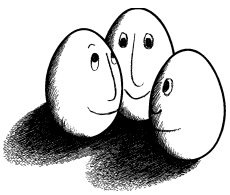
Masterarbeit

Ressourcenschonung bei der Auswertung von Materialeigenschaften in ternären Systemen durch adaptive Messungen

David Arnu



Masterarbeit
Fakultät Informatik
Technische Universität Dortmund

Dortmund, 12. August 2014

Betreuer:

Prof. Dr. Katharina Morik
Dipl.-Inform. Hendrik Blom

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

1.1 Themengebiet

Die Suche nach neuen Materialien mit verbesserten oder bislang unbekannten Eigenschaften ist ein wichtiges Forschungsgebiet mit dem einige drängende Probleme unserer Gesellschaft gelöst werden können. Leichte, aber zugleich robuste Werkstoffe können helfen den Treibstoffverbrauch von Fahrzeugen zu reduzieren; effizienter arbeitende Halbleiter können den Energieverbrauch elektronischer Geräte senken und zugleich den Bedarf an seltenen Rohstoffen; verbesserte Stromspeicher sind ein wichtiger Schritt für die Energieversorgung der Zukunft. Aber auch in der Pharmaindustrie gibt es bislang unerforschte Wirkstoffe und chemische Verbindungen die unsere Gesundheit positiv beeinflussen können. Dies sind nur einige wenige Beispiele, die zeigen wie sehr die Materialforschung unseren Alltag beeinflussen kann.

Doch es gibt eine Vielzahl von Problemen, welche die Materialforschung zu bewältigen hat. Die große Anzahl von Einflussfaktoren bei der Materialerzeugung ist dabei eines der entscheidendsten. So können moderne Stahllegierungen rund zehn verschiedene Zusatzstoffe enthalten, um zum Beispiel die Härte oder Korrosionsbeständigkeit zu beeinflussen. Außerdem können physikalische Einflussfaktoren, wie Temperatur oder Druck bei der Verarbeitung, eine große Rolle spielen.

Diese Vielzahl von Faktoren sorgt auf den ersten Blick für ein sehr chaotisches System, bei dem nur kleine Änderungen der Eingangsparameter zu großen Unterschieden im fertigen Produkt führen. Selbst wenn der Einfluss der Parameter verstanden ist, so ist eine exakte Modellierung meist sehr schwierig, insbesondere da sich viele reale Systeme nicht linear und stetig verhalten, sondern es oftmals komplexe Zusammenhänge gibt und Veränderungen sprunghaft auftreten, wenn bestimmte Schwellwerte überschritten werden.

1.2 Wie kann die Datenanalyse hier helfen?

Es ist offensichtlich, dass die manuelle Auswertung mehrerer hundert verschiedener Materialzusammensetzungen nicht sehr effizient ist, zumal für hochdimensionale Merkmale Zusammenhänge nur schwer zu erkennen sind.

Das Erkennen von Zusammenhängen, sowie die Verarbeitung großer Datenmengen, zählen zu den klassischen Aufgaben der Informatik. Somit ist die kombinatorische Materialforschung auch ein interdisziplinäres Feld in dem die verwendeten Verfahren vielfältig

sind und aus unterschiedlichen Bereichen der Datenanalyse und Statistik stammen können. Ein wichtiges Hilfsmittel zur Entdeckung von Mustern und der Kategorisierung der Proben ist das Clustering, wie es zum Beispiel bei Long et al. [19] eingesetzt wird. Es werden auch Verfahren zur Dimensionreduktion, wie die Hauptkomponentenanalyse , eingesetzt, um wichtige Einflussfaktoren zu extrahieren. Auch Regressionsverfahren zur Vorhersage von Materialeigenschaften [6] werden genutzt.

1.3 Probleme der kombinatorischen Materialforschung

Bei der automatisierten Verarbeitung der Daten für die kombinatorische Materialforschung sind einige Schwierigkeiten zu beachten. So ist zunächst gilt es die Aggregation und Speicherung der Messdaten sinnvoll zu verwalten. Dies umfasst das Anbinden der Messgeräte an die sonstige Infrastruktur und die sichere Speicherung der Daten in einem geeigneten Datenbankformat. Wie eine solche Infrastruktur aussehen kann beschreibt Meguro [22] anhand eines Internet-basierten Interface für Dünnschichtbibliotheken.

Ein weiteres Problem ist das generelle Volumen der erzeugten Daten. So nehmen zum Beispiel die Daten der XRD-Messungen für 342 Messpunkte einige Megabyte Speicherplatz in Anspruch; zwar sind dies Größenordnungen, die moderne Hardware bewerkstelligen kann, aber für ältere Computer, oder besonders speicherintensive Analyseverfahren, kann dies bereits zu einem Problem werden.

Beim Erstellen von Modellen und Prognosen muss der Messfehler, der bei physikalischen Prozessen niemals ganz ausgeschlossen werden kann, ebenfalls berücksichtigt werden. So kann es bei der Bestimmung der chemischen Zusammensetzung leicht zu Abweichungen von bis zu $2at.\%$ kommen (wobei Stoffgemische in der Literatur in der Regel auf bis zu $0,1at.\%$ genau angegeben werden). Bei den später noch entscheidenden XRD-Messungen können die gemessenen Intensitäten ebenfalls um bis $\pm 0.6°$ verschoben sein, bei einer Messgenauigkeit von etwa $0,02°$. Eine beim Winkel $\theta = 40°$ gemessene Intensität kann in Wahrheit somit irgendwo im Bereich von $\theta = 39,4° - 40,6°$ liegen. Dadurch ergibt sich bei 2000 bis 4000 gemessenen Winkeln je Messpunkt das Problem einzelne Maximalwerte exakt zu identifizieren. Auch die Zuordnung zwischen der Lage der Intensitätsspitzen für zwei verschiedene Messpunkte wird dadurch erschwert. Es stellt sich die Frage, ob diese zwei Spitzen eine gemeinsame physikalische Ursache haben oder verschiedene Ursachen haben. Die Entwicklung der Spektren muss dabei auch über die drei Dimensionen des ternären Systems verfolgt werden, was die Darstellung und Interpretation zusätzlich erschwert.

Ein Problem bei der Entwicklung und Erprobung neuer Verfahren ist, dass meist noch keine Erkenntnisse über die untersuchten Systeme vorliegen. Das heißt aus Sicht der Datenanalyse gibt es keine oder nur selten gelabelte Datensätze; es ist also schwer vorherzusagen welche Daten nun interessant sind und welche nicht.

1.4 Zielsetzung

In dem oben beschriebenen Prozess der Entwicklung neuer Materialien steht am Anfang die Identifizierung von interessanten Bereichen, für welche eine eingehendere Untersuchung lohnenswert erscheint.

Das Verfahren der Dünnschichtbibliotheken ermöglicht es sehr viele neue Materialkombinationen automatisch zu erstellen und zu erfassen. Jedoch festzulegen, welche Bereiche für eine genauere Untersuchung in Frage kommen, ist schwierig, da sich der Begriff *interessant* nicht einfach an einzelnen Variablen festlegen lässt und man im Zweifel auf Domänenwissen von Experten zurückgreifen muss, um passende Entscheidungen zu treffen, zum Beispiel um widersprüchliche Daten als Messfehler zu erkennen.

Die Informatik kann hier unterstützend tätig werden. Eines der Probleme der derzeitigen Arbeitsprozesse ist, dass sehr viele aufwendige Messungen getätigt werden müssen, um anschließend aus vielen ähnlichen Daten einige wenige heraus zu filtern. Dies kostet Arbeitszeit. da dieses Filtern, trotz aller Hilfsmittel, ein größtenteils manueller Prozess bleibt. Der Zeitaufwand ist ein wichtiger Faktor, denn gerade die Bestimmung der XRD-Spektren dauert im Verhältnis sehr lange und blockiert dabei nur begrenzt zur Verfügung stehende Ressourcen in Form von teuren Geräten. Im Vergleich dauert die Bestimmung der chemischen Zusammensetzung mittels EDX nur ein paar Stunden, ebenso wie die Messung des elektrischen Widerstandes; während die vollständige XRD-Messung eines Wafers mit 342 Messpunkten rund 4 Tage dauern kann.

Wegen der langen Laufzeit der XRD-Messung birgt gerade dieser Arbeitsschritt ein großes Einsparungspotential, welches in dieser Arbeit untersucht werden soll. Dabei sollen Erkenntnisse aus der Analyse der bereits vorhandenen oder einfacher zu erhaltende Messdaten genutzt werden, um die XRD-Messungen gezielt für einige wenige Punkte durchzuführen und dabei möglichst alle auftretenden Materialphasen anhand ihrer charakteristischen XRD-Spektren zu finden. Die hierfür notwendige Modellbildung, die jedem Messpunkt ein möglichst ähnliches XRD-Spektrum zuordnet, ist ein weiteres nützliches Selektionsmerkmal, um weitergehende Untersuchungen auf relevante Bereiche zu beschränken.

Zu ermitteln, mit welchen Verfahren diese als relevant erachteten Messpunkte ausgewählt werden können und wie entsprechende Modelle erstellt werden ist die Zielsetzung dieser Arbeit.

1.5 Verlaufsplan der Messungen

Analyse der vorhandenen Messdaten Anhand der Verteilung der chemischen Zusammensetzung und deren Verhältnis zu den bereits bekannten Materialeigenschaften soll eine erste Einteilung der Messdaten erfolgen. Hierfür wird zum einen die Annahme getroffen, dass Materialproben mit ähnlicher chemischer Zusammensetzung auch ähnliche Eigenschaften aufweisen, andererseits sind aber gerade jene Bereiche, in denen diese

Lokalität verletzt wird, möglicherweise besonders interessant. So kann zunächst zum Beispiel mittels Cluster-Verfahren der Wafer in Bereiche mit ähnlichem Verhalten eingeteilt werden. Zusätzlich können Regressionsverfahren eingesetzt werden, um Punkte zu finden, die sich deutlich anders verhalten als ihre Nachbarschaft vermuten lassen würde. Dies könnte entweder auf Messfehler oder im idealen Fall auf Phasenübergänge hindeuten.

Auswahl der ersten Stichproben Anhand der so ermittelten Verteilung, beziehungsweise der daraus erstellten Einteilung, können die ersten Punkte bestimmt werden, für welche die XRD-Spektren gemessen werden sollen. Diese sollen nach Möglichkeit bereits möglichst viele verschiedene Formen der XRD-Spektren erfassen, damit im Anschluss entschieden werden kann, wo weitere Messungen sinnvoll sind.

An dieser Stelle muss außerdem evaluiert werden, wie die Stichprobengüte gemessen werden kann. Für die vorliegenden Datensätze kann dies über einen Vergleich der ausgewählten Spektren mit allen erhobenen Spektren erfolgen, wobei die genutzten Kriterien in dieser Arbeit untersucht werden.

Auswertung der XRD-Spektren Nachdem die ersten Messungen vorliegen gilt es anhand dieser nun bekannten Spektren zu entscheiden, welche weiteren Punkte für eine Messung in Frage kommen. Diese zweite Entscheidungsphase erscheint sinnvoll, da aus den XRD-Spektren weitere Erkenntnisse gewonnen werden können.

Mit der erneuten Auswahl von Messpunkten sollen nach Möglichkeit zwei Fragen beantwortet werden: ob es möglicherweise weitere Bereiche gibt , die charakteristische XRD-Spektren aufweisen und ob mit weiteren Messungen die Grenzübergänge zwischen den Bereichen genauer erfasst werden können.

Die Auswahl der Messpunkte erfolgt dabei nach Kriterien des *active-learning* beziehungsweise den Methoden des *optimal sensor-slacement*. Die Ansätze beider Verfahren sind ähnlich; es wird versucht aus vorhandenen Informationen zu entscheiden, welche weiteren Punkte ausgewertet werden sollen. Die Entscheidungskriterien können unterschiedlich gewählt werden. So können Punkte bevorzugt werden, für die die Unsicherheit des bisherigen Modells am größten ist oder die sich am stärksten von den übrigen Messpunkten unterscheiden.

Welche Kriterien hier sinnvoll sind und wie die zusätzlichen Informationen der XRD-Spektren verarbeitet werden können muss noch untersucht werden.

Abschließende Modellbildung In der letzten Phase soll allen Punkten, für die keine XRD-Messung vorliegt, ein bekanntes Spektrum zugewiesen werden, welches dem unbekannten Spektrum am ähnlichsten ist. Schwierig ist es, hier neben den von Anfang an bekannten Größen auch die neuen Informationen aus den XRD-Spektren mit einfließen zu lassen. Die zu lösende Optimierungsaufgabe lautet demnach: Reduziere den Fehler bei der Zuordnung von Punkten mit unbekannten XRD-Spektren zu bekannten Spek-

tren, so dass die zugewiesenen Bereiche tatsächlich die größte Ähnlichkeit (z.B. kleinste DTW-Distanz) mit dem ausgewerteten XRD-Spektrum haben.

Eine zu beachtende Nebenbedingung wird sein die tatsächliche Anzahl der XRD-Messungen möglichst gering zu halten, ohne dass die Qualität der Vorhersagen deutlich sinkt. Es wird notwendig sein zwischen der Genauigkeit der Vorhersagen und der benötigten Zeit abzuwägen.

1.6 Einordnung in das Problemfeld

Mit einer automatisierten Vorauswahl von Bereichen für die eine detaillierte Suche sinnvoll ist, kann die Bearbeitung von neuen Materialbibliotheken deutlich beschleunigt werden.

Die Reduktion der XRD-Messungen spart Zeit auf den Messgeräten, einer begrenzt zur Verfügung stehenden Ressource. Außerdem kann das Datenvolumen deutlich reduziert werden, da weniger Messungen ausgewertet werden müssen. Der Vorteil ist hier vor allem die Reduktion von redundanten Informationen (durch sehr viele ähnliche XRD-Spektren), wodurch relevantere Ergebnisse leichter auszumachen sind. Das gesparte Datenvolumen ist zwar langfristig auch von Belang; mit einer nicht komprimierten Größe von einigen Megabyte sind die Messdaten im Vergleich zu anderen physikalischen Prozessen, die leicht mehrere Gigabyte an Daten produzieren können (Astronomie, Teilchenphysik), jedoch überschaubar.

Es ist vorstellbar, dass sich hier gefundene Ergebnisse und Methoden leicht auf weitere Bereiche der Materialforschung übertragen lassen. So könnten auch kompliziertere Systeme mit mehr Komponenten oder noch mehr bekannten Einflussfaktoren gemessen werden. Auch können andere Verfahren anstatt der XRD-Analyse verwendet werden.

1.7 Übersicht der folgenden Kapitel

Im nächsten Kapitel erfolgt eine allgemeine Einführung in die kombinatorische Materialforschung. Es werden einige physikalische Grundlagen erläutert und der Prozess zur Erzeugung von Dünnschichtbibliotheken beschrieben. Die Analyseverfahren EDX- und XRD-Messung werden erklärt. Zusätzlich enthält das Kapitel eine ausführliche Beschreibung der in dieser Arbeit verwendeten Datensätze.

Im dritten Kapitel wird ein Blick auf verschiedene theoretische Ansätze, wie eine effiziente Datenauswertung erfolgen kann, geworfen. In Kapitel 4 werden die verwendeten Algorithmen und Methoden beschrieben. In Kapitel 5 wird der Ablauf der adaptiven Messungen und wie dadurch die Anzahl der notwendigen Messungen reduziert werden kann beschrieben. In Kapitel 6 wird evaluiert, wie mittels bereits vorhandener Daten möglichst relevante Stichproben bestimmt werden können, für die eine Messung der XRD-Spektren sinnvoll erscheint. Im Anschluss werden in den Kapiteln 7 und 8 anhand

dieser Stichproben zwei Verfahren untersucht, welche die vorgeschlagene Methodik der adaptiven Messungen umsetzen. Im Abschluss bietet Kapitel 9 eine Zusammenfassung und Beurteilung der Ergebnisse der Arbeit und gibt einen Ausblick auf weitere Forschung und noch offene Fragen.

2 Hintergrund Materialforschung

2.1 Überblick

Die kombinatorische Materialforschung ist ein modernder Ansatz, dessen Ziel die systematische Erzeugung und Auswertung von neuen Materialzusammensetzungen ist.

Ein wichtiges Hilfsmittel sind Dünnschichtbibliotheken von Materialien. Dabei werden auf einem Trägermaterial, zum Beispiel eine Silizium-Scheibe (ähnlich der Basis für elektronische Schaltungen), dünne Schichten der zu untersuchenden Materialien in unterschiedlicher Dicke aufgetragen, so dass an jedem Punkt der Scheibe eine andere Zusammensetzung entsteht. Motiviert ist dieses Verfahren dadurch, dass sich die Materialien im mikroskopischen Bereich der Dünnschichten - Schichtdicken von einigen hundert Nanometern sind üblich - hinreichend ähnlich verhalten wie makroskopische und homogene Proben derselben Materialzusammensetzung und Fertigungsbedingungen. Der Vorteil dieser Materialbibliotheken ist, dass sie *am Stück* produziert und ausgewertet werden können. Das verringert das Auftreten von systematischen Fehlern [10] bei der Auswertung, wie verschiedene äußere Einflüsse, falsche Parametereinstellung oder schlicht verwechselte Proben. Außerdem ist es deutlich zeit- und kosteneffizienter eine Materialbibliothek zu erzeugen als mehrere hundert Proben einzeln zu fertigen (ein Prozess, der das Schmelzen, Vermischen und abschließende Abkühlen der Proben erfordert).

Für die Auswertung der Materialbibliotheken kommen Hochdurchsatzverfahren zum Einsatz, welche die schnelle Analyse der vielen erzeugten Proben ermöglichen. Hier kommen Aspekte aus der Datenverarbeitung und -analyse zum Tragen. So gilt es zunächst mitunter recht große Datenmengen zu speichern und zu verwalten [22]. Desweiteren sind Verfahren nötig, welche diese großen und zum Teil hochdimensionalen Daten aufbereiten, filtern und visualisieren können [30].

Die kombinatorische Materialforschung ist ein im höchsten Maße aktuelles Themengebiet an dem aktiv geforscht wird. Eine komplette Übersicht aller Verfahren und Forschungsvorhaben würde den Umfang einer Einleitung bei weitem überschreiten, aber die Übersichtsartikel von Green [10] und Maier [21] bieten einen guten Einstiegspunkt. Forschungsgebiete sind unter anderem die Pharmazeutik [32] oder die Entwicklung von Formgedächtnissmetallen [36, 11].

Dass dieses Themengebiet auch abseits der wissenschaftlichen Gemeinde interessiert verfolgt wird, zeigt die Förderung durch verschiedene Institutionen, zum Beispiel durch die US-Regierung mit der *Materials Genome Initiative [16],* welche mit einen Etat von hundert Millionen US-Dollar den Prozess der Entwicklung von neuen Materialien bis zu

deren Markteinführung beschleunigen soll, womit auch ganz klar wirtschaftliche Interessen ausgedrückt werden.

2.2 Allgemeiner Arbeitsablauf in der Materialforschung

Die Arbeitsabläufe in der kombinatorischen Materialforschung lassen sich in etwa folgendermaßen skizzieren:

Es gibt eine initiale Vermutung über eine zu untersuchende Materialmischung. Diese Vermutung kann auf vorherige Experimente oder auf allgemeinen Wissensstand beruhen. Auch kann auf Aussagen zurückgegriffen werden, welche durch theoretische Vorüberlegungen entstanden sind. So erstellt zum Beispiel Drautz [7] Modelle um die magnetischen Eigenschaften von Metalle vorherzusagen.

Anschließend wird eine Dünnschichtbibliothek für diesen Bereich erzeugt und ausgewertet. Die eingesetzten Verfahren können je nach Zielsetzung variieren; einige Techniken zur Analyse der Zusammensetzung (EDX) und der Struktur (XRD) werden am Ende dieses Abschnittes vorgestellt. Die Zielsetzung ist im Allgemeinen aber die gleiche, nämlich interessante, vom Rest der Probe abweichende Bereiche zu identifizieren.

Werden interessante Bereiche identifiziert, so ist die Kategorisierung nicht abgeschlossen, sondern die Proben müssen noch hinsichtlich der gewünschten Materialeigenschaften eingehender untersucht werden. So geht es unter anderem um die genaue Spezifikation der Kennlinien der Materialphasen. Dies ist ein aufwendigerer Prozess, bei dem zum Beispiel für eine Materialprobe der elektrische Widerstand bei unterschiedlichen Temperaturen gemessen wird; Zarnetta [36] diese Verfahren sehr ausführlich. Ziel dabei ist möglichst genau das Verhalten des Materials zu dokumentieren. Dies ist eine durchaus umfassende Aufgabe, die viele Versuchs- und Messreihen erfordert. Eine anschauliche Dokumentation dieser Arbeitsschritte bietet die Dissertation von Zarnetta [35], in der die Entwicklung von Formgedächstnissmetallen ausführlich beschrieben wird.

2.3 Physikalische Grundlagen

Materialphasen Bei den meisten Materialien ordnen sich im festen Zustand die Atome in bestimmten Gittermustern an. Diese Struktur sorgt unter anderem für charakteristische Eigenschaften des jeweiligen Materials. Aus der stabilen Gitterstruktur des Kohlenstoffs folgt die enorme Härte eines Diamanten, während frei bewegliche Elektronen in Metallen für deren gute elektrische Leitfähigkeit sorgen.

Werden zwei oder mehr chemische Elemente vermischt, können sich die verschiedenen Atome zu gemeinsamen Gitterstrukturen verbinden. Diese Gitterstrukturen werden auch Materialphasen genannt. Welche Materialphase eine Materialmischung annimmt hängt von unterschiedlichen Faktoren ab. So kann zum Beispiel das rasche Abkühlen, dafür

sorgen, dass sich eine bestimmte Phase ausbildet, die beim langsamen Abkühlen nicht vorliegen würde. Dieser Effekt wird unter anderem beim Härten von Stahl ausgenutzt.

Weitere Beispiele und ausführliche theoretische Hintergründe bieten Lehrbücher der Werkstoffkunde wie zum Beispiel von Gottstein [9].

Herstellung von Dünnschichtbibliotheken Dünnschichtbibliotheken können mittels Kathodenzerstäubungsverfahren, auch *sputtern* genannt, hergestellt werden. Dieses Verfahren beruht darauf, dass mit Hilfe eines Edelgases unter hoher Energie Atome des Ausgangmaterials herausgelöst und auf ein Trägermaterial aufgetragen werden können. Edelgase, wie Argon, werden verwendet, weil sie nur schwach mit anderen Materialien reagieren.

Die herausgelösten Atome bilden anschließend eine dünne Schicht auf einem Trägermaterial, dem Wafer. Durch die Verwendung von Blenden kann die Expositionsdauer der Scheibe für unterschiedliche Bereiche variiert werden, so dass verschieden dicke Schichten entstehen.

Dieser Prozess kann mit anderen Materialien und unter anderen Blendeneinstellungen wiederholt werden, so dass die Schichtdicken in einem anderen Gradienten verlaufen. Wird der Wafer nun erhitzt, so dass aufgetragenen Materialien mit einander verschmelzen, erhält man eine fertige Kollektion, oder Bibliothek, mit verschiedenen Mischungsverhältnissen der aufgetragenen Stoffe.

Eine genaue Beschreibung dieses Prozesses und der dafür verwendeten Geräte bietet Ludwig et al. [20].

