. Drei Korrelationsordnungen sind in Abschnitt 5.3.1 vorberechnet. Eine Korrelation von einem Parameterpaar wird als Korrelation der 2. Ordnung bezeichnet. Eine Korrelation von einem Parametertripel ist die Korrelation der 3. Ordnung. Eine Korrelation von einem Parametertupel (4 Parameter) ist die Korrelation der 4. Ordnung. Beim Mausklick auf einen **Z**-Parameter nämlich auf einen dunkelblauen Knoten in dem Bipartiten Graph entstehen mehrere neue Verbindungsknoten.

In Abbildung 5.12(b) lassen sich die fünf besten Korrelationen der 2. Ordnung zwischen dem **Z**-Parameter JRandom und einem Parameter aus anderen Kategorien anzeigen. Der Parameter JRandom ist im NLG als ein dunkelblauer Knoten mit seinem Parameternamen visualisiert. Beim Mausklick behalten die betroffenen Knoten nämlich JRandom und seine fünf am besten korrelierten Partnerparameter ihre ursprüngliche Farben. Die irrelevanten Knoten werden ausgegraut. Die entstandenen Korrelationen sind mit Verbindungsknoten zu präsentieren. In Abbildung 5.12(b) sind die rosa Verbindungsknoten mit absteigenden Sättigungen dargestellt. Die Verbindungsknoten enthalten die Ordnungsanzahl, den Korrelationswert und die korrelierten Parameternamen. Entsprechend sind die Verbindungsknoten aufgrund dem Wert der Korrelationen in einer absteigender Reihenfolge sortiert. Die Verbindungsknoten mit dem größeren Korrelationswert stehen höher als die mit dem niedrigeren Wert. Wenn die Verbindungsknoten den gleichen Korrelationswert haben, werden sie untereinander zusammengesetzt.
Die Anzeige der Korrelationen der 2. Ordnung ist durch die Interaktionen der Systemsteuerung in Abbildung 5.10 **(a)** realisiert. Die Korrelationsordnung kann entweder durch den Schieberegler oder

durch die Radiobuttons im Bereich correlation order geändert werden. Der Wert von dem Drop-Down-Box im Bereich correlation number ist die Anzahl der zu zeigenden Korrelationen der 2. Ordnung. Analog lassen sich die Korrelationen der 3. beziehungsweise der 4. Ordnung bei der Selektion des Parameters JRandom generieren. Die Korrelationsordnungen sind in Abbildung 5.10 **(b)** und Abbildung 5.10 **(c)** dargestellt. Der Wert von dem Drop-Down-Box ermöglicht fünf besten Korrelationen der 3. beziehungsweise der 4. Ordnung. Die Visualisierungen in Abbildung 5.12(c) und Abbildung 5.12(d) zeigen die entsprechenden Korrelationen. Die Verbindungsknoten für die Korrelationen der 3. Ordnung sind in orange visualisiert. Die Verbindungsknoten für die Korrelationen der 4. Ordnung sind in grün visualisiert.
Die Interaktionen sind dynamisch. Wie in Abbildung 5.10 **(d)** gezeigt können die Korrelationsordnung und die Anzahl der Korrelationen vor oder nach der Selektion eines Parameters geändert werden. Mit einer festgelegten Ordnung oder Anzahl der Korrelationen ist es möglich, andere **Z**-Parameter jederzeit zu wählen. Die Korrelationen werden sofort geändert und aktualisiert.

In der MDS-Ansicht kann der Nutzer die Übersicht der Daten anschauen. In der NLG-Ansicht kann der Nutzer die Korrelationen der Parameter anschauen. Eine Kombination von den beiden Ansichten wird folgend in der Ansicht der Parallelen Koordinaten dargestellt. In dieser kann der Nutzer sowohl eine Übersicht über die Datenpunkte als auch die hochdimensionalen Strukturen zwischen den Parametern gewinnen.

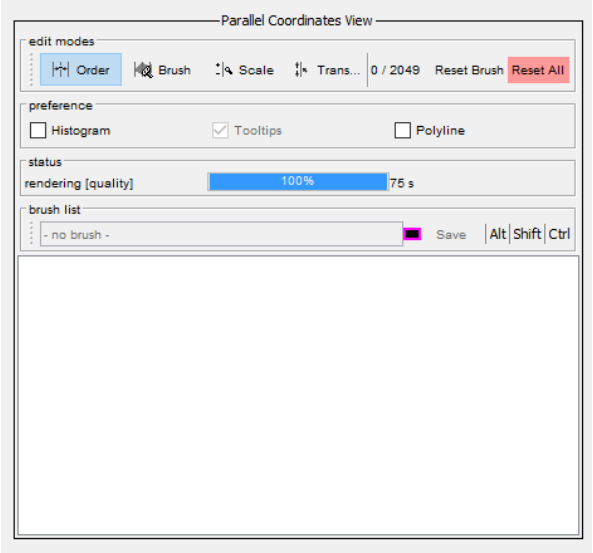


Abbildung 5.11: Steuerungsbereich für die Ansicht der Parallelen Koordinaten

5.4 Parallele Koordinaten

Die Ansicht der Parallelen Koordinaten (PK-Ansicht) dient als eine Ergänzung für die Visualisierungen mithilfe der Technik MDS und NLG. Dabei kann der Nutzer nach der Analyse mit der MDS-Ansicht und der NLG-Ansicht die ausführliche Struktur der Daten anschauen. Somit kann der Nutzer ein besseres Verständnis für die wichtigen Parameter bekommen.

5.4.1 Koordinierte Visualisierung

Die erforderlichen Daten zur Visualisierung der Parallelen Koordinaten lassen sich in Abschnitt 5.1 im Data Parser generieren. Aus dem Datenmodell nach dem Parsing ist eine HashMap mit Parameterwerten und Dateneinträgen abgeleitet. Um die ähnlichen Visualisierungen bei der MDS-Ansicht und bei der NLG-Ansicht zu ermöglichen, werden die Parameter für Parallele Koordinaten ebenfalls anhand ihrer Kategorien gruppiert werden. Die Farbe der Linienzüge haben die Farbe der Datenpunkte in der MDS-Ansicht.
In Abbildung 5.13 ist die Übersicht der Parallelen Koordinaten zu sehen. Die Dateneinträge sind als Linienzüge visualisiert. Die Farbzuweisung der Linienzüge entspricht der Farbzuweisung der Datenpunkte bei der MDS-Ansicht in Abbildung 5.2. Die Farben der Linienzüge weisen auf die Eigenschaften der Dateneinträge hin. Die bläuliche Linienzüge beziehen sich auf die fehlerfreien Testfälle, die rötlichen Linienzüge vertreten die fehlerhaften Testfälle. Die grauen Linienzüge sind neutrale Testfälle.
Die vertikalen Achsen der Parallelen Koordinaten vertreten die Parameter der Daten. Beim NLG werden die Parameter anhand ihrer Kategorien gruppiert, sie haben verschiedene Farben zur Unterscheidung der Kategorien. Die Parameter bei der Parallelen Koordinaten haben die gleichen Kategorien beziehungsweise die gleichen Farben. In Abbildung 5.13(a) und (b) sind die Parameter unterhalb der PK-Ansicht mit den gleichen Farben der Knoten wie in Abbildung 5.7 dargestellt. Die gelben Achsen in Abbildung 5.13(a) sind die **B**-Parameter. Die grünen Achsen sind die **E**-Parameter. Die dunkelblauen Achsen sind die **D**-Parameter. In Abbildung 5.13(b) sind die **Z**-Parameter als hellblaue Achsen dargestellt.

5.4.2 Interaktionsmöglichkeiten

CTEDS bietet die PK-Ansicht folgende Interaktionen an, mit denen die Analyse der Daten veranschaulicht werden. Die Interaktionen sind in dem Steuerungsbereich in Abbildung 5.11 zu sehen. Im Bereich edit modes sind vier Funktionen verfügbar. Die Anordnung der Achsen ist möglich durch die Funktion Order. Teilmenge der Linienzüge zu wählen ist möglich durch die Funktion Brush. Die Funktionen Scale und Tranlate ändern die Skalierung und die Translation der Achsen. Im Bereich preference sind Darstellung der Histogramme, Hervorhebung der Linienzüge möglich. Wenn die Funktion Brsuh aktiviert ist, können die selektierten Teilmengen als Brush in der Brush-Liste (brush list) gespeichert werden.

Brushing

Die Abbildung 5.14 zeigt die Interaktion der Brushing-Funktion an. Beim Drücken der linker Maustaste kann man entlang einer Achse beliebig viel Linienzüge wählen, indem man die Maus senkrecht nach unten oder nach oben zieht. Die ausgewählten Linienzüge werden entsprechend hervorhoben. Die nicht ausgewählten Linienzüge werden ausgegraut. Die Brushing-Funktion hat die gleiche Auswirkung wie die Lasso-Selektion bei der MDS-Ansicht in Abschnitt 5.2.4. Die Brushing-Funktion bei Parallelen Koordinaten veranschaulicht im Vergleich zu der MDS die Orientierung der Linienzüge beziehungsweise die Zusammenhänge der Parameter.

