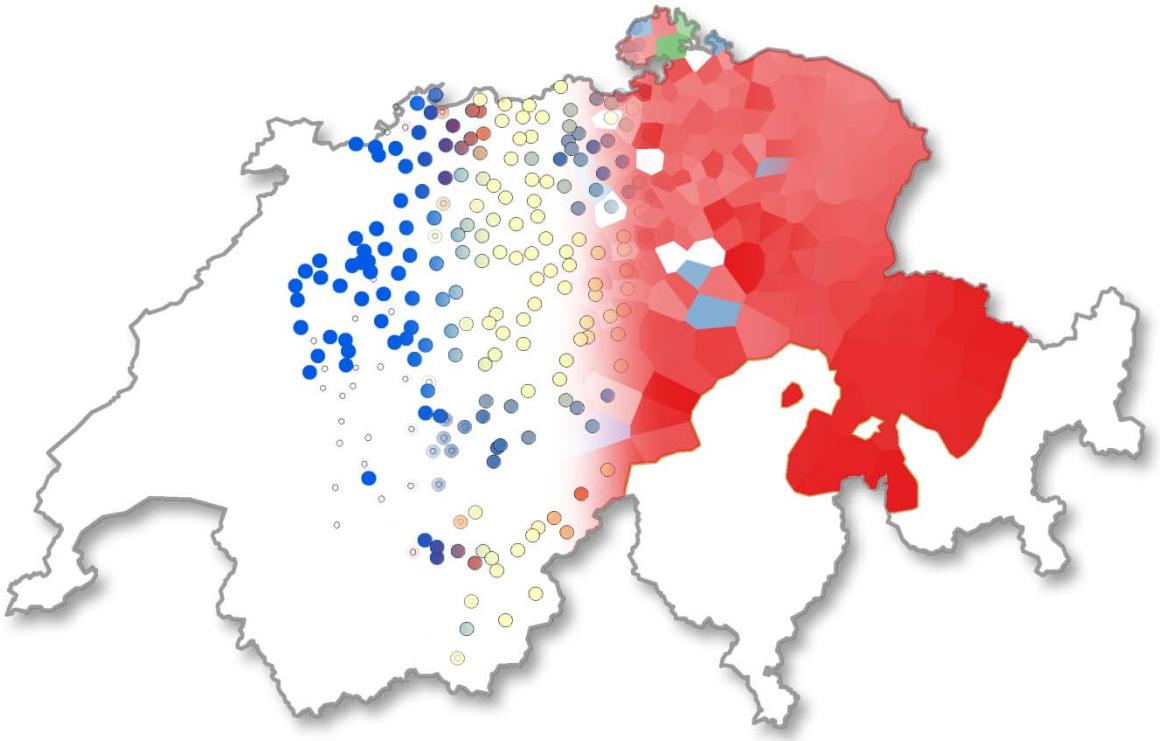
Visualisierung und geostatistische Analyse mit Daten des Syntaktischen Atlas der Deutschen Schweiz (SADS)

Pius Sibler

05-716-378



Masterarbeit GEO 511

Abteilung GIS, Geographisches Institut der Universität Zürich

Betreuung & Fakultätsmitglied Geographisches Institut: Prof. Dr. Robert Weibel

Betreuung Deutsches Seminar: Prof. Dr. Elvira Glaser & Gabriela Bart

Eingereicht am 29. April 2011

Danksagung

Ich möchte mich an dieser Stelle bei meinem Betreuer Prof. Dr. Robert Weibel für die Inspiration und die zeitintensive Unterstützung während dieser Arbeit bedanken.

Ebenfalls möchte ich mich meinen Mitbetreuerinnen am Deutschen Seminar Prof. Dr. Elvira Glaser für die spannenden Diskussionen und im Speziellen Gabriela Bart für die sprachwissenschaftliche Beratung danken.

Pia Bereuter danke ich für ihre umfangreiche Bereitstellung von Hintergrundmaterial sowie Tipps und Tricks zur R-Umgebung.

Christof Baumgartner und Marco Serraino möchte ich für die willkommenen Ablenkungen im WG-Leben und die Unterstützung bei kleinen Unsicherheiten und Fragen danken.

Der Lektorin Anne-Marie Sibler-Bertschy danke ich für die aufgebrachten Stunden des Aufspürens orthographischer und grammatikalischer Verbrechen.

Ein ganz besonderer Dank gilt Carolin Lerch, die mich in Zeiten der Unsicherheit mit bekräftigenden Worten beruhigt hat und bei der ich stets Geborgenheit und Wärme in suchenden und herausfordernden Zeiten erfahren durfte.

Meine Eltern haben mir die Freiheit gegeben, mein Wunschstudium anzutreten und mich darin stets unterstützt, wofür ich mich hier ganz herzlich bedanken möchte.

Zusammenfassung

Der Syntaktische Atlas der Deutschen Schweiz (SADS) ist ein Projekt zur Erfassung der Dialektsyntax in der Deutschschweiz. Bisher existieren daraus lediglich Punktkarten. Es bestehen zudem Vermutungen über die räumliche Verbreitung von syntaktischen Phänomenen, welche noch nie quantitativ analysiert wurden. Ausgewählte Fragen zu den linguistischen Phänomenen Finalanschluss, Komparativ und Artikelverdoppelung aus dem SADS sind die Datengrundlage dieser Arbeit. Ziel ist einerseits das Erstellen von dialektsyntaktischen Flächenkarten und andererseits die geostatistische Untersuchung von Hypothesen über die räumliche Verteilung der untersuchten linguistischen Phänomene. Mit der Bildung von Auftretensintensitäten der syntaktischen Varianten an den Befragungsorten konnte die Methode von Rumpf et al. (2009) erfolgreich angewendet werden. Sie nutzt die Kernel Density Estimation (KDE) zur flächenhaften Aufbereitung der Phänomene und teilt das Untersuchungsgebiet in nach Intensität abgestufte Flächen mit dominanten Varianten ein. In der geostatistischen Analyse konnten Strukturkenngrössen zu Kompaktheit, Homogenität und Komplexität der Karten und Varianten einen Überblick über die räumliche Charakteristik der einzelnen Fragen geben. Moran's I, Getis-Ord G_i^*, Semivariogramme und die Trendoberflächenanalyse bilden die geostatistischen Methoden, mit denen räumliche Hypothesen für die drei Phänomene überprüft werden konnten. Die verwendeten Verfahren dieser Arbeit können in Zukunft auf weitere Phänomene ausgedehnt werden.

Abstract

The Syntactic Atlas of German-speaking Switzerland (Syntaktischer Atlas der Deutschen Schweiz (SADS)) is an atlas project aiming to collect dialect-syntactic data in German-speaking Switzerland. So far, only point maps have been generated out of the SADS data basis. Assumptions about the distribution of syntactic phenomena over space do exist but they have never been analysed quantitatively. Selected questions on three syntactic phenomena provide the data framework of this thesis. The goal is on one hand to create areal syntactic dialect maps and on the other to use geostatistical methods to test spatial hypotheses about the distribution of the examined linguistic phenomena. Using relative intensities of syntactic variants at the sampled places the method of Rumpf et al. (2009) has been implemented successfully. It uses Kernel Density Estimation (KDE) to transform linguistic phenomena into areas with dominant variants which are graduated by their intensity of occurence. In the geostatistical analysis part, structural characteristics about compactness, complexity and homogeneity have proven useful to provide an overview of the spatial characteristics of the specific questions. The geostatistical methods to test the spatial hypotheses include Moran's I, Getis-Ord G_i^*, Trend Surface Analysis and semivariogram models. The methods provided in this thesis can be extended to further phenomena in the future.

Inhalt

Abbildungen

Tabellen

Abkürzungen

BFS	Bundesamt für Statistik
BFS-Nr	Gemeindenummern des Bundesamtes für Statistik
df	DataFrame (Datentyp in R)
GP	Gewährsperson
GIW	Gewichteter Identitätswert
HD	Hamming-Distanz
KDE	Kernel Density Estimation
KSDS	Kleiner Sprachatlas der deutschen Schweiz
LD	Levenshtein-Distanz
RIW	Relativer Identitätswert
SADS	Syntaktischer Atlas der deutschen Schweiz
SDS	Sprachatlas der deutschen Schweiz
SBS	Sprachatlas von Bayrisch-Schwaben
spdf	SpatialPointsDataFrame (Datentyp in R)
Swisstopo	Bundesamt für Landestopographie
TA	Trendoberflächenanalyse

Teil I: Einleitung und Hintergrund

1. Einleitung

1.1. Motivation

Man stelle sich eine alltägliche Szene vor. Ein redseliger Zürcher und eine schüchterne Bernerin versuchen zusammen einen leckeren Zwiebelkuchen zu backen. Der Zürcher bietet an, den Teig der „Weihe" zuzubereiten, seine Kochpartnerin solle doch währenddessen die „Böle" schälen. Die Bernerin, welche etwas Gefallen am Zürcher gefunden hat, versucht ihr Nichtvermögen einer sinnvollen Deutung des eben Gesagten zu unterdrücken und meint, sie bearbeite den Boden des „Chueche", er könne ja damit beginnen, die „Zibele" zu hacken. Sogleich stürzen sich beide auf den Teig, ihre Hände berühren sich. Er umarmt sie, gibt ihr ein „Chüssli", sie erwidert mit einem „Müntschi".