EDX-Messungen Mittels der Energiedispersiven Röntgenspektroskopie (EDX) kann die chemische Zusammensetzung einer Probe bestimmt werden. Diese Methode beruht darauf, dass die Probe durch Elektronenbeschuss zum Emittieren von Röntgenstrahlung gebracht wird. Jedes chemische Element sendet dabei Strahlung in einer charakteristischen Wellenlänge aus. Anhand der gemessenen Intensitäten kann die Zusammensetzung der Probe ermittelt werden. Je nach Art des Stoffgemisches werden jedem Messpunkt zwei, drei oder mehr Werte für die jeweilige Konzentration der vorkommenden Materialien zugewiesen, wobei die Nebenbedingung gilt, dass alle Werte zusammen $100 at.\%$ ergeben.

Als Messfehler wird mit Abweichung von bis zu $2\,at.\%$ gerechnet. Dies ist eine sehr allgemeine Fehlerschranke; durch Kalibrierung mit Referenzsystemen kann die Genauigkeit auf weniger als $1\,at.\%$ gesenkt werden, wie zum Beispiel von Hamann [11] und Loebel [18] gezeigt wird.

XRD-Analyse Das Verfahren der Röntgendiffraktometrie (engl. *X-Ray Diffraction* oder kurz XRD) beruht darauf, dass mittels Röntgenstrahlung die Kristallstruktur einer Materialprobe untersucht werden kann. Dies liegt daran, dass Röntgenstrahlung mit einer Wellenlänge zwischen 1 nm ($10^{-9}\,m$) und 1 pm ($10^{-12}\,m$) unterschiedlich stark

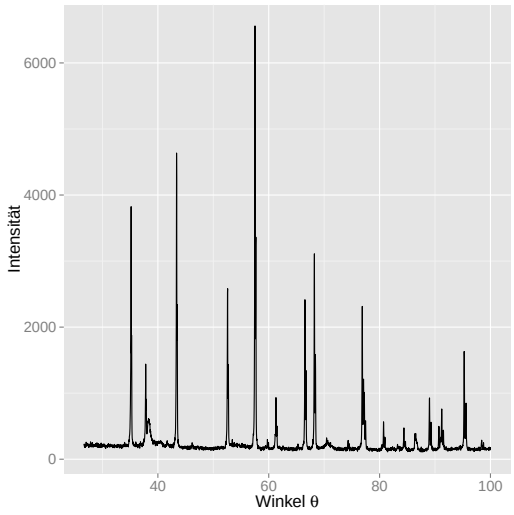


Abbildung 2.1: Beispiel eines XRD-Spektrums für einen Messpunkt des TiTaAl-Datensatzes

mit der Gitterstruktur der Atome interagiert, deren Abstand zueinander in der selben Größenordnung liegt. Für eine genaue Untersuchung wird ein Messpunkt aus vielen verschiedenen Winkeln θ geröntgt und die Intensität der reflektierten Strahlung gemessen. Somit ergibt sich ein charakteristisches Spektrum mit einigen besonders starken Ausprägungen für bestimmte Winkelbereiche, wie in Abbildung 2.1 zu sehen. Aus der Lage, der Intensität und auch aus der Breite der auftretenden Spitzenwerte können genaue Rückschlüsse über die atomare Struktur des untersuchten Materials gezogen werden. Der genaue physikalische Hintergrund wird in diesen beiden Lehrbüchern von Suryanarayana und Norton [31] und Hammond [12] anschaulich und ausführlicher, als es an dieser Stelle möglich ist, erläutert.

Zu beachten ist allerdings, dass die Vermessung einer Probe recht zeitaufwendig ist, da sehr viele verschiedene Winkel betrachtet werden müssen. So geben Takeuchi et al. [15] für die komplette Vermessung einer Dünnschichtbibliothek mit 535 Messpunkten und dem Winkelbereich $\theta = 20° - 75$ eine Dauer von etwa fünfeinhalb Tagen an. Diese sehr lange Dauer ist der Grund, warum in dieser Arbeit der Schwerpunkt darauf gelegt wird, mit deutlich weniger XRD-Messungen, einen ähnlich umfassenden Überblick über eine Materialbibliothek zu erhalten als dies eine vollständige Betrachtung aller Messpunkte ermöglicht.

2.4 Datenbeschreibung

Die untersuchten ternären Systeme sind

- Kupfer-Nickel-Zink (CuNiZn)

	min *at.*% A	max *at.*% A	min *at.*% B	max *at.*% B	min *at.*% C	max *at.*% C
CuNiZn	2,38	89,56	0,00	51,01	2,31	55,01
NiCrRe	1,15	96,55	0,00	92,61	0,00	93,34
TiCoW	0,00	38,86	21,36	93,35	0,00	50,96
TiNiCu	1,39	70,75	1,25	78,42	2,60	90,86
TiTaAl	38,80	83,20	10,40	57,10	1,70	12,70
TiTaRe	30,70	78,00	10,20	54,50	7,60	25,40

(a) Chemische Zusammensetzungen

	min Ω	max Ω	min d	max d	min ρ	max ρ	RGB	Schichtdicke	fehlende Werte
CuNiZn	0,03	0,07	235,4	743,1	83,31	852,22	ja	ja	0
NiCrRe	0,17	1,65	-	-	-	-	ja	nein	4
TiCoW	0,12	0,80	-	-	-	-	ja	nein	4
TiNiCu	0,13	7,87	202,8	706,5	33,5	3418,8	nein	ja	15
TiTaAl	0,53	2,08	497,2	1085,0	291,8	1225,4	nein	ja	0
TiTaRe	0,51	0,97	725,0	1549,5	192,3	503,3	nein	ja	0

(b) Materialeigenschaften

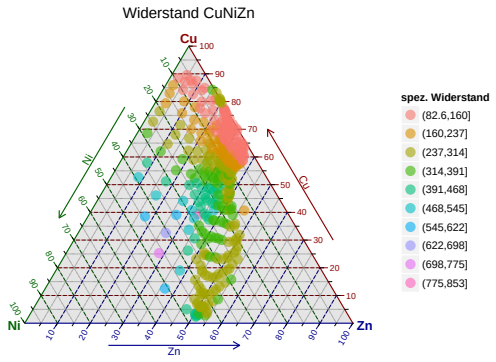
	min θ	max θ	min Int	max Int	Median Int	arithm. Mittel	#Messungen	$\theta_i - \theta_{i+1}$
CuNiZn	26,70	79,99	292	3064000	869	5934	2030	0,026
NiCrRe	30,00	84,99	123	1056000	375	893	4189	0,013
TiCoW	30,00	84,99	2340	993100	8170	10720	4124	0,013
TiNiCu	26,70	79,99	29	64020	107	281	2030	0,026
TiTaAl	26,70	99,99	96	14210	199	282	2792	0,026
TiTaRe	26,70	99,99	123	758300	335	1390	2792	0,026

(c) Intensitäten der gemessenen XRD-Spektren

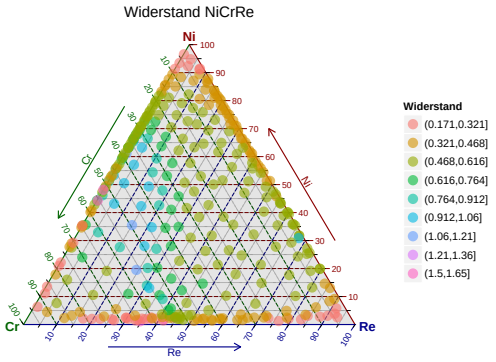
Tabelle 2.1: Übersicht der vorhandenen Datensätze

- Nickel-Chrom-Rhenium (NiCrRe)
- Titan-Cobalt-Wolfram (TiCoW)
- Titan-Nickel-Kupfer (TiNiCu)
- Titan-Tantal-Aluminium (TiTaAl)
- Titan-Tantal-Rhenium (TiTaRe)

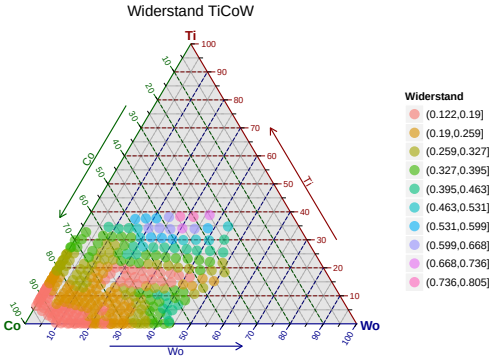
Die Tabelle 2.1 gibt einen ausführlichen Überblick über die im folgenden beschriebenen Messwerte. In Abbildung 2.2 sind die Datensätze als ternäre Diagramme über die chemische Zusammensetzung dargestellt. In diesen Diagrammen gilt, dass jeder Punkt im Dreieck die Nebenbedingung erfüllt, dass die Summe der drei Koordinatenachsen genau 100% ergibt. In den Spitzen des Dreiecks liegen die Punkte mit jeweils der reinen Konzentration eines Elements.



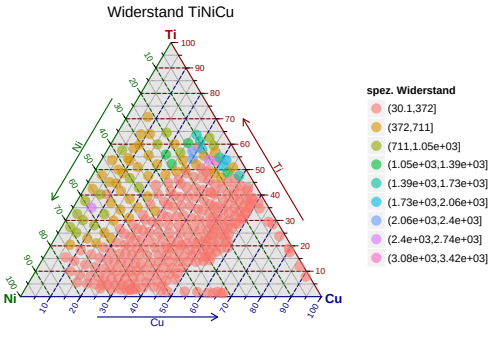
(a) CuNiZn



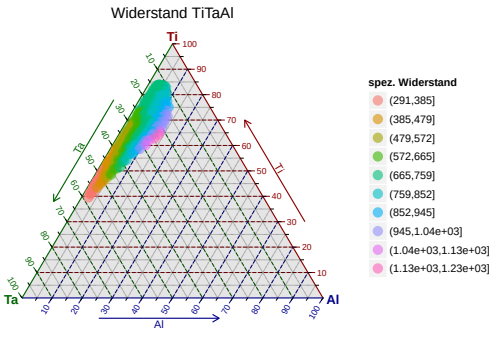
(b) NiCrRe



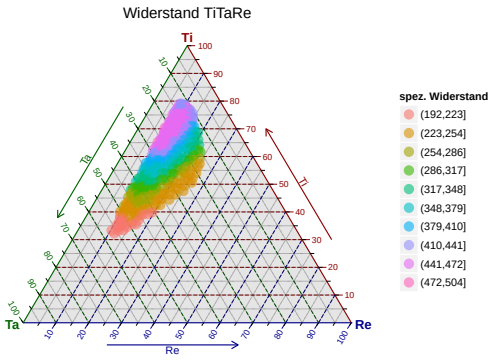
(c) TiCoW



(d) TiNiCu



(e) TiTaAl



(f) TiTaRe

Abbildung 2.2: Ternäre Diagramme der untersuchten Datensätze. Gezeigt wird die Verteilung der chemischen Zusammensetzung und der elektrische Widerstand

2.4.1 Untersuchte Merkmale

Chemische Zusammensetzung Die chemische Zusammensetzung gibt an in welchem Mischungsverhältnis die einzelnen Elemente der Materialprobe vorkommen. Für Dünnschichtbibliotheken liefert sie eine eindeutige Identifizierung eines Messpunktes, da der gesamte Wafer unter gleichen Umweltbedingungen gefertigt ist. In anderen Szenarien ist die Zuordnung nicht unbedingt eindeutig, zum Beispiel wenn eine Probe unter verschiedenen Temperatur untersucht wird.

Die genaue Bestimmung der chemischen Zusammensetzung ist bei Dünnschichtverfahren schwierig, da verschiedene Einflussfaktoren darauf einwirken können. Theoretisch könnten die Maschinenparameter der *sputter*-Kammern genaue Auskunft über den aufgetragenen Gradienten der Materialien geben. In der Praxis wird dies aber durch mechanische Einflüsse gestört. Auch je nach verwendeten Materialien kann die aufgetragene Konzentration variieren, da manche Stoffe für dieses Verfahren weniger gut geeignet sind. Deswegen ist die nachträgliche Bestimmung der Zusammensetzung für jeden Messpunkt mittels des EDX-Verfahrens notwendig.

Elektrischer Widerstand Der elektrische Widerstand ist eine wichtige Kenngröße bei der Beschreibung von Materialeigenschaften. Die elektrische Leitfähigkeit hängt unter anderem von der atomaren Struktur des untersuchten Materials ab. Wie sich der elektrische Widerstand für bestimmte Mischungsformen von Materialien verhält ist in vielen Fällen bekannt [25], weswegen davon abweichende Beobachtungen ein möglicher Indikator für neue Materialphasen sein können. Auf einer Waferscheibe erfolgt eine einzelne Widerstandsmessung dadurch, dass mehrere Punktkontakte am Messpunkt aufgesetzt werden und der zwischen ihnen fließende Strom gemessen wird. Dieses Verfahren wird von Thienhaus et al. [33] beschrieben und dauert für einen gesamten Wafer lediglich einige Stunden. Die so gemessenen Werte sind der normale elektrische Widerstand (Ω), welcher auch von der Schichtdicke an der untersuchten Stelle abhängt. Ist die Schichtdicke bekannt kann der spezifische Widerstand ($\mu\Omega cm$) berechnet werden, welcher als Materialkonstante unabhängig von der Dicke des Materials ist. Für die untersuchten Datensätze liegen für NiCrRe und TiCoW keine Angaben zur Schichtdicke vor, weswegen der spezifische Widerstand nicht berechnet werden kann. Da es allerdings einen deutlichen Zusammenhang zwischen den beiden Widerstandswerten gibt, werden im Folgenden beide gleichbedeutend verwendet und nur vom elektrischen Widerstand gesprochen.

RGB-Farbwerte Für einige der Datensätze liegen zusätzlich Informationen über die Farbe der Materialprobe in Form von RGB-Werten vor. Dafür wurde die Aufnahme des Wafers in ihre einzelnen Intensitätswerte der Farben Rot, Grün und Blau zerlegt. Die Intensitäten haben eine Werte, die von 1 bis 256 reichen.

Schichtdicke Die Dicke der aufgetragenen Materialschicht ist eine weitere Messgröße, die Aussagen über die Struktur des angefertigten Wafers ermöglicht. So ist sie für die Berechnung der Materialkonstante des spezifischen Widerstandes notwendig. Ohne die

Schichtdicke kann nur der gemessene Widerstand als Größe verwendet werden, welcher durch eine unterschiedlich dick aufgetragene verzerrt sein kann.

Für die Beurteilung der XRD-Spektren kann die Schichtdicke auch als Vergleichswert hinzugezogen werden. Denn falls die Schichtdicke deutliche Ausreißer zeigt, kann dies ein guter Indikator für mögliche Materialfehler sein, welche wiederum die Konfidenz für getroffene Schlussfolgerungen senken können.

Für die Auswertung ist hierbei schwierig, dass nicht für alle Datensätze auch die Schichtdicke bekannt ist.

Übersicht der bekannten Schichtdicken:

- CuNiZn: Es sind deutlich unterschiedliche Bereiche erkennbar. Von einer Seite des Wafers zur zur gegenüberliegenden Seite gibt es einen deutlichen Anstieg. Das Maximum der Schichtdicke ist etwa dreimal so groß wie das Minimum.
- TiNiCu: Die ist Schichtdicke relativ gleichmäßig, jedoch gibt es an den Rändern eine deutliche Absenkung. Eine erkennbare Verdickung verläuft am der rechten Seite des Wafer. Das Maximum etwa dreimal so groß wie das Minimum.
- TiTaAl und TiTaRe: Der Anstieg der Schichtdicke verläuft gleichmäßig in einer Richtung. Die maximale Schichtdicke ist etwa doppelt so groß wie die minimale.

2.4.2 Beschreibung der XRD-Spektren

Gemessene Winkel Die gemessenen Winkel variieren je nach ternärem System; eine Übersicht findet sich in Tabelle 2.1c.

Auffällige Messwerte Bei den Datensätzen CuNiZn, TiNiCu und TiTaRe fällt ein extrem hoher Pegel im Bereich um $\theta = 69° - 70°$ auf. Die gemessene Intensität ist dort um etliche Größenordnungen höher als für die übrigen Winkel; außerdem weisen alle Punkte des ternären Systems dort diese entsprechend hohen Werte auf (mit gewissen Schwankungen). Dieser Ausschlag in den XRD-Spektren stammt von den Siliziumkristallen des Trägermaterials und nicht von der aufgetragenen Materialschicht. Deswegen ist es sinnvoll diesen Bereich für folgende Untersuchungen auszuschließen, um die Ergebnisse nicht unnötig zu verzerren. Für den späteren Einsatz bei unbekannten ternären Systemen können derartige Probleme während der Untersuchung festgestellt werden, zum Beispiel wenn alle Messungen einen Spitzenwert im gleichen Winkelbereich aufweisen, der um mehrere Größenordnungen die übrigen Messungen übersteigt.

CuNiZn Wie in Tabelle 2.1c zu sehen variieren die Intensitäten für diesen Datensatz sehr stark, was sich in der großen Differenz zwischen arithmetischem Mittel und Median zeigt. Die bekannten, konstant hohen Spitzenwerte im Bereich $\theta = 69° - 70°$, weisen mit einer maximalen Intensität von etwa 3.000.000 und einem Mittelwert von etwa 150.000 deutlich größere Werte auf als es bei den anderen untersuchten Datensätzen der Fall ist.

Es gibt weitere sehr starke Ausschläge, welche aber auf wenige Messungen beschränkt sind und deshalb nicht einfach im Vorfeld entfernt werden können. Auffällig sind zum Beispiel mehrere hohe Ausschläge $\theta = 43° - 44°$, welche Intensitäten zwischen $100.000 - 300.000$ haben. Diese können nicht im Vorfeld als Mess- oder Materialfehler angesehen werden, da sie an verschiedenen Stellen und nicht alle direkt am Rand des Wafers auftreten. So gibt es für $\theta = 43, 90°$ einen einzelnen großen Spitzenwert bei $Cu_{30,7}Ni_{19,9}Zn_{49,3}$, für $\theta = 43, 56°$ hingegen treten die starken Ausschläge genau am Rand des Wafers auf und sind wahrscheinlich nur bedingt aussagekräftig.

Weitere Winkel die stärkere Ausprägungen aufweisen sind bei $\theta = \{61, 7° ; 74, 8°\}$ zu finden, welche auf dem Wafer auch ein ähnliches Gebiet abdecken. Im Bereich $\theta = 66° - 72°$ zeigt sich ein Plateau mit konstant erhöhten Werten, welches bis zu den nächsten Intensitätsspitzen bei $\theta = 74°$ reicht.

NiCrRe Die Übersicht der Verteilung im NiCrRe-Datensatz in Tabelle 2.1c zeigt eine ähnlich hohe maximale Intensität der XRD-Spektren wie beim CuNiZn-Datensatz; allerdings lässt die geringe Abweichung von Median und arithmetischem Mittel darauf schließen, dass diese Ausreißer weniger häufig vorkommen.

Die größten Intensitätswerte beschränken sich auf den Bereich um $Ni_{1,8}Cr_{15,4}Re_{81,8}$ und $\theta = 40, 6°$ und enthalten etwa ein Dutzend zusammenhängender Messpunkte. Diese sind in etwa um ein zehnfaches größer als die nächsten gemessenen Spitzenwerte

Es treten insgesamt rund ein Dutzend weiterer Spitzen in den Spektren auf, deren Intensitäten sich je nach Messpunkt deutlich unterscheiden können. Einzelne markante Bereiche sind $\theta = 41, 5°$ mit drei ausgeprägten Spitzenwerten im Bereich um $Ni_{24,7}Cr_{21,9}Re_{53,1}$.

Dass die XRD-Spektren für diesen Datensatz eine starke Varianz aufweisen ist nicht verwunderlich, da die Messpunkte fast das gesamte ternäre Systeme umfassen und somit größere Unterschiede zu erwarten sind. Eine Übersicht über das ternäre System findet sich bei Slyusarenko et al. [29], in der auch mehrere Materialphasen aufgezeigt werden.

Aufgrund der starken Varianz der Spektren und der chemischen Zusammensetzung stellt dieser Datensatz einen guten Vergleich zu den TiTaRe- und TiTaAl-Daten dar.

TiCoW Zunächst fällt auf, dass die minimale, sowie die durchschnittliche Intensität für dieses ternäre System höher sind als bei den übrigen untersuchten Datensätzen.

Die größte Intensität wird für den Bereich bei $\theta = 40, 8°$ angenommen. Auf dem Wafer nehmen die Messungen, die in diesem Winkel eine starke Intensität aufweisen, einen recht großen Bereich mit rund 60 Messpunkten ein. Auffällig ist hierbei, dass sowohl vier Punkte im Inneren dieses Bereiches, sowie ein einzelner Punkt am Rand des Wafers, die stärksten Intensitäten aufweisen. Da die Schichtdicke für diesen Wafer nicht vorliegt, kann ein Materialfehler zunächst nicht ausgeschlossen werden.

Eine weitere auffällige Intensitätsspitze findet sich um $\theta = 43°$, welche auch nur für einen kleinen abgegrenzten Bereich von Punkten auf dem Wafer eine deutlich höhere Intensität

aufweist. Dieser Bereich ist besonders interessant, da er sich in ganz ähnlicher Form in den Widerstandsmessungen zeigt, wie im folgenden Abschnitt näher erläutert wird.

Die Spitzenwerte im Bereich um $\theta = 57,5°$ fallen dadurch auf, dass die Lage der dazugehörigen Messpunkte auf dem Wafer stark variiert. So liegen die meisten Spitzenwerte für diese Messungen im Bereich des Wafers mit einer hohen Kobaltkonzentration; allerdings verlagern sich die Spitzenwerte ab $\theta = 57,43°$ sprunghaft in den Bereich der höchsten Titankonzentration.

Außerdem fallen in der Betrachtung die Winkel um $\theta = 78,2°$ auf, da diese lediglich für zwei Messpunkte auffällig höhere Intensitäten aufweisen.

TiNiCu Für diesen Datensatz ist auch eine Filterung der Störgrößen sinnvoll, weswegen der Winkelbereich zwischen $\theta = 69° - 70°$ ausgelassen wird.

Die Spektren der einzelnen Messpunkte unterscheiden sich nicht sehr stark von einander; es zeigt sich eine deutliche Spitze im Bereich $\theta = 75,4°$. In den Winkelbereichen direkt unter- und oberhalb der Siliziumspitze zeigen sich ebenfalls deutlich erhöhte Intensitäten. Es fällt auf, dass das Intensitätsniveau in diesem Bereich konstant höher ausfällt, so dass sich ein deutliches Plateau in den Spektren abzeichnet. Es könnte sich um weitere Interferenzen des Silizium-Wafers handeln

Weitere Spitzenwerte sind im Bereich $\theta = 42,5° - 44°$ zu beobachten, wobei diese in ihrer Intensität nicht stark abheben und eher über die gesamte Breite erhöhte Intensitätswerte aufweisen, nicht einzelne markante Spitzenwerte.

Ansonsten gibt es auch hier die Ausreißer am Rand des Wafers um $\theta = 38,2°$. Die bekannte Formgedächtnissphase des TiNiCu-Systems [34] ist auf Anhieb nicht an den XRD-Spektren zu erkennen. Der vorhandene Datensatz deckt diesen bekannten Phasenraum jedoch auch nur in geringen Teilen ab.

TiTaAl Die XRD-Spektren zeigen, wie bei TiTaRe, auch hier über den gesamten Wafer ein sehr ähnliches Muster. Es zeigen sich etwa zehn größere Ausschläge in den Bereichen um $\theta = \{35,2° ; 38,2° ; 43,4° ; 52,6° ; 57,4° ; 66,5° ; 68,5° ; 76,9° ; 95,1°\}$, mit Intensitäten von etwa 1000 bis 6500, wobei im Bereich $\theta = 57,4°$ die größten Intensitäten auftreten. Für den Bereich um $\theta = 38,2°$ gibt es zehn besonders auffällige Ausreißer, da diese jedoch exakt am Rand des Wafers liegen und gleichzeitig bei vollkommen unterschiedlichen Materialkonzentrationen, handelt es sich hierbei wahrscheinlich um Messfehler.