Anordnung der Achsen

Die Abbildung 5.15 zeigt die Interaktion der Anordnung der Achsen an. Die Funktion ist entscheidend für die Suche nach Strukturen in den Daten. Die Korrelationen der Parameter in Parallelen Koordinaten werden paarweise dargestellt, es ist daher wichtig, die Achsen richtig anzuordnen. Somit haben die weit voneinander stehenden Achsen die Möglichkeit, die Korrelation der Parameter zu zeigen.
Auf der linken Seite der Abbildung 5.15 befinden sich zwei Achsen A_1 und A_2 ganz weit. Die zwei Achsen sind jeweils mit einem Rechteck in Magenta markiert. Es ist schwer die Korrelation zwischen den beiden Achsen zu analysieren. Mit der Anordnungsfunktion ist die Anzeige der Korrelation zwischen den beiden Achsen möglich. Beim Drücken der linken Maustaste zieht man die linke Achse A_1 horizontal nach rechts, so dass die A_1 sich direkt rechts von der Achse A_2 befindet. Nach der Anordnung ist die Korrelation zwischen den beiden Achsen A_2 und A_1 zu betrachten. Anhand dem Muster der Linienzüge zwischen den beiden Achsen ist die Korrelation relativ negativ.

Darstellung der Histogramme und Hover über einem Linienzug

Die Abbildung 5.16 zeigt die Darstellung der Histogramme. Mit der Aktivierung der Funktion der Histogramme werden die Verteilung der Parameterwerte jeder Achse als Histogramm angezeigt. Das ist wichtig für die Skalierung der Achsen. Abbildung 5.17 zeigt die Funktion der Hervorhebung eines Linienzugs. Die Parameterwerte des Linienzug werden als Tooltip angezeigt.

5.5 Interaktionen zwischen Ansichten

In Abschnitt 4.2.2 wurde der Arbeitsablauf vom `CTEDS` vorgestellt. Durch die Interaktionen zwischen der MDS-Ansicht und der NLG-Ansicht kann der Nutzer verschiedene Projektionen bezüglich der multivariaten Daten ermöglichen. Die Ansichten hängen sich dadurch zusammen.
Die Abbildung 5.18 zeigt die Voraussetzung für die Verbindung der Ansichten. Der Nutzer muss zuerst ein Cluster durch die Lasso-Selektion (siehe Abschnitt 5.2.4) erzeugen. Die Datenpunkte des Clusters sind in Abbildung 5.18(a) unterhalb des orangen Polygons zu sehen. In dem Steuerungsbereich der MDS-Ansicht (siehe Abbildung 5.18(b)) sind die Buttons nach der Generierung des Clusters aktiviert. Die Tool-Tip-Anzeige zeigt die Funktion des Buttons To NLG. Mit dem Button kann das Cluster in einen Bipartiten Graph in der NLG-Ansicht umgewandelt werden.
Die Abbildung 5.19 zeigt die Eigenschaften des umgewandelten NLG aus dem Cluster. In Abbildung 5.19(a) befinden sich zwei Node-Link-Graphen nebeneinander. Der linke NLG mit dem weißen Hintergrund ist der vollständige Graph mit allen Testfällen der Daten. Der rechte NLG mit dem hellblauen Hintergrund ist der Untergraph aus dem Cluster. Die Testfälle des Untergraphen sind nur eine Teilmenge von den ganzen Testfällen. Der Unterschied zwischen den beiden NLG ist deutlich gezeigt. Mit den gleichen Parametern sind die zehn besten Korrelationen zwischen **Z**-Parametern und **D**-Parametern in beidem Graphen anders dargestellt. Die Korrelationen in dem linken NLG stammen aus einem **Z**-Parameter. Die Korrelationen in dem rechten NLG stammen aus zwei **Z**-Parametern, wobei ein **Z**-Parameter der gleiche **Z**-Parameter im linken NLG ist.
Der Untergraph aus dem Cluster besitzt die gleichen Interaktionen bezüglich der NLG-Ansicht. Mit

einem Mausklick auf einen **Z**-Parameter in dem Untergraph werden die Korrelationen angezeigt. In Abbildung 5.19(b) sind die sechs besten Korrelationen der 2. Ordnung zwischen **JRandom** and einem anderen Parameter in beidem NLG zu sehen. Mit der Teilmenge der Testfälle sind die Korrelationen der 2. Ordnung in dem Untergraph anders wie die in dem vollständigen Graph. Analog sind die Korrelationen der 3. und 4. Ordnungen auf die gleiche Weise zu sehen.
Mit der Clusterspeicherung in der Liste können mehrere vorhandenen Cluster in Node-Link-Graphen umgewandelt werden. Somit können die Cluster und die erzeugten Untergraphen miteinander verglichen und analysiert werden.
Wenn ein **Z**-Parameter in der NLG-Ansicht ausgewählt ist, können der Parameter und seine dargestellten korrelierten Parameter zusammen in der PK-Ansicht gezeigt werden. Der Nutzer kann in der PK-Ansicht die zweidimensionalen Strukturen zwischen den Parametern anschauen und somit spezifische Merkmale bei den Parametern entdecken.

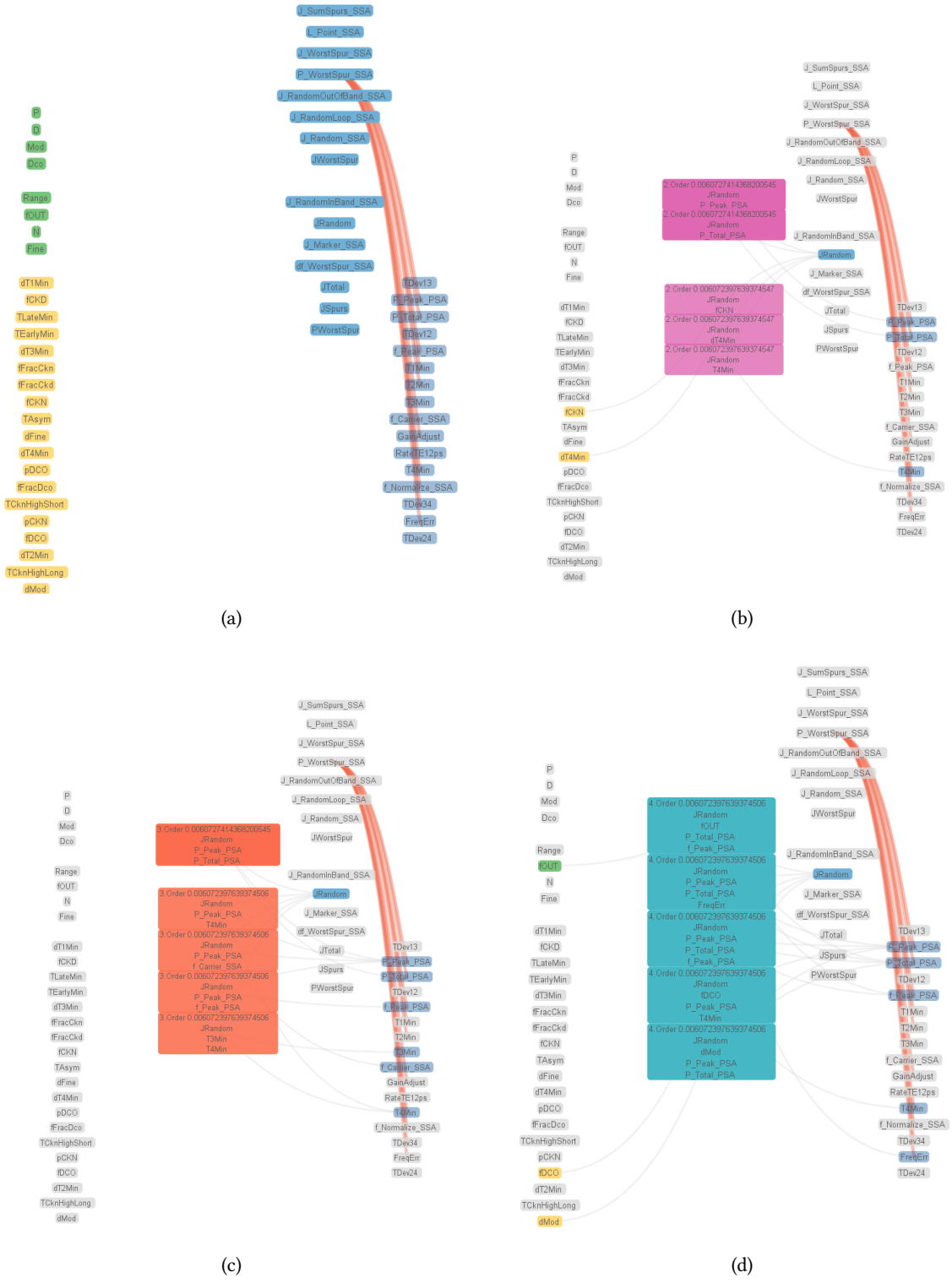
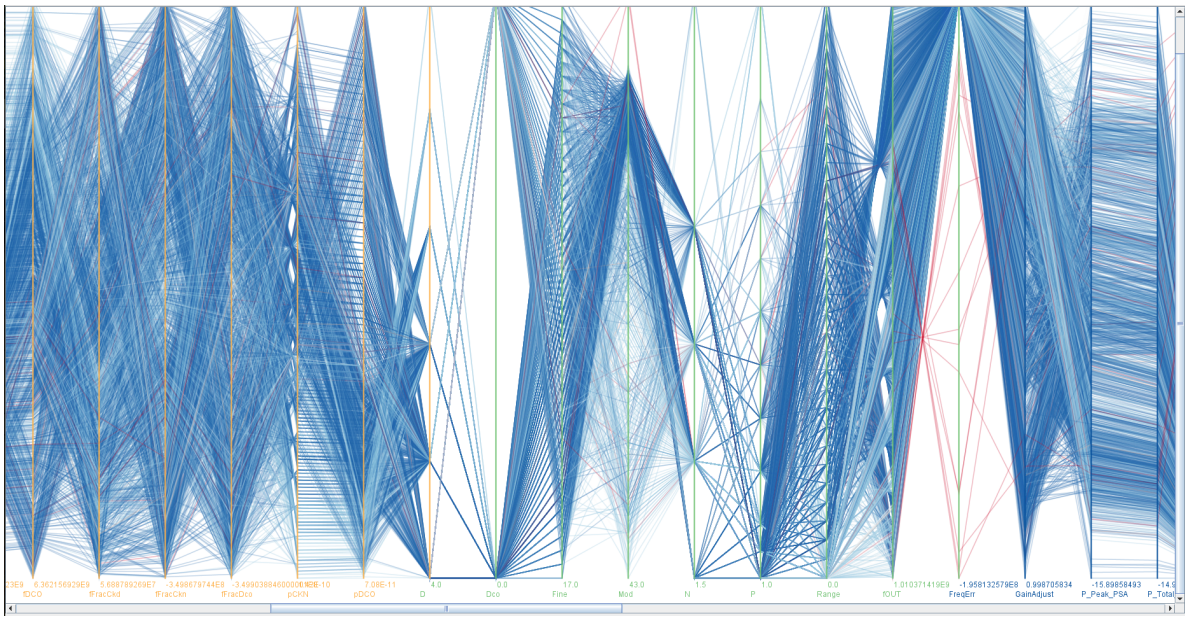
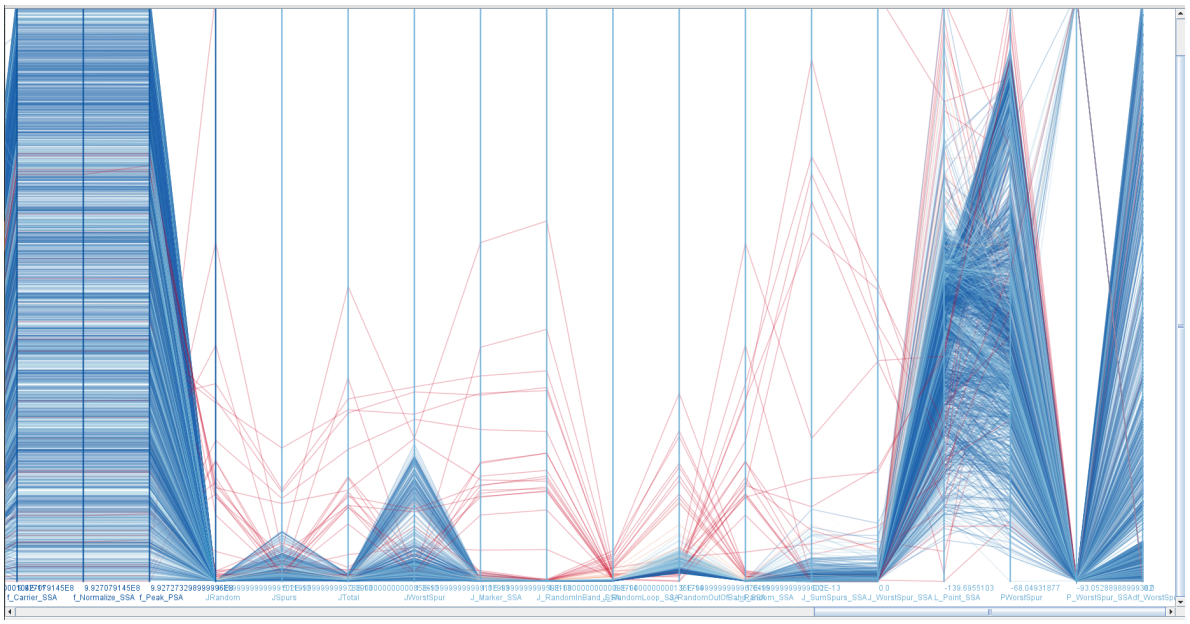