Kommunikationsschwierigkeiten aufgrund verschiedener Dialekte in der Deutschschweiz sind alltäglich und ein Dauergesprächsthema. Die vielen Besuche auf der Homepage „Chochichäschtliorakel"[1], wo aufgrund von gewählten Lautkombinationen von ein paar Wörtern bestimmt wird, woher die eingebende Person kommt, oder die Beliebtheit des neu herausgegebenen *Kleinen Sprachatlas der deutschen Schweiz* (Christen et al. 2010) zeigen das Interesse der Öffentlichkeit am Variantenreichtum der Schweizerdeutschen Dialekte. Die Beispiele zeigen auch, dass Sprache und Raum zusammengehören.

Für die Sprachwissenschaft ist die Verteilung von Dialekten interessant, da sie Vielfalt und Verschiedenheit der Sprache widerspeigelt. Im *Sprachatlas der deutschen Schweiz* (SDS) (Hotzenköcherle et al. 1962-2003) wurden in akribischer Arbeit unzählige Phänomene aus den Bereichen Wortschatz, Laut- und Formenlehre verortet und als Punktkarten wiedergegeben.

Allerdings bezieht sich das Interesse, sowohl von der Wissenschaft, wie auch von der Öffentlichkeit, meist auf Phänomene des Wortschatzes und der Lautung. Die Syntax, sprich die Wortfolge und der Aufbau von Sätzen, wird selten erwähnt. Sie ist laut Glaser (2008) nach wie vor ein „Stiefkind der Dialektologie" (Schwarz 1950:118).

Zurzeit wird am Deutschen Seminar der Universität Zürich unter der Leitung von Prof. Dr. Elvira Glaser an der Fertigstellung einer Erweiterung des SDS gearbeitet, dem Syntaktischen Atlas der Deutschen Schweiz (SADS)[2]. Für eine erweiterte Auswertung der SADS-Daten wurde die Abteilung Geographische Informationssysteme des Geographischen Instituts für eine interdisziplinäre Zusammenarbeit angefragt. Für die beiden Forschungsrichtungen ergeben sich interessante Synergien. So kann die Linguistik vom reichen Methodenschatz zur Erstellung von Karten und zur Analyse räumlicher Phänomene profitieren. Umgekehrt bilden die Daten des SADS eine höchst interessante Datengrundlage für das Erweitern des Verständnisses von geistes- und sozialwissenschaftlichen Phänomenen mit Raumbezug.

Die vorliegende Masterarbeit bildet das erste Produkt dieser hoffentlich fruchtbaren Zusammenarbeit.

[1] Chochichästliorakel: http://dialects.from.ch/, Zugriff: 22.4.2011
[2] SADS: http://www.ds.uzh.ch/dialektsyntax/, Zugriff: 22.4.2011

1.2. Problemstellung:

1.2.1. Ziele

Diese Masterarbeit verfolgt zwei Hauptziele. Erstens sollen reproduzierbare Möglichkeiten aufgezeigt werden, wie mithilfe von Methoden und Programmen der Geographischen Informationswissenschaft aus syntaktischen Dialektdaten **Flächenkarten** erstellt werden können. Ausgegangen wird von ausgewählten linguistischen Phänomenen, die im SADS-Projekt erfasst wurden. Anhand dieser soll eine Verfahrensweise gefunden werden, welche die Untersuchung von weiteren Phänomenen ermöglicht.

Zweitens wird versucht, Methoden der Geostatistik dafür zu verwenden, Rückschlüsse über die räumliche Verteilung der vorgegebenen Phänomene bilden zu können. Dabei sollen räumliche Hypothesen abgeleitet werden, die einerseits aus den von Linguisten geäusserten Grundhypothesen über die **räumlichen Charakteristiken** der untersuchten Phänomene stammen und andererseits aus den gewonnenen Eindrücken der Flächenkarten des ersten Teils. Anschliessend soll die Anwendung von geostatistischen Methoden helfen, die Hypothesen quantitativ zu überprüfen.

1.2.2. Forschungsfragen

Die Hauptziele können weiter in untergeordnete Forschungsfragen umgewandelt werden.

Hauptziel 1: Erstellen von syntaktischen Flächenkarten

- Wie lassen sich syntaktische Daten in Flächenkarten umwandeln und welche Methoden der Geoinformationswissenschaft sind dazu geeignet?

Bisher wurden nur Punktekarten aus den im SADS-Projekt gewonnenen Daten erstellt. Es wird an einzelnen Orten eine Angabe über die dort vorherrschenden Dialektverhältnisse gemacht. Dies entspricht der langjährigen Tradition der Sprachgeographie und ihrer Sprachatlanten. Sprache ist aber als eine räumlich kontinuierliche Erscheinung zu verstehen, womit sich eine flächenhafte Repräsentation anbietet. Dazu müssen die an den Punkten erfassten Syntaxdaten zu einer Fläche interpoliert werden. Computergestützte Verfahren helfen hierbei. Dies wurde bereits vereinzelt für andere grammatische Bereiche vorgenommen, in der Syntax hat dies aber nach wie vor Seltenheitswert. Ziel ist es, anhand ausgewählter Phänomene, eine Methodik zu entwickeln, die für weitere Phänomene und Fragen wiederverwendet werden kann.

Die grosse Herausforderung ist das nominale Skalenniveau, mit welchem die syntaktischen Daten erfasst wurden. Eine Operationalisierung von Dialektunterschieden mittels einer linguistischen Distanz, beispielsweise über die Levenshtein-Distanz (LD), ist für phonetische Daten noch direkt möglich. Bereits bei lexikalischen Daten wird dies schon um einiges schwieriger, da gewisse Wörter in verschiedenen Dialekten mitunter komplett unterschiedlich sein können.

Gibt es Ansätze in der Geostatistik, welche mit nominalskalierten Daten umgehen können? Oder kann das Skalenniveau der Daten so geändert werden, dass syntaktische Unterschiede zwischen zwei untersuchten Gebieten mit einer messbaren Distanz quantifiziert werden können?

- Welche Vor- und Nachteile bilden syntaktische Flächenkarten gegenüber herkömmlichen Punktkarten?

Sofern Flächenkarten erstellt werden können, wie sind diese Karten gegenüber den herkömmlichen einzuschätzen? In welchen Bereichen sind sie überlegen und wo haben sie Nachteile?

Hauptziel 2: Beurteilung von räumlichen Zusammenhängen in der Deutschschweizer Dialektsyntax mit geostatistischen Methoden

- Welche geostatistischen Methoden helfen, Aussagen über die räumliche Verteilung von syntaktischen Phänomenen zu machen?

Verschiedene Hypothesen zur räumlichen Verteilung von syntaktischen Phänomenen existieren. Weitere könnten aus der Interpretation der für das erste Hauptziel erstellen Flächenkarten hinzukommen. Können herkömmliche geostatistische Verfahren auf Syntaxdaten angewendet werden, um Hypothesen über

solche geographische Verteilungen zu untermauern? Welche Methoden sind besser geeignet, welche weniger?

- Sind in den untersuchten Daten räumliche Abhängigkeiten erkennbar?

Konkret auf die untersuchten Phänomene bezogen, soll untersucht werden, ob geostatistische Verfahren räumliche Zusammenhänge nachweisen und quantitativ deren Ausprägung zeigen können.

1.3. Aufbau der Arbeit

Die Masterarbeit besteht aus vier Hauptteilen (Abbildung 1-1). Die Forschungsfragen und Ziele der Arbeit (Kapitel 1) und ein Hintergrundkapitel (Kapitel 2) mit Überblick über den aktuellen Forschungsstand in der Dialektgeographie und einer Beschreibung der Datengrundlage bilden zusammen den einleitenden Teil (I) der Arbeit.

Der erste grosse Hauptteil (II) befasst sich mit der Bildung von Flächenkarten. Darin ist zuerst in Kapitel 3 und 4 beschrieben, wie die verwendeten Methoden funktionieren und wie sie auf die SADS-Daten angewendet wurden. Danach werden die Flächenkarten präsentiert (Kapitel 5) und diskutiert (Kapitel 6).

Die in Teil II aufbereiteten Daten werden im dritten Hauptteil einer geostatistischen Analyse basierend auf raumbildenden linguistischen Hypothesen (Kapitel 7) unterzogen. Die dafür eingesetzten Verfahren werden in Kapitel 8 zunächst theoretisch und anschliessend mit Blick auf die Implementierung auf die SADS-Daten beschrieben. Die geostatistischen Resultate werden in Kapitel 9 gezeigt und in Kapitel 10 besprochen.

Die beiden Teile II und III sind grösstenteils getrennt voneinander zu betrachten. Teil IV (Kapitel 11) zieht Schlüsse aus den Resultaten der beiden Hauptteile und gibt einen Ausblick auf eine mögliche, zukünftige, Forschung auf dem behandelten Gebiet.