TiTaRe Ohne die extremen Ausschläge des Siliziums lassen sich leichter Strukturen in den XRD-Spektren erkennen. Zunächst fällt auf, dass sich die Spektren in ihrer Form weniger stark unterscheiden als bei den anderen ternären Systemen. Dies ist wahrscheinlich darin begründet, dass die Materialzusammensetzung weniger Varianz aufweist, ähnlich wie auch bei TiTaAl; insbesondere die Rhenium-Konzentration variiert kaum.

Die stärksten Ausschläge gibt es bei den Winkelbereichen $\theta = \{39° ; 61,5°; 71 ; 75,5° ; 96°\}$, wobei die Intensitäten für $\theta = 96°$ deutlich niedriger und weniger auf einen einzelnen Winkel beschränkt sind.

Für diesen Datensatz ist zu vermuten, dass es bei allen Messpunkten um ähnliche Ausprägungen der gleichen Materialphase oder mehrere Mischphasen handelt. Durch die starke Ähnlichkeit der einzelnen XRD-Spektren ist dieser Datensatz wahrscheinlich ein gutes Referenzsystem für die Genauigkeit des späteren Gesamtsystems. Außerdem können aufgrund der ähnlichen Konzentration von zwei Materialien Vergleiche mit dem TiTaAl-System gezogen werden.

Zusammenfassende Betrachtung der XRD-Spektren Vergleichend alle XRD-Spektren zu betrachten ist aufgrund der großen Datenmenge schwierig. Für einen vollständigen Überblick müssten die Spektren aller 342 Messpunkte verglichen werden und nach Möglichkeit noch die chemische Zusammensetzung angezeigt werden, zum Beispiel über die Position im ternären System. Eine Möglichkeit einen Überblick über den allgemeinen Aufbau der XRD-Spektren zu bekommen, ist es die Messwerte der Messpunkte zusammenzufassen. Somit erhält man nur noch ein Spektrum, welches zum Beispiel die maximale beobachtete Intensität zeigt. Damit können in einer Darstellung alle Winkelbereiche mit Intensitätsspitzen erkannt werden. In Abbildung 2.3 sind die aggregierten XRD-Spektren für die untersuchten Datensätze zu sehen. Die deutlichen Unterschiede in den gemessenen Intensitäten fallen insbesondere für TiTaAl und TiTaRe stark auf, da in dieser Darstellung einzelne Spitzenwerte kaum zu erkennen sind. Im Vergleich zeigt die Betrachtung der Spektren für diese beiden Datensätze alleine, dass es durchaus markante Intensitätsspitzen gibt (siehe Abbildung 2.4).

2.4.3 Charakterisierung eines interessanten Bereiches am Beispiel von Titan-Kobalt-Wolfram

Der TiCoW-Datensatz ist ein gutes Beispiel, um zu zeigen, wie ein auffälliger Bereich innerhalb eines ternären Systems gestaltet sein kann. Der aufgezeigte Bereich zeigt eine deutliche lokale Änderung im elektrischen Widerstand, die sowohl im ternären Diagramm (Abb. 2.2), als auch in den Koordinaten des Wafers (Abb. 2.5) gut zu erkennen ist. Aus diesem Grund wird der Datensatz in den folgenden Kapitel immer wieder aufgegriffen, um zu überprüfen, ob dieser Bereich von den untersuchten Methoden angemessen berücksichtigt wird. Dies ist natürlich kein formaler Beweis, aber in Ermangelung von eindeutig gekennzeichneten Datensätzen eine zuverlässige Kontrollmöglichkeit.

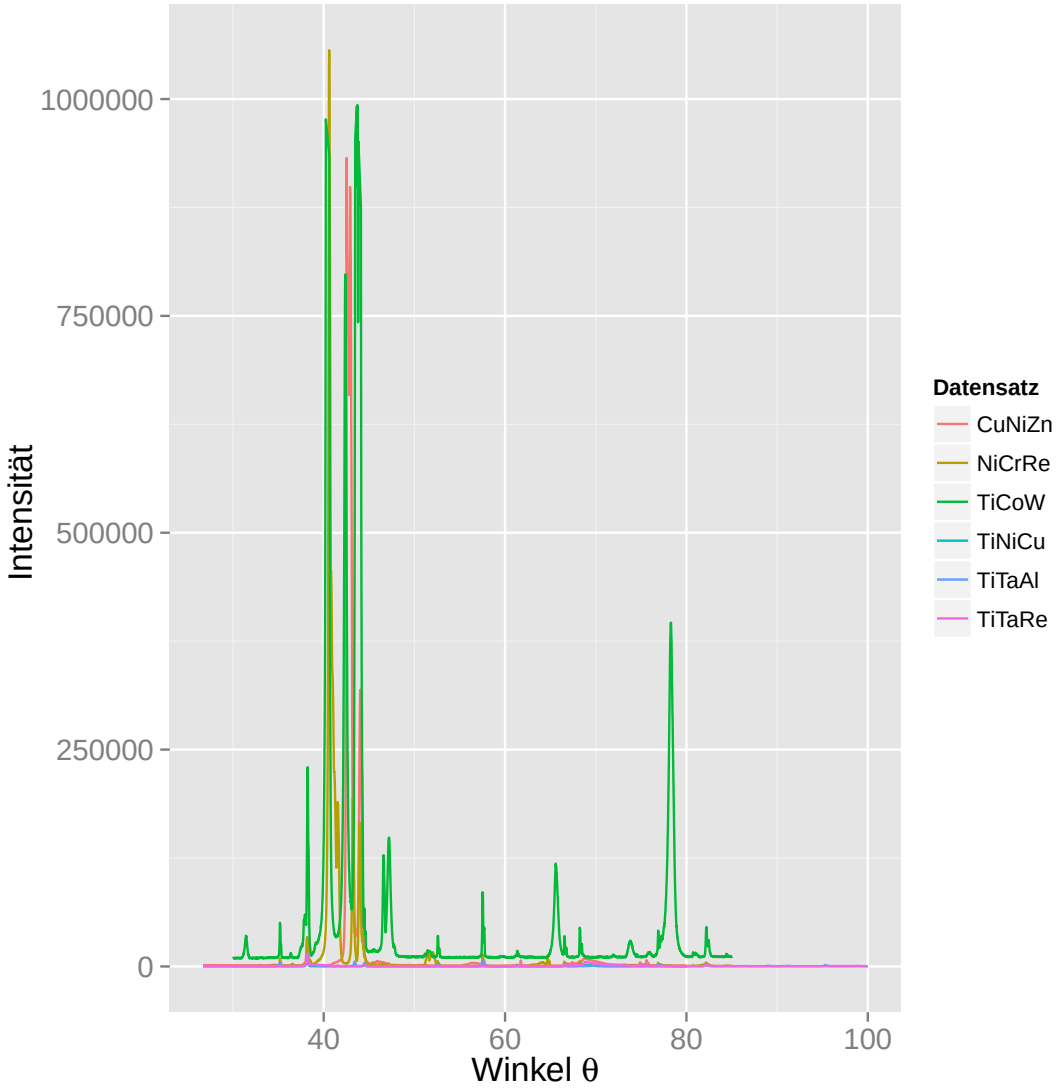


Abbildung 2.3: Übersicht der aggregierten maximalen Intensitäten der XRD-Spektren. Für jeden Datensatz wurde jeweils das Maximium über alle 342 Messpunkte für jeden Winkel gemessen. Die Messungen im Bereich $\theta = 69° - 70°$ wurden für CuNiZn, TiNiCu und TiTaRe entfernt

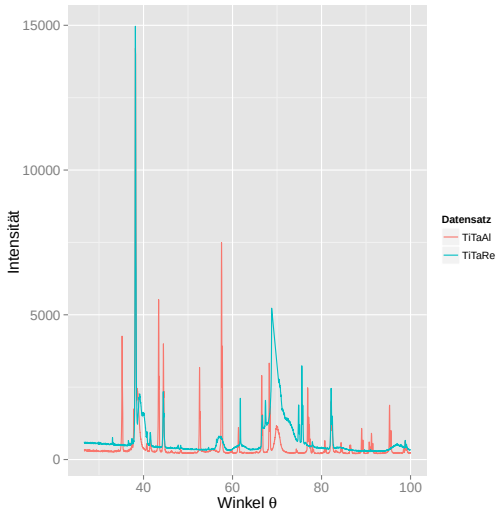
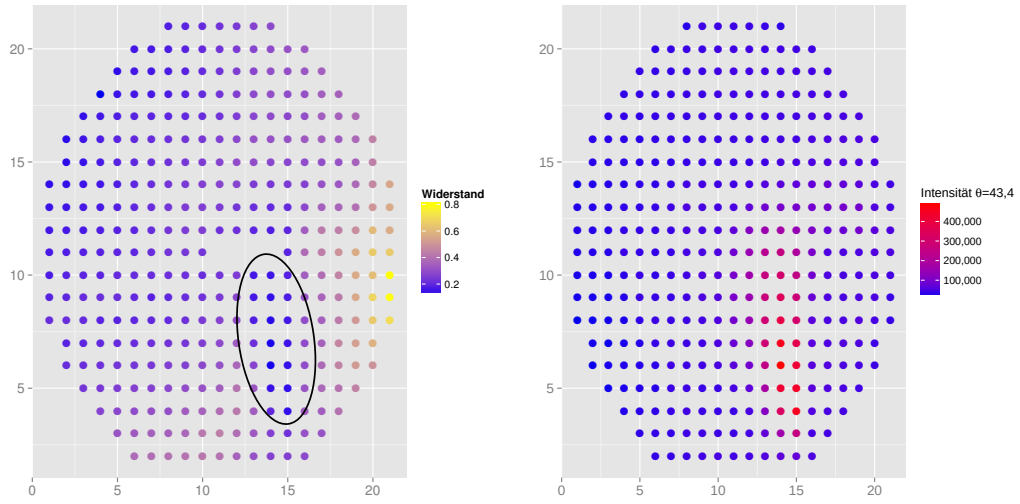


Abbildung 2.4: Übersicht der aggregierten maximalen Intensitäten der XRD-Spektren für TiTaAl und TiTaRe



(a) elektrischer Widerstand

(b) Intensität der XRD-Messung für $\theta = 43,41°$

Abbildung 2.5: TiCoW-Datensatz, mit markiertem Bereich der interessanten Phase.

3 Kostenorientierte Lernverfahren

Für viele reale Prozesse, aber auch bei theoretischen Prozessen, ist ein entscheidendes Kostenmaß die Anzahl von benötigten Funktionsauswertungen. Dabei kann es sich um komplizierte Berechnungen handeln oder um aufwendige physikalische Messungen. Verschiedene Disziplinen der Informatik, aber auch der Wirtschaftswissenschaften im Bereich der *Operations Research*, beschäftigen sich mit der Frage, wie die Anzahl der notwendigen Operationen reduziert werden kann. Ist die Zielfunktion bekannt, so lässt sich mittels mathematischer Verfahren, wie Gradientmethode oder dem Simplex-Algorithmus zum Lösen von linearen Gleichungssystemen, ein minimaler Funktionswert bestimmen. Bei einer unbekannten Zielfunktion gibt es direkte Suchverfahren oder populationsbasierte Verfahren, wie die evolutionären Algorithmen und deren Weiterentwicklung wie Räuber-Beute-Algorithmen oder schwarmbasierte Verfahren.

All diesen Verfahren ist gemein, dass sie versuchen mit möglichst geringen Kosten eine möglichst gute, nicht zwingend optimale, Lösung zu finden. Beim Maschinellen Lernen ist der Ansatz meist ein anderer. Hier ist das Ziel zuverlässig, ausgehend von einer Menge bekannter Beispiele, Aussagen über neue, unbekannte Beispiele zu treffen. Dabei wird die initiale Trainingsmenge als gegeben angesehen. Allerdings gilt auch hier die Prämisse, dass Funktionsauswertungen teuer sind. Die der Auswertung der XRD-Spektren ist jedoch gerade ein Beispiel, bei dem versucht wird, mit möglichst wenig Messungen, also Funktionsauswertungen, verlässliche Aussagen zu treffen. Aus diesem Grund werden einige Verfahren, die diesen Aspekt berücksichtigen, vorgestellt.

3.1 Active Learning

Die Idee beim *active learning* ist, dass das Lernverfahren in der Lage ist zu entscheiden, für welche Punkte eine Funktionsauswertung erfolgen soll. Diese Entscheidung kann zum Beispiel dadurch beeinflusst werden, welche Punkte im Modell die größte Unsicherheit aufweisen oder welche Punkte wahrscheinlich den größten Informationsgewinn liefern. Eine Grundvoraussetzung für dieses Vorgehen ist, dass es eine Orakel-Funktion gibt, die in der Lage ist für den gewünschten Punkt ein Labelzu ermitteln.

Einen sehr guten Überblick über die Entwicklung und verschiedene Verfahren des *active learning* bieten die Arbeiten von Settles [26, 27]. Ein Ansatz wird hier aufgeführt, um beurteilen zu können, inwieweit sich das Verfahren auf die Messung der XRD-Spektren übertragen lässt, beziehungsweise welche Probleme dabei auftreten.

Auswahl nach Unsicherheit (uncertainty sampling) Dieser verbreitete und in der Regel einfach umzusetzende Ansatz beruht darauf, dass Punkte von der Orakel-Funktion ausgewertet werden, für die der Klassifikator am unsichersten ist. Gibt es zum Beispiel bereits eine Hypothese wie die Trennlinie zwischen zwei Klassen verläuft, so würde diese Strategie Punkte auswählen, welche möglichst nah an dieser liegen. Punkte, die weit von der Trennlinie liegen, sind hingegen wahrscheinlich recht sicher zu klassifizieren und insofern weniger informativ.

Für probabilistische Modelle würde entsprechend die Auswahl danach erfolgen, für welche Punkte die Zuordnungswahrscheinlichkeit am uneindeutigsten ist. Wieder bezogen auf ein Zweiklassenproblem, wie zum Beispiel die Logistischen Regression, sind dies Punkte für die die Wahrscheinlichkeit einer Klasse zugeordnet zu werden $0,5$ beträgt.

Der Nachteil der Methode der geringsten Konfidenz ist, dass nur die Informationen für die wahrscheinlichste Klasse berücksichtigt werden, was bei mehr als zwei Klassen dazu führt, dass relevante Informationen, wie die Verteilung der übrigen Klassen, verloren gehen. Verfahren, die diesen Nachteil umgehen, berücksichtigen stattdessen die Entropie nach Shannon [28] als Maß für die Unsicherheit.

Neben dem *uncertainty sampling* gibt es noch weitere Ansätze, welche zum Beispiel eine Menge von Modellen verwalten. Die Entscheidung, welche Punkt ausgewählt werden, wird durch den Vergleich der Ergebnisse der Modelle getroffen. Ein Kriterium wäre solche Punkte auszuwählen, für die die Modelle unterschiedliche Klassen vorschlagen.

3.2 Verwandte Methoden

Design of Experiments In der Statistik, insbesondere im Bereich der Versuchsplanung [24], gibt es auch Bemühungen verlässliche Aussagen über die Güte von Stichproben treffen zu können. Damit einher geht die Überlegung, welche Punkte des Messraums die zugrundeliegenden statistischen Prozesse gut beschreiben können. Diese Überlegungen werden unter dem Begriff des *Optimal Experimental Design* zusammengefasst. Eine wichtige Informationsquelle hierfür ist die Inverse der Kovarianzmatrix des untersuchten statistischen Prozesses, die Fischersche Informationsmatrix [4], deren Analyse Aufschluss darüber gibt, welche Punkte für das statistische Experiment ausgewählt werden.

Optimal Sensor Placement Für Probleme, die eine räumliche Struktur haben, beispielsweise die Messung der Temperatur in einem Gebäude, ist es wünschenswert mit möglichst wenig Messpunkten verlässliche Ergebnisse zu erhalten. Hierbei liegt der Schwerpunkt weniger auf der korrekten Klassifizierung von Bereichen, sondern vielmehr auf der möglichst vollständigen Überdeckung des Messbereiches. Meistens liegen den verwendeten Messverfahren physikalische Prozesse zugrunde, so dass eine geeignete mathematische Modellierung möglich ist. Ein Beispiel für eine derartige Modellierung zeigen Krause et al. [17] mittels gaußschen Prozessen.

3.3 Übertragung auf die Auswertung von XRD-Spektren

Die Ansätze des *active learning* decken sich stark mit der gestellten Aufgabe die Verteilung der XRD-Spektren auf einem Wafer effizient zu charakterisieren. Die Herangehensweise des Optimal Sensor Placement deckt sich mit der Forderung möglichst wenige Messungen vorzunehmen, während mittels *active learning* Punkte für eine Messung vorgeschlagen werden können, für welche die Unsicherheit der Vorhersagen am größten ist. Die Schwierigkeit bei der Übertragung beruht allerdings vor allem auf der Struktur der XRD-Spektren. Anstatt eines eindimensionalen Merkmals, wie etwa der Klassenzugehörigkeit, bestehen die XRD-Spektren aus hochdimensionalen Messreihen. Selbst eine Charakterisierung der Spektren in typische Ausprägungen, um damit einzelne Klassen zu generieren, ist nicht ohne Weiteres möglich, da für jede Probe von Materialmischungen die XRD-Spektren vollkommen andere Formen und Wertebereiche aufweisen können.

Dennoch ist es möglich einige der Ansätze dieser Verfahren auf das gegebene Problemfeld zu übertragen. In Kapitel 7 wird gezeigt wie, mittels der DTW-Distanz zwischen bekannten XRD-Spektren, die Distanzen in deren Umfeld geschätzt werden können. Diese Abschätzungen bilden zusammen eine Gruppe von Modellen und aus deren gemeinsamen Informationen wird versucht den Verlauf der DTW-Distanzen für unbekannte Messpunkte zu extrapolieren. Diese Abschätzung wird genutzt, um Punkte für weitere XRD-Messungen vorzuschlagen.

In Kapitel 8 wird ebenso eine Gruppe von Modellen erzeugt, jedoch über einzelne Bereiche der XRD-Spektren, welche zusätzlich auf lediglich zwei Ausprägungen reduziert werden. Damit ist es möglich weitere Punkten anhand der Unsicherheit der erstellten Modelle auszuwählen.

4 Verwendete Methoden

Dieses Kapitel liefert einen Überblick über die Methoden, welche in den folgenden Kapiteln verwendet werden. Dabei handelt es sich um die beiden Clustering-Algorithmen *k-means* und DBSCAN, Regression mittels linearen Modellen, Klassifikation mittels Support-Vector-Maschinen, sowie Dynamic-Time-Warping- und Hamming-Distanz als Distanzmaße. Diese Verfahren sind bereits in der Software RapidMiner [23] vollständig implementiert und wurden in dieser Form auch für alle Berechnungen dieser Arbeit verwendet. Aus diesem Grund liegt der Schwerpunkt des Kapitels auf der allgemeinen Funktionsweise der Verfahren und weniger auf den Details der Implementierung oder genauen algorithmischen Umsetzung. Einige der folgenden Verfahren wurden bereits in einer vorangegangen Studienarbeit [1] verwendet, in der es um die Zusammenhänge zwischen Materialeigenschaften und XRD-Spektren ging.

4.1 Clustering-Verfahren

Clustering-Verfahren versuchen in einer gegebenen Menge von Merkmalen, ähnliche Bereiche zu Identifizieren und zu einer gemeinsamen Gruppe, *Cluster*, zusammenzufassen.

Ein Vorteil von Clustering-Verfahren ist, dass sie keine vorgegebenen Klassen benötigen. Es gibt dabei zwei Ansätze und zwar partitionierende Verfahren und hierarchische Verfahren.

Partitionierende Algorithmen versuchen selbstständig die Daten nach Ähnlichkeit in Gruppen einzuteilen. Dies kann sehr nützlich sein, um überhaupt einen Überblick über die Struktur der vorhandenen Daten zu bekommen.

Hierarchische Algorithmen hingegen versuchen die vorhandenen Daten so zu ordnen, dass eine Baumstruktur entsteht. In den jeweiligen Ebenen des Baums werden ähnliche Elemente mit zunehmender Genauigkeit der Elemente zusammengefasst, wobei die Blätter des Baumes die einzelnen Beispiele sind.

k-means Der *k-means*-Algorithmus [14] ist einer der bekanntesten Clustering-Algorithmen. Der Ansatz beruht darauf, dass der Merkmalsraum in k feste Cluster, mit $k \in \mathbb{N}$, eingeteilt wird. Zu jedem Cluster gibt es einen zentralen Punkt, der Zentroid genannt wird. Alle Beobachtungen werden nun dem jeweils nächstgelegenen Zentroiden zugewiesen, wodurch die Einteilung der Daten festgelegt wird. Als Distanzfunktion wird in der Regel der Euklidische Abstand genutzt.

Der Algorithmus arbeitet nun iterativ und versucht in jeder Iteration die Qualität der Einteilung zu verbessern. Dafür wird der Mittelwert der jeweiligen Cluster bestimmt und als neuer Zentroid festgelegt. Im nächsten Iterationsschritt werden die Zuordnung der Cluster und die Zentroiden neu berechnet. Als initiale Zentroiden können beliebige Punkte des Merkmalsraums oder auch zufällig ausgewählte Beobachtungen verwendet werden. Die Iterationen verlaufen nach dem folgenden Schema:

1. Berechne für jeden Punkt den nächstgelegenen Zentroid und weise ihm diesen zu.
2. Berechne k neue Klassenzentren anhand der neuen Klasseneinteilung.
3. Falls das Abbruchkriterium noch nicht erfüllt ist, beginne wieder mit Schritt 1.

Es können verschiedene Abbruchkriterien verwendet werden. Am einfachsten ist es eine maximale Anzahl von Iterationen zu wählen, was den Vorteil hat, dass zugleich die maximale Laufzeit bekannt ist. Eine andere Variante betrachtet die Lage der Klassenzentren; falls sich diese nicht mehr hinreichend stark ändern, bricht der Algorithmus ab.

DBSCAN Den Nachteil der festgelegten Anzahl von Clustern umgeht der dichtebasierte DBSCAN-Algorithmus von Ester [8]. Ein weiterer Vorteil des DBSCAN-Algorithmus ist, dass die Form der gefundenen Cluster beliebig sein kannn und nicht im Vorfeld durch die Distanzfunktion wie beim *k-means*-Clustering bestimmt wird.

Das Grundprinzip des DBSCAN-Algorithmus beruht darauf um jeden Beispielpunkt herum eine ϵ-Umgebung zu definieren und eine minimale Anzahl *minPts* von Punkten festzulegen, die in dieser Umgebung liegen sollen. Mit Hilfe dieser Werte lassen sich Kernpunkte definieren, in deren ϵ-Umgebung mindestens die geforderte Menge von Punkten liegt. Ein Punkt p gilt von einem Punkt q aus als *density-reachable*, wenn er in der ϵ-Umgebung von q liegt und q ein Kernpunkt ist. Daraus abgeleitet gelten zwei Punkte p und q als *density-connected*, wenn es eine Reihe von weiteren Punkten o gibt, über die p und q *density-reachable* sind.

Aus dieser Definition lässt sich nun ein Clustering-Verfahren ableiten, in dem für jeden Cluster gelten muss, dass alle Punkte in einem Cluster miteinander *density-connected* sein müssen und es keine weiteren Punkte gibt, für die diese Bedingung erfüllt ist, die Menge also vollständig ist.