Abbildung 5.12: Die 5 besten Korrelationen zwischen einem Zielparameter und anderen Parametern in verschiedenen Ordnungen: (a) Ausgangsgraph ohne Selektion eines Zielparameters; (b) Korrelationen der 2. Ordnung zwischen JRandom und einem anderen Parameter; (c) Korrelationen der 3. Ordnung zwischen JRandom und Parameterpaaren; (d) Korrelationen der 4. Ordnung zwischen JRandom und Parametertripeln.



(a)



(b)

Abbildung 5.13: Übersicht der Parallele Koordinaten: 5.13(a) zeigt die erste Hälfte der Parallelen Koordinaten mit den Parametern drei Kategorien; 5.13(b) zeigt die andere Hälfte der Parallelen Koordinaten mit den Parametern letzter Kategorie.

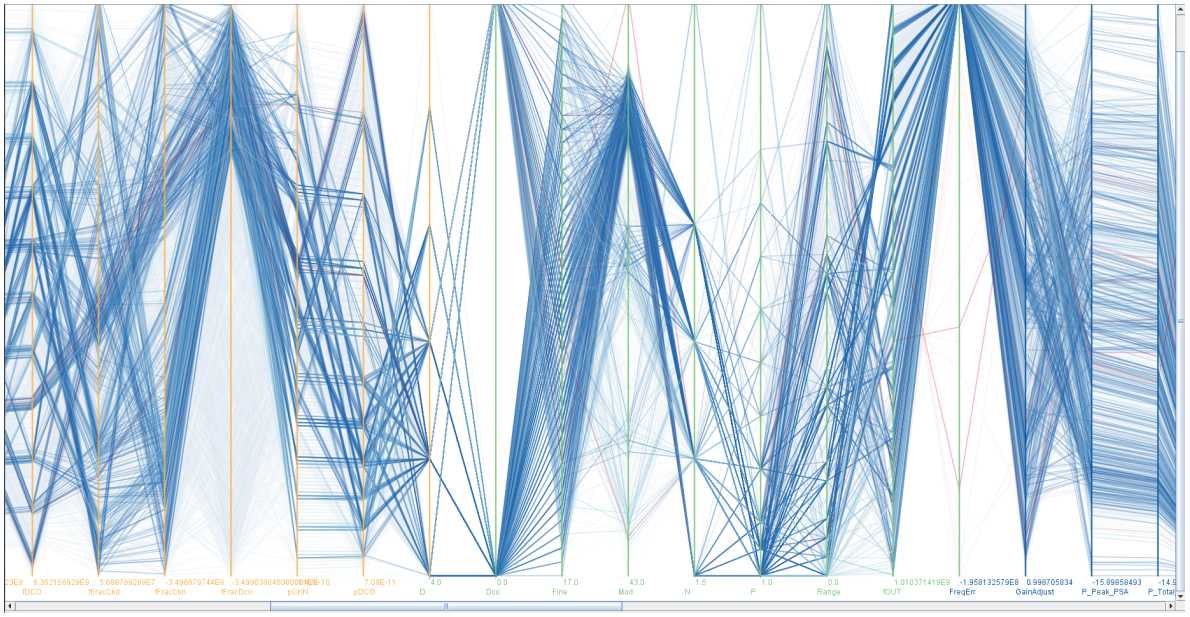


Abbildung 5.14: Brushing-Funktion für die Selektion der Linienzüge der Parallelen Koordinaten.

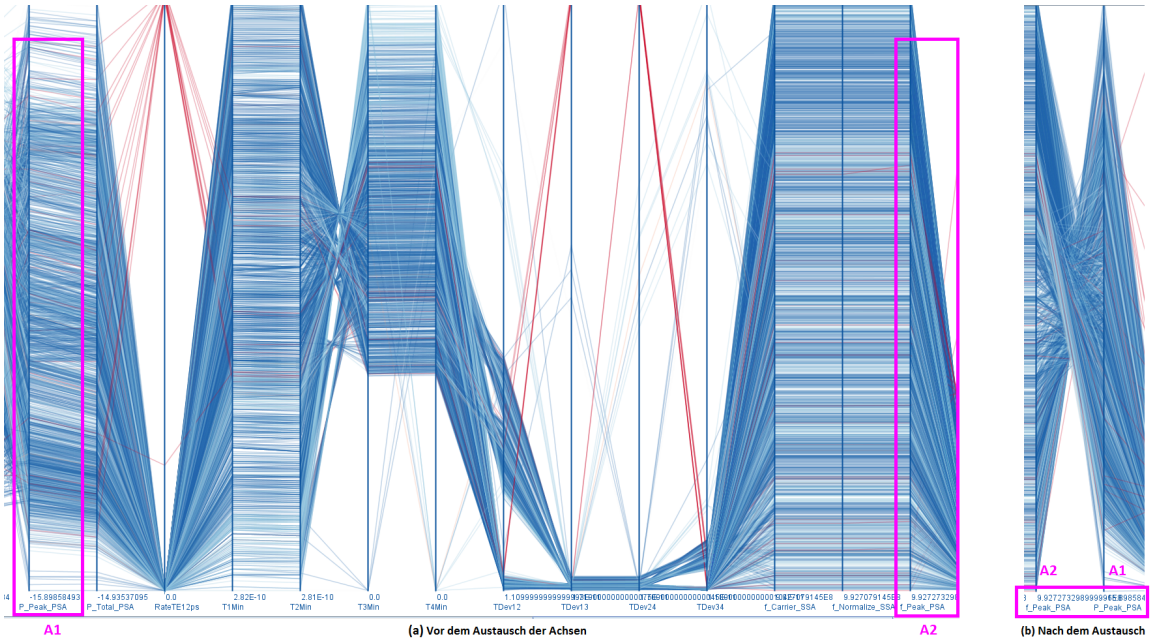


Abbildung 5.15: Anordnung der Achsen: (a) zwei Achsen markiert von einem Rechteck in Magenta sind weit voneinander; (b) die markierten Achsen sind mit der Anordnungfunktion nebeneinander sortiert