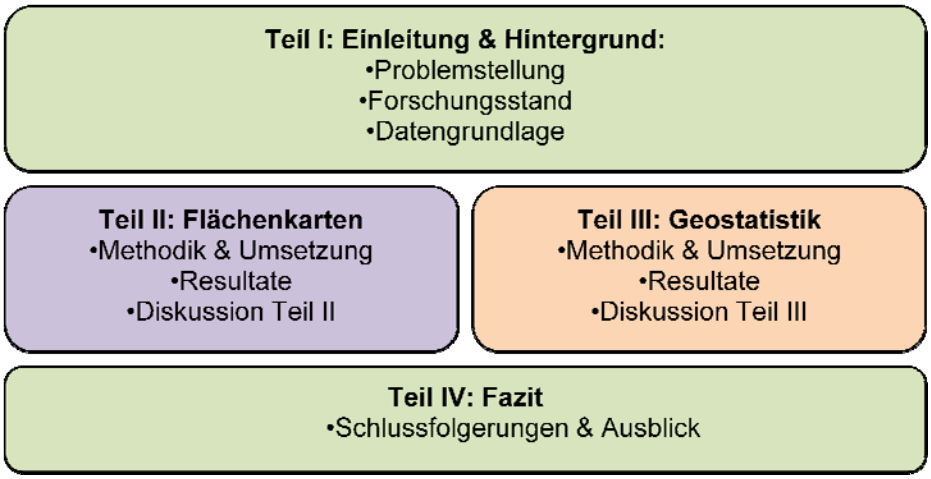


Abbildung 1-1: Die vier Hauptteile der Arbeit

2. Hintergrund

In diesem Kapitel werden die wissenschaftlichen Grundlagen, in welche diese Arbeit eingebettet ist, erläutert. Nach einer Klärung der häufig verwendeten Begriffe (2.1), wird der aktuelle dialektgeographische Forschungsstand präsentiert (2.2). Die Datengrundlage des SADS ist in Abschnitt 2.3 beschrieben und die daraus für die Arbeit entnommenen Fragen im letzten Abschnitt (2.4).

2.1. Begriffe

Im Folgenden sollen einige Begriffe, die für die Arbeit zentral sind, hervorgehoben werden. Um Zweideutigkeiten und Missverständnissen vorzubeugen, wird versucht, die Begriffe so klar wie möglich einzuschränken. Dies kann bedeuten, dass nicht der gesamten Tragweite eines Begriffs Rechnung getragen wird, hilft aber andererseits die Verständlichkeit des Geschriebenen zu erhöhen. Da die Arbeit von einem Hintergrund in der Geoinformationswissenschaft ausgeht, wird hier das Hauptaugenmerk auf Begriffe der Sprachwissenschaften gerichtet.

Dialekt und Hochsprache

So intuitiv und einfach die beiden Begriffe im alltäglichen Umgang verwendet werden, so schwer tut sich die Linguistik, diese zu definieren. Löffler (2003) nennt als übergeordnete Gemeinsamkeit aller Definitionsversuche die „relative Unselbständigkeit des Begriffs Dialekt/Mundart […]. Dialekt steht immer in einer komplementären Beziehung zu einer […] Bezugsgrösse, meist der übergeordneten Hochsprache“ (Löffler 2003:3). Der Unterschied von Dialekt und Hochsprache kann nach verschiedenen Kriterien vorgenommen werden, wobei in dieser Arbeit jenes der räumlichen Ausdehnung verwendet wird: Dialekt gilt als orts- und raumgebunden, die Hochsprache als überörtlich. Diese Arbeit befasst sich ausschliesslich mit den Deutschschweizer Dialekten und behandelt die übergeordnete Sprache, das Hochdeutsche, nicht. Auf der Ebene von Sprachkarten können Dialekte auch als Synonym für die Untersuchungsorte verwendet werden.

Phonetik, Lexik, Morphologie, Syntax

Phonetik bzw. die Phonologie ist mit dem deutschen Begriff der Lautlehre gleichzusetzen, Lexik mit jenem des Wortschatzes, die Morphologie im linguistischen Zusammenhang mit der Wortbildung und die Syntax mit dem Satzbau. Sie sind alle dem Überbegriff der grammatischen Beschreibung untergeordnet und bilden so genannte grammatische Bereiche (Löffler 2003). Tabelle 2-1 zeigt Beispiele der vier Untergruppen in der Deutschschweiz. Die Beispiele sind dem „Kleinen Sprachatlas der deutschen Schweiz“ (KSDS) von Christen et al. 2010 entnommen.

	Fragestellung	**Hochdeutsch**	**Deutschschweizer Dialektbeispiel**
Phonetik	„Wie wird ein Wort ausgesprochen?“	Rücken	Rugge, Rügge, Rigge, Rügg (KSDS, Karte 84)
Lexik	„Welche Worte werden für einen Begriff verwendet?“	Kuss	Müntschi, Schmutz, Kuss, Chuss (KSDS, Karte 8)
Morphologie	„Aus welchen Teilen ist das Wort aufgebaut?“	Bruder/Brüder	Brüeder/Brüedere, Brüeder/Brüeder, Brueder/Brüeder (KSDS, Karte 113)
Syntax	„Wie werden die Worte angeordnet?“	Gehen lassen	Gaa laa, la gaa (KSDS, Karte 120)

Tabelle 2-1: Die vier grammatischen Bereiche mit Beispielen aus dem KSDS (Christen et. al 2010)

Linguistische Variante

Eine linguistische Variante wird hier mit einer linguistischen Variablen gleichgesetzt. Zentral für diese Arbeit ist der Begriff der syntaktischen Variablen und folgt der Definition von Spruit (2006: 494), wonach eine „Form oder eine Wortfolge in einem syntaktischen Kontext, in welchem sich zwei Dialekte unterscheiden können“ umschrieben ist.

Linguistisches Phänomen

Als linguistisches Phänomen werden im vorliegenden Fall syntaktische Konstruktionen bezeichnet, die jeweils durch verschiedene Varianten realisiert werden können.

Linguistische vs. geographische Distanz

Die linguistische Distanz ist der Grundstein der Dialektometrie (siehe 2.2.2), welche Unterschiede zwischen Dialekten quantitativ zu messen versucht. Es gibt verschiedene Methoden, um diese linguistische Distanz zu messen. Die geographische Distanz bezieht sich in dieser Arbeit auf die euklidische Distanz, welche der Luftdistanz zwischen zwei Orten entspricht.

Thiessen-Polygone

Voronoi-Diagramme, auch Thiessen-Polygone genannt, beschreiben eine in der geographischen Informationswissenschaft weit verbreitete, einfache Interpolation einer Punktgruppe (Burrough & McDonnell 1998). Aus jedem Punkt wird eine Teilfläche gebildet. Das Innere dieser Gebiete beschreibt alle Orte, bei welchen die euklidische Distanz zum enthaltenen Punkt kleiner als zu allen anderen Punkten ist (Boots 1999). Sie werden für die Darstellung von Sprachkarten vor allem seit Goebl (1982) verwendet und bieten eine Möglichkeit, Punktdaten in Flächenform darzustellen.

Akzeptanz und Intensität

Mit Akzeptanz ist hier der relative Anteil einer Variante an allen möglichen Varianten gemeint. Werden an einem Ort beispielsweise zwei Varianten gleich oft akzeptiert, resultiert ein Akzeptanzwert von 0.5 pro Variante. Von Intensitäten wird gesprochen, wenn Akzeptanzwerte interpoliert werden.

Dominante vs. präferierte Varianten

Als dominant wird hier eine Variante beschrieben, die an einem Untersuchungsort häufiger als „akzeptiert“ bezeichnet wurde als alle anderen.

Die präferierte Variante beschreibt die von Gewährspersonen als am natürlichsten bewertete.

2.2. Forschungsstand

Für eine ausführliche Auseinandersetzung zur Beziehung von Sprache und Raum sei auf das gleichnamige zweiteilige Werk „Language and Space“ der Herausgeber Auer & Schmidt (2010) bzw. Lameli et al. (2010) verwiesen. Eine gute Übersicht über die Entwicklung der interdisziplinären Zusammenarbeit von Linguistik und Geographie, ausgehend von der Darstellung von dialektologischen Phänomenen in der Dialektgeographie über die Dialektometrie bis hin zur syntaktischen Mikrovariation, ist in Kapitel 1.2 der Doktorarbeit von Spruit (2008) oder im Aufsatz von Pickl & Rumpf (unveröffentlicht) zu finden. Die wichtigsten Eckpunkte werden im Folgenden nochmals erläutert.

2.2.1. Dialektgeographie

Die Marburger Schule

Die traditionelle Dialektgeographie begann mit Adelbert von Keller im 18. Jahrhundert. Er gilt als Vater der Idee, eine Sprachkarte zu zeichnen, welche Gebiete nicht nur zwischen den groben Standard-Sprachräumen, sondern innerhalb eines Sprachraums aufzeigen sollte (Schrambke 2010). Er rief in einem Brief um „Mitwirkung zur Sammlung schwäbischen Sprachschatzes“ (Keller 1855: 9) auf. Er hatte die Absicht, Eigenschaften der Sprache, welche nicht in der Schriftsprache vorhanden sind, zu sammeln. Dies beschränkte sich schliesslich auf lexikalische, phonologische und morphologische Daten, welche er in

Form einer Übersetzungsaufgabe indirekt und damit schriftlich an 320 Schulen im damaligen Württemberg sammeln liess (Schrambke 2010).