Diese kurze Erklärung zeigt bereits, dass die Anzahl der Cluster nicht festgelegt ist und nur indirekt durch die Wahl von ϵ und die Anzahl von notwendigen Punkten *minPts* beeinflusst wird.

Lineare Modelle Lineare Modelle versuchen einen einfachen funktionalen Zusammenhang zwischen mehreren unabhängigen Einflussvariablen $x_1, \ldots, x_n \in \mathbb{R}$ und einer von diesen abhängigen Zielvariable $y \in \mathbb{R}$ herzuleiten, um so ein Regressionsproblem zu lösen. Für diesen Zusammenhang werden, wie der Name andeutet, lediglich lineare Parameter verwendet, wodurch sich die Gleichung des linearen Modells formulieren lässt als

$$y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_n x_{i,n} + \epsilon_i \,,$$

für alle Beispiele $i = 1 \dots m$. Ziel ist es den Fehlerverktor $\epsilon \in \mathbb{R}^m$ zu minimieren. Die Modellgleichung lässt sich mit Hilfe der Methode der Kleinsten-Quadrate [13] lösen, welche unter den Annahmen der statistischen Unabhängigkeit aller Einflussvariablen x_i und eines normalverteilten Fehlervektors ϵ sogar optimal ist.

Neben verlässlichen statistischen Aussagen ist der Vorteil von linearen Modellen, dass sie einfach zu berechnen und ihre Ergebnisse gut interpretierbar sind. Der vermeintliche Nachteil, dass lediglich lineare Zusammenhänge modelliert werden, kann durch geeignete lineare Transformationen der Einflussvariablen ausgeglichen werden. So kann durch das Produkt von zwei Einflussvariablen $x_k * x_l$ die Wechselwirkung auf y modelliert werden.

Support-Vector-Maschinen Ein binäres Klassifikationsproblem besteht aus einer Menge von m Beispielen $x_i \in \mathbb{R}^n$ mit $i \in \{1 \dots m\}$, wobei jedem Beispiel genau eine von zwei möglichen Klassen y zugeordnet wird. Ziel ist es nun eine beschreibende Funktion zu finden, die jedem Beispiel die korrekte Klasse zuordnet. Eine mögliche Herangehensweise ist die Stützvektor-Methode (SVM), die versucht eine trennende Hyperebene zwischen die beiden Klassen zu legen. Falls das Klassifikationsproblem linear separierbar ist, so kann es theoretisch beliebig viele solcher Trennebenen geben. Die SVM konstruiert aber die Trennebene unter dem Aspekt, dass sie den größten Abstand zu allen bekannten Beispielen hat. Dieses Vorgehen ist dadurch motiviert, dass so das Risiko einer Fehlklassifikation für neue unbekannte Beispiele minimiert wird.

In der Praxis ist ein perfekt zu trennendes Problem eher selten, weswegen die Bedingung der korrekten Klassifikation für alle Beispiele aufgelockert werden kann. Dabei sind Fehlklassifikationen erlaubt, werden jedoch mit einem konstanten Faktor gewichtet. Bei der Suche nach einer optimalen Lösungen werden die Lösungen mit zu vielen Fehlern schlechter bewertet.

Ein weiteres wichtiges Hilfsmittel für SVMs ist die Verwendung einer Kernel-Funktion, die den Merkmalsraum der Beispiele in einen höherdimensionalen Raum transformiert. Durch diese Transformation ist es nun möglich komplexere Trennfunktionen als nur eine lineare Hyperebene zu verwenden. Praktikabel ist dieses Verfahren dadurch, dass für die Berechnung der optimalen Lösung nur das Skalarprodukt der Beispielvektoren betrachtet wird und die Komplexität des Bildraumes somit keine Rolle spielt.

In dieser Arbeit wird der C-SVM-Algorithmus von Boser et al.[3] genutzt, welcher unter anderem in der LIBSVM-Bibliothek [5] implementiert ist und in RapidMiner zur Verfügung steht. Beide Arbeiten bieten auch eine ausführlichere Beschreibung der SVM-Methode und des zu lösenden Optimierungsproblems.

4.2 Distanzmaße

Dynamic-Time-Warping Zwei Wertereihen $X := (x_1, \dots, x_n)$ und $Y := (y_1, \dots, y_m)$ mit $m, n \in \mathbb{N}$ können auf unterschiedliche Weisen mit einander verglichen werden. Ein einfaches Ähnlichkeitsmaß ist die Summe des Euklidischen Abstands für alle Wertepaare

(x_i, y_i). Dabei gibt es jedoch zwei Probleme, zum einen müssen die beiden Reihen die gleiche Länge haben und zum anderen werden strukturelle Ähnlichkeiten nicht berücksichtigt. Am Beispiel der XRD-Spektren veranschaulicht hätte ein Spektrum, welches nur eine Intensitätsspitze im Bereich $\theta = 40°$ hat, den gleichen Abstand zu zwei Spektren mit Intensitätsspitzen bei entweder $\theta = 42°$ oder $\theta = 75°$, obwohl die Ähnlichkeit zum ersten Spektrum deutlich größer ist.

Ein Distanzmaß, welches dieses Problem umgeht, ist das Dynamic-Time-Warping (DTW) [2]. Der Ansatz dieses Verfahrens beruht darauf, dass die Ähnlichkeit zwischen jedem Paar von Punkten aus X und Y berechnet wird und in einer Kostenmatrix $C \in \mathbb{R}^{n \times m}$ eingetragen wird. Beginnend mit dem ersten Paar (x_1, y_1) wird nun ein zusammenhängender Pfad bis zum letzten Paar (x_n, y_m) gesucht, welcher in der Summe die geringsten Kosten aufweist. Die Summe aller Kosten des gefunden Pfades wird als DTW-Distanz bezeichnet. Für das obige Beispiel bedeutet dieses Verfahren, dass die DTW-Distanz zwischen dem ursprünglichen Spektrum und dem Spektrum mit der Intensitätsspitze bei $\theta = 42°$ kleiner ist als die DTW-Distanz zum zweiten Spektrum. Diese Tatsache macht die DTW-Distanz als Ähnlichkeitsmaß für XRD-Spektren interessant, da der Winkel θ, für den ein Spitzenwert beobachtet wird, für verschiedene Messungen leicht verschoben sein kann.

Hamming-Abstand Der Hamming-Abstand gibt den Unterschied zwischen zwei gegebenen Zeichenketten an. Für zwei Zeichenketten X und Y über ein gleiches Alphabet und mit der selben Länge n ist der Hamming-Abstand definiert als die Summe der Stellen $i = 1 \dots n$ für die gilt $x_i \neq y_i$. Für binäre Vektoren lässt sich der Hamming-Abstand mittels der XOR-Operation leicht ermitteln.

Für mehr als zwei Zeichenketten kann ebenfalls ein Hamming-Abstand berechnet werden, zum Beispiel über den minimalen oder maximalen Abstand zwischen allen Kombinationen von Paaren.

5 Adaptive Messungen

Ein gängiger Ansatz in der Materialforschung ist es zunächst eine vollständige Messung für alle Punkte des zu untersuchenden Wafers vorzunehmen. Dabei werden viele redundante Daten erzeugt, die wieder verworfen werden, da die betreffenden Bereiche für die jeweilige Fragestellung nicht relevant sind. Erst mit der Identifizierung und Abgrenzung von interessanten Bereichen beginnt die systematische Erfassung von Kennwerten für genau diesen Bereich. Diese Arbeit erfordert oftmals wiederholte Experimente, die Erzeugung und Messung von Materialeigenschaften unter verschiedenen Umweltbedingungen, wie Temperatur oder Druck. An dieser Stelle ist die Datenanalyse weniger von Nutzen, da vor allem die Erfahrung und das Expertenwissen der Materialforscher gefordert ist.

Aber die Vorauswahl und das Identifizieren von interessanten Bereichen kann mittels adaptiver Messungen unterstützt und automatisiert werden, so dass weniger unnötige Daten ausgewertet müssen.

5.1 Schematischer Verlauf der adaptiven Messungen

Als Grundlage für sämtliche Untersuchungen dienen eindimensionale Messdaten, welche im Vorfeld mit geringem Zeitaufwand erhoben werden können. Diese Messdaten werden auf dem Wafer in einem Raster mit N Messpunkten erfasst; für die hier untersuchten Datensätze gilt $N = 342$. Diese Daten beinhalten die chemische Zusammensetzung, die den Messpunkt eindeutig charakterisiert, den elektrischen Widerstand und weitere problemspezifische Messwerte, wie zum Beispiel die Farbe der Materialprobe. Eine mögliche Messanordnung ist in Abbildung 5.1 zu sehen und die einzelnen Schritte werden im Folgenden erklärt.

Initiale Stichprobe Aufgrund der Analyse dieser vorhandenen Messdaten werden n_i Punkte für eine initiale Stichprobe ausgewählt, für welche die noch unbekannten XRD-Spektren gemessen werden sollen. Ziel der Analyse ist es, Bereiche aufzuzeigen, in denen sich die Messpunkte ähnlich verhalten. Der elektrische Widerstand ist hierfür ein guter Indikator, denn er hängt stark von der Konzentration der vorhandenen chemischen Elemente ab. Veränderungen in der atomaren Struktur der Kombination von zwei oder mehr Elementen können den Wert zusätzlich beeinflussen.

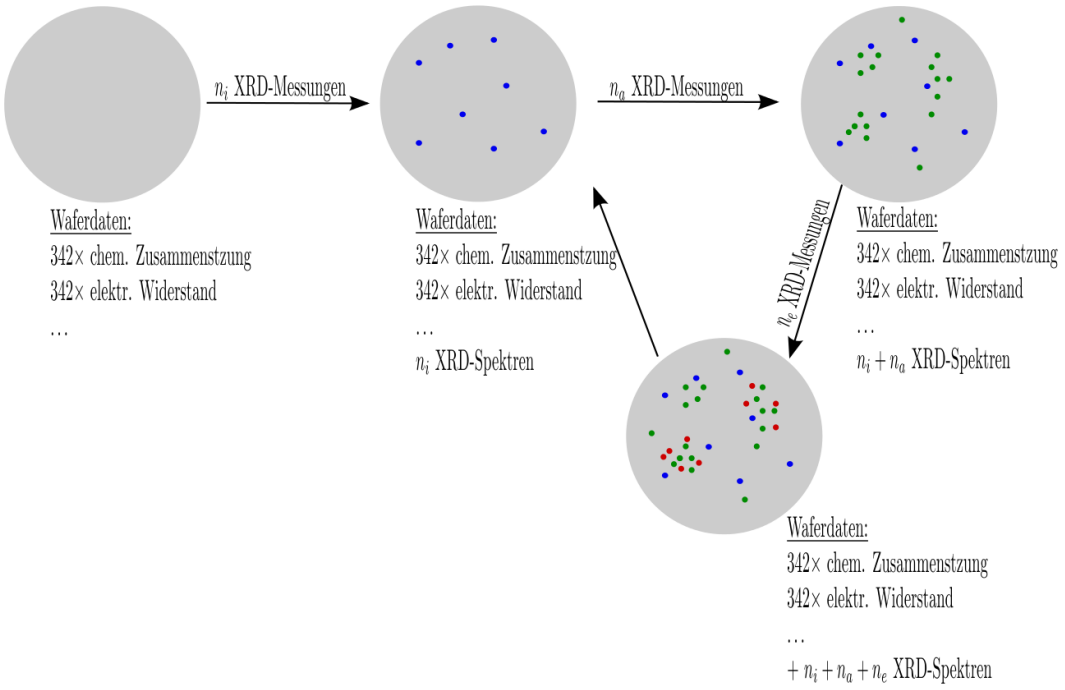


Abbildung 5.1: Skizzierte Ablauf der adaptiven Messungen

Ziel der initialen Stichprobe ist es mit lediglich n_i Messpunkten eine aussagekräftigere Repräsentation des Wafers zu wählen als es eine reine Zufallsauswahl könnte. Insbesondere sollen die gemessenen XRD-Spektren möglichst vielfältig sein.

Approximationsgestützte Messungen Mittels der n_i Datenpunkte ist es möglich eine erste Approximation für die Struktur der $N - n_i$ unbekannten XRD-Spektren zu erstellen. Die Untersuchung der Approximation kann nun aufzeigen welche Messpunkte auf ein interessantes Gebiet hindeuten. Anhand dieser Untersuchung können n_a neue Messpunkte ausgewählt werden, für die die XRD-Spektren gemessen werden.

Messungen aufgrund des Approximationsfehlers In einem letzten Schritt bietet der Vergleich der tatsächlichen Messwerte mit den approximierten Werten die Möglichkeit die Messungen weiter auf die Umgebung von Punkten einzuschränken, für welche die Approximation sehr ungenau ist. Während die Motivation für die ersten beiden Messdurchgänge explorativ war, um interessante Gebiete zu entdecken, liegt der Fokus in diesem Schritt darin den Übergangsbereich zwischen Gebieten, die sich vorhersagbar verhalten, und den wahrscheinlich interessanten, weil abweichenden, Gebieten abzugrenzen. Dafür werden nochmals n_e Messpunkte ausgewählt.

Weitere Iterationen Es besteht nun die Möglichkeit, die Schritte nach der initialen Stichprobe in einer weiteren Iterationen erneut zu durchlaufen. Damit könnte die Genauigkeit der Vorhersagen noch weiter verbessert werden. Je nach Aufgabenstellung und verfügbaren Ressourcen könnte eine maximale Anzahl von XRD-Messungen oder die Qualität eines erstellten Modells für alle Messpunkte als Abbruchkriterium dienen.

5.2 Praktische Anwendung

Der Vorteil dieser Modellierung der Problemstellung liegt darin, dass einzelne Teile des Verfahrens verändert oder problemspezifisch angepasst werden können. Die für die initiale Stichprobe verwendeten Daten können sich unterscheiden oder es kann durch bereits zuvor gesammelte Informationen ein Schwerpunkt auf bestimmte Bereiche des Wafers gelegt werden. Es besteht außerdem die Möglichkeit diese Verfahren vollkommen autonom laufen zu lassen, da sich die Auswahl der Messpunkte über eine Schnittstelle an die für die XRD-Messungen verantwortliche Maschine weiterleiten lässt.

Andererseits erschweren die vielen Variationspunkte eine systematische Untersuchung und Optimierung der angewendeten Verfahren. Allein durch die Festlegung der Anzahl der Messpunkte n_i , n_a und n_e entstehen exponentiell viele neue Kombinationen. Aus diesem Grund können in dieser Arbeit auch nur exemplarisch einige Methoden untersucht werden, wobei das vergleichsweise abgeschlossene Problem der initialen Stichprobe ausführlich untersucht wird. Die beiden in Kapitel 7 und 8 beschriebenen Verfahren zur Approximation und Auswertung der XRD-Spektren zeigen zwei verschiedene Ansätze, wie weiter vorgegangen werden kann.

6 Initiale Stichprobe

Das Ziel dieser Arbeit ist eine Messanordnung zu erstellen, welche die möglichen Ausprägungen der XRD-Spektren für ein gegebenes ternäres System möglichst gut erfasst, ohne dabei für sämtliche Messpunkte eine vollständige Messung der XRD-Spektren durchzuführen. Für diesen Zweck sollen in einem Vorverarbeitungsschritt die vorhandenen Daten zu dem ternären System analysiert werden, um möglichst informative Messpunkte bereits im Vorfeld auszuwählen.

Das Problem bei der Selektion ist, dass jedes neue ternäre System sich vollkommen anders verhalten kann als zuvor untersuchte Systeme. Aus diesem Grund kann auf zuvor erkannte Zusammenhänge anderer Systeme zurückgegriffen werden, wie es bei anderen Lernaufgaben oftmals der Fall ist. Deswegen muss jeder neue Datensatz als vollkommen neue Aufgabe betrachtet werden.

Ziel der initialen Stichprobe ist es zunächst die Messdaten in Gruppen einzuteilen, die bezüglich einer oder mehrerer bereits bekannter Materialeigenschaften ein ähnliches Verhalten aufweisen. Die Hypothese ist, dass Punkte die hinsichtlich ihrer chemischen Zusammensetzung und anderer Materialeigenschaften ähnlich sind, auch ähnliche XRD-Spektren aufweisen. Diese Begründung beruht unter anderem auf den Ergebnissen der zuvor erstellten Studienarbeit [1]. Darin wurde gezeigt, dass bezüglich eines festen Startpunktes auf dem Wafer die meisten Punkte mit einer vorgegebenen DTW-Distanz ein kompaktes und zusammenhängendes Gebiet um diesen Punkt herum bilden. Für diesen Zweck wurde eine Breitensuche von diesem Punkt ausgehend gestartet, welche abbrach, wenn die DTW-Distanz in einer gegebenen Richtung über einem Schwellwert lag. Dabei ergab sich für die drei auch hier verwendeten ternären Systeme CuNiZn, NiCrRe und TiCoW, dass im Durchschnitt über zwei Drittel aller ähnlichen Punkte durch dieses Verfahren gefunden werden.

Es gilt zu überprüfen welche Verfahren geeignet sind, um die Messpunktes des Wafers in geeignete Gruppen einzuteilen. Für diesen Zweck werden die in Kapitel 4 vorgestellten Clustering-Algorithmen und lineare Regression verwendet. Es gilt auch zu klären wie die Güte der erzeugten Partitionierung gemessen werden kann.

Anschließend wird überprüft welche der angewendeten Verfahren tatsächlich eine möglichst repräsentative Stichprobe der XRD-Spektren liefern.

6.1 Partitionierung der Daten

Zunächst muss untersucht werden welche Methoden geeignet sind die vorhandenen Messdaten in ähnliche Bereiche zu unterteilen. Die hier vorgestellten Methoden sind keineswegs die einzig denkbaren Wege dieses Problem zu lösen, bieten jedoch verschiedene Lösungsansätze, die es zu betrachten gilt.

DBSCAN Die erste Methode ist ein dichtebasiertes Clustering mit Hilfe des DBSCAN-Algorithmus, um zusammenhängende Bereiche mit einem ähnlichen Widerstandswert zusammenzufügen. Der Vorteil dieses Verfahrens beruht darauf, dass die Anzahl der Cluster im Vorfeld nicht bekannt sein muss, sondern sich aus den gegebenen Daten ergibt. Außerdem sind die Cluster nicht an eine feste Form gebunden, sondern passen sich der Verteilung der Beobachtungen an.

Es bedarf dennoch einer guten Anpassung der Parameter des Algorithmus, um eine adäquate Beschreibung der Daten zu erhalten. Die Einstellung der Parameter wird dadurch erschwert, dass keine annotierten Datensätze als Trainingsmenge vorliegen.

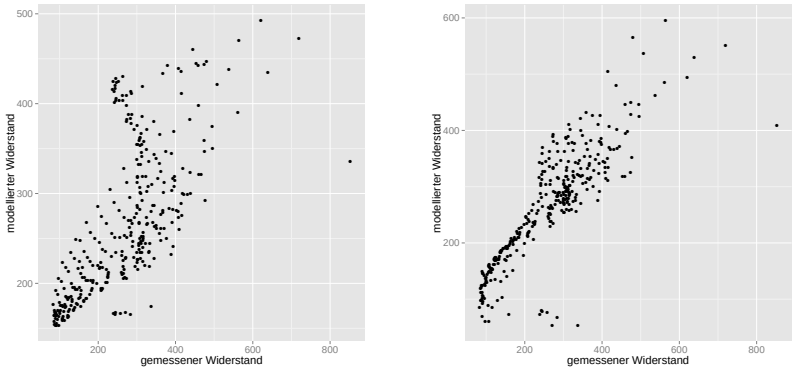
k-means Eine weitere Variante ist das Clustering mittels *k-means*. Einer der Vorteile dieses Algorithmus ist, dass es bis auf die Anzahl der Cluster k keine weiteren Parameter gibt. Außerdem kann k so gewählt werden, dass es mit der Anzahl der Größe der initialen Stichprobe n_i übereinstimmt und somit aus jedem Cluster auch eine Stichprobe ausgewählt wird.

Stichprobe aus Wertebereichen Eine weitere einfache Möglichkeit ist die Einteilung des Widerstandes, als relevantes Unterscheidungskriterium, in i äquidistante Intervalle ,somit ist nicht jedes Intervall gleich dicht besetzt. Die Partitionierung ergibt sich dann aus der Verteilung der Intervalle. Der Vorteil hierbei ist natürlich die sehr einfache Umsetzung dieses Verfahrens. Ein deutlicher Nachteil ist jedoch, dass die Einteilung nur eine einzige Materialeigenschaft berücksichtigt und die Partitionierung nicht unbedingt zusammenhängende Bereiche auf dem Wafer erzeugt.

Regressionsverfahren Die Motivation für diesen Ansatz ist, dass wahrscheinlich ein starken Bezug zwischen Materialeigenschaften und der chemischen Zusammensetzung besteht. Als einfaches Beispiel kann ein ternäres System mit Kupfer betrachtet werden. Da Kupfer ein guter elektrischer Leiter ist, ist die Annahme berechtigt, dass bei einem hohem Kupferanteil der Widerstand niedriger ist.

Natürlich werden bei einer solchen vereinfachten Annahme auch die Wechselwirkungen zwischen den Elementen nicht berücksichtigt, was zu vergleichsweise ungenauen Modellen führt. Werden die Wechselwirkungen jedoch in das lineare Modell mit aufgenommen und folgende Modellgleichung für drei chemischen Elemente A, B und C verwendet

$$y = \beta_0 + \beta_1 A + \beta_2 B + \beta_3 C + \beta_4 AB + \beta_5 AC + \beta_6 BC + \epsilon\,, \tag{6.1}$$



(a) Ohne Wechselwirkung der Materialeigenschaften

(b) Mit Wechselwirkung der Materialeigenschaften, wie in Formel 6.1

Abbildung 6.1: Vergleich der linearen Modelle mit und ohne Wechselwirkungen für den CuNiZn-Datensatz

so verringert sich der Modellfehler deutlich, wie Abbildung 6.1 veranschaulicht. Als interessante Punkte hinsichtlich ihrer XRD-Spektren könnten sich solche Punkte erweisen, bei denen die modellierte Materialeigenschaft nicht gut durch das Modell beschrieben wird. Denn dies könnte ein Hinweis darauf sein, dass sich dort eine besondere Materialphase gebildet hat, deren Eigenschaften sich vom restlichen ternären System unterscheiden.

Am Beispiel der TiCoW-Daten lässt sich dieses Verhalten recht gut beobachten. Die Abweichung des durch das Modell vorhergesagtem Widerstands und der tatsächlichen Messung ist für den ausgewiesenen Bereich der vermuteten Materialphase besonders groß, wie Abbildung 6.2 zeigt.

Aus diesen Überlegungen lässt sich das folgende Auswahlverfahren ableiten. Nach der Bildung des linearen Modells werden jene Punkte ausgewählt, für die der Modellfehler am größten ist. Der Vorteil dieser Methode ist die einfache Modellbildung mit nur wenigen Parametern und dass der Modellfehler direkt als Indikator für interessante Punkt mit einfließt.

Ein gravierender Nachteil ist allerdings, dass es keine Berücksichtigung der räumlichen Verteilung der auffälligen Punkte gibt. Bilden zum Beispiel alle Ausreißer einen zusammenhängenden Cluster im ternären System, so wird dieser überproportional stark vermessen und andere Bereiche des Wafers werden gar nicht berücksichtigt.

Kombination von Regression- und Clustering-Verfahren Der Nachteil der Auswahl ausschließlich anhand des Regressionsfehlers kann ausgeglichen werden, wenn diese Information mit den partitionierenden Verfahren kombiniert wird.