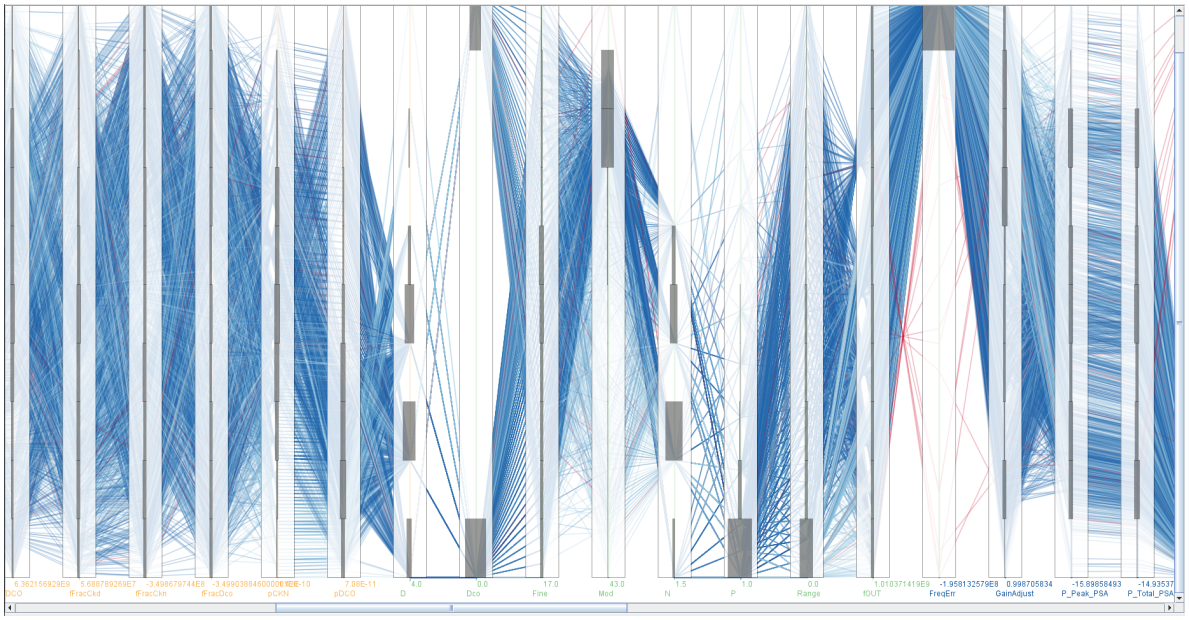


Abbildung 5.16: Die Histogramme zeigen die Verteilung der Parameterwerte

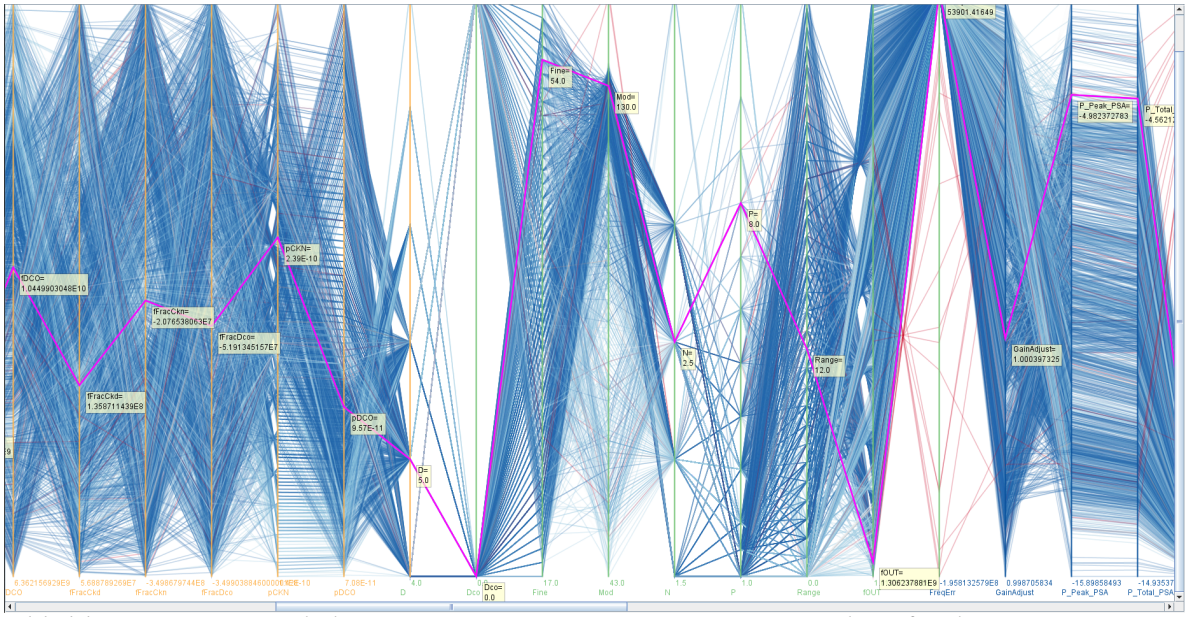
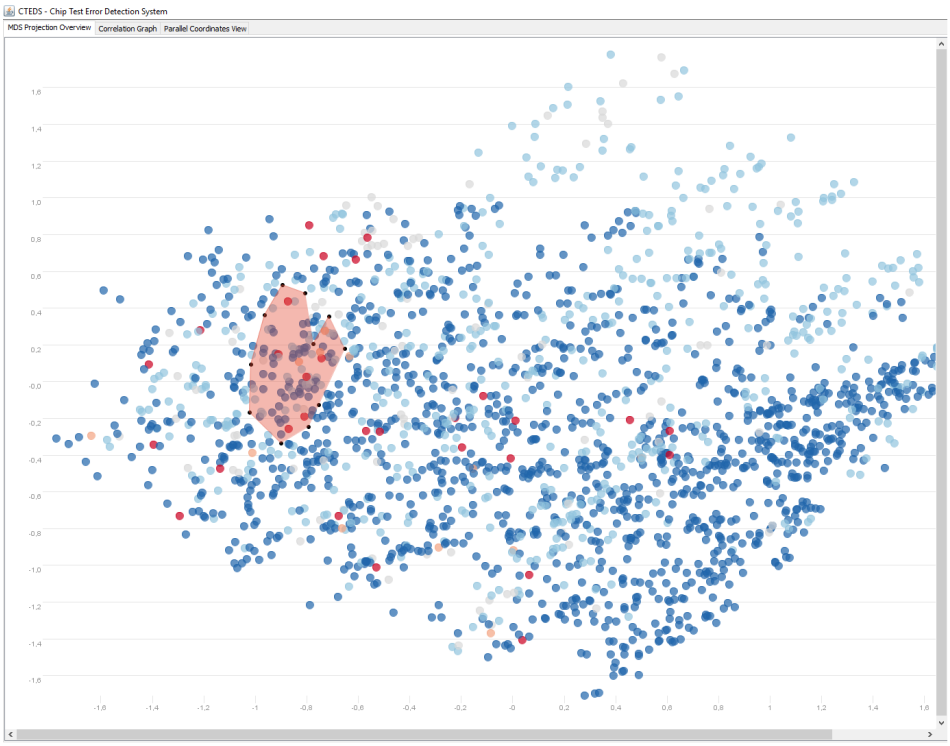
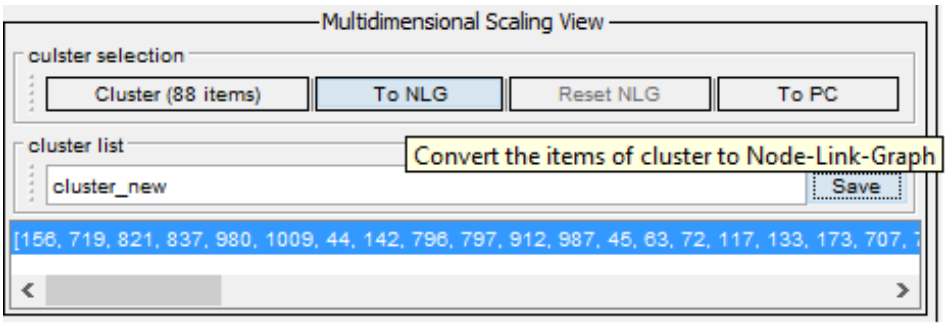


Abbildung 5.17: Hervorhebung eines Linienzugs in Magenta mit Tooltips für die Parameterwerte des Datenpunkts.