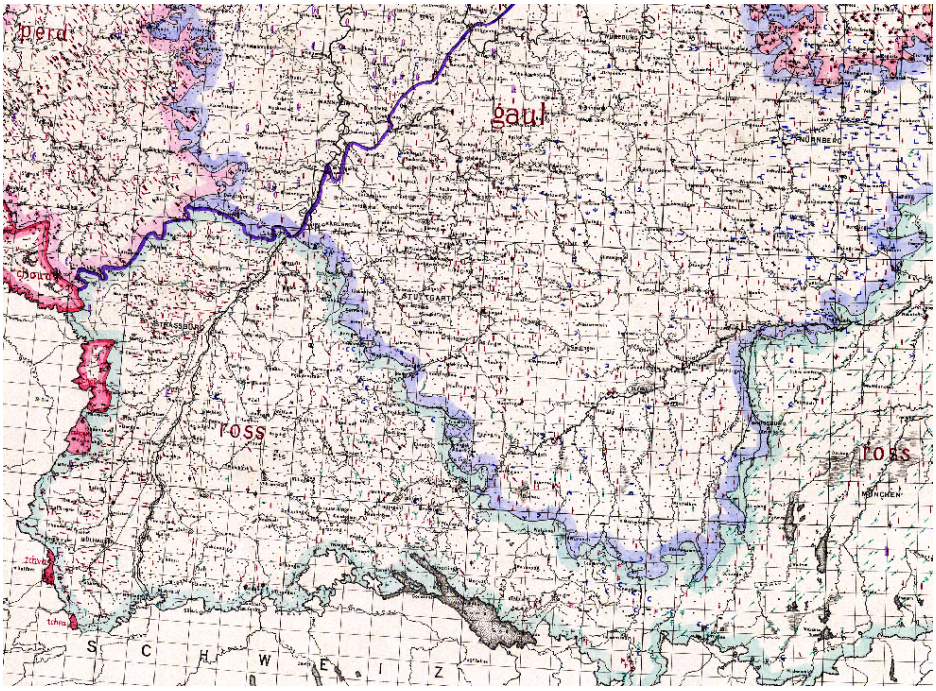


Abbildung 2-1: Ausschnitt in Süddeutschland aus der Lautkarte "Pferde" aus dem Sprachatlas des Deutschen Reichs mit handgezeichneten Isoglossen (nach Digitaler Wenkeratlas diwa)

Die ersten handgezeichneten Karten wurden von seinem Doktoranden Georg Wenker 1878, basierend auf eigenen Untersuchungen erstellt (Wenker 1877). Sie bestanden zuerst aus 42 und wurden für spätere Arbeiten auf 38 bzw. 40 Sätze, die so genannten Wenkersätze, angepasst (diwa)[1]. Sie sollten hauptsächlich phonologische, wie auch einige morphologische Informationen über die Sprache an den Untersuchungsorten liefern (Schrambke 2010). Diese Sätze verteilte er an Lehrer, welche sie von Schülern an Volksschulen übersetzen liessen, er wählte folglich, wie vor ihm Keller, eine indirekte Erhebungsmethode. Wenker veröffentlichte mit diesen Informationen den ersten Sprachatlas überhaupt, den „Sprach-Atlas der Rheinprovinz nördlich der Mosel sowie des Kreises Siegen" und daraus später den auf ein grösseres Gebiet ausgedehnte „Sprachatlas von Nord- und Mitteldeutschland" (Wenker 1881). Darin werden die verschiedenen Sprachgebiete durch handgezeichnete Grenzlinien unterschieden, so genannte „Isoglossen". Diese entstehen durch das Vergleichen von sprachlichen Merkmalen benachbarter Regionen. Ein Beispiel aus Wenkers "Sprachatlas des Deutschen Reichs" ist in Abbildung 2-1 zu sehen.

Die Württemberger Schule

Die Untersuchung von sprachlicher Mikrovariation, der Verteilung von Dialekten (Spruit et al. 2009), hat eine lange Tradition. Karl Haag (1898) wagte den ersten Versuch, Unterschiede zwischen Dialekten räumlich mithilfe einer direkten, mündlichen, Datenerhebung zu erfassen. Er war auch der erste, welcher die Notwendigkeit einer Gewichtung der linguistischen Unterschiede forderte. Er bildete mehrere Isoglossen für verschiedene linguistische Phänomene und erhielt dadurch eine linguistische Distanz (Abbildung 2-2). Je weniger Merkmale zwei Dialekte, sprich zwei Regionen, gemeinsam haben, umso grösser ist deren linguistische Distanz und dadurch umso dicker die Grenzlinie dazwischen (Rumpf et al. 2009).

[1] diwa: Digitaler Wenker Atlas: www.diwa.info, Zugriff: 19.4.2011

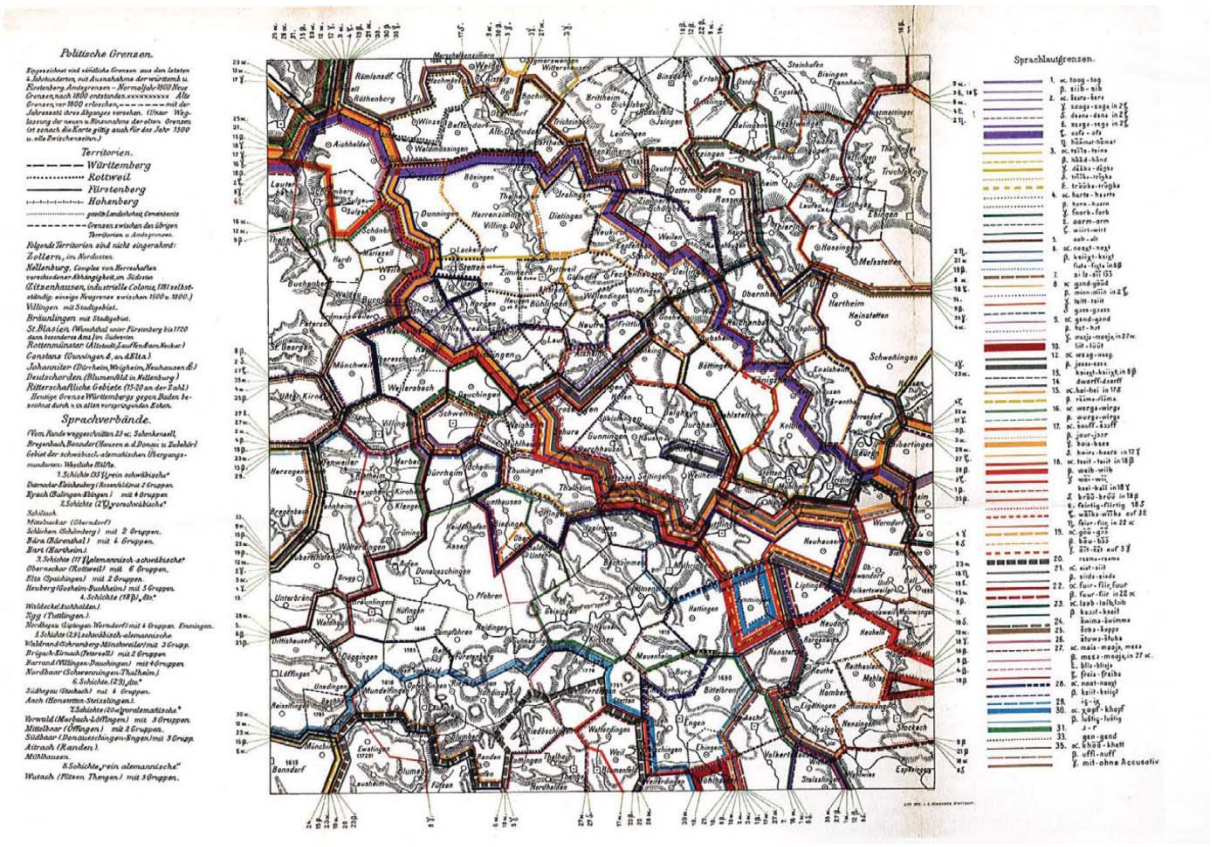


Abbildung 2-2: Dialektkarte von Haag (1898) mit mehreren eingezeichneten Isoglossen

2.2.2. Dialektometrie

Jean Séguy (1973) prägte den Begriff der Dialektometrie, welche quantitativ Unterschiede zwischen Dialekten misst, indem eine linguistische Distanz berechnet wird (Nerbonne & Kretzschmar 2003). Dafür wurden von ihm aus einem Datensatz verschiedener Dialektpunkte Paare gebildet. Deren Distanz entspricht den aufsummierten Dialektvarianten, welche sich unterscheiden.

Salzburger Schule

Hans Goebl (1982), Begründer der Salzburger Schule, war der erste, welcher computergestützte Verfahren zur Messung linguistischer Distanz verwendete. Indem alle Dialekte mit allen verglichen werden, entsteht eine Distanzmatrix, welche als Ausgangslage für weitere Berechnungen dient. Goebl erfand mehrere Masse zur Berechnung von linguistischen Distanzen und gilt als Schwergewicht in der dialektometrischen Forschung (Nerbonne & Kretzschmar 2003). Er empfahl weiter das Werkzeug der Clusteranlayse, um Dialekte zu gruppieren (bspw. Goebl 2006). Erstmals wurde auch verbreitet mit Voronoi-Diagrammen als Darstellungsgrundlage gearbeitet, womit Flächenkarten erzeugt werden konnten. Ansätze dazu sind aber bereits bei Haag zu sehen (Abbildung 2-2).