Der Vorhersagefehler kann als Gewicht für die Stichprobenauswahl der anderen hier aufgeführten Verfahren genutzt werden. Anschaulich bedeutet dies, dass die Punkte, für

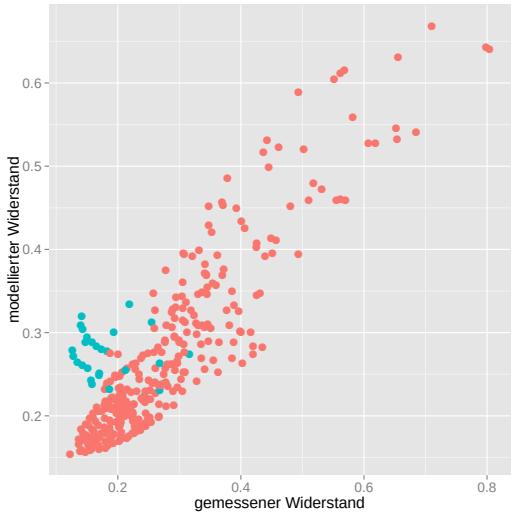


Abbildung 6.2: Regressionsfehler des linearen Modells. Der interessante Phasenbereich für $\theta = 43,4°$ ist farblich gekennzeichnet

die das Regressionsverfahren einen großen Fehler hatte, bei der zufälligen Auswahl aus einem Cluster oder Wertebereich eine höhere Wahrscheinlichkeit haben ausgewählt zu werden. Der Vorteil hierbei ist, dass der Vorhersagefehler nicht als einziges Kriterium herangezogen wird und auch sichergestellt wird, dass nicht nur Punkte aus lokal begrenzten Bereichen ausgewählt werden. Andererseits stellt die zufällige Auswahl sicher, dass Punkte mit einem geringen Vorhersagefehler nicht vollständig ausgeschlossen werden.

Die Gewichte werden berechnet, indem jeder Regressionsfehler e_i für einen Cluster c_j, durch die Summer aller Fehler $\sum_{e_i \in c_j} e_i$ geteilt wird. Damit gelten für die Gewichte die Bedingungen eines Wahrscheinlichkeitsmaßes, nämlich dass alle Wahrscheinlichkeiten $p_i \geq 0$ sind und $\sum p_i = 1$ gilt.

In Clustern, in denen die Fehler sehr ähnlich sind, ist die Auswahl somit annähernd gleichverteilt und in Clustern, in denen nur wenige Punkte einen großen Regressionsfehler haben, werden diese fast sicher ausgewählt.

Zufallsstichprobe Die einfachste Methode ist es einfach zufällig n_i Punkte aus dem vorhandenen Datensatz zu ziehen. Die Stichprobe sollte gleichverteilt sein, so dass jeder Punkt mit gleicher Wahrscheinlichkeit gezogen werden kann. Ziehen mit Zurücklegen ist nicht sinnvoll, da das erneute Messen eines XRD-Spektrums keine neuen Informationen liefert. Es fließt keinerlei vorhandene Information über das ternäre System in den Zufallsprozess ein.

Der Vorteil dieses Verfahrens ist seine einfache Umsetzung und dass keine vorherigen Messungen notwendig sind. Daraus folgt auch, dass Messfehler oder andere Störgrößen die Auswahl nicht beeinflussen.

Andererseits ist gerade die Hypothese, dass durch die Verwendung von zusätzlichen Informationen, die Auswahl positiv beeinflusst werden kann. Aus diesem Grund dient das Verfahren auch als untere Schranke für die Beurteilung der anderen Verfahren; diese sollten bessere Ergebnisse liefern.

6.2 Beurteilung der Clustering-Algorithmen

6.2.1 Datenvorbereitung

Gerade für Clustering-Algorithmen ist es wichtig, dass die Daten aufbereitet werden, um gute Ergebnisse zu erzielen. Das wichtigste Ziel ist es der betrachteten Materialeigenschaft eine größere Gewichtung zu geben als der chemischen Zusammensetzung, da diese die relevante Zielgröße ist. Da die Clustering-Verfahren die Messgrößen als Koordinaten im Euklidischen Raum betrachten, ist die unterschiedliche Skalierung hier der entscheidende Faktor. Denn während die chemischen Materialkonzentrationen zwischen 0 und 100 $at.\%$ liegen, ist der elektrische Widerstand in der vorliegenden Einheit von $\mu\Omega cm$ deutlich kleiner, mit Werte kleiner 1, während der gemessenen Widerstand wiederum in Ohm (Ω) gemessen wird und Werte deutlich über 100 annehmen kann.

Die Gewichtung wird durch eine geeignete Transformation der Daten erreicht. Als sinnvoll erwies sich die chemische Zusammensetzung auf den Bereich von 0 bis 1 zu normieren, wobei die Transformation für jedes Element einzeln durchgeführt wird. Es wäre auch denkbar, die tatsächlichen Prozentwerte zu nutzen; da jedoch die Verteilung der Konzentrationen für die vorliegenden Proben nicht gleichmäßig ist, könnte dies zu einer ungewollten Verzerrung des Einflusses der Komponenten führen. Anschaulich betrachtet, ohne zusätzliches Wissen vorauszusetzen, könnte die Variation eines Elementes um 10%ige genau so starken Einfluss haben wie eine 90%ige Variation für ein anderes Element. Insbesondere können solche Unterschiede dann auftreten, wenn der Suchraum des ternären Systems bereits eingeschränkt ist, wie zum Beispiel für die vorliegenden TiTaAl- oder TiTaRe-Datensätze.

Der Widerstand wird ebenfalls auf einen festen Bereich normiert, um die abweichenden Größenordnungen im Wertebereich auszugleichen. Da der Widerstand jedoch als besonders relevant betrachtet wird, könnte eine stärkere Gewichtung des Widerstandes sinnvoll sein, um hier Unterschiede stärker zu betonen als in der Materialzusammensetzung, weswegen der Effekt von verschiedenen Normierungen untersucht wird.

6.2.2 Gütekriterium für die Clustering-Algorithmen

Üblicherweise werden für die Beurteilung von Clustering-Verfahren die Ergebnisse auf einer bereits annotierten Testmenge verglichen. Für die vorliegenden Daten liegt eine solche jedoch nicht vor.

	k-means	DBSCAN
k	$\{4 \dots 20\}$	
ϵ		$\{0,01\,,\ 0,1\,,\ 0,2\,,\ 0,5\,,\ 1\,,\ 2\,,\ 5\,,\ 10\}$
minPts		$\{4 \dots 10\}$
Normierung Widerstand	$0 - \{1 \dots 10\}$	$0 - \{1 \dots 10\}$

Tabelle 6.1: Suchraum für die Parameter der Clustering-Algorithmen

Cluster-Varianz Ein sinnvolles Beurteilungskriterium ist die Varianz der relevanten Materialeigenschaften innerhalb eines Clusters. Das Ziel ist Bereiche mit durchaus unterschiedlicher Größe zu finden, in denen die Materialeigenschaft möglichst ähnlich ist. Dies lässt sich als folgende Optimierungsaufgabe formulieren

$$\min \frac{1}{k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \frac{1}{n} (x_{ij} - \bar{x}_i)^2 \,,$$

also der Minimierung der Varianz der n_i Elemente im i-ten Cluster für alle k Cluster.

Normierung der Cluster-Varianz Für die Berechnung der Cluster-Varianz kann ebenfalls eine Normierung verwendet werden. Angenommen es gibt drei Cluster in denen jeweils die niedrigen, mittleren und großen Widerstandswerte liegen so wird der absolute Fehler für den Cluster mit den kleinen Werten, bei gleichem relativen Fehler, kleiner ausfallen als für den Cluster mit den großen Werten. Um dies auszugleichen, kann vor der Varianzberechnung der Widerstand innerhalb jedes einzelnen Clusters noch einmal auf den Bereich 0 bis 1 normiert werden; der Effekt dieser Normierung wird im nächsten Abschnitt bewertet.

6.2.3 Auswertung

In Tabelle 6.2 sind die besten gefundenen Werte für die durchschnittliche Cluster-Varianz aufgeführt. Der untersuchte Suchraum wird in Tabelle 6.1 angegeben; für *k-means* ergeben sich daraus 160 und für DBSCAN 1280 mögliche Parameterkombinationen. Für jede Parameterkombination werden 20 Iterationen durchgeführt.

Ein Problem bei der Optimierung der Cluster-Varianz ist, dass eine stärkere Gewichtung des Widerstandes den absoluten Wert der Varianz erhöht. Die Varianz ist die Summe der quadrierten Abweichungen. Wenn nun die absoluten Werte der Abweichungen um einen konstanten Faktor vergrößert werden, so vergrößert sich die beobachtete Varianz um den quadrierten Faktor.

Durch die erneute Normierung innerhalb der Cluster auf Widerstandswerte zwischen 0 und 1 kann dieser Zusammenhang abgeschwächt werden. Die Einteilung der Cluster erfolgt somit unter der stärkeren Gewichtung des Widerstandes, während für die Berechnung der Cluster-Varianz die normierten Werte verwendet werden. Die Normierung der Varianz vergrößert aber die durchschnittliche Varianz innerhalb der Cluster. Dies liegt

Parameter	k-means	k-means normiert	DBSCAN	DBSCAN normiert
CuNiZn	0,003	0,008	0,011	0,008
NiCrRe	0,014	0,034	0,01	0,012
TiCoW	0,003	0,011	0,009	0,007
TiNiCu	0,003	0,014	0,005	0,012
TiTaAl	0,002	0,007	0,009	0,004
TiTaRe	0,003	0,007	0,041	0,008

Tabelle 6.2: Durchschnittliche *within*-Cluster-Varianz nach der Parameter-Optimierung der in Tabelle 6.1 aufgeführten Parameter

	k-means		**DBSCAN**			
	k	Gew. Wid.	ϵ	min points	Gew. Wid.	Anz. Cluster
CuNiZn	20	1	0,1	11	1	5
NiCrRe	19	1	5	4	1	10
TiCoW	18	1	0,1	19	1	4
TiNiCu	19	1	0,1	20	1	4
TiTaAl	20	1	0,1	16	1	5
TiTaRe	20	1	0,2	20	1	2

(a) Ohne Normierung der Cluster-Varianz

	k-means		**DBSCAN**			
	k	Gew. Wid.	ϵ	min points	Gew. Wid.	Anz. Cluster
CuNiZn	20	10	0,1	4	10	15
NiCrRe	6	1	10	8	9	2
TiCoW	17	10	0,1	5	10	13
TiNiCu	18	10	0,5	4	10	4
TiTaAl	20	10	0,1	4	10	22
TiTaRe	19	9	0,2	5	10	17

(b) Mit Normierung der Cluster-Varianz

Tabelle 6.3: Optimierte Parameterwerte

(a) DBSCAN mit normierter Cluster-Varianz für NiCrRe, keine sinnvolle Partitionierung gefunden

(b) DBSCAN ohne normierte Cluster-Varianz für TiCoW, sämtliche Unterschiede Bereich rechts-unten werden nicht widergegeben

Abbildung 6.3: Beispiele für schlechte Clustering-Resultate trotz Parameteroptimierung

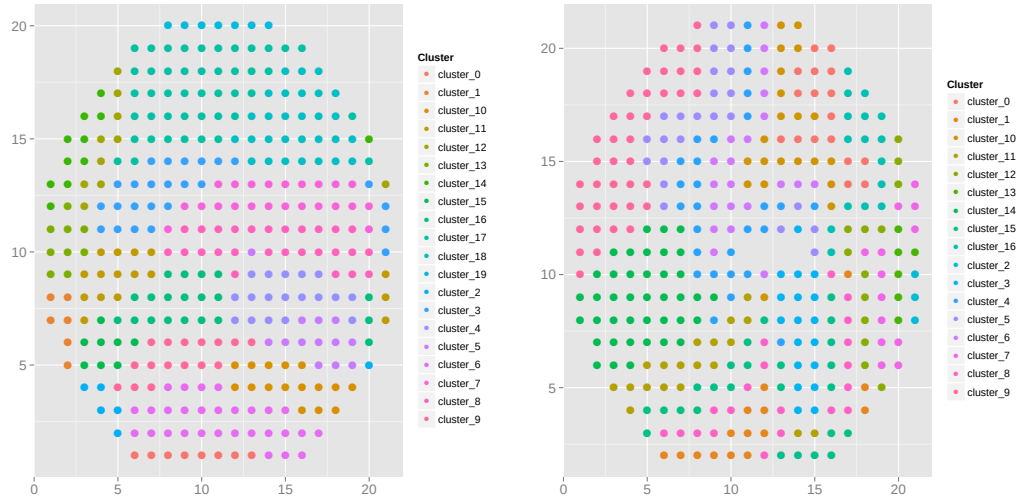
daran, dass zuvor die maximale Merkmalsausprägung für alle bis auf einen Cluster kleiner 1 war, nun aber für alle Cluster 1 ist. Der untersuchte Wertebereich und damit der absolute Fehler steigt also leicht, womit der direkte Vergleich der Varianzen in Tabelle 6.2 nicht möglich ist.

6.2.4 Beurteilung der Normierung

Betrachtet man die gefunden Partitionierung auf den Wafer-Koordinaten, so fallen einige Probleme auf. Der DBSCAN-Algorithmus liefert trotz des Bewertungskriteriums und der Parametersuche nicht immer gute Ergebnisse, sondern sehr oft degenerierte Lösungen. So zeigt die Abbildung 6.3 zwei Beispiele in denen große Bereiche mit unterschiedlichen Widerstandswerten nicht genauer unterteilt werden.

Der *k-means*-Algorithmus ist hingegen deutlich robuster, denn es kommt zu keinen degenerierten Lösungen. Die Wahl von k kann mit der Anzahl der initialen Messungen n_i gekoppelt werden. Bei DBSCAN müsste zunächst die Cluster-Anzahl und -Größe berücksichtigt werden um zu entscheiden, aus welchen Clustern Elemente betrachtet werden.

Für *k-means* scheint die Variante mit der normierten Cluster-Varianz die besseren Ergebnisse zu liefern, da sie etwas genauer die tatsächliche Struktur auf dem Wafer wiedergibt, wie Abbildung 6.4 zeigt.



(a) k-means mit normierter Cluster-Varianz für CuNiZn

(b) k-means mit normierter Cluster-Varianz für TiCoW

Abbildung 6.4: Beispiele für gute Clustering-Resultate nach Parameteroptimierung

6.3 Bewertung der initialen Stichprobe

6.3.1 Maximale DTW-Distanz

Ein erster Ansatz ist die maximale DTW-Distanz zwischen den XRD-Spektren der Stichprobe mit der aller Punkte zu vergleichen. Dieser Wert kann mit der maximalen DTW-Distanz des gesamten Datensatzes verglichen werden. Motiviert ist dies dadurch, dass zumindest überprüft werden kann, ob die beiden unähnlichsten Spektren in der Stichprobe enthalten sind. In diesem Fall sollte die Abweichung zwischen Stichprobe und vollständigen Daten möglichst gering sein.

Problematisch dabei ist jedoch, dass sehr viel Information über die Spektren verloren geht. Es wird lediglich ein Merkmal, nämlich die DTW-Distanz, von je zwei XRD-Spektren berücksichtigt, wie die Verteilung der übrigen Spektren ist bleibt außen vor. Dies zeigt sich spätestens dann, wenn im Vergleich zwischen der zufälligen Auswahl und der Auswahl per vorheriger Klasseneinteilung, die reine Zufallsauswahl gleiche und zum Teil bessere Ergebnisse liefert.

Aus diesem Grund scheint der Ansatz der maximalen DTW-Distanz für eine Beurteilung der Stichprobengüte nicht geeignet und wird im Folgenden nicht weiter untersucht.

6.3.2 Aggregation aller XRD-Spektren

Für die Charakterisierung der XRD-Spektren sind die Spitzenwerte in der Intensität am aussagekräftigsten, da sie auf bestimmte Materialphasen hindeuten. Somit ist ein mögliches Bewertungskriterium, ob in einer Stichprobe die maximalen Werte in ähnlicherweise repräsentiert sind wie in der gesamten Messung. Als einfache Aggregation hierfür kann die Maximums-Norm auf jeden einzelnen Messwinkel angewendet werden. Zwar wird bei dieser Methode die Verteilung der XRD-Spektren nicht berücksichtigt, also welche Form von Muster besonders oft vorkommt, aber dies spielt für die Entdeckung von interessanten Materialphasen eine untergeordnete Rolle, da es wichtiger ist möglichst alle Ausprägungen zu finden.

Bereits bei der Beschreibung der XRD-Spektren in Kapitel 2 wurden die maximalen Werte für alle Messpunkte und alle Winkel zu einem einzigen Spektrum zusammengefasst, um charakteristische Merkmale der untersuchten Systeme zu zeigen. Diese Methodik kann auch eingesetzt werden, um einen Vergleich zwischen der Stichprobe und den gesamten Wafer-Daten zu ermöglichen. Dafür werden die XRD-Spektren der Stichprobe auf die gleiche Weise aggregiert wie die für den gesamten Datensatz. Die daraus resultierenden zwei Spektren können nun wiederum mittels der DTW-Distanz verglichen werden.

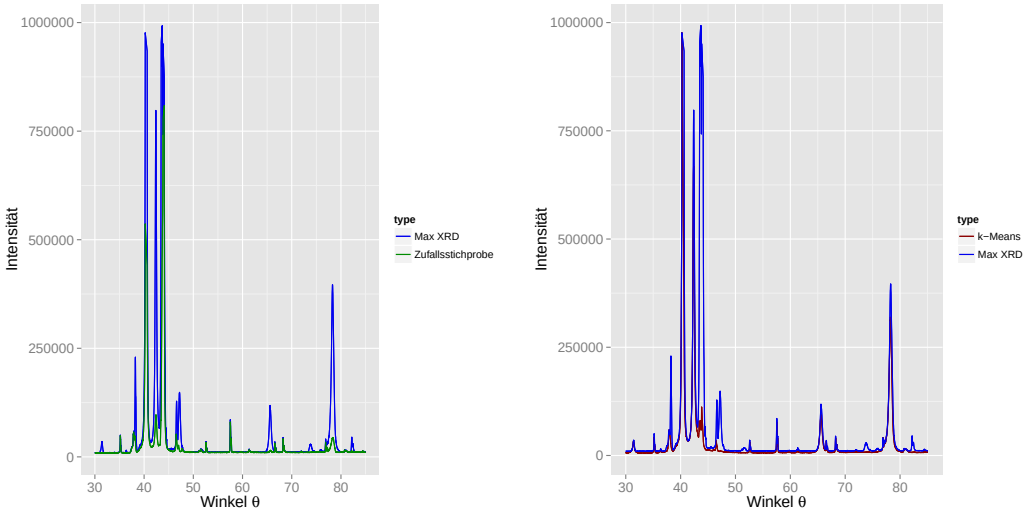
Ist die DTW-Distanz gering, so ist dies ein Indiz dafür, dass in den XRD-Spektren der Stichprobe die gleichen charakteristischen Spitzenwerte vorkommen wie im gesamten Datensatz. Fehlt ein Spitzenwert oder hat nur eine sehr geringe Intensität, so ist die DTW-Distanz entsprechend größer und die Qualität der Stichprobe geringer. Anschaulich wird dieser Unterschied in Abbildung 6.5 gezeigt.

Die Ergebnisse für die einzelnen Verfahren sind in der Abbildung 6.6 zu sehen. Da die Größenordnung der DTW-Distanzen für die unterschiedlichen Materialmischungen verschieden ist, wurden die Distanzen jeweils auf Werte zwischen 0 und 1 normiert. Man kann deutlich sehen, dass lediglich das gewichtete *k-means*-Verfahren bei steigender Stichprobengröße eine deutliche Verkleinerung der Differenzen zeigt. Das deutet darauf hin, dass diese Methode tatsächlich in der Lage ist eine kleine Menge von Messpunkten zu finden, welche in ihren maximalen Intensitäten den gesamten Wafer-Daten sehr ähnlich ist.

6.4 Zwischenergebnis und Auswahl der weiter verwendeten Methoden

Der Vergleich der aggregierten XRD-Spektren ist eine sinnvolle Methode um die Güte der Stichproben zu beurteilen, auch wenn der Nachteil besteht, dass für den Vergleich die Spektren aller Messpunkte bekannt sein müssen. Für die Beurteilung der initialen Stichprobe liefert diese Methode dennoch sinnvolle Erkenntnisse.

Die Einteilung mittels *k-means*-Clustering liefert zusammen mit der Gewichtung durch den Fehler der linearen Regression die besten Ergebnisse hinsichtlich der Ähnlichkeit der



(a) Zufällige Stichprobe, die DTW-Distanz zwischen den Spektren beträgt 28689

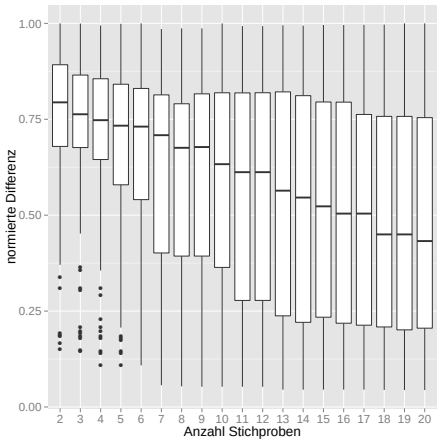
(b) Stichprobe ermittelt mit gewichtetem *k-means*, die DTW-Distanz zwischen den Spektren beträgt 19186

Abbildung 6.5: Vergleich der aggregierten XRD-Spektren zweier Stichproben der Größe 20 mit dem aggregierten Spektrum des gesamten Wafers für TiCoW

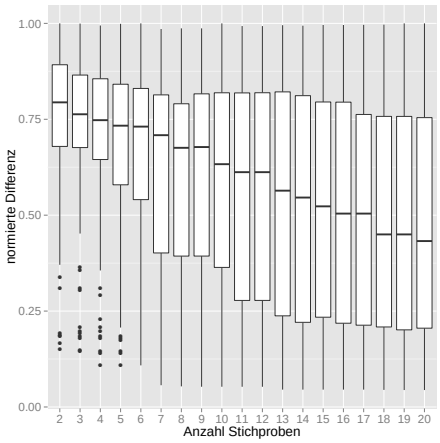
aggregierten XRD-Spektren. Aus diesem Grund ist dieses Verfahren ein guter Kandidat für die weiteren Untersuchungen. Die Ergebnisse der reinen Zufallsstichprobe waren vergleichsweise gut und diese Methode sollte für alle weiteren Untersuchungen als Referenz dienen, denn es ist wünschenswert, dass eine gerichtete Auswahl signifikant bessere Ergebnisse liefert als eine reine Zufallsauswahl.

Der DBSCAN-Algorithmus hat sich als nicht robust genug erwiesen, um in einem automatisierten Prozess verwendet zu werden. Die Ergebnisse für die einfache Einteilung des Widerstandes in äquidistante Klassen waren auch deutlich schlechter, als die des *k-means*-Algorithmus.

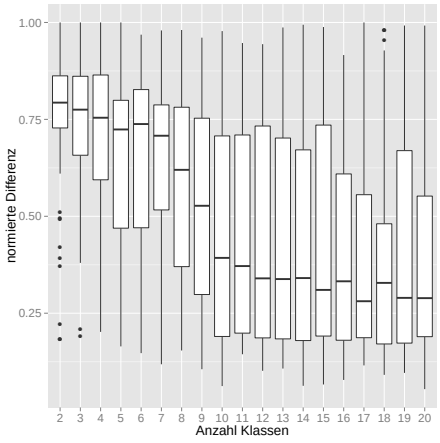
Aus diesem Grund werden in den folgenden Kapiteln nur noch die reine Zufallsstichprobe und die Kombination aus *k-means*-Clustering und Gewichtung durch das lineare Modell als Verfahren für die initiale Stichprobe verglichen.