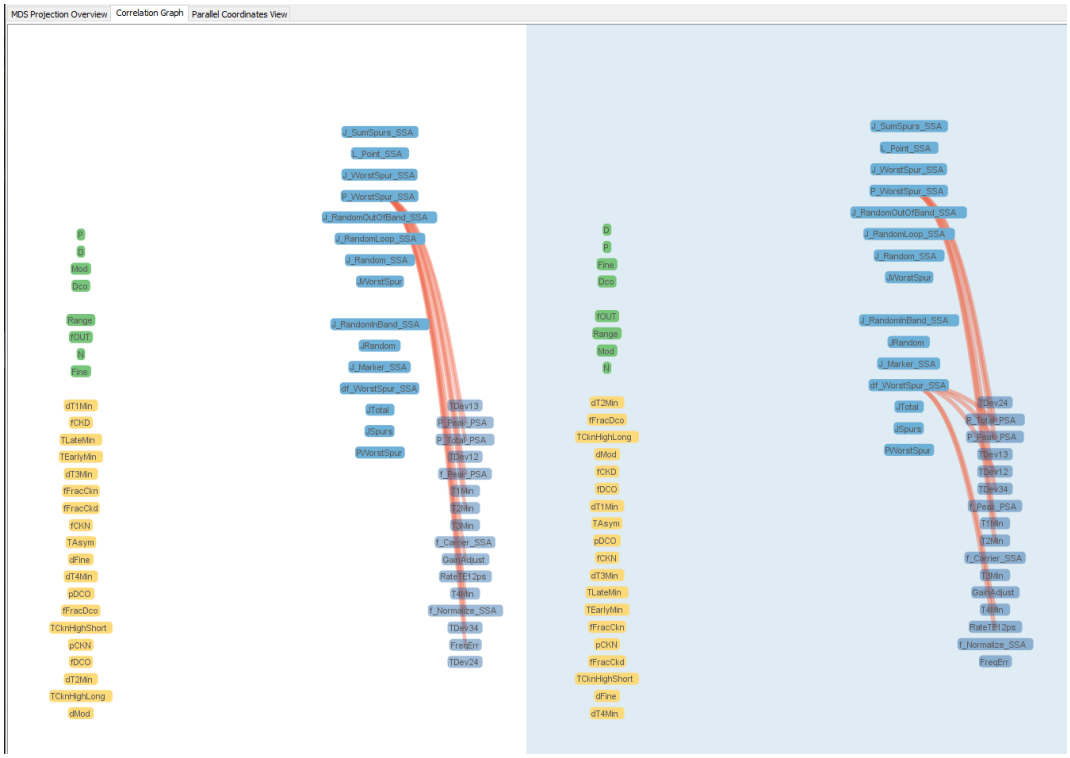


(a)

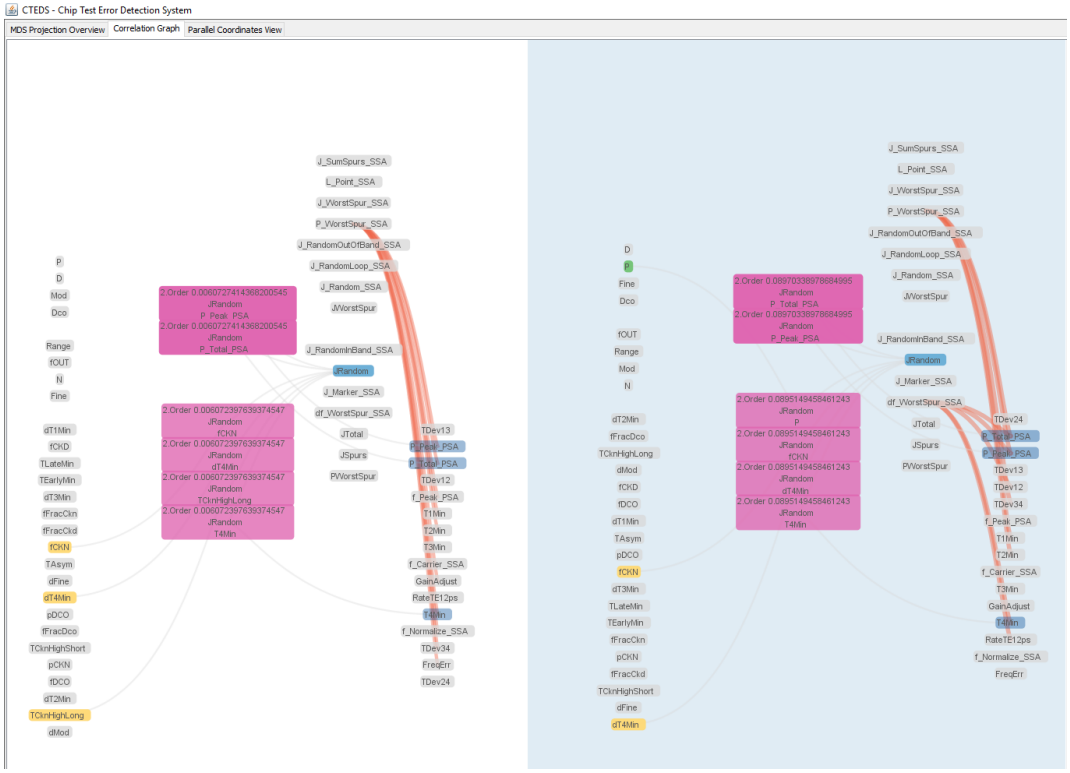


(b)

Abbildung 5.18: (a) Lasso-Selektion eines Clusters der Datenpunkte bei der MDS-Ansicht; (b) Interaktionen der MDS-Ansicht: Das Cluster kann in einen Node-Link-Graph umgewandelt werden und in der NLG-Ansicht gezeigt werden.



(a)



(b)

Abbildung 5.19: (a) NLG-Ansicht mit dem vollständigen Node-Link-Graph (links) und einem Untergraph (rechts). Der Untergraph ist von dem Cluster in der MDS-Ansicht generiert; (b) Die 6 besten Korrelationen der 2. Ordnung zwischen JRandom and anderem Parameter in beidem NLG.

6 Auswertung

In diesem Kapitel wird das umgesetzte Systems CTEDS anhand von einem geeigneten Datensatz ausgewertet. Zuerst wird der verwendete Testdatensatz für die Auswertung des Systems in Abschnitt 6.1 vorgestellt. Anschließend wird ein Anwendungsfall in Abschnitt 6.2 durchgeführt. Dabei werden die Ergebnisse des Systems vorgestellt.

6.1 Testdatensatz

Für die Auswertung des umgesetzten Systems CTEDS wird hochdimensionaler Testdatensatz verwendet. Der Datensatz ist ein umfangreicher realer Datensatz, der von den Experten der Hardwareindustrie angeboten wird. Der Testdatensatz enthält insgesamt 2049 Testfälle für die Chip- und Schaltkreisüberprüfung, die jeweils aus 87 Parametern bestehen. Die ersten 5 Parameter jedes Testfalls sind die Parameter des Haushalts, welche keinen Einfluss auf die multivariate Datenanalyse ausüben. Deswegen werden die Haushaltsparameter in dem System komplett ignoriert. Die restlichen 82 Parameter werden anhand der Kenntnisse der Experten in vier Kategorien unterteilt, die unterschiedliche Auswirkungen auf die Datenanalyse haben. Die Kategorien der Parameter sind wie folgt definiert:

- Eingangsparameter (**E**) sind die Eingabe einer Chipüberprüfung, die eine Teilmenge von dem gesamten Parameterraum für Chipüberprüfung sind.
- Berechnete Parameter (**B**) sind die aus den Eingangsparametern berechneten Parameter. Die Parameter und die **E**-Parameter werden zusammen als Eingabeparameter betrachtet.
- Gemessene Debug-Informationen (**D**) sind die Messungen der Debug-Informationen eines Testfalls der Chipüberprüfung.
- Zielparameter (**Z**) sind ein Teil von den gemessenen Informationen, die fehleranfällig sind. Die Parameter kennzeichnen fehlerhafte und fehlerfreie Testfälle. Die Entdeckung der Ursachen der Zielparameter ist das Ziel des System CTEDS. Die **Z**-Parameter und die **D**-Parameter werden zusammen als Ausgabeparameter betrachtet.

In dem Testdatensatz sind für jeden Testfall insgesamt 19 **E**-Parameter, 21 **B**-Parameter, 25 **D**-Parameter sowie 16 **Z**-Parameter vorhanden. Unter den gesamten Parametern sind 11 davon konstante Parameter, deren Einträge konstante Werte sind. Um die Präzision der Berechnung der Korrelationen zwischen Parametern zu erhöhen, werden die konstanten Parameter bei der Berechnung der Korrelationen entfernt.
Da das System zur Entdeckung der Fehlerquellen der fehlerhaften Testfälle entwickelt ist, werden die **Z**-Parameter bezüglich deren Korrelationen zu anderen Parametern hauptsächlich betrachtet

und analysiert. Unter den **Z**-Parametern ist der Parameter JTotal das hauptsächliche fehleranfällige Parameter, in dem die restlichen fehleranfälligen **Z**-Parameter enthalten sind. Prinzipiell besteht JTotal aus JRandom und JSpurs, welche jeweils weitere Unterparameter der Kategorie **Z**-Parameter enthalten. Hierzu soll das System CTEDS sich entweder auf das **Z**-Parameter JRandom oder auf das **Z**-Parameter JSpurs konzentrieren, um die Fehlerquellen der fehleranfälligen Parameter festzustellen. Nach der Anforderung der Experten analysiert das System CTEDS das **Z**-Parameter JRandom als ein Fokusparameter, mit dem die Fehlerquellen untersucht werden. Da der Parameter JRandom ein gemessenes Parameter in jedem Testfall der Chipüberprüfung ist, hat JRandom 2049 Einträge mit einem Maximum $1,16 \times 10^{-9}$ und einem Minimum $1,46 \times 10^{-12}$. Die Standardabweichung der Werte von JRandom beträgt $3,73869 \times 10^{-11}$. Die Testfälle mit JRandom von Wert größer als 1×10^{-11} werden als fehlerhafte Testfälle bezeichnet, während die fehlerfreien Testfälle mit JRandom von Wert kleiner als 2×10^{-12} sind. Die Unterscheidung der Testfälle anhand JRandom ist in Tabelle 6.1 aufgelistet.