Groninger Dialektometrie

Den nächsten grossen Schritt in der Dialektometrie machten John Nerbonne und Wilbert Heeringa (Nerbonne & Heeringa 1997). Sie verwendeten die erstmals von Kessler (1995) für linguistische Zwecke genutzte Levenshtein-Distanz (LD) als Messwert für die linguistische Distanz und bauten diese aus. Bei dieser Distanz werden zwei Begriffe miteinander verglichen und errechnet, wie viele Änderungen es braucht, um ein Wort in ein anderes umzuwandeln. Die drei möglichen Änderungsformen sind **Einfügen, Löschen** und **Substituieren** (Heeringa 2004). Basierend auf phonetischer Umschrift können so Distanzen zwischen zwei phonetischen Varianten innerhalb eines Phänomens errechnet werden. Abbildung 2-3 zeigt als Beispiel die Umwandlung von zwei Laut-Varianten des englischen Begriffs „afternoon“ und die daraus resultierende LD. Zudem verwendet die Groninger Forschergruppe um Nerbonne Datenreduktionstechniken wie Multi-Dimensional Scaling, Clusteranalysen und Faktoranalysen, um Dialektverteilungen zu visualisieren (z.B. Heeringa 2004; Nerbonne & Heeringa 2009).

æəftənɐn	delete ə	1
æftənɐn	insert r	1
æftərnɐn	subst. ɐ/u	1
æftərnɐn		
		3

Abbildung 2-3: Levenshtein-Distanz zwischen zwei phonetischen Varianten von „afternoon“: Durch einmal löschen, einmal einfügen und eine Ersetzung wird die eine Variante in die andere umgewandelt (nach Heeringa 2004: 124)

Neue Dialektometrie von Rumpf et al.

Alle diese dialektometrischen Verfahren richten das Hauptaugenmerk auf die linguistische Distanz und streben eine globale Einteilung der Untersuchungsgebiete in Dialektregionen an. Angestossen von Jonas Rumpf und Simon Pickl gelingt einer Forschergruppe aus Ulm und Augsburg in einem neuen Ansatz, mithilfe von Kernel-Dichteschätzungen, eine fein strukturierte Einteilung der Dialektvariation, welche auch die geographische Distanz einbindet (Rumpf et al. 2009; Rumpf et al. 2010).

In den bisherigen Arbeiten der Dialektometrie und der sprachlichen Mikrovariation wurden vor allem phonetische (z.B. Kessler 2005; Heeringa 2004; Nerbonne 2009) und zum Teil auch lexikalische Phänomene (z.B. Nerbonne & Kleiweg 2003) einer Sprache untersucht. Obwohl seit Mitte der neunziger Jahre ein verstärktes Interesse für syntaktische Phänomene aufgekommen ist, zum Beispiel im SAND[2]-Projekt und dem dieser Arbeit zu Grunde liegenden SADS (Kortmann 2010), sind diese verglichen mit den anderen Phänomenklassen erst in wenigen Arbeiten aufgegriffen worden (bspw. Barbiers et al. 2002; Spruit 2008).

Diese Arbeit nutzt im ersten Hauptteil (Teil II) schwerpunktmässig die von Rumpf et al. (2009) vorgeschlagenen Methoden zur Generierung von Flächenkarten und wendet sie auf syntaktische Daten an. Im zweiten Hauptteil (Teil III) wird dann versucht, mithilfe von geostatistischen Methoden Aussagen über die räumliche Verteilung von syntaktischen Phänomenen zu gewinnen. Diese Kombination von Linguistik und Geostatistik ist bisher noch sehr wenig verbreitet, obwohl Lee & Kretzschmar (1993) bereits vor fast 20 Jahren die Vorteile einer solchen Zusammenarbeit hervorgehoben haben.

2.3. Datengrundlage: Der Syntaktische Atlas der Schweiz (SADS)

Initiiert im Jahr 1935, ist in den Jahren 1962 bis 2003 mit dem Sprachatlas der Schweiz (SDS) ein enormes Werk erschienen, das die Verteilung von Dialekt-Phänomenen im Deutschschweizer Sprachraum aufzeigt. Obwohl ursprünglich geplant, werden syntaktische Begebenheiten nur sehr untergeordnet behandelt (Glaser 1997; Bucheli & Glaser 2002). Um dieses Defizit zu beheben und den Atlas um syntaktische Phänomene zu erweitern, wurde im Jahr 2000 im Rahmen eines Nationalfondsprojektes eine gross angelegte Befragung gestartet, welche über 3000 Personen aus 383 verschiedenen Orten (Abbildung 2-4) umfasst (Bucheli & Glaser 2002). Die Daten werden aufbereitet als Punktkarten im Syntaktischen Atlas der Deutschen Schweiz (SADS) erscheinen. Damit ist eine umfangreiche Sammlung von räumlich referenzierten Sprachdaten vorhanden, welche nicht nur das Erforschen von linguistischen Eigenheiten der Deutschschweizer Dialekte erlaubt, sondern auch deren räumliche Verteilung und Visualisierung.

[2] SAND: A Syntactic Atlas of the Dutch Dialects:
http://www.meertens.knaw.nl/projecten/sand/sandeng.html, Zugriff: 22.4.2011

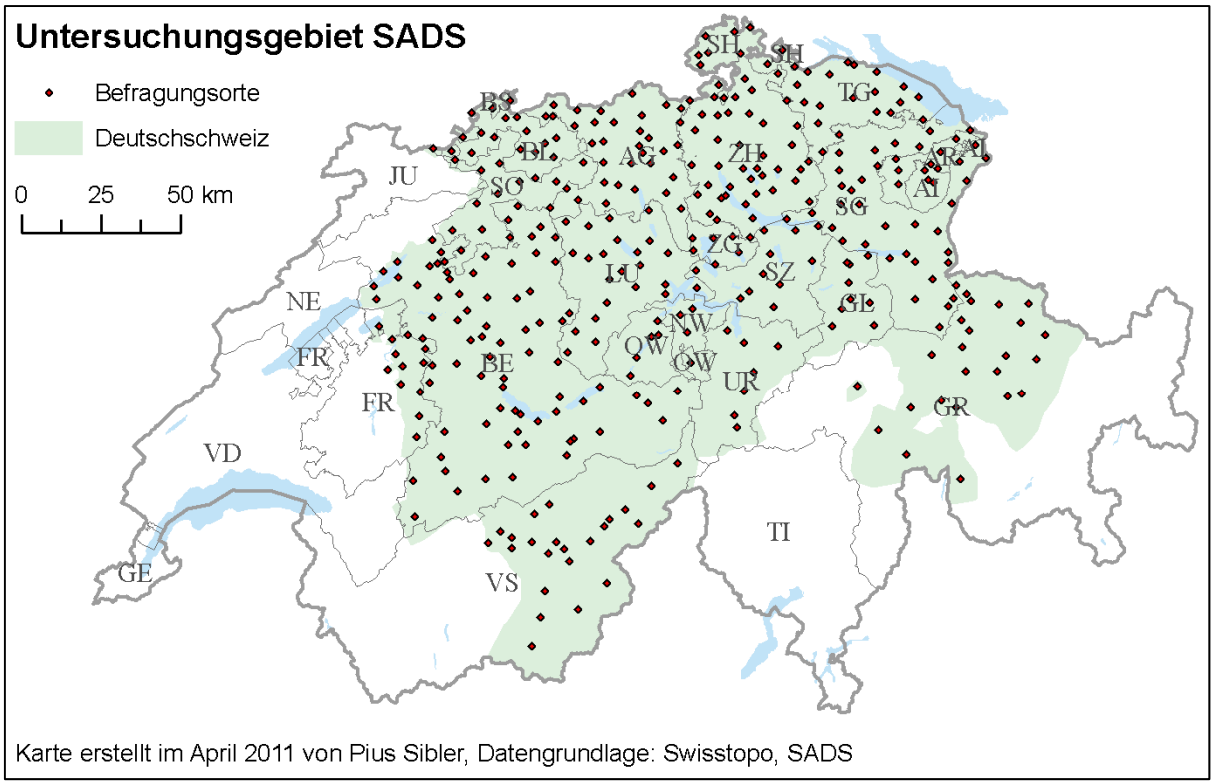


Abbildung 2-4: Im SADS-Projekt enthaltene Orte mit befragten Personen

Eckdaten zum SADS

Für den SADS wurden vier Fragebogen erstellt. Sie werden im Folgenden mit den römischen Ziffern I-IV repräsentiert. Die Fragebogen wurden im Abstand von zwei bis sechs Monaten verteilt, um eine Überforderung der Befragten zu vermindern. Ein weiterer Hintergedanke war, auf allfällig erkannte Schwachstellen früherer Fragebogen reagieren zu können, um Verbesserungen in weiteren Serien anzubringen (Bucheli Berger 2008).