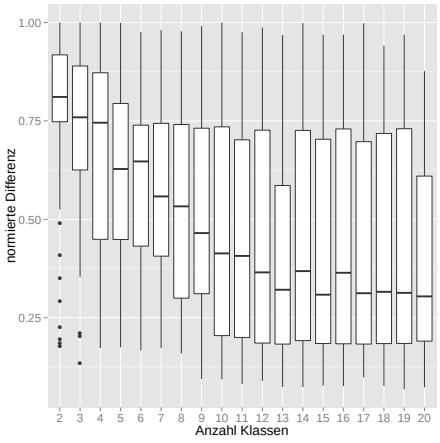
(a) Zufällige Auswahl



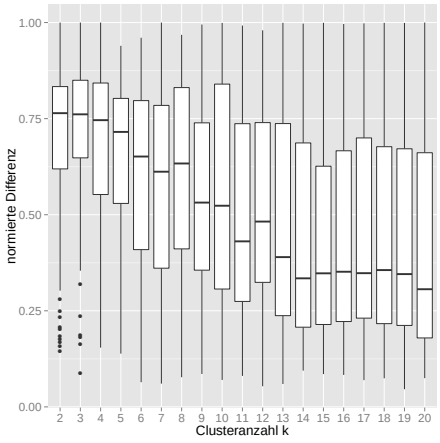
(b) Zufällige Auswahl, gewichtet mit Regressionsfehler



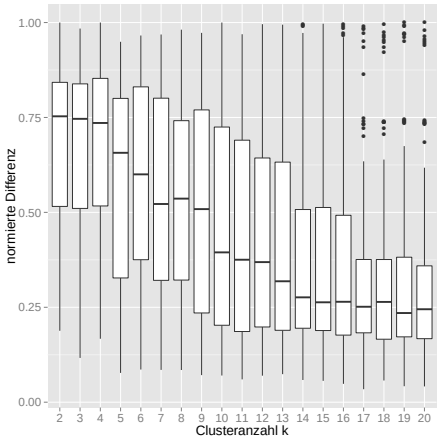
(c) Einteilung in Klassen



(d) Einteilung in Klassen, gewichtet mit Regressionsfehler



(e) k-means-Clustering



(f) k-Means-Clustering, gewichtet mit Regressionsfehler

Abbildung 6.6: Boxplots über die Entwicklung der DTW-Distanzen zwischen den aggregierten XRD-Spektren der Stichproben und der vollständigen Wafer

7 Auswertung der XRD-Spektren mittels Approximation der DTW-Distanzen

7.1 Motivation

Nachdem für eine erste Stichprobe von Messpunkten die XRD-Spektren bestimmt wurden ist nun das Ziel die darin enthaltenen Informationen zu nutzen,m um daraus Erkenntnisse über die Verteilung der Spektren für den gesamten Messraum abzuleiten. Diese Erkenntnisse sollen genutzt werden, um weitere Messpunkte zu bestimmen, deren Auswertung letztendlich genügend Daten liefern soll, um für alle verbliebenen Punkte des Messraumes verlässliche Aussagen über deren XRD-Spektren zu treffen.

In der Modellsprache des *active learning* entspricht der Vorgang der XRD-Analyse der Orakel-Funktion. Die erste Stichprobe dient dazu eine repräsentative Sammlung von Labels - hier XRD-Spektren - zu finden, mit welcher der Prozess startet, um für möglichst sinnvolle Punkte die Orakel-Funktion zu befragen. Die schwierige Aufgabe besteht nun darin zu ermitteln in welcher Form die XRD-Spektren bei der Auswahl der Messpunkte helfen können.

Die Schwierigkeit beruht darauf, dass die XRD-Spektren keine eindimensionalen Merkmale sind. Wäre dies der Fall, so könnte gesagt werden das gemessene XRD-Spektrum gehört zu einer bestimmten Materialphase α oder die eine gemessene Intensität hat den Wert x.

Stattdessen haben die XRD-Spektren eine hochdimensionale Struktur, deren eindeutige Zuordnung schwierig ist. Nichtsdestotrotz bleibt die Motivation des *active learning* bestehen nur solche Punkte zu messen, die besonders informativ sind. Eine Methode, um den Informationsgehalt für die XRD-Spektren zu quantifizieren, wird in diesem Kapitel untersucht.

7.2 Aus den XRD-Spektren ableitbare Informationen

Intensität der Spektren Bei der Beschreibung der vorhandenen Datensätze fällt auf, dass die Intensitätsniveaus der einzelnen ternären Systeme starke Unterschiede aufweisen können.

Deshalb sind die gemessenen Intensitätswerte von Interesse, da sie dabei helfen können mögliche Spitzenwerte zu identifizieren und generell einen Überblick über die Verteilung der Messwerte zu bekommen.

Aus diesem Grund ist es grundsätzlich nützlich einfache Kennwerte über die XRD-Spektren zu sammeln. Diese können sowohl pro Messpunkt, als auch pro Messwinkel oder Winkelbereich betrachtet werden. Interessant sind dabei vor allem maximale Werte, um Ausreißer oder Spitzenwerte zu identifizieren, Durchschnittswerte und Varianzen.

DTW-Distanz zwischen den Messwerten Ein sinnvolles Maß für die Ähnlichkeit zwischen zwei XRD-Spektren stellt die DTW-Distanz dar. Große DTW-Distanzen deuten auf veränderte Materialstrukturen mit anderen Phasenausprägungen hin und eine genauere Untersuchung der dazwischen liegenden Punkte könnte lohnenswert sein. Im Gegensatz dazu kann eine kleine Distanz auf einen statischen Bereich mit nur wenigen Änderungen hindeuten, in dem weitere Messungen nur wenig neue Erkenntnisse bringen, wobei bei diesen Bedingungen auch die räumliche Distanz auf dem Wafer, bzw. die Unterschiede in der chemischen Zusammensetzung zu berücksichtigen sind.

Erschwert wird diese Betrachtungsweise dadurch, dass die DTW-Distanz nur ein Maß zwischen genau zwei Messpunkten ist und daraus keine direkten Aussagen über benachbarte Punkte getroffen werden können. Bei n Messpunkten gibt es somit $n * (n - 1)$ verschiedene Kombinationen, wobei die DTW-Distanz ein symmetrisches Maß ist.

7.3 Schätzung der DTW-Distanzen mittels Kegelauswahl

Mittels der Kennwerte der XRD-Spektren verlässliche Vorhersagen oder Heuristiken zu erstellen ist schwierig, vor allem da die untersuchten Spektren sich je nach Materialprobe stark unterscheiden können. Deswegen wird nur die DTW-Distanz als Informationsquelle untersucht.

Nach der initialen Stichprobe sind für n_i Punkte des Wafers die XRD-Spektren bekannt. Die relevante Frage ist nun wie die bekannten XRD-Spektren genutzt werden können, um die nächsten Punkte zu bestimmen, für die eine Messung der XRD-Spektren sinnvoll scheint. Wie in Kapitel 3 beschrieben ist der gängige Ansatz Modelle zu erstellen, welche für die noch nicht ausgewerteten Punkte deren Klassenzugehörigkeit oder Funktionswert berechnen und dann Punkte auszuwählen, die das Modell am stärksten verbessern würden oder für welche die größte Unsicherheit besteht.

Genau diese Zuweisungsfunktion fehlt aber noch und ist wahrscheinlich auch schwierig zu finden, insbesondere wenn sie für verschiedene ternäre Systeme gelten soll. Deshalb wird eine alternative Herangehensweisen erprobt.

Eine Möglichkeit ist es aus den bekannten DTW-Distanzen die Ähnlichkeit zwischen bekannten und noch nicht gemessenen Spektren zu schätzen. Wie in der Studienarbeit gezeigt haben dicht zusammen liegende Punkte auch häufig eine geringe DTW-Distanz zueinander; die XRD-Spektren wechseln also meistens nicht spontan ihre charakteris-

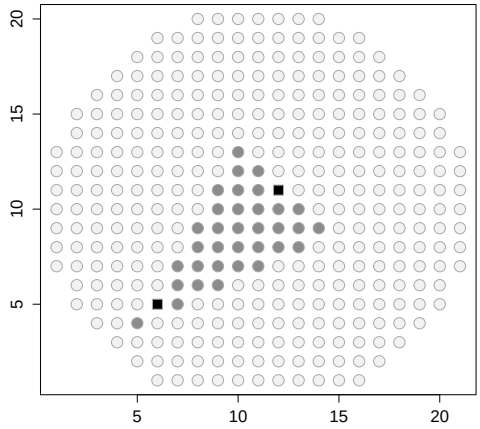


Abbildung 7.1: Überdeckter Bereich zwischen Start- und Endpunkt (schwarz)

tischen Ausprägungen. Unter dieser Annahme kann zunächst einmal die DTW-Distanz für Punkte im Umfeld von zwei gemessenen Punkten geschätzt werden.

Das Umfeld der Messpunkte wird modelliert indem zwischen beiden bekannten Punkten jeweils ein kegelförmiger Bereich des Wafers ausgewählt wird, auf dessen Abschlusskante der andere Messpunkt liegt. Diese Auswahl spiegelt grob den Informationsverlauf der DTW-Distanz wieder. In der Nähe des Endpunktes liegen mehr Punkte in der Kegelauswahl und für diese Punkte ist wahrscheinlich auch eine bessere Abschätzung der DTW-Distanz möglich. Durch die Symmetrie der DTW-Distanz entstehen so um die beiden ausgewählten Punkte Bereiche, für die DTW-Distanz wahrscheinlich ähnlich zu der bekannten Messung ist.

Abbildung 7.1 zeigt wie ein solcher Kegel zwischen zwei Punkten auf dem Wafer verläuft. Die Kegelfläche wird durch die Veränderung der Steigung der Verbindungsgeraden zwischen den beiden Messpunkten berechnet. Dadurch überdecken Kegel zwischen zwei weit entfernen Messpunkten einen breiteren Bereich als nah beieinander liegende Punkte. Es ist zu beachten, dass auf dem diskreten Raster die Berechnung der Kegel zwangsläufig zu Rundungsfehlern führt, weswegen in dem obigen Beispiel auch ein Punkt jenseits des Startpunktes noch mit in der Auswahl liegt.

Die DTW-Distanz für jeden Punkt im Auswahlgebiet kann auf verschiedene Arten geschätzt werden. Die bekannte DTW-Distanz wird als Referenzgröße herangezogen und diese wird anhand der Ähnlichkeit des neuen Punktes mit dem jeweiligen Endpunkt gewichtet. Die Idee ist, dass Punkte in dem Kegel, die nah am Startpunkt liegen, wahrscheinlich auch eine kleine DTW-Distanz haben. Punkte die nah am Endpunkt des Kegels liegen haben eine ähnliche DTW-Distanz wie dieser. Die Nähe kann einfach durch die Manhattan-Distanz auf dem Wafer ausgedrückt werden oder über den Euklidischen Abstand der chemischen Zusammensetzung.

Approximationsmethoden zum Schätzen der DTW-Distanzen Die folgenden drei Methoden werden für die Schätzung der DTW-Distanz $\widehat{DTW}_k$ eines Punktes k genutzt

- Euklidischer Abstand: $\widehat{DTW}_k = \frac{DTW - Distanz}{||x_s - x_z||_2} * ||x_s x_k||_2$
- normierter Euklidischer Abstand: $\widehat{DTW}_k = \frac{DTW - Distanz}{||x_s - x_z||_2} * ||x_s x_k||_2$, die chemische Zusammensetzung ist auf $(0 - 1]$ normiert, indem durch das jeweilige Maximum dividiert wird
- Manhattan-Distanz: $\widehat{DTW}_k = \frac{DTW - Distanz}{||x_s - x_z||_1} * ||x_s x_k||_1$, die 1-Norm bezieht sich auf die X/Y-Koordinaten des Wafers,

wobei x_s und x_z Start- und Endpunkt des Kegels bezeichnen. Der Euklidische Abstand bezieht sich auf die Distanz zwischen x_s und x_z im ternären System, während sich die Manhattan-Distanz auf die X/Y-Koordinaten der Punkte auf dem Wafer bezieht. Die Normierung des Euklidischen Abstandes soll, wie bei der initialen Stichprobe in Kapitel 6, Unterschiede in der Verteilung der chemischen Elemente ausgleichen.

7.4 Auswertung der Approximation der DTW-Distanz

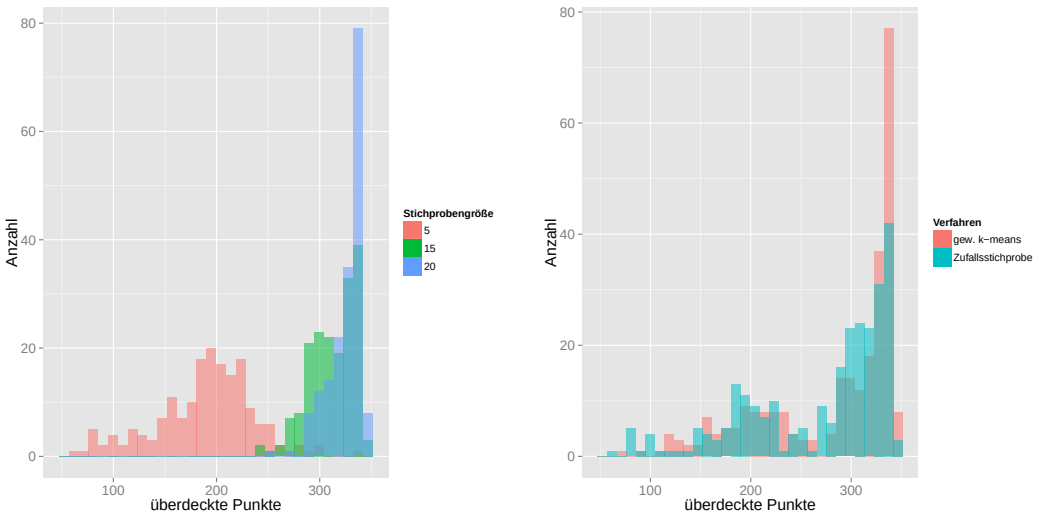
Zunächst gilt es zu klären wie groß der Fehler der geschätzten DTW-Distanzen für die von den Kegeln überdeckten Punkte ist.

Für die Größe wird $n_i = 5, 15, 20$ getestet und es werden je 15 Stichproben mit dem gewichteten *k-means*-Algorithmus und der reinen Zufallsauswahl als Kontrollgruppe berechnet. Es ergeben sich also für die sechs Materialproben insgesamt 540 untersuchte Stichproben.

Überdeckte Punkte Im besten Fall liegen alle 342 Punkte unter mindestens einem Kegel; im schlechtesten Fall sind es lediglich 57 Punkte. Der Durchschnitt ist bei $275,5$ überdeckten Punkten, wobei die Verteilung deutlich rechtssteil ist. Wie Abbildung 7.2 zeigt ist die Überdeckungsrate mit mehr Kegeln auch deutlich größer. Die Unterteilung nach Selektionsstrategie zeigt keine großen Unterschiede, wobei die Auswahl mit *k-means* leicht bessere Ergebnisse liefert.

Relativer Fehler der approximierten DTW-Distanzen Um ein vergleichbares Maß für die Qualität der zugewiesenen DTW-Distanzen für die nicht gemessenen XRD-Spektren zu haben, ist es notwendig das Fehlermaß auf eine gemeinsame Bezugsgröße zu normieren, da die absoluten Werte der DTW-Distanzen zwischen einzelnen Punkten deutliche Unterschiede aufweisen können. Der relative Fehler zwischen einer Beobachtung $\tilde{x}$ und dem wahren Wert x berechnet sich durch

$$|\frac{\tilde{x} - x}{x}|$$



(a) Unterteilung nach Stichprobengröße

(b) Unterteilung nach Stichprobenverfahren

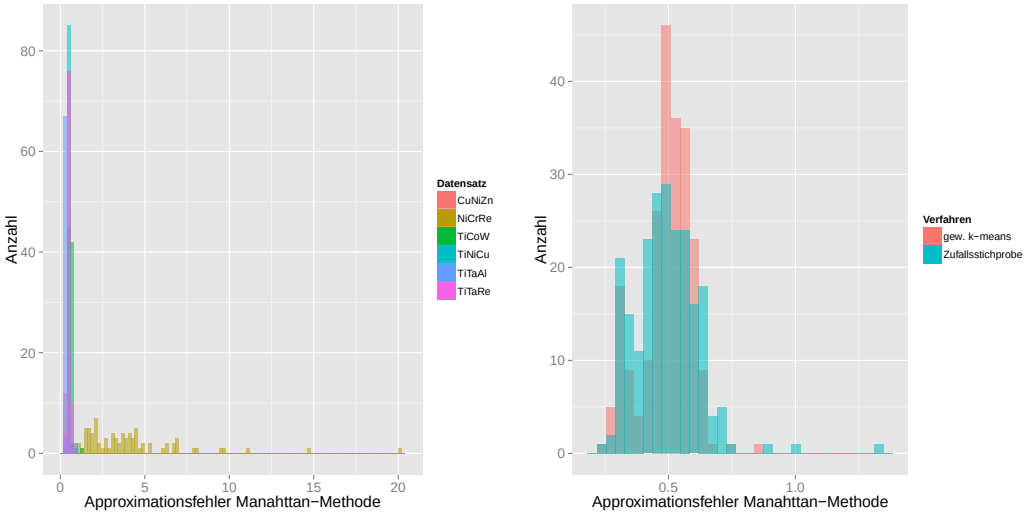
Abbildung 7.2: Verteilung der durch die Kegel der initialen Stichprobe überdeckten Punkte

Für die vorliegenden Datensätze ist es möglich diesen Fehler zu berechnen, da die korrekte DTW-Distanz zwischen zwei Messpunkten bekannt ist. In einem realen Szenario ist dies nicht möglich, da die XRD-Spektren nicht vollständig vorliegen.

Für alle drei Approximationsmethoden liegt der durchschnittliche relative Fehler etwa bei 1. Allerdings zeigt das Histogramm in Abbildung 7.3a, dass der Großteil der Fehlerwerte unter 1 liegt und es einige deutliche Ausreißer mit Fehlerwerten über 10 gibt. Es zeigt sich, dass alle Beispiele mit Fehlerwerten größer 2 aus dem NiCrRe-System stammen. Wird dieses aus der Berechnungsgrundlage entfernt, so verringert sich der durchschnittliche Fehler auf etwa 0,5 für alle drei Methoden.

Aufgrund dieser sehr starken Abweichung wird für die folgenden Betrachtungen das NiCrRe-System entfernt. Als weitere Einschränkung wird nur die Manhattan-Methode näher betrachtet, da sie im Durchschnitt die besten Ergebnisse liefert, zumindest ohne die Verzerrung durch die extremen Ausreißer der NiCrRe-Daten.

Die Auswahlstrategie hat keinen eindeutig positiven Einfluss auf den Fehler der DTW-Distanzen. Das arithmetische Mittel und der Median für die Zufallsauswahl sind 0,493 und 0,492 und für die Auswahl mit gewichtetem *k-means* sind die Werte 0,493 und 0,502. Die Verteilung der Fehler (siehe Abbildung 7.3b) zeigt aber Unterschiede, denn die Fehlerwerte der Zufallsauswahl streuen deutlich stärker, sind jedoch auch kleiner, während sich die Fehlerwerte der *k-means*-Auswahl deutlich um den Durchschnittswert zentrieren, wobei beide Verteilungen bimodal sind. Eine mögliche Erklärung für diese Beobachtung könnte darin begründet sein, dass bei der reinen Zufallsauswahl die Manhattan-Distanz



(a) Aufgeteilt nach Datensatz

(b) Ohne NiCrRe, aufgeteilt nach Stichprobenverfahren

Abbildung 7.3: Verteilung der relativen Fehler für die Approximation der DTW-Distanz mit Manhattan-Approximation

zwischen den ausgewählten Startpunkten im Schnitt etwas geringer ist, nämlich 12,2 im Vergleich zu 12,9.

Aus den Ergebnissen folgt, dass die weiteren Untersuchungen auf die DTW-Approximation mittels Manhattan-Distanz und einer Stichprobengröße von 20 eingeschränkt werden.

7.5 Informationsextraktion aus der Kegelüberdeckung

Die Approximation der DTW-Distanzen mittels der Kegelauswahl ergibt zwei Informationsquellen: die DTW-Distanzen und die Überdeckung des Wafers durch die Kegel. Diese Informationen sollen, wie in den Schritten zuvor, genutzt werden, um Gebiete zu erkennen, in denen weitere Messungen am sinnvollsten erscheinen. Sinnvoll in diesem Zusammenhang sind solche Punkte, für die entweder noch wenige oder keine Informationen vorliegen, oder solche für die die bisherigen Annahmen eine besonders große Unsicherheit aufweisen.

Maximale DTW-Distanz Punkte mit einer großen DTW-Distanz zueinander haben sehr unterschiedliche XRD-Spektren. Das heißt in der Folgerung, dass ihre Materialphasen sich voneinander unterscheiden, woraus weiter folgt, dass von Eigenschaften des einen Materials nicht ohne weiteres auf Eigenschaften des anderen Materials geschlossen werden kann. Es ist also zu vermuten, dass je größer die DTW-Distanz ist, desto ungenauer ist Vorhersage für diesen Punkt.

Die Distanzen eines einzelnen Kegels sind zunächst jedoch kein guter Indikator, da im vorherigen Schritt die initiale Stichprobe darum bemüht war möglichst unterschiedliche Punkte auszuwählen. Interessanter ist somit die mittlere DTW-Distanz zu allen bekannten Punkten.

Betrachtet wird die Summe aller approximierten DTW-Distanzen pro Punkt. Da allerdings nicht jeder Punkt von gleich vielen Kegeln überdeckt wird, kann diese Summe verzerrt sein. Zum Beispiel könnte ein mittig auf dem Wafer liegender Punkt zwar kleine Distanzen aufweisen, wird aber wahrscheinlich von sehr vielen Kegeln überdeckt werden und dadurch in der Summe eine größere Distanz haben als ein Punkt am Rand mit zwar einer im Verhältnis sehr viel größeren DTW-Distanz, der aber nur von einem Kegel überdeckt wird.

Aus diesem Grund erscheint eine Normierung der Summe durch die Anzahl der überdeckenden Kegel sinnvoll, wodurch sich folgende Formel ergibt:

$$\widetilde{dtw_i} = \frac{\sum_{j=1}^{n} dtw_{ij}}{\#\ddot{u}berdeckte\,Kegel}\,.$$

Anzahl überdeckender Kegel Das obige Auswahlkriterium folgt dem Paradigma möglichst widersprüchliche Punkte auszuwählen, um sie zu bestimmen. Ein anderer vorgestellter Ansatz war jene Punkte auszuwählen, für die noch keine oder möglichst wenige Informationen vorliegen.

Eine solche Auswahl erfolgt über die Anzahl der überdeckenden Kegel, jedoch ohne Berücksichtigung der approximierten DTW-Distanzen.

Bevorzugt sollen jene Punkte ausgewählt werden, welche nur von einer bestimmten Anzahl von Kegeln überdeckt werden. Untersucht werden zwei Varianten und zwar die Punkte, die unter einem Kegel und unter den meisten Kegeln liegen.

Für die nur unter einem Kegel liegenden kann zunächst angenommen werden, dass man über sie nur minimale Informationen hat und ihre Analyse wahrscheinlich besonders viele neue Informationen liefert. Ein Nachteil dieser Argumentation ist, dass diese Punkte in der Regel nahe am Anfang oder Ende eines Kegels liegen, also wahrscheinlich ähnlich zu bereits bekannten Messpunkten sind.