Wert von JRandom	**Eigenschaften der Testfälle**
$[\max, \quad 1 \times 10^{-11})$	fehlerhaft
$[1 \times 10^{-11}, \quad 4 \times 10^{-12})$	weniger fehlerhaft
$[4 \times 10^{-12}, \quad 2,5 \times 10^{-12})$	neutral
$[2,5 \times 10^{-12}, \quad 2 \times 10^{-12})$	weniger fehlerfrei
$[2 \times 10^{-12}, \quad \min]$	fehlerfrei

Tabelle 6.1: Die Eigenschaften der Testfälle sind anhand JRandom festgelegt. [,) weist auf ein Intervall der Werte von JRandom hin, welches den rechten Randwert ausschließt. [,] ist ein Intervall, das beide Randwerte einschließt.

6.2 Anwendungsfall

Um das System auszuwerten wird ein Anwendungsfall mit dem in Abschnitt 6.1 vorgestellten Datensatz durchgeführt.

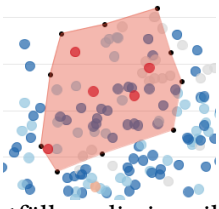


Abbildung 6.1: Cluster 1 enthält 50 Testfällen, die jeweils fehlerhaft, neutral, weniger fehlerfrei, fehlerfrei sind.

Zuerst kann der Nutzer die MDS-Ansicht betrachten. Dabei ist eine Übersicht der 2049 Testfälle vorhanden (siehe Abbildung 5.2). Der Nutzer kann von der Übersicht ausgehen, dass die Testfälle auf einem Streudiagramm projiziert werden. Die meisten Testfälle sind in der Übersicht fehlerfrei und streuen relativ gleichmäßig im Streudiagramm. Die fehlerhaften Testfälle, die in rot dargestellt sind, streuen meistens auf der linken Seite des Streudiagramms. Um die fehlerhaften Testfälle weiter zu

analysieren kann der Nutzer ein Cluster generieren. Der Nutzer verwendet die Lasso-Selektion und generiert ein Cluster mit 50 Testfällen (siehe Abbildung 6.1). Das Cluster enthält vier Testfalltypen: fehlerhaft, neutral, weniger fehlerfrei und fehlerfrei. Durch Klicken auf den Button To NLG in der MDS-Ansicht kann der Nutzer das ausgewählte Cluster in einen Bipartiten Graph umwandeln, der in der NLG-Ansicht dargestellt wird.

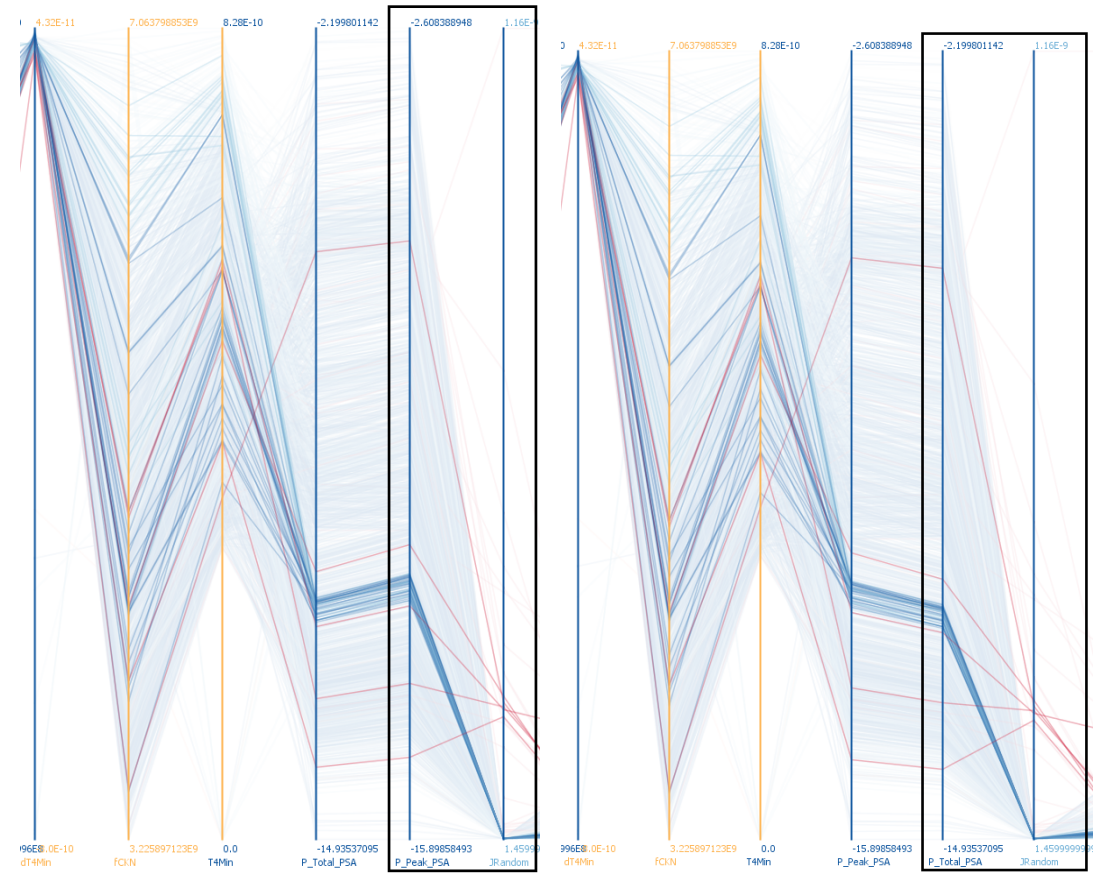
Durch Auswählen der Registerkarte Correlation Graph in der Benutzeroberfläche kann der Nutzer von der MDS-Ansicht zu der NLG-Ansicht wechseln. Der aus dem Cluster generierte Untergraph befindet sich in der NLG-Ansicht mit dem hellblauen Hintergrund (siehe Abbildung 6.2(a)). Um wichtige Korrelationen zwischen einem Parameter-Tupel bei dem Untergraph zu gewinnen, kann der Nutzer auf einen **Z**-Parameter klicken. Von der Übersicht in der MDS-Ansicht kann der Nutzer ausgehen, dass der **Z**-Parameter JRandom für fehlerhafte Testfälle verantwortlich ist. Durch Klicken auf JRandom bei dem Untergraph sind fünf Verbindungsknoten in rosa generiert (siehe Abbildung 6.2(b)). Die Verbindungsknoten zeigen die fünf besten Korrelationen zwischen JRandom und einem anderen Parameter des Untergraphen. Der Nutzer kann damit feststellen, dass zwei **D**-Parameter P_Peak_PSA und P_Total_PSA jeweils am besten mit JRandom korreliert sind. Die Korrelationen jeweils zwischen JRandom und dT4Min, JRandom und fCKN sowie JRandom und T4Min sind die zweit besten Korrelationen.
Nach der Entstehung der Korrelationen werden die relevanten Parameter des Untergraphen in der PK-Ansicht hervorhoben (siehe Abbildung 6.3). Um eine ausführliche Struktur zwischen JRandom und den fünf Parametern zu erkennen, kann der Nutzer durch Auswählen der Registerkarte Parallel Coordinates View zu der PK-Ansicht wechseln. Der Nutzer kann durch die Interaktion Order in der PK-Ansicht die Achsen beliebig anordnen, damit die Parameter jeweils mit JRandom eine zweidimensionale Struktur bilden können. In Abbildung 6.3 ((a) bis (e)) kann der Nutzer die Strukturen der Linienzüge zwischen JRandom und dem jeweiligen Parameter erkennen.
Um weitere interessante Parameter-Tupel bezüglich JRandom zu finden, kann der Nutzer wieder zu der NLG-Ansicht zurückgehen. Dabei kann der Nutzer die Korrelationsordnung in der NLG-Ansicht auf die 3. Ordnung ändern. Dadurch sind die fünf besten Korrelationen der 3. Ordnung in Abbildung 6.2(c) zu sehen. Dabei kann der Nutzer ebenfalls feststellen, dass das Parameterpaar (P_Peak_PSA, P_Total_PSA) mit JRandom am besten korreliert ist. Analog kann der Nutzer die Korrelationen der 4. Ordnung generieren. Fünf beste Korrelationen beziehungsweise die Parameter-Tupel sind in Abbildung 6.2(d) zu erkennen. Für jede Änderung der Korrelationen werden die Darstellungen der Parameter und Testfälle in der PK-Ansicht entsprechend aktualisiert. Der Nutzer kann jederzeit zu der PK-Ansicht wechseln, um die Strukturen genauer zu analysieren.

Anhand des Anwendungsfalls kann man feststellen, dass `CTEDS` eine geeignete interaktive Darstellungsplattform sowohl für die Darstellung allgemeiner Dateninformationen als auch für die Darstellung spezifischer Beziehungen innerhalb der Daten bietet. Dem Nutzer wird dadurch ermöglicht, wichtige Korrelationen zwischen Fehlern und Parametern zu erkennen.
Jedoch muss bei der MDS-Ansicht beachtet werden, dass es Projektionsfehler gibt. Durch mathematische Berechnung einer Distanzmatrix wird die Koordinatenmatrix der Multidimensionalen Skalierung generiert, bei der die Koordinaten nicht immer den richtigen Distanzen zwischen zwei Testfällen entsprechen. Hierbei ergibt die Übersicht keine vollständigen Informationen über die Datenpunkte und ihre Parameter sondern nur eine Approximation. Der Hinweis auf die Projektionsfehler soll in der Übersicht mit Interaktionen ermöglicht werden.