Befragt wurden 3185 Gewährspersonen (GP) an 383 Orten in der Schweiz, welche sich an der Verteilung der Orte im SDS orientierten. Um Problemen, wie der Vermischung durch Migration, vorzubeugen, wurde konservativ befragt. Dies bedeutet, dass die Informanten möglichst im befragten Ort aufgewachsen sein mussten sowie auch mindestens ein Elternteil von ihnen[3].

Pro Ort wurden zwischen drei und 26 Personen befragt, welche aus allen Alters- und Sozialschichten stammten (Bucheli Berger 2008). Die exakten Verteilungen unterscheiden sich von Frage zu Frage, da gewisse Antworten unbrauchbar waren und entfernt werden mussten. Aus diesem Grund unterscheidet sich auch die Gesamtpersonenzahl pro Frage. Der Median der untersuchten Personen liegt zwischen fünf und sechs GP pro Untersuchungsort. Abbildung 2-5 zeigt, wie viele Gewährspersonen pro Ort für die im nachfolgenden Kapitel vorgestellten zehn untersuchten Fragen berücksichtigt wurden.

[3] Projektübersicht zum SADS: http://www.ds.uzh.ch/dialektsyntax/eckdaten.html, Zugriff: 22.4.2011

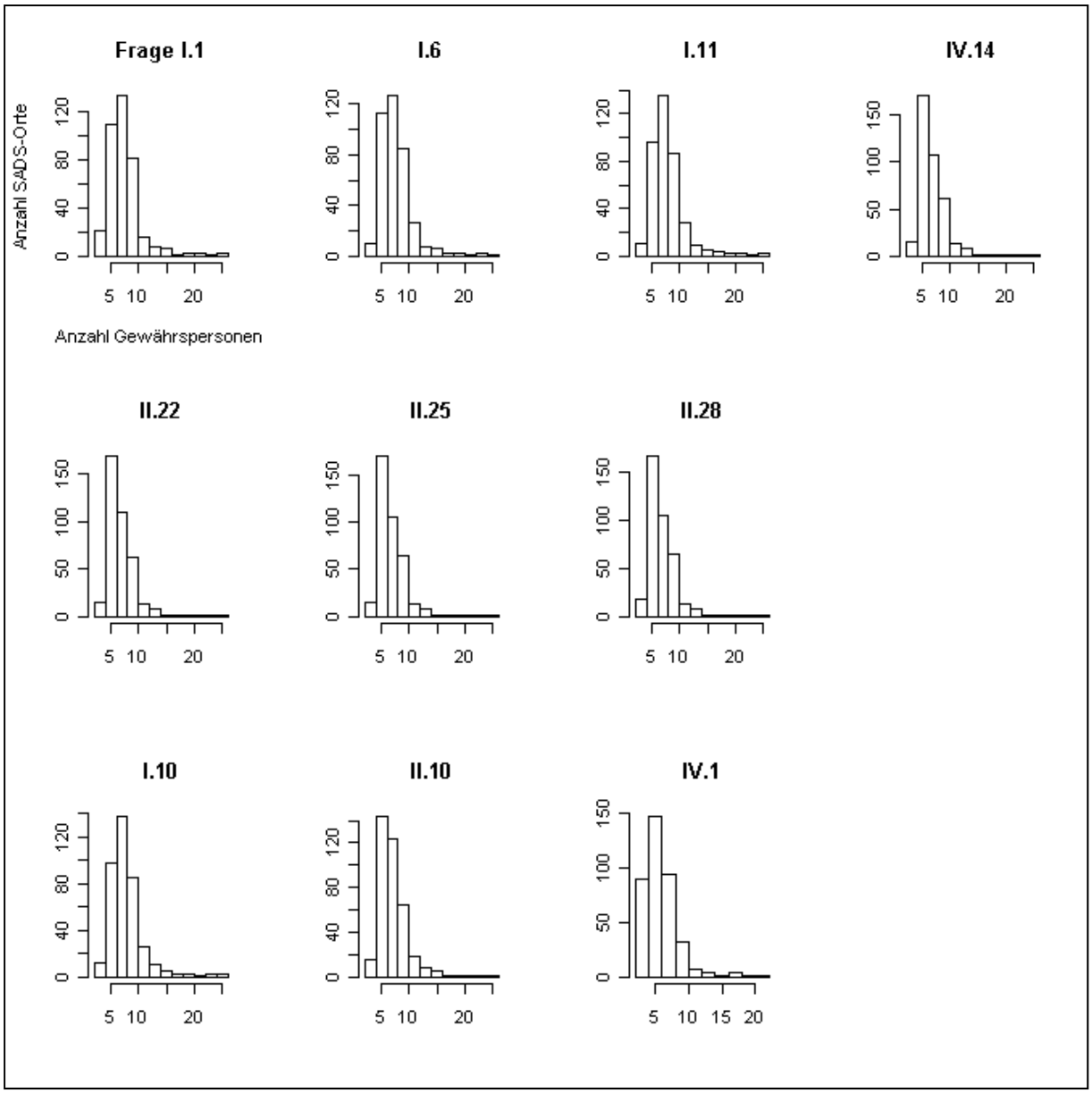


Abbildung 2-5: Histogramme der Anzahl befragten Personen pro Untersuchungsort für die zehn untersuchten Fragen der Phänomene Finalanschluss (obere Zeile), Komparativ (Mitte) und Artikelverdoppelung (unten).

Insgesamt sind im SADS-Datensatz über 50 syntaktische Phänomene abgedeckt. Um die Aussagekraft zu erhöhen, sind oft mehrere Fragen zu demselben Phänomen gestellt worden. Es handelt sich um **Übersetzungsfragen**, bei denen ein hochdeutscher Satz in Dialekt übersetzt werden musste, um **Ergänzungsfragen**, bei denen der Beginn eines Satzes gegeben ist und von den GP vervollständigt werden musste und um **Ankreuzfragen**, bei welchen aus einer Auswahl von Antworten erstens die akzeptierten Varianten und zweitens die präferierte Antwort angekreuzt werden konnten. Um das Ausfüllen für die GP etwas angenehmer zu gestalten und um die Deutungsmöglichkeit einzuschränken, wurde um die Fragen oftmals eine kleine Geschichte erzählt (Bucheli & Glaser 2002). Abbildung 2-6 gibt eine Übersicht mit Beispielen zu den verschiedenen Fragetypen.

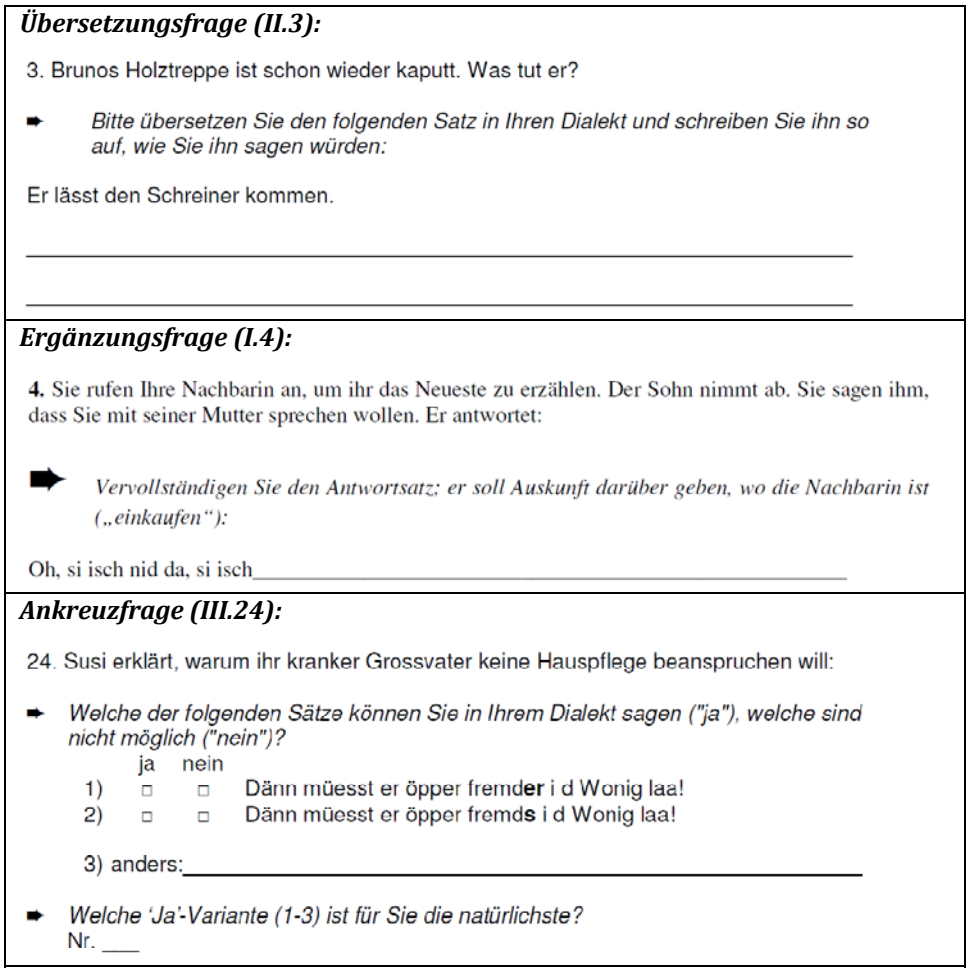


Abbildung 2-6: Die drei angewendeten Fragetypen mit Beispielen aus den SADS-Fragebogen

Da die Antworten von den GP selbst eingetragen werden konnten, liegt eine indirekte Befragung vor, welche gemäss Bucheli Berger (2008) in den meisten untersuchten Fragen zu hinreichend brauchbaren Daten führte.