Es ist zu vermuten, dass die Punkte, die von den meisten Kegeln überdeckt werden, deutlich weniger informativ sind als jene, die nur von wenigen Punkten überdeckt werden; zur Kontrolle werden aber auch diese untersucht.

Auswahl der nächsten Messpunkte Um den Einfluss des Informationsgewinn durch die Kegelapproximation zu ermitteln, werden neue Stichproben nach den oben beschriebenen Verfahren gezogen. Die Größe wird dabei auf $n_a \in \{25, 30, 35, 40, 45, 50\}$ erhöht. Zum Vergleich wird eine reine Zufallsstichprobe des gesamten Wafers, sowie eine Stichprobe nach dem gewichteten *k-means*-Verfahren der gleichen Größe herangezogen.

Die Hoffnung ist, dass die zusätzliche Information hilft zuverlässiger weitere interessante XRD-Spektren zu identifizieren. Da das Gesamtziel außerdem eine Zuordnung der unbekannten XRD-Spektren zu dem jeweils ähnlichsten bekannten Spektrum ist, kann diese Auswahl auch helfen die Übergangsbereiche zwischen charakteristischen Mustern besser zu erfassen. Eine Konsequenz daraus könnte sein, dass der Vergleich der maximalen Intensitäten keine besseren Ergebnisse liefert, da das Verfahren der Kegelauswahl kompaktere Bereiche des Wafers abdeckt als die anderen Stichprobenverfahren.

Die Auswahl der Kegelpunkte erfolgt nach den zuvor beschriebenen Verfahren:

- Punkte die von einem Kegel überdeckt werden,
- Punkte die unter den meisten Kegeln liegen ,
- Punkte, für die die summierte DTW-Distanz geteilt durch die Anzahl der überdeckenden Kegel maximal ist.

Natürlich kann nicht gewährleistet werden, dass die Anzahl der Punkte, die das untersuchte Kriterium erfüllen groß genug ist, um eine Stichprobe der gewünschten Zielgröße zu ermöglichen. Ist dies nicht der Fall, so wird die Stichprobe um weitere Punkte aus der nächsten Kategorie ergänzt. Gibt es also nur m Punkte die unter einem Kegel liegen mit $m < n_a$, so werden $n_a - m$ Punkte die unter zwei Kegeln liegen hinzugezogen. Interessant ist, dass es nur einen Fall gibt, bei dem die Gesamtzahl der von Kegeln überdeckten Punkte kleiner als 50 ist. In so einem Fall muss die Auswahl mit weiteren zufälligen Punkten ergänzt werden.

Bei dieser Auswahlstrategie ist zu erwarten, dass Punkte ausgewählt werden, die in dem Selektionsschritt zuvor bereits ausgewählt wurden und deren XRD-Spektren bereits bekannt sind. In der Praxis würden diese Punkte natürlich nicht noch einmal gemessen werden. Um einen Vergleich mit anderen Auswahlverfahren zu erlauben, ist zunächst jedoch eine gleich große Stichprobe sinnvoll. Für die späteren Verfahren wird dies berücksichtigt und es werden nur Punkte ausgewählt, deren XRD-Spektren noch nicht bekannt sind, so dass nach diesem Schritt insgesamt $n_i + n_a$ XRD-Spektren gemessen wurden.

Verteilung der ausgewählten Messpunkte je nach Auswahlstrategie Die Frage, ob es bestimmte charakteristische Häufungen der ausgewählten nächsten Messpunkte gibt, lässt sich mit ja beantworten, wobei sich einige grobe Richtlinien ableiten lassen. Zwar gelten für die einzelnen Materialien andere Schwerpunkte, aber bei einer Stichprobengröße von 50 und 15 Wiederholungen je Verfahren, liegen genügend verschiedene Proben vor, um erste Verallgemeinerungen zu treffen, beziehungsweise auf eine feinschichtige Unterteilung zu verzichten und alle Auswahlmengen für ein Verfahren zusammenzufassen.

Werden die Punkte ausgewählt, die von keinem Kegel bedeckt werden, so ergibt sich eine starke Häufung von Messpunkten an den Randbereichen der Wafer. Dies ist nicht verwunderlich, da hier die wenigsten Überschneidungen zwischen den initialen Messpunkten liegen. Im Allgemeinen ist dies kein wünschenswertes Verhalten, da die Wahrscheinlich-

keit von Material- oder Messfehlern im Randbereich deutlich höher ist. Können diese aber, zum Beispiel durch eine vorherige Untersuchung des Wafers, ausgeschlossen werden, so kann unter Umständen ein Schwerpunkt bei den Randbereichen auch relevante Ergebnisse liefern, so wie das vorliegende NiCrRe-System gerade am Rand einige charakteristische Bereiche aufweist. Im Gegensatz dazu ist für die Auswahl nach der größten Anzahl von Überdeckungen eine starke Häufung im zentralen Bereich des Wafers zu sehen.

Zum Vergleich liefert die reine Zufallsstichprobe eine gleichmäßige Überdeckung der ausgewählten Messpunkte und auch beim gewichteten *k-means*-Verfahren fallen zwar noch einige Häufungspunkte auf, diese sind aber auch nicht sehr ausgeprägt, was wahrscheinlich an der großen Cluster-Anzahl k liegt.

Beuerteilung und weiteres Vorgehen Es lassen sich einige Eindrücke festhalten, auch wenn diese eher subjektiver Natur sind und nur durch Beispiele belegt werden können. Diese Beobachtungen ergeben folgende Schlüsse:

- Die Abschätzung der DTW-Distanzen für Punkte innerhalb eines Kegels ist mit der einfachen Approximation mittels der Manhattan-Distanz am besten. Dies ist ein unerwartetes Ergebnis, da die Materialkonzentration eigentlich mehr Informationen über das Materialverhalten liefert. Andererseits ist es aber auch gut möglich, dass solche linearen Zusammenhangsannahmen bei einem eindeutig nicht linearen System zu viele falsche Annahmen treffen und somit die einfachere These zuverlässigere Ergebnisse liefern kann.
- Eine größere Menge an initialen Messpunkten n_i und mehr ausgewählte Kegelpunkte n_a führen zu kleineren Approximationsfehler. Von daher werden für die folgenden Untersuchungen nur noch die beiden größten Mengen betrachtet. Die minimale notwendige Menge von Messungen für möglichst gute Ergebnisse zu bestimmen ist ein eigenständiges Pareto-Optimierungsproblem, welches mit den bisherigen Erkenntnissen noch nicht zu lösen ist. Im Gegenzug erhöht sich die Übersichtlichkeit der weiteren Untersuchungen enorm, wenn die Anzahl der Parameterkombinationen verringert wird.
- Die Verteilung der ausgewählten Kegelpunkte ist für keine der Methoden optimal. Eine niedrige Anzahl der Überdeckungen führt zu einem Übergewicht von Randpunkten und bei der maximalen Überdeckung liegen alle ausgewählten neuen Messpunkte im Zentrum, ein Ergebnis, welches zum einen auch mit deutlich weniger Aufwand erreichbar ist und zum anderen auch nur wenig Informationen aus den ersten XRD-Messungen ableitet.
- Auch die maximale DTW-Distanz führt zu einer sehr starken Konzentration am Randbereich, was vermutlich wiederum daran liegt, dass zwischen zwei Punkten, die an gegenüberliegenden Rändern des Wafers liegen, die DTW-Distanz erwartungsgemäß groß ist. Für weit entfernt liegende Punkte werden aber auch sehr große Kegel aufgespannt, so dass es viele Überschneidungen von Kegeln mit großer approximierter DTW-Distanz gibt. Die Verlässlichkeit solch großer Abschätzungen

ist wahrscheinlich eher gering. Als Konsequenz daraus wird das Kegelmodell noch einmal verbessert, wie im folgenden Abschnitt beschrieben wird.

7.6 Verbesserung der Kegelapproximation durch Beschränkung des maximalen Abstands

Eine Konsequenz aus den obigen Erfahrungen ist es den maximalen Abstand von zwei Punkten, zwischen denen ein Kegel berechnet werden soll, einzuschränken. Dies hat zur Folge, dass lokale Annahmen über die Verteilung und Entwicklung der DTW-Distanzen zwischen zwei Punkten stärker berücksichtigt werden und weniger ungenaue Annahmen getroffen werden müssen.

Bei der ersten Approximation der DTW-Distanzen und der ersten Auswahl von Messpunkten wird eine maximale Manhattan-Distanz von 14 gewählt; dies entspricht der Hälfte der maximalen Distanz zwischen zwei Wafer-Punkten für das vorliegende Raster von 342 Messpunkten. Da die initiale Stichprobenmenge mit 20 Punkten noch recht klein ist, erscheint eine zu große Einschränkung nicht sinnvoll, da sonst zu wenig Vergleichspunkte betrachtet werden. Die durchschnittliche Manhattan-Distanz zwischen zwei Stichprobenpunkten ist bei allen 90 Experimenten $12,86$ und etwa 61% aller betrachteten Paare von Messpunkten haben eine Manhattan-Distanz kleiner oder gleich 14. Somit bleibt gleichermaßen eine ausreichend große Auswahlmenge erhalten und wahrscheinlich können besonders uninformative Paare herausgefiltert werden.

7.7 Iterative Verbesserung des Modells

Ein wichtiger Bestandteil beim *active learning* ist dynamisch neue Informationen zu verarbeiten, um zu bestimmen, welche Elemente von der Orakel-Funktion ausgewertet werden. Bei der Auswahl von Messpunkten mittels der Kegelfunktion entstehen neue Informationen dadurch, dass neue XRD-Spektren gemessen werden und somit neue DTW-Distanzen bestimmt werden können.

Bei dem hier beschriebenen Versuchsablauf werden nach den initialen 20 Messungen durch die Kegelapproximation 50 neue Messpunkte ausgewertet. Nun ist es möglich die tatsächliche DTW-Distanz zwischen den neuen Punkten und denen der initialen Stichprobe zu berechnen. Diese DTW-Distanz kann mit dem zuvor approximierten Wert verglichen werden.

Ist die Abweichung zwischen approximierter und tatsächlicher DTW-Distanz besonders groß, so ist dies ein Hinweis, dass der angenommene lineare Verlauf der Ähnlichkeit der XRD-Spektren für diesen Punkt weniger stark zutrifft. Physikalisch bedeutet dies wahrscheinlich, dass für diesen Punkt ein weiterer Spitzenwert in den XRD-Spektren auftritt, oder, im Vergleich zu den betrachteten Spektren, nicht mehr vorkommt. Diese Änderung in den XRD-Spektren charakterisiert die Forderung möglichst *interessante*

Gebiete des ternären Systems zu entdecken und neben möglichst allen Formen von XRD-Spektren auch deren Übergänge zu erfassen.

Punkte deren tatsächliche DTW-Distanz stark von der approximierten DTW-Distanz abweichen können als Indikatoren für eine unerwartete Änderung der XRD-Spektren aufgefasst werden. Weitere Messungen im Umfeld sind also wahrscheinlich informativer als Messungen in Gebieten, in denen die Vorhersage und somit der lineare Verlauf der DTW-Distanzen zu stimmen scheinen. Die neu gewonnenen Informationen aus den Messungen können also in einem weiteren Iterationsschritt genutzt werden, um gezielt weitere Messpunkte auszusuchen.

Die vorherige Kegelapproximation zeigt, dass ein großer Abstand zwischen zwei betrachteten Punkten problematisch sein kann. Somit ergeben sich bereits zwei weitere Parameter, welche bei der erneuten Auswahl von Messpunkten variiert werden können. Im Folgenden wird lediglich eine Kombination getestet, um die Funktionsweise des Verfahrens zu evaluieren.

Für den Punkt, dessen approximierte DTW-Distanz zu einem der 20 initialen Messpunkte am stärksten von der tatsächlichen DTW-Distanz abweicht, wird eine neuer Kegel zwischen dem ursprünglichen Startpunk,t der für die Abschätzung genutzt wurde, und dem ausgewählten Punkt gelegt. Damit wird die Stichprobenmenge, aus der ein neuer Messpunkt bestimmt wird, auf den lokalen Bereich um die Abweichung eingeschränkt. Die maximale Manhattan-Distanz, die zwischen den beiden Punkten liegen darf, beträgt 7. Durch den recht geringen Abstand wird versucht den Suchraum auf Punkte einzuschränken, bei denen der Fehler auf tatsächlichen lokalen Änderungen beruht.

Sofern nicht bereits für alle Punkte dieses Kegels die XRD-Spektren bekannt sind, wird ein zufälliger noch nicht ausgewerteter Punkt aus dem Kegel ausgewählt. Sind bereits alle Spektren innerhalb des Kegels bekannt, so wird der Punkt mit der nächst größeren Abweichung betrachtet. Das Verfahren endet, wenn 30 neue Punkte ausgewählt wurden, wobei, um den Suchraum und somit die Laufzeit zu beschränken, maximal 60 Kegel betrachtet werden. Sind anschließend noch nicht genügend neue Punkte ausgewählt worden, so werden die fehlenden Punkte zufällig aus der gesamten Menge der noch nicht ausgewerteten Punkte bestimmt.

Um die Anzahl n_e der ausgewerteten XRD-Spektren zu variieren, kann die Anzahl der Punkte, die pro berechnetem Kegel ausgewählt werden, erhöht werden. So werden auch die Varianten bei denen zwei oder drei Punkte je Kegel ausgewählt werden untersucht. Die Anzahl der in diesem Schritt ausgewählten Messpunkte ist somit $n_e = 30, 60, 90$, wodurch sich eine Gesamtzahl von 100, 130 oder 160 ausgewerteten XRD-Spektren ergibt. Bei 342 Punkten auf den Wafern entspricht das einer Abdeckung von 29%, 38% bzw. 47% aller Punkte.

7.8 Abschließende Modellbildung

7.8.1 Einfache Zuordnung per weiterer Kegelmodelle

Um für alle nicht ausgewerteten Punkte des Wafers eine Aussage über ihr XRD-Spektrum treffen zu können, lässt sich das wahrscheinlich ähnlichste Spektrum der bekannten Punkte bestimmen und so eine Zugehörigkeitsfunktion erstellen.

Für diesen Zweck wird wiederum zwischen allen nah beieinander liegenden bekannten Messpunkten eine Kegelapproximation bestimmt. Die maximale Manhattan-Distanz zwischen zwei Messpunkten wird auf 7 festgelegt, um wieder die Auswahl auf eine lokale Umgebung einzuschränken und vor allem die Anzahl der zu berechnenden Kegel deutlich zu reduzieren, da deren Anzahl mit der Anzahl der Messpunkte exponentiell wächst So kommen bei 30 neuen Messpunkten 2.970 potentielle neue Kegel hinzu, bei 60 Messpunkten sind es 7.740 und bei 90 bereits 14.310.

Die Beschränkung des Abstandes kann allerdings nur einen kleinen Teil der potentiellen Kegel im Vorfeld filtern, da aufgrund der Auswahlstrategie der neuen Messpunkte bereits die meisten entsprechend dicht zusammen liegen sollten. Andererseits ist diese Berechnung im realen System pro Datensatz nur ein einziges Mal notwendig, weswegen eine etwas höhere Rechenzeit akzeptabel ist.

Nach der Berechnung der Kegel wird jedem Punkt mit unbekanntem XRD-Spektrum der ausgewertete Messpunkt zugeordnet, zu dem die approximierte DTW-Distanz am kleinsten ist.

Überprüfung der erstellten Modelle Für die Überprüfung der getroffenen Zuordnung können zwei verschiedene Fehlermaße betrachtet werden.

Als direktes Fehlermaß kann der relative Fehler zwischen der approximierten DTW-Distanz und der geringsten DTW-Distanz unter allen anderen 340 Punkten betrachtet werden. Dieser sollte möglichst gering sein, um die ursprüngliche Zielsetzung allen Punkten mit unbekanntem Spektrum ein möglichst ähnliches Spektrum zuzuweisen zu erfüllen. Durch die Verwendung der approximierten DTW-Distanz ergibt sich eine zweite Kontrollgröße, denn es wird sowohl die Güte der Approximation, als auch die Güte der Zuordnung gemessen.

7.8.2 Verteilung der Messpunkte

Zunächst kann überprüft werden, ob die Auswahl der Punkte für welche die XRD-Spektren bestimmt werden sinnvoll erscheint und möglicherweise interessante Gebiete gut abgedeckt werden. Außerdem sind einige der verwendeten Auswahlmethoden nicht deterministisch, so dass sich die Ergebnisse für verschiedene Durchläufe nochmals von einander unterscheiden können. In Abbildung 7.4, sind Beispiele für alle untersuchten

Datensätze zu sehen. Zwei Ergebnisse werden genauer betrachtet, um die Ergebnisse zu interpretieren.

CuNiZn Bei der betrachteten Auswahl für den CuNiZn-Datensatz (Abbildung 7.4a) fallen mehrere Dinge auf. Zunächst einmal sind die initialen Stichproben ausgewogen über den gesamten Wafer verteilt. Hingegen konzentrieren sich die Punkte der Kegelauswahl stark im oberen rechten Bereich des Wafers. Auf der einen Seite ist eine derart starke Konzentration nicht unbedingt wünschenswert, andererseits liegen genau in diesem Gebiet Messpunkte deren XRD-Spektren sowohl für $\theta = 44,8°$, als auch für $\theta = 73,7°$ erkennbar erhöhte Intensitäten aufweisen. Insofern ist dies ein Indiz dafür, dass die prinzipielle Idee die Stichprobenauswahl an den approximierten XRD-Distanzen zu orientieren in der Lage ist, auffällige Bereiche besser zu entdecken. Die Punkte der zweiten Iteration, welche anhand der Abweichungen zwischen tatsächlichen und approximierten DTW-Distanzen bestimmt werden, sind weniger eindeutig einzuordnen. Hier ist es aufgrund der hohen Konzentration von ausgewerteten Punkten in einem Bereich sehr wahrscheinlich, dass in den aufgespannten Kegeln zwischen den Punkten mit größtem Vorhersagefehler und zugehörigem Startpunkt nur bereits ausgewertete Punkte liegen und so die Auswahl mit zufälligen Punkten aufgefüllt wird.

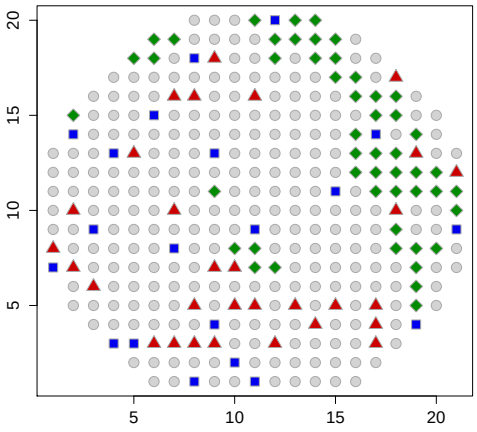
TiCoW Die Auswahl des TiCoW-Datensatz zeigt ebenfalls sehr gute Ergebnisse. Bereits die Startpunkte zeigen eine leichte Gewichtung um den interessanten Bereich im unteren Teil des Wafers, (vgl. Abb. 2.5) ohne andere Teile vollkommen unberücksichtigt zu lassen. In diesem Beispiel konzentrieren sich die Punkte der ersten Iteration weniger stark auf nur ein Gebiet, es ist aber eine Häufung im Bereich des Wafers zu erkennen, für den auch der Widerstand die größte Varianz aufweist. Deutlicher als im ersten Beispiel zeigt sich die Rolle des zweiten Iterationsschrittes, der deutlich die Bereiche der auffälligen Widerstandswerte und XRD-Spektren weiter abdeckt. Dies sind auch die Bereiche in denen die XRD-Spektren für $\theta = 40,8°$ und $\theta = 43°$ die auffälligen Intensitätsspitzen haben. Die Punkte am äußersten Rand sind in den meisten Fällen wiederum die Folge einer zufälligen Auswahl, da dort wahrscheinlich keine Kegelüberdeckung vorliegt.

7.8.3 Zuordnung von Punkten mit unbekannten XRD-Spektren

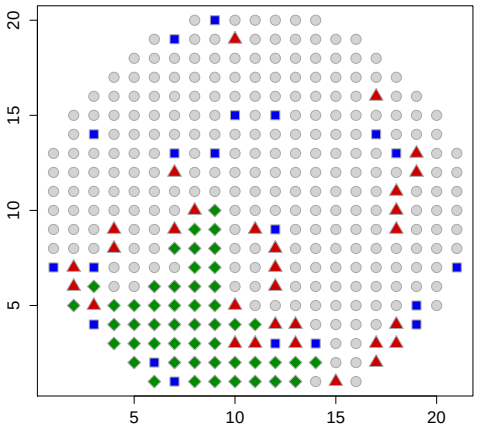
Neben dem Aufzeigen von interessanten Bereichen war es auch das Ziel für die Punkte für die keine XRD-Messung erfolgte, ein möglichst ähnliches bekanntes Spektrum zuzuordnen, um so eine Charakterisierung aller Messpunkte zu erreichen.

Der relative Fehler zwischen der approximierten DTW-Distanz und der tatsächlich geringsten Distanz unter allen übrigen 341 Messpunkten kann bestimmt werden. Dies gibt ein Maß für die allgemeine Güte der Approximation.

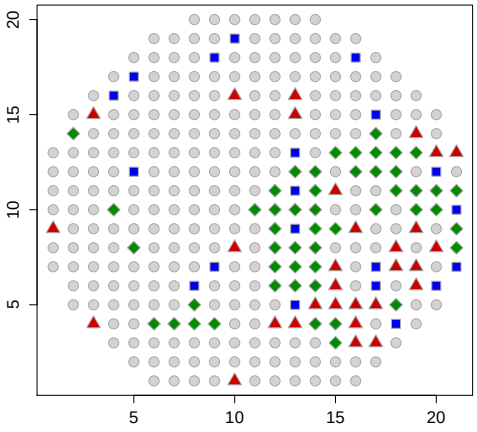
In beiden Fällen ist der Fehler allerdings ziemlich groß, mit einem durchschnittlichen relativen Fehler von $1,63$ für adaptive Messungen und einem Fehler von $1,18$ für reine Zufallsstichproben. Die Tabelle 7.1 gibt die Ergebnisse für die einzelnen Materialien an.