(a) (b)

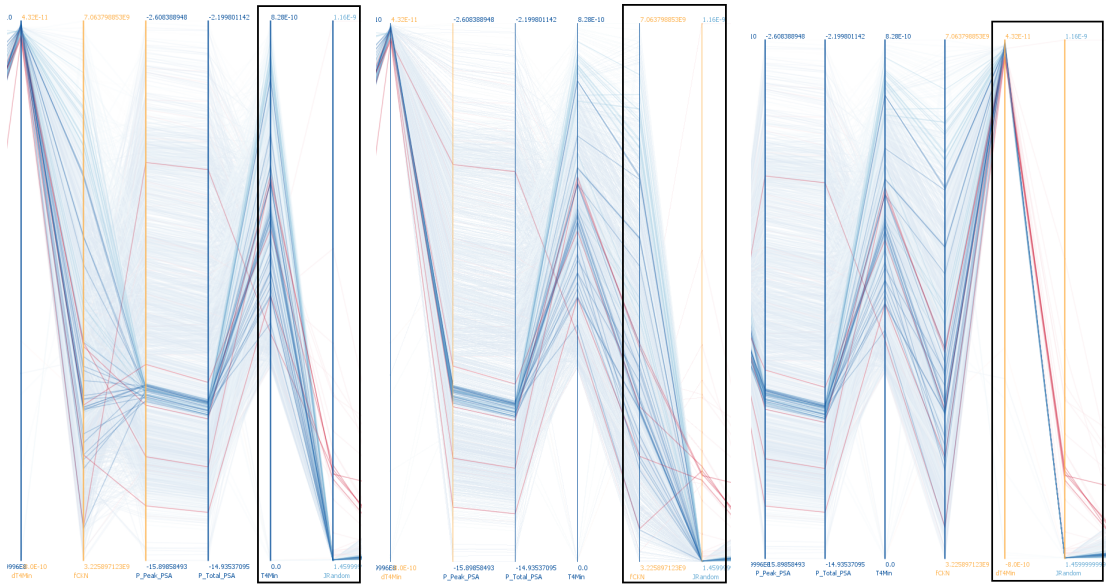
(c) (d)

Abbildung 6.2: Cluster 1 mit 50 Testfällen wird in einen Untergraph umgewandelt. Er ist in der NLG-Ansicht dargestellt. (a): Ausgangsgraph des Clusters; (b) - (d): Die fünf besten Korrelationen der verschieden Korrelationsordnungen zwischen JRandom und anderen Parametern.



(a) P_Peak_PSA und JRandom

(b) P_Total_PSA und JRandom



(c) T4Min und JRandom

(d) fCKN und JRandom

(e) dT4Min und JRandom

Abbildung 6.3: Die Testfälle von dem Cluster 1 und die fünf am besten mit JRandom korrelierten Parameter der 2. Korrelationsordnung in dem Untergraph werden nach dem Auswählen des JRandom in der PK-Ansicht hervorhoben. Die irrelevanten Testfälle sind dabei ergraut.

7 Zusammenfassung und Ausblick

In Rahmen dieser Masterarbeit ist ein Visual-Analytics-System `Chip Testing Error Detection System (CTEDS)` zur multivariaten Datenanalyse entwickelt worden. Das System ist eine Java-Anwendung basierend auf einer Java-Bibliothek für Informationsvisualisierung und geeignet für die Analyse der großen CSV- oder TSV-Dateien in Hochdimensionen. Die Eingangsdaten für das umgesetzte System sind ein multivariater Datensatz mit mehreren Testfällen der Chipüberprüfungen aus der Hardwareindustrie. Für jeden Testfall sind eine große Menge von Parametern in dem Datensatz vorhanden, die einen hochdimensionalen Parameterraum bilden. Das Ziel des Systems ist die Feststellung der Fehlerquellen der in den Testfällen aufgetauchten Fehler. Dabei kommt die Analyse der Zusammenhänge zwischen Fehlern und entsprechenden Parametern der Daten eine hohe Bedeutung zu. Eine erfolgreiche Vorbearbeitung der multivariaten Rohdaten ist im dem System vorausgesetzt für zutreffende grafische Darstellungen der Datenelemente. Daher sind visuelle Präsentationen der Daten und geeignete Interaktionstechniken wichtige Grundlagen des Systems, die dem Nutzer erlauben, große Mengen von Informationen auf einmal zu erforschen und zu verstehen.

Das System besteht aus drei visuellen Ansichten, die jeweils eine Visualisierung und entsprechende Interaktionstechniken anbieten. Die erste Ansicht der Multidimensionalen Skalierung beschäftigt sich mit der Darstellung der Übersicht über die Testfälle des Datensatzes. Die Visualisierung der Multidimensionalen Skalierung ist ein Scatterplot, wo die Testfälle als Punkte dargestellt werden. Durch einen klassischen Algorithmus der Multidimensionalen Skalierung ist eine Koordinatenmatrix zu erzeugen. Die Koordinaten bestimmen die projizierten Positionen der Datenpunkte in dem Scatterplot. Es ist anschaulich zu sehen, wie sich die projizierten Testfälle aufgrund ihrer Ähnlichkeiten in einem zweidimensionalen Raum befinden.
Da es Fehler in den Testfällen gibt, werden die projizierten Datenpunkte farblich unterschieden. Die fehlerhaften Testfälle und fehlerfreien Testfälle werden anhand von einem Zielparameter der Daten in gegensätzlichen Farben einer divergierenden Farbpalette abgebildet. Durch die spezielle Farbzuweisung sind die verschiedenen Testfälle übersichtlich zu sehen. Mögliche Cluster der Datenpunkte in dem Scatterplot sind dadurch einfach zu erkennen.
Eine interaktive Bruhsing-Funktion ermöglicht die Abspeicherung der Cluster in dem System. Die abgespeicherten Cluster können interaktiv in andere Visualisierungen umgewandelt werden.

Die Ansicht des Node-Link-Graphen ist die zweite visuelle Ansicht des Systems. Bei der Ansicht handelt es sich um die Darstellung der Datenparameter und ihre Korrelationen. Da der Node-Link-Graph als ein Bipartiter Graph voreingestellt wird, befinden sich die Parameter als Knoten des Graphen in einer vertikalen Linie auf der linken und rechten Seite des Bipartiten Graphen. Keine Kanten werden innerhalb von den Parametergruppen dargestellt, um die Unordnung vieler Kanten zu vermeiden. Die Knoten werden farblich unterschieden, damit die Kategorien der Parameter sofort erkannt werden können. Die Interaktionen bei der Ansicht ermöglichen eine visuelle Präsentation der Korrelationen zwischen den Parametern in unterschiedlichen Korrelationsordnungen. Die Korrelationen unter

Parameter-Tupeln mit Anzahl der Parameter von 2 bis 4 sind bei der Datenvorbearbeitung berechnet. Durch die Interaktion der Korrelationsanzeige werden die am besten korrelierten Parameter-Tupel bezüglich eines Parameters hervorhoben und die Korrelationen werden ebenfalls als zusätzliche Knoten dargestellt. Dadurch ist es möglich, die Fehlerquellen der Fehler zu analysieren. Solche Korrelationen der Parameter-Tupel überwinden die Einschränkungen üblicher Korrelationsanalyse bei früheren Verfahren wie das System aus [ZMZM15], bei dem nur paarweise Korrelationen von zwei einzelnen Parametern gezeigt werden. Die Korrelationen in verschiedenen Ordnungen bieten eine Möglichkeit an, die potenziellen Fehlerquellen zu entdecken.

Die dritte Ansicht ist für die Visualisierung der Parallelen Koordinaten gedacht. Mit der werden die hochdimensionalen Strukturen und multivariaten Daten übersichtlich dargestellt. Diese Ansicht dient als eine Ergänzung und eine Erweiterung für die Visualisierungen der MDS und des Node-Link-Graphen. Bei den Parallelen Koordinaten werden sowohl die Datenpunkte in der MDS-Ansicht als auch die Parameter in der NLG-Ansicht ausführlich visualisiert. Um die Zusammenhänge deutlich zu zeigen haben die visuellen Elementen in den Parallelen Koordinaten entsprechende Farbzuweisungen wie die in anderen zwei Ansichten. Somit können die in den beiden anderen Ansichten verborgenen Informationen in der Ansicht der Parallelen Koordinaten entdeckt werden.

In der abschließenden Auswertung des Systems wird ein Anwendungsfall durchgeführt, damit die Funktionen des Systems überprüft werden können. Die Visualisierungen und vorhandenen Interaktionen bieten eine gute interaktive Darstellungsplattform sowohl für die Darstellung allgemeiner Dateninformationen als auch für die Darstellung spezifischer Beziehungen innerhalb der Daten.
Aus dem Anwendungsfall gehen einige wichtige Aspekte hervor, die für eine Weiterentwicklung und Verbesserung berücksichtigt werden sollten.