Datenbank

Die mit dem oben beschriebenen Verfahren erhobenen Informationen wurden danach in einer FileMaker Datenbank abgelegt. Diese Datenbank ist nach Fragen der vier Fragebogen geordnet. Für jede Frage wurde eine Tabelle erstellt, wobei diese nach Gewährpersonen geordnet sind. Jedes Tupel entspricht darin einer befragten Person. Jede Person ist mit einer BFS-Nummer versehen, damit sie einem der SADS-Untersuchungsorte zugeordnet werden kann. Für jede Variante, die in den Antworten vorgekommen ist, gibt es ein logisches Attribut, welches den Wert „1" enthält, falls diese Variante akzeptiert wurde und „0" falls nicht. Bei den Fragetypen, bei denen eine natürliche Variante angegeben werden konnte, ist noch ein zusätzliches Feld aufgeführt, welches einen, manchmal auch mehrere, Codes für die präferierten Varianten enthält. Tabelle 2-2 zeigt vereinfacht ein Objekt der SADS-Datenbank.

Frage
GP-Nummer (Primärschlüssel)
BFS-Nr
Variante 1
Variante 2
⋮
Variante X
natürliche Variante (optional)

Tabelle 2-2: Vereinfachte Entität „Frage" in der SADS Datenbank

Im Zuge des SADS-Projektes sind bereits verschiedene Beiträge, Masterarbeiten und auch Dissertationen entstanden. Eine vollständige Liste der bisher erschienenen Publikationen ist auf der Projekthomepage[4] vorzufinden.

2.4. Untersuchte Phänomene und deren Klassierung

Die Sammlung des SADS beinhaltet über 100 verschiedene Fragen zu mehr als 50 Phänomenen. Diese Fragen alle in den Rahmen einer Masterarbeit einzubeziehen macht wenig Sinn. Deshalb wurden in Absprache mit Linguistinnen drei Phänomene daraus gewählt, welche in den folgenden Unterkapiteln erläutert werden.

Längst nicht alle Phänomene, die im SADS untersucht wurden, sind wissenschaftlich ausgewertet. Die drei gewählten Phänomene sind jedoch alle bereits in Publikationen behandelt worden und es liegt deshalb ein gewisser theoretischer Hintergrund vor, der konsultiert werden kann. Aus diesem Grund sind bereits Karten vorhanden, die einige Fragen der behandelten Phänomene abbilden. Sie sind jeweils als Punktkarten vorliegend.

Es ist zudem zu betonen, dass anhand der Phänomene Methoden zur Kartengenerierung sowie Methoden aus der Geostatistik getestet werden sollen. Die Arbeit hat nicht zum Hauptziel, linguistische Aussagen zu machen, sondern ist eingebettet in die Geographische Informationswissenschaft und hat deshalb vornehmlich methodischen und explorativen Charakter. Sie hat nicht den Anspruch, allumfassend die Untersuchungsergebnisse des SADS mit einzubeziehen. Ausgehend von den Resultaten einer kleinen Auswahl von Fragen soll mit dieser Arbeit eine Basis geschaffen werden, um weitere Karten und Analysen herstellen zu können.

Der Entscheid für die drei ausgewählten Phänomene fiel auch aufgrund der Vermutung, dass sie verschiedene räumliche Verteilungen besitzen. Die Ausarbeitung der Klassierungen der Varianten wurde freundlich unterstützt von Gabriela Bart und ist abgesprochen mit Prof. Dr. Elvira Glaser. Es wurden teilweise mehrere Antworten einer Variante zugewiesen, andere Antworten wurden nicht berücksichtigt, da sie nicht dem untersuchten Phänomen unterzuordnen sind.

Eine Zusammenfassung aller berücksichtigten Phänomene mit den dazugehörigen Klassierungen und Varianten ist in der Phänomentabelle im Anhang A zu finden. Darin, wie auch im weiteren Verlauf der Arbeit, werden **Antworten** und **Klassen** unterschieden. Antworten beziehen sich auf ein konkretes Antwortfeld in der SADS-Datenbank und Klassen umfassen mehrere zu einer Variante gehörende Antworten.

2.4.1. A: Finalanschluss

Bei der Wahl und Position des Anschlussmittels für Finalsätze wird untersucht, welches Wort einen Finalsatz einleitet und wo in der Satzstruktur dieses Wort gesetzt wird.

Das erste ausgewählte Phänomen wird anhand von vier Fragen aus dem ersten und vierten Fragebogen untersucht. Sie umfassen eine Übersetzungsfrage (I.1), eine Ergänzungsfrage (I.6) und zwei Ankreuzfragen (I.11 & IV.14). Die räumliche Grundhypothese lautet „Es gibt eine West-Ost-Verteilung“ der Varianten. Seiler ergänzt, die Verteilung der Varianten könnte sich entlang einer „schiefen Ebene“ (Seiler 2005: 331/332) bewegen.

Ausgangspunkt für die Klassierung bildet die erste Frage des ersten Fragebogens, welche bereits von Seiler (2005) intensiv behandelt wurde. Darin werden jeweils nur die Varianten *für (…zu)* und *zum (…zu)* kartiert. Diese beiden Varianten kommen in den Antworten der Gewährspersonen am häufigsten vor und bilden in dieser Arbeit die kleinere Klassierung. Sie wird im Folgenden als „eingeschränkte Klassierung“ bezeichnet. Zu dieser Klassierung kommt zusätzlich noch eine erweiterte Klassierung, welche weitere Varianten des Anschlussmittels für Finalsätze enthält.

[4] Publikationen aus dem SADS-Datensatz: http://www.ds.uzh.ch/dialektsyntax/publikationen.html, Zugriff: 22.4.2011

Die weiteren drei Fragen zum Finalanschlussmittel wurden der Struktur der ersten Frage folgend klassiert.

2.4.2. B: Komparativ

Das Phänomen „Komparativ" behandelt Vergleichswörter, sprich jene Wörter, die verwendet werden, um einen Vergleich zwischen zwei Dingen darzulegen. Für eine genaue Erklärung sei auf Friedli (2005) verwiesen.

Die räumliche Grundhypothese heisst: „Es gibt eine dominante Variante, die überall vorkommt und dazu einzelne Varianten, die kleinere Areale bilden."

Die drei Ankreuzfragen III.22, III.25 und III.28 befassen sich alle mit diesem Phänomen. Es wird auch hier eine Einteilung in zwei Klassierungen vorgenommen, wobei sich die eingeschränkte Klassierung nach den Hauptvarianten *als*, *wie*, *weder* und *wa(n)* nach Friedli (2005) orientiert. Die erweiterten Varianten enthalten weitere Antworten (Bsp. *weder + Zusatz*, siehe Anhang A), die gegeben wurden.

2.4.3. C: Artikelverdoppelung

Artikelverdoppelung bezeichnet eigentlich nicht das ganze Phänomen, welches in den Ankreuzfragen I.10 und II.10 sowie in der Übersetzungsfrage IV.1 erfragt wird. Genauer behandelt die letzte Gruppe von Fragen die Stellung des indefiniten Artikels in der adverbiell erweiterten Nominalphrase (Steiner 2005; 2006 & im Druck). Die Artikelverdoppelung bildet lediglich eine Möglichkeit, die im Schweizerdeutschen vorkommen kann, ab. Hinzu kommen noch der vor- und der nachgestellte Artikel. Wird im Folgenden vom Phänomen als Ganzem gesprochen, wird der Verständlichkeit wegen nur der Begriff der Artikelverdoppelung verwendet.

Das Phänomen C dient als Vergleich zu den Phänomenen A und B, da die Grundhypothese hier lautet, dass keine Areale gebildet werden.

Da in den Antworten nur drei verschiedene Varianten vorkommen, wird auf eine Unterscheidung zwischen erweiterter und Grundklassierung verzichtet.

Teil II: Visualisierung von syntaktischen Phänomenen in der Deutschschweiz

3. Methodik zur Erstellung der Flächenkarten

Das erste Kapitel des ersten Hauptteils der Arbeit befasst sich mit der Methodik, die benötigt wird, um Flächenkarten nach dem Verfahren von Rumpf et al. (2009) zu erstellen. Abbildung 3-1 ist eine grobe Skizze der übergeordneten Arbeitsschritte, die von den erfassten Antworten in der SADS-Datenbank bis zu den kartographischen Endresultaten benötigt werden. Zudem ist darin auch enthalten, welche Programme dafür verwendet werden.