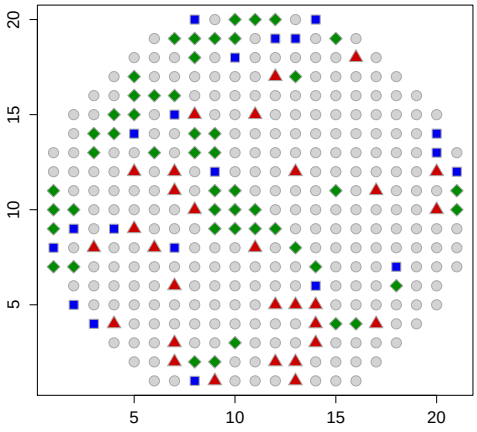
(a) CuNiZn



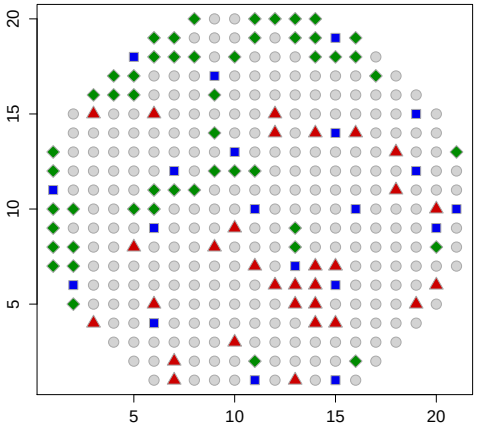
(b) NiCrRe



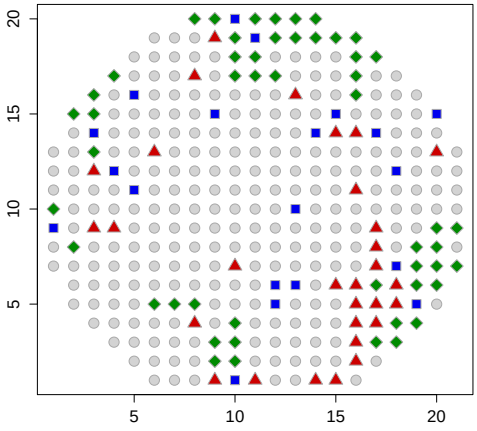
(c) TiCoW



(d) TiNiCu



(e) TiTaAl



(f) TiTaRe

Abbildung 7.4: Ausgewertete Messpunkte. 20 initiale Messpunkte k-means (blau, □); zweite Auswahl: 50 Punkte mit max. appr. DTW-Distanz (grün, ♢); dritte Auswahl: je 1 zufälliger Punkt aus dem Kegel zwischen Punkten mit größtem Vorhersagefehler und entsprechendem Startpunkt (rot, △)

Materialien	Kegelauswahl	zufällige Auswahl
CuNiZn	$2,16$	$1,12$
NiCrRe	$1,75$	$2,22$
TiCoW	$1,65$	$0,62$
TiNiCu	$1,43$	$0,95$
TiTaAl	$0,55$	$0,55$
TiTaRe	$2,70$	$1,66$

Tabelle 7.1: Durchschnittlicher relativer Fehler der Zuordnung der unbekannten XRD-Spektren für $n = 100$ gemessene XRD-Spektren. Verglichen wird die approximierte DTW-Distanz mit der tatsächlich geringsten DTW-Distanz aller Messpunkte

Für mehr ausgewertete XRD-Spektren reduziert sich der durchschnittliche Fehler für $n = 130$ auf $1,67$ und für $n = 160$ auf $1,14$,für die adaptiven Messungen auf $0,77$, sowie $0,69$ für die Zufallsstichproben.

Die recht großen Fehler sind ein deutlicher Hinweis, dass für die Zuordnung von unbekannten XRD-Spektren eine bessere Modellbildung erforderlich ist. Dies ist wünschenswert, um die die beobachten guten Ergebnisse der Stichprobenauswahl auch quantitativ widerspiegeln zu können.

8 Modellbildung durch Diskretisierung

8.1 Segmentierung der XRD-Spektren

Der zuvor beschriebene Ansatz beruht stark auf der DTW-Distanz als Änhlichkeitsmaß und dem Versuch die untersuchten XRD-Spektren auf ein charakteristisches und vergleichbares Merkmal zu reduzieren. Ein Nachteil dieses Verfahrens ist, dass die XRD-Spektren gar nicht direkt betrachtet werden. So können zum Beispiel Messfehler, wie der extreme Ausschlag des Trägermaterials in den Spektren, durch die Reduktion einen Wert verschleiert werden. Insgesamt fehlt der DTW-Distanz oder auch anderen möglichen Distanzmaßen die Möglichkeit zwischen verschiedenen charakteristischen Spektren zu unterscheiden.

Um ein Spektrum zu beschreiben sind, vor allem die charakteristischen Spitzenwerte ein auffälliges Merkmal. Vor allem für eine erste und schnelle Analyse, sind die exakte Lage, welche bedingt durch den großen Messfehler bereits nur ungenau angegeben werden werden, und die genaue Intensität eher zweitrangig. Die Charakterisierung für welche Bereiche überhaupt ein Spitzenwert vorliegt von größerem Interesse.

Aus diesem Grund ist es naheliegend zu versuchen die XRD-Spektren in ihren beiden Dimensionen zu komprimieren und so den zu bearbeitenden Merkmalsraum zu reduzieren. Die Anzahl der betrachteten Winkel lässt sich reduzieren indem die Intensitäten für zusammenhängende Intervalle von Winkeln zusammengefasst werden. Wie die gemessenen Intensitäten für ein solches Intervall aggregiert werden kann variabel gestaltet werden, aber die durchschnittliche oder größte Intensität ist eine naheliegende Wahl.

Gleichermaßen lassen sich die Intensitäten auch weiter vereinfachen indem nicht mehr der genaue Wert, welcher zudem für jedes ternäre System eine andere Größenordnung aufweist, betrachtet wird. Die stärkste Reduktion wäre eine rein binäre Entscheidung, ob ein Spitzenwert vorliegt oder nicht.

Wie mit einer solchen Modellierung das Problem der adaptiven Messung der XRD-Spektren gelöst werden kann soll an dieser Stelle aufgezeigt werden. Eine vollständige Analyse des Verfahrens und insbesondere eine genaue Auswertung der möglichen Parameter und der Ergebnisse war im Rahmen dieser Arbeit nicht möglich.

8.2 Versuchsaufbau

Durch die oben beschriebene Diskretisierung werden die XRD-Spektren auf einen binären Vektor reduziert. In diesem Vektor steht eine 1 dafür, dass in diesem Winkelbereich ein Spitzenwert vorliegt, und eine 0 dafür, dass die gemessenen Intensitäten unterhalb des Schwellwertes für einen Spitzenwert liegen.

Zwar erhöht eine solche Reduktion die Übersichtlichkeit enorm, eine direkte Modellierung der Werte eines binären Vektors ist aber nicht weniger schwierig als die der gesamten Messreihe von XRD-Intensitäten. Jedoch ist es möglich für jedes Winkelintervall ein unabhängiges Klassifikationsmodell zu erstellen, welches versucht zu entscheiden, ob in diesem Intervall ein Spitzenwert zu erwarten ist oder nicht.

Um die Qualität dieses Ansatzes zu untersuchen, wird der folgende Versuchsaufbau gewählt:

- Je 50 Messwinkel θ werden zu einem Intervall zusammengefasst. Abhängig von der Genauigkeit der Messungen entspricht das bei den vorliegenden Daten einer Intervallbreite von $0,65°$ bei NiCrRe und von $1,3°$ für alle anderen Datensätze und ergibt je nach gemessenen Winkelbereich zwischen 40 und 80 Intervalle.
- Schwellwert für einen Spitzenwert ist der Median aller beobachteten Intensitäten. In der Praxis ist dieser Wert im Vorfeld nicht bekannt und müsste zum Beispiel aus einer initialen Stichprobe oder vorherigen Experimenten abgeleitet werden.
- Als Klassifikationsverfahren wird die C-SVM aus der LIBSVM-Bibliothek verwendet. Als Klassifikationsmerkmale werden alle zuvor erhobenen Merkmale (chemische Zusammensetzung, elektrischer Widerstand, RGB-Farbwert) verwendet.
- Die SVM wird mit einer Teilmenge aller 342 Messpunkte trainiert, wobei jeweils für jedes Intervall unabhängig eine Parameteroptimierung durchgeführt wird.
- Als Trainingsmengen werden folgende vier Teilmengen erstellt: Rein zufällige Stichproben mit einem Umfang von 20 und 100 Messpunkten, Stichproben mittels gewichtetem *k-means*-Verfahren, die auch für das initiale Sampling verwendet wurden und 100 Messpunkte, die nach der finalen Iteration der Kegelapproximation ausgewählt wurden. Von allen Mengen werde 15 unterschiedliche Iterationen getestet.

Bei der Diskretisierung der XRD-Spektren kann es vorkommen, dass für keinen der Punkte der Stichprobe oder aber für alle der Schwellwert überschritten wird. In diesem Fall ist auch keine Modellbildung möglich, da nur Beispiele einer Kategorie vorliegen. Diese Fälle werden erkannt und nicht berücksichtigt, um die Beurteilung der erstellten Modelle nicht zu beeinträchtigen. In Tabelle 8.1 ist der durchschnittliche Vorhersagefehler der Modelle über alle sechs Datensätze und die 15 Iterationen aufgeführt und zusätzlich der Anteil der nicht berücksichtigten Bereiche ohne Modelle. Zunächst fällt auf, dass der Vorhersagefehler für die zufälligen Stichprobenverfahren etwas höher ist als für die adaptiven Verfahren. Interessant ist der Anteil der ausgewerteten Intervalle, welcher für die zufälligen Stichprobenverfahren deutlich höher ist. Eine mögliche Ursache könnte sein,

Verfahren	rel. Fehler	ausgewertete Intervalle
Zufallsstichprobe, Größe 20	0,25	0,35
gewicht. *k-means*, $k = 20$	0,22	0,55
Zufallsstichprobe, Größe 100	0,19	0,60
fin. Kegelapproximation, 100	0,17	0,39

Tabelle 8.1: Vorhersagefehler der C-SVM für die diskretisierten XRD-Spektren und Anteil der Bereiche für die ein Modell erstellt werden konnte

dass die adaptiven Verfahren genau die Messpunkte auswählen, in denen Spitzenwerte vorliegen und somit häufiger nur positive Beispiele in der Trainingsmenge sind.

8.3 Mögliche Verbesserung

Adaptive Verbesserung Ähnlich wie bei der Kegelapproximation wäre es wünschenswert anhand einer initialen Stichprobe möglichst geschickt neue Messpunkte zu bestimmen. Die hier gewählte Modellierung und das verwendete Klassifikationsverfahren bieten hier zwei neue Ansätze.

Die C-SVM liefert neben einer Prognose der Klasse auch ein Konfidenzmaß für diese Aussage. Die Konfidenz ergibt sich aus dem Abstand, den der beobachtete Punkt von der trennenden Hyperebene hat. Punkte, die sehr weit von der trennenden Hyperebene entfernt liegen, besitzen auch eine große Konfidenz, da es unwahrscheinlich ist, dass für diesen Punkt die getroffene Klassifikation falsch ist.

Mit diesem Maß können also gezielt Punkte mit einer niedrigen Konfidenz ausgewählt werden, um an dieser Stelle die Unsicherheit des gesamten Modells zu reduzieren. Der genaue Umgang mit diesem Konfidenzmaß kann variiert werden und im Bereich des *active-learnings* werden einige Varianten beschrieben.

Neben der Konfidenz des Klassifaktors, welche bei der Verwendung einer anderen Methode nicht direkt gegeben sein muss, ist auch eine Auswahl über die Ähnlichkeit der Verteilung der Spitzenwerte möglich. Betrachtet man die binären Vektoren der Vorhersagemodelle, so ist die Hamming-Distanz eine mögliche Größe, um zu entscheiden für welche Punkte eine Messung der XRD-Spektren sinnvoll sein könnte. Punkte, für welche die prognostizierte Hamming-Distanz zu den bekannten XRD-Spektren besonders groß ist, haben vermutlich besonders stark abweichende XRD-Spektren und kommen somit für eine weitere Messung in Frage.

Ein denkbarer Versuchsaufbau wäre, dass zunächst eine initiale Stichprobe gezogen wird und anhand dieser Werte, wie oben geschildert, die binären Vektoren für alle nicht ausgewerteten Punkte bestimmt werden. Aus dieser Menge werden die Punkte ausgewählt, deren Vektoren die größte Hamming-Distanz zu den bekannten Werten haben. Dabei könne die mittlere Distanz oder der größte einzelne Wert als Auswahlkriterium dienen. In einem letzten Schritt werden im Umfeld der Punkte deren tatsächlichen Vektoren von

den Klassifikationen der Modelle am stärksten abweichen, nochmals weitere Messpunkte bestimmt.

Modellierung von Abhängigkeiten Die hier beschriebenen Verfahren agieren unabhängig voneinander. Das heißt, es gibt keinen Informationsaustausch zwischen den Modellen. Es ist aber auch denkbar Wechselwirkungen und Abhängigkeiten zwischen den Intervallen zu berücksichtigen. So könnten zum Beispiel Graphen genutzt werden, um die Wahrscheinlichkeit, dass in einem bestimmten Intervall ein Spitzenwert auftritt, von den Beobachtungen in anderen Intervallen abhängig zu machen.

9 Zusammenfassung und Ausblick

Die Zielsetzung dieser Arbeit war zu evaluieren, ob es möglich ist durch adaptive Messungen den Zeitaufwand bei der Auswertung von Dünnschichtbibliotheken zu reduzieren. Die angewendeten Verfahren sollten in der Lage sein aus der Analyse von bereits vorhandenen Messdaten zu entscheiden für welche Messpunkte eine kostenintensive XRD-Messung am informativsten ist.

Um dies zu realisieren, wurde in Kapitel 5 ein Verlaufsschema entwickelt, welches dem Konzept des *active learning* entspricht und in der Lage ist iterativ gewonnene Informationen auszuwerten und neue Messpunkte auszuwählen. Der Aufbau dieses Schemas ist dabei so modular gewählt, dass einzelne Komponenten austauschbar und erweiterbar sind.

Der erste Teil des Verlaufsschemas, eine repräsentative Stichprobe zu ermitteln, wurde in Kapitel 6 untersucht. Es zeigte sich, dass die Gruppierung der Messpunkte mittels *k-means*-Clustering eine gute Methode ist, um möglichst informative Punkte für eine XRD-Messung auszuwählen. Neben einer geeigneten Normierung der Daten konnte außerdem gezeigt werden, dass die Kombination aus Clustering- und Regressions-Verfahren eine gute Herangehensweise ist, um Bereiche mit unerwartetem Materialverhalten zu erkennen und für eine genauere Untersuchung auszuwählen. Dies ist bereits ein wichtiger erster Schritt für eine beschleunigte Auswahl von Messpunkten. Mit dem Vergleich der aggregierten maximalen Intensitäten wurde eine Möglichkeit gefunden die Ergebnisse mehrerer XRD-Messungen kompakt darzustellen. Auch kann mit der Aggregation die Güte einer Stichprobe bestimmt werden.

Die iterative Bestimmung von weiteren Messpunkten mittels der Approximation der DTW-Distanzen in Kapitel 7 konnte zeigen, dass diese Auswahlmethode in der Lage ist sowohl die verschiedenen Bereiche einer Materialprobe abzudecken, als auch gezielt Übergangsbereiche mit großer Dynamik in den XRD-Spektren genauer zu erkunden. Das Ziel einer abschließenden Modellbildung, um alle nicht ausgewerteten Messpunkten mittels bekannter XRD-Spektren zu klassifizieren, konnte mit diesem Ansatz allerdings nicht zufriedenstellend erfüllt werden.

Die Methode die XRD-Spektren auf diskrete Wertebereiche zu reduzieren und nur zu unterscheiden, ob in einem bestimmten Winkelintervall ein Spitzenwert vorliegt oder nicht, scheint viel versprechend, da die Vorhersagegenauigkeit auch bei nur wenigen Trainingsbeispielen recht hoch ist. Die Möglichkeiten mittels mehrerer Klassifikations-SVMs eine Konfidenz für die Vorhersagegenauigkeit zu erstellen bietet eine gute Grundlage, um weitere Ansätze aus dem *active learning* zu übernehmen und dieses Verfahren weiter zu untersuchen.

Ausblick Soweit zu beurteilen ist die Anwendung von adaptiven Messungen ein neuer Ansatz in der kombinatorischen Materialforschung und bietet somit Potential für eine Vielzahl von weiteren Untersuchungen. Die Approximationsverfahren für die DTW-Distanz können noch weiter verfeinert werden, insbesondere in Hinblick auf die Kategorisierung der unbekannten XRD-Spektren. Außerdem wären weitere Datensätze, insbesondere solche mit bekannten Materialphasen, interessant, um den Eindruck der bisherigen Ergebnisse sowohl quantitativ, als auch qualitativ zu bestätigen. Außerdem könnten auch andere Distanzmaße als die DTW-Distanz untersucht werden.

Das Verfahren der Diskretisierung der XRD-Spektren konnte leider nicht vollständig umgesetzt werden. Die bisherigen Ergebnisse motivieren jedoch, diesen Ansatz weiter zu verfolgen.

Literaturverzeichnis

[1] David Arnu. Zusammenhangsanalyse von Materialeigenschaften und Röntgenbeugung (XRD) für ternäre System. Studienarbeit, TU Dortmund, 2014. `http://www-ai.cs.uni-dortmund.de/PublicPublicationFiles/arnu_2014a.pdf` , letzter Zugriff: 12. August 2014.

[2] Donald J. Berndt and James Clifford. Using dynamic time warping to find patterns in time series. In *KDD workshop*, pages 359–370. Seattle, WA, 1994.

[3] Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152. ACM, 1992.

[4] Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical Science*, pages 273–304, 1995.

[5] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

[6] Rossana Dell'Anna, Paolo Lazzeri, Roberto Canteri, Christian J. Long, Jason Hattrick-Simpers, Ichiro Takeuchi, and Mariano Anderle. Data Analysis in Combinatorial Experiments: Applying Supervised Principal Component Technique to Investigate the Relationship Between ToF-SIMS Spectra and the Composition Distribution of Ternary Metallic Alloy Thin Films. *QSAR & Combinatorial Science*, 27(2):171–178, 2008.

[7] Ralf Drautz and D. G. Pettifor. Valence-dependent analytic bond-order potential for magnetic transition metals. *Phys. Rev. B*, 84:214114, Dec 2011.

[8] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *KDD*, volume 96, pages 226–231, 1996.

[9] Günter Gottstein. *Physikalische Grundlagen der Materialkunde*. Springer, 2014.

[10] Martin L. Green, Ichiro Takeuchi, and Jason R. Hattrick-Simpers. Applications of high throughput (combinatorial) methodologies to electronic, magnetic, optical, and energy-related materials. *Journal of Applied Physics*, 113(23), 2013.

[11] S Hamann, ME Gruner, S Irsen, J Buschbeck, C Bechtold, I Kock, SG Mayr, A Savan, S Thienhaus, E Quandt, et al. The ferromagnetic shape memory system Fe-Pd-Cu. *Acta Materialia*, 58(18):5949–5961, 2010.

[12] C. Hammond. *The Basics of Crystallography and Diffraction*. International Union of Crystallography texts on crystallography. OUP Oxford, 2009.

[13] Joachim Hartung, Bärbel Elpelt, and Karl-Heinz Klösener. *Statistik: Lehr- und Handbuch der angewandten Statistik*. Oldenbourg Verlag, 2005.

[14] T. Hastie, R. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning*. Springer, july 2003.

[15] I. Takeuchi and C. J. Long and O. O. Famodu and M. Murakami and J. Hattrick-Simpers and G. W. Rubloff and M. Stukowski and K. Rajan. Data management and visualization of x-ray diffraction spectra from thin film ternary composition spreads. *Review of Scientific Instruments*, 76(6):062223, 2005.

[16] Tom Kalil and Cyrus Wadia. Materials genome initiative: A renaissance of american manufacturing, 2011. http://www.whitehouse.gov/blog/2011/06/24/materials-genome-initiative-renaissance-american-manufacturing , letzter Zugriff: 12. August 2014.

[17] Andreas Krause, Ajit Singh, and Carlos Guestrin. Near-optimal sensor placements in Gaussian processes: Theory, efficient algorithms and empirical studies. *The Journal of Machine Learning Research*, 9:235–284, 2008.

[18] R Löbel, S Thienhaus, A Savan, and A Ludwig. Combinatorial fabrication and high-throughput characterization of a Ti-Ni-Cu shape memory thin film composition spread. *Materials Science and Engineering: A*, 481:151–155, 2008.

[19] C.J. Long, J. Hattrick-Simpers, M. Murakami, R.C. Srivastava, I. Takeuchi, V.L. Karen, and X. Li. Rapid structural mapping of ternary metallic alloy systems using the combinatorial approach and cluster analysis. *Rev Sci Instrum*, 78(7):072217, 2007.

[20] Alfred Ludwig, Robert Zarnetta, Sven Hamann, Alan Savan, and Sigurd Thienhaus. Development of multifunctional thin films using high-throughput experimentation methods. *International Journal of Materials Research*, 99(10):1144, 2008.

[21] Wilhelm F Maier, Klaus Stöwe, and Simone Sieg. Kombinatorische und Hochdurchsatz-Techniken in der Materialforschung. *Angewandte Chemie*, 119(32):6122–6179, 2007.

[22] S. Meguro, T. Ohnishi, M. Lippmaa, and H Koinuma. Elements of informatics for combinatorial solid-state materials science. *Measurement Science and Technology*, 16(1):309, 2005.

[23] Ingo Mierswa, Michael Wurst, Ralf Klinkenberg, Martin Scholz, and Timm Euler. Yale: Rapid prototyping for complex data mining tasks. In Lyle Ungar, Mark Craven, Dimitrios Gunopulos, and Tina Eliassi-Rad, editors, *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 935–940, New York, NY, USA, August 2006. ACM.

[24] H. Petersen. *Grundlagen der Statistik und der statistischen Versuchsplanung.* Number Bd. 1 in Grundlagen der Statistik und der statistischen Versuchsplanung. Ecomed, 1991.

[25] D.D. Pollock. *Electrical Conduction in Solids: An Introduction.* American Society for Metals, 1985.

[26] Burr Settles. Active Learning Literature Survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009. aktualierte Fassung: 26.01.2010.

[27] Burr Settles. Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 6(1):1–114, 2012.

[28] Claude Elwood Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001.

[29] E.M. Slyusarenko, A.V. Peristyi, E. Yu Kerimov, M.V. Sofin, and D Yu Skorbov. Ternary systems of nickel and rhenium with transition metals. *Journal of alloys and compounds*, 264(1):180–189, 1998.

[30] Changwon Suh, Simone C Sieg, Matthew J Heying, James H Oliver, Wilhelm F Maier, and Krishna Rajan. Visualization of high-dimensional combinatorial catalysis data. *Journal of combinatorial chemistry*, 11(3):385–392, 2009.

[31] C. Suryanarayana and G. Norton. *X-Ray Diffraction: A Practical Approach.* Artech House Telecommunications. Springer, 1998.

[32] Nicholas K. Terrett, Mark Gardner, David W. Gordon, Ryszard J. Kobylecki, and John Steele. Combinatorial synthesis - the design of compound libraries and their application to drug discovery. *Tetrahedron*, 51(30):8135 – 8173, 1995.

[33] Sigurd Thienhaus, Sven Hamann, and Alfred Ludwig. Modular high-throughput test stand for versatile screening of thin-film materials libraries. *Science and Technology of Advanced Materials*, 12(5):054206, 2011.

[34] R Zarnetta, D König, C Zamponi, A Aghajani, J Frenzel, G Eggeler, and A Ludwig. R-phase formation in Ti_{39} Ni_{45} Cu_{16} shape memory thin films and bulk alloys discovered by combinatorial methods. *Acta Materialia*, 57(14):4169–4177, 2009.

[35] Robert Zarnetta. Combinatorial development and discovery of ternary and quaternary shape memory alloys, 2010.

[36] Robert Zarnetta, Ryota Takahashi, Marcus L Young, Alan Savan, Yasubumi Furuya, Sigurd Thienhaus, Burkhard Maaß, Mustafa Rahim, Jan Frenzel, Hayo Brunken, et al. Identification of Quaternary Shape Memory Alloys with Near-Zero Thermal Hysteresis and Unprecedented Functional Stability. *Advanced Functional Materials*, 20(12):1917–1923, 2010.

Erklärung

Hiermit erkläre ich, David Arnu, die vorliegende Masterarbeit mit dem Titel *Ressourcenschonung bei der Auswertung von Materialeigenschaften in ternären Systemen durch adaptive Messungen* selbständig verfasst und keine anderen als die hier angegebenen Hilfsmittel verwendet, sowie Zitate kenntlich gemacht zu haben.

Dortmund, 12. August 2014