Ausblick

Die bei der Auswertung gewonnenen Erkenntnisse erlauben eine gezielte Verbesserung des Systems zur multivariaten Datenanalyse in hohen Dimensionen. Das System ist zurzeit spezifisch für einen realen Datensatz von Advantest umgesetzt worden. Eine mögliche Erweiterung wäre, das System generisch weiterzuentwickeln, um andere multivariaten Datensätze zu analysieren. Ein generisches System zur Analyse und zur Untersuchung der Fehlerquellen wird hilfreich für die Erhöhung der Überprüfungsqualität in der Hardwareindustrie sein. Die Entscheidung der Kategorien der Parameter können in Zukunft auch dynamisch erstellt werden, dass die Parameter beliebig zugeordnet werden können.
Die Visualisierungen und Interaktionen des Systems können weiter verbessert werden. Für die Übersichtsvisualisierung können neben der Multidimensionalen Skalierung andere Algorithmen zur Dimensionsreduktion wie Hauptkomponentenanalyse (PCA) [SWG87] , Diskriminanzfunktion (LDA) [DHS00] und Kanonische Korrelation (CCA) [JW88] eingesetzt werden. Somit können die Projektionsfehler beseitigt werden. Die Visualisierung der Übersicht gilt zurzeit für die Datenpunkte. Es wäre möglich, dies auf Parameter zu erweitern. Dazu können in Zukunft zusätzliche interaktive Visualisierungen wie Heatmaps und Histogramme genutzt werden, die eine Übersicht über die Parameter verschaffen.
Der Node-Link-Graph lässt sich ebenfalls erweitern. Verschiedene Graphenlayouts wie Force-Directed

Graph [BEGT12] können verwendet werden. Die Darstellung der Korrelationen bezüglich eines Zielparameters ist zurzeit statisch. Mehr Interaktionen wie Hover, Mausklick und Mausziehen können sich in die Verbindungsknoten integrieren, damit die relevanten Parameter deutlich angezeigt werden. Die Knoten des Node-Link-Graphen können ebenfalls interaktiver erstellt werden. Beispielsweise kann man die Knoten sichtbar oder unsichtbar erstellen, damit man sich auf wichtigere Knoten konzentrieren kann.
Schlussendlich sollen die erkannten potenziellen Fehlerquellen jederzeit in geeigneten grafischen Darstellungen abgespeichert werden und somit bearbeitet werden können. Dies kann den Benutzern erlauben, die Zusammenhänge innerhalb von den Fehlerquellen durch Visualisierung und Interaktionen noch besser zu analysieren und weitere Erkenntnisse zu gewinnen.

Literaturverzeichnis

[BCQF10] S.-H. Bae, J. Y. Choi, J. Qiu, G. C. Fox. Dimension Reduction and Visualization of Large High-dimensional Data via Interpolation. In *Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing*, HPDC '10, S. 203–214. ACM, 2010. (Zitiert auf Seite 20)

[BEGT12] M. J. Bannister, D. Eppstein, M. T. Goodrich, L. Trott. Force-Directed Graph Drawing Using Social Gravity and Scaling. In *Graph Drawing*, Band 7704 von *Lecture Notes in Computer Science*, S. 414–425. Springer, 2012. (Zitiert auf Seite 65)

[CC08] M. A. A. Cox, T. F. Cox. Multidimensional Scaling. In *Handbook of Data Visualization*, S. 315–347. Springer Berlin Heidelberg, 2008. (Zitiert auf den Seiten 20 und 21)

[CMS99] S. K. Card, J. D. Mackinlay, B. Shneiderman, Herausgeber. *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann Publishers Inc., 1999. (Zitiert auf den Seiten 8 und 18)

[Com94] P. Comon. Independent Component Analysis: a new concept? In *Signal Processing Vol. 36, Nr. 3*, S. 287–314. 1994. (Zitiert auf Seite 20)

[DHS00] R. O. Duda, P. E. Hart, D. G. Stork. *Pattern Classification (2Nd Edition)*. Wiley-Interscience, 2000. (Zitiert auf Seite 64)

[Fri09] M. Friendly. The history of the cluster heat map. *The American Statistician*, 2009. (Zitiert auf Seite 22)

[Gre07] M. Greenacre. Correspondence Analysis in Practice. *Chapman and Hall/CRC*, 2007. (Zitiert auf den Seiten 19 und 25)

[Har69] F. Harary. *Graph Theory*. Addison-Wesley Series in Mathematics. Addison Wesley, 1969. (Zitiert auf Seite 44)

[HCL05] J. Heer, S. K. Card, J. Landay. Prefuse: A Toolkit for Interactive Information Visualization. In *ACM Human Factors in Computing Systems (CHI)*, S. 421–430. 2005. (Zitiert auf den Seiten 8, 27 und 28)

[Hov16] J. van den Hoven. Clustering with optimised weights for Gower's metric. *Thesis at University of Amsterdam*, 2016. (Zitiert auf Seite 37)

[JJ10] S. Johansson, J. Johansson. Visual Analysis of Mixed Data Sets Using Interactive Quantification. *SIGKDD Explor. Newsl.*, 11(2):29–38, 2010. (Zitiert auf den Seiten 21, 24 und 27)

[JJJ08] S. Johansson, M. Jern, J. Johansson. Interactive Quantification of Categorical Variables in Mixed Data Sets. In *2008 12th International Conference Information Visualisation*, S. 3–10. 2008. (Zitiert auf den Seiten 21 und 25)

[JW88] R. A. Johnson, D. W. Wichern, Herausgeber. *Applied Multivariate Statistical Analysis*. Prentice-Hall, Inc., 1988. (Zitiert auf Seite 64)

[KMSZ06] D. A. Keim, F. Mansmann, J. Schneidewind, H. Ziegler. Challenges in Visual Data Analysis. In *Proceedings of the Conference on Information Visualization*, IV '06, S. 9–16. IEEE Computer Society, Washington, DC, USA, 2006. (Zitiert auf den Seiten 8, 16 und 17)

[KMSZ09] D. Keim, F. Mansmann, A. Stoffel, H. Ziegler, Herausgeber. *Visual Analytics*. Springer US, 2009. (Zitiert auf den Seiten 8 und 16)

[Mac02] D. J. C. MacKay. *Information Theory, Inference & Learning Algorithms*. Cambridge University Press, 2002. (Zitiert auf Seite 25)

[Mel87] M. Mellinger. Multivariate data analysis: Its methods. *Chemometrics and Intelligent Laboratory Systems*, 2:29–36, 1987. (Zitiert auf Seite 19)

[MW14] J.-D. F. T. v. L. J. J. v. W. B. Z. Michael Wybrow, Niklas Elmqvist. Interaction in the Visualization of Multivariate Networks. In *Multivariate Network Visualization*, S. 97–125. Springer International Publishing, 2014. (Zitiert auf Seite 18)

[Pic09] C. Pich. Applications of Multidimensional Scaling to Graph Drawing. *PhD Thesis at University of Konstanz*, 2009. (Zitiert auf den Seiten 10 und 38)

[RC94] R. Rao, S. K. Card. The Table Lens: Merging Graphical and Symbolic Representations in an Interactive Focus + Context Visualization for Tabular Information. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '94, S. 318–322. ACM, New York, NY, USA, 1994. (Zitiert auf Seite 25)

[Ren12] A. C. Rencher. Methods of multivariate analysis. Wiley, 2012. (Zitiert auf den Seiten 10, 18 und 19)

[Rob07] J. C. Roberts. State of the Art: Coordinated & Multiple Views in Exploratory Visualization. In *Proceedings of the Fifth International Conference on Coordinated and Multiple Views in Exploratory Visualization*, CMV '07, S. 61–71. IEEE Computer Society, 2007. (Zitiert auf Seite 12)

[RS00] S. T. Roweis, L. K. Saul. Nonlinear Dimensionality Reduction by Locally Linear Embedding. Science, 2000. (Zitiert auf Seite 20)

[SB03] B. Shneiderman, B. B. Bederson. *The Craft of Information Visualization: Readings and Reflections*. Morgan Kaufmann Publishers Inc., 2003. (Zitiert auf Seite 17)

[SDMT16] J. Stahnke, M. Dörk, B. Müller, A. Thom. Probing Projections: Interaction Techniques for Interpreting Arrangements and Errors of Dimensionality Reductions. *IEEE Trans. Vis. Comput. Graph.*, 22(1):629–638, 2016. (Zitiert auf den Seiten 8, 21, 22 und 26)

[Sei95] R. Seide. The upper bound theorem for polytopes: an easy proof of its asymptotic version. *Computational Geometry*, 5(2):115–116, 1995. (Zitiert auf Seite 24)

[Shi12] M. A. Shiker. Multivariate Statistical Analysis. *British Journal of Science*, 6, 2012. (Zitiert auf Seite 19)

[Shn96] B. Shneiderman. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In *Proceedings of the 1996 IEEE Symposium on Visual Languages*, VL '96, S. 336–. IEEE Computer Society, Washington, DC, USA, 1996. (Zitiert auf Seite 17)

[SKNP04] M. Sips, D. A. Keim, S. C. North, C. Panse. Visual Data Mining in Large Geospatial Point Sets. *IEEE Computer Graphics and Applications*, 24:36–44, 2004. (Zitiert auf Seite 15)

[SWG87] K. E. S. Wold, P. Geladi. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2:35–72, 1987. (Zitiert auf den Seiten 19, 20 und 64)

[Tim02] N. H. Timm. Applied Multivariate Analysis. Springer-Verlag, 2002. (Zitiert auf Seite 19)

[WM16] B. Wang, K. Mueller. The Subspace Voyager: Exploring High-Dimensional Data along a Continuum of Salient 3D Subspaces. *CoRR*, abs/1603.04781, 2016. (Zitiert auf den Seiten 8, 21, 23 und 27)

[ZMZM15] Z. Zhang, K. T. McDonnell, E. Zadok, K. Mueller. Visual Correlation Analysis of Numerical and Categorical Data on the Correlation Map. *IEEE Transactions on Visualization and Computer Graphics*, 21(2):289–303, 2015. (Zitiert auf den Seiten 8, 21, 24, 25, 27 und 64)

Erklärung

Ich versichere, diese Arbeit selbstständig verfasst zu haben. Ich habe keine anderen als die angegebenen Quellen benutzt und alle wörtlich oder sinngemäß aus anderen Werken übernommene Aussagen als solche gekennzeichnet. Weder diese Arbeit noch wesentliche Teile daraus waren bisher Gegenstand eines anderen Prüfungsverfahrens. Ich habe diese Arbeit bisher weder teilweise noch vollständig veröffentlicht. Das elektronische Exemplar stimmt mit allen eingereichten Exemplaren überein.

Ort, Datum, Unterschrift