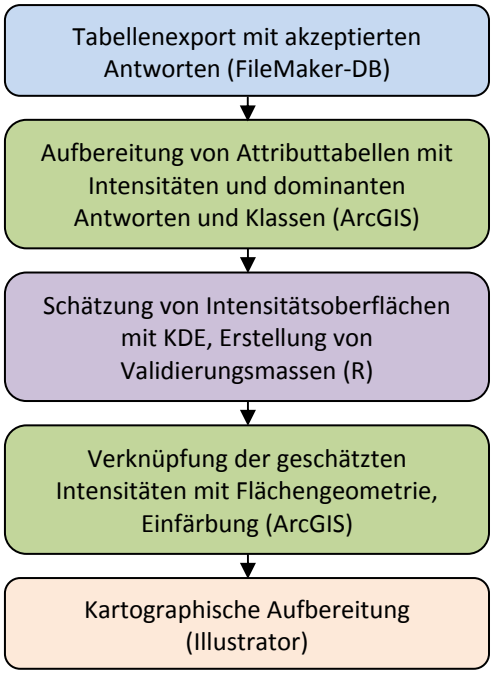


Abbildung 3-1: Flussdiagramm der Arbeitsschritte zur Erstellung von Flächenkarten

Die folgenden drei Abschnitte orientieren sich grob an diesem Ablauf und gehen detailliert auf das Vorgehen von der Tabellenaufbereitung der SADS-Daten (3.1) über die Abgrenzung des Untersuchungsgebiets (3.2) hin zur Umwandlung in Flächenkarten mithilfe der zentralen KDE (3.3) ein. Andere denkbar gewesene Methoden zur Kartenaufbereitung werden im letzten Abschnitt (3.4) erwähnt.

3.1. Aufbereitung der SADS-Daten

Die Originaldaten aus den vier SADS-Fragebogen sind in einer FileMaker-Datenbank vorliegend. Die bearbeiteten Daten entstammen einer Version vom 4. April 2010. Daraus wurden Tabellen im dBASE-Format (.dbf) exportiert und in ESRI ArcGIS und MS Excel aufbereitet für den weiteren Gebrauch in einem geographischen Informationssystem. Neben den Daten des SADS-Projektes sind zusätzlich Geodaten der Landestopographie Schweiz (Swisstopo), sowie Bevölkerungsdaten des Bundesamtes für Statistik verwendet worden. Tabelle 3-1 zeigt die verwendeten externen Datenquellen.

Daten	Datenquelle	Datenbeschreibung	Stand
SwissNames Ortschaften	Swisstopo	Punktdatensatz mit 190‘000 georeferenzierten Ortsnamen, entsprechend der Namen der LK 1:25‘000. Details unter: http://www.swisstopo.admin.ch/internet/swisstopo/de/home/products/landscape/toponymy.html, Zugriff: 19.4.2011	2007
Administrative boundaries (Vec200)	Swisstopo	Flächendatensatz, der die Schweiz nach den politischen Ebenen gliedert. Genaue Beschreibung des aktuellsten Datensatzes 2010 unter: http://www.swisstopo.admin.ch/internet/swisstopo/de/home/products/landscape/vector200.html, Zugriff: 19.4.2011	2006
Volkszählung 2000	BFS	In der Volkszählung wurden nach Wohngebäuden eingeteilt, verschiedene Merkmale über die Bevölkerung und die Gebäude aufgenommen. Detaillierte Beschreibung unter : http://www.bfs.admin.ch/bfs/portal/de/index/dienstleistungen/geostat/datenbeschreibung/volks-__gebaeude-2.html, Zugriff: 19.4.2011	2000

Tabelle 3-1: Zusätzlich zum SADS-Datensatz verwendete Daten mit Datenherkunft und Ursprungsjahr

3.1.1. Tabellen-Export aus der SADS-Datenbank

In der FileMaker-Benutzeroberfläche können Daten in so genannten „Layouts“ aufbereitet werden. Im SADS-Datensatz wurde für jede Frage ein solches Layout eingerichtet. Darin sind für die GP neben den akzeptierten Antworten zusätzliche Attribute wie deren BFS-Nummer des Wohnorts, Gemeinde- und Kantonsnamen und weitere enthalten. Für den Export muss eine Auswahl getroffen werden. Ziel ist es, die Export-Tabelle möglichst schlank zu halten und dabei trotzdem keine wichtigen Informationen zu verlieren. Dazu gehören im konkreten Fall sicher die Antworten, welche in der gewählten Klassierung vorkommen. Weiter ist die BFS-Nummer zentral, da sie später als Bindeglied zu weiteren Datensätzen wirkt und nach diesen schliesslich gruppiert wird. Neben der Wahl der Attribute muss auch sichergestellt werden, dass nur Antworten verwendet werden, die vollständig sind und den Anforderungen der SADS-Forscher entsprechen. Dafür ist in der Datenbank ein eigenes Feld „OK“ vorgesehen. Hat dieses den Wert 0, so darf diese Gewährsperson nicht berücksichtigt werden. Beispiele von unbrauchbaren Antworten sind Übersetzungen in Schriftsprache oder Fremdhilfe beim Ausfüllen (Bucheli Berger 2008).

Für den Export wurden alle Gewährspersonen mit nicht gültigen Antworten, sprich jene mit „0“-Werten in der OK-Spalte, abgewählt. Danach konnte in einem Dialog ausgesucht werden, welche Attribute exportiert werden sollen. Exportiert wurden alle Antworten, die in den in Anhang A nachschlagbaren Klassierungen Eingang finden. Zudem wurden die Nummern der Gewährspersonen exportiert. Der Export der BFS-Nummer ist mit diesem Vorgehen nicht direkt möglich, weshalb diese Information erst später hinzugefügt wurde. Alle anderen Felder der Datenbank wurden beim Export weggelassen. Eine Ausnahme bildet die Frage I.6, bei welcher das „SONST“-Feld auch exportiert wurde, weil darin mit der Nummer „37“ eine Antwort vorkommt, die in der Klassierung der Phänomene Einzug hält.

Als Exportformat wurde eine .dbf-Tabelle gewählt, da diese in einem kurzen Test, verglichen mit der in FileMaker implementierten .csv-Exportfunktion, die auch denkbar gewesen wäre, besser weiterverarbeitet werden konnte. Nach Export dieser Tabellen wurde die FileMaker-Datenbank für das weitere Vorgehen nicht mehr gebraucht.

3.1.2. Tabellenaufbereitung im Texteditor

Nachdem die gewünschten Daten aus der Original-Datenbank in .dbf-Tabellen abgelegt wurden, mussten diese für die weitere Anwendung mit einer BFS-Nummer versehen und so aufbereitet werden, dass sie in einem Geographischen Informationssystem weiterverwendet werden können und schliesslich auch für automatisierte Datenanalysen geeignet sind.

Im ersten Schritt mussten die Felder gesäubert und in einheitliche Datenformate umgewandelt werden. dBASE-Tabellen können in der verwendeten Excel-Version 2007 zwar geöffnet, jedoch nicht als solche

abgespeichert werden, weshalb sie als .csv abgelegt wurden. Dieses Format ist ein Text-Format, was den entscheidenden Vorteil hat, in einem Texteditor lesbar und bearbeitbar zu sein. Dadurch konnten die Exporttabellen manuell für die Weiterverwendung in ArcGIS aufbereitet werden. Ziel war eine Tabelle, die pro GP nur noch „0"- und „1"-Einträge für die Antwortfelder enthält und damit die Akzeptanz der zur Verfügung stehenden Antworten darstellt. Verschiedene Anpassungen waren dazu nötig.

Trennzeichen

Das von Excel verwendete Trennzeichen, das Semikolon, kann von ArcGIS nicht erkannt werden, weshalb dieses als erstes durch ein Komma ersetzt wurde. Die Frage I.6 ist hier ein Sonderfall, da im SONST-Feld Kommata enthalten sind, welche durch ein anderes Sonderzeichen ersetzt werden mussten, bevor im csv die Trennzeichen von Semikolons in Kommata umgewandelt werden durften.

Felder mit leeren Einträgen

In den Tabellen existieren teilweise leere Zellen, die durch einen „0"-Eintrag ersetzt werden mussten. Dies konnte, wie die Ersetzung der Semikolon-Zeichen, mit einem einfache „Suchen-Ersetzen"-Befehl im Texteditor umgesetzt werden. Einen Spezialfall bilden leere Einträge, die sich im letzten Feld befinden. Um diese zu ersetzen, benötigt der verwendete Editor eine erweiterte „Suchen-Ersetzen"-Funktion, die auch Zeilenumbrüche unterstützt. Im verwendeten Texteditor (Notepad++)[1] wurde damit nach [`,\r`] gesucht und mit [`,0\r`] ersetzt.

Damit die Konsistenz gewahrt wird, musste das „SONST"-Feld der Frage I.6 angepasst werden. Felder, die die Nummer 37 enthalten, wurden durch eine „1" ersetzt, alle anderen mit einer „0".

3.1.3. Tabellenaufbereitung im Geographischen Informationssystem

Die nach einzelnen GP geordneten Tabellen mit Einzelantworten mussten nun in nach BFS-Nummern aggregierte Tabellen umgewandelt werden. Diese enthalten relative Häufigkeitswerte für die einzelnen Antworten und schliesslich auch für die Klassierung nach verschiedenen Varianten. Dazu wurden die .csv-Daten, deren Aufbereitung im vorangegangenen Unterabschnitt (3.1.2) beschrieben ist, in einem Geographischen Informationssystem, hier ArcGIS 9.3, geöffnet. Auf dem auf der Software-CD auffindbaren Dokument *Tabellenaufbereitung GIS.pdf* im Ordner *1_Tabellenaufbereitung* wird die Aufbereitung einer Tabelle am Beispiel der Frage I.1 im Detail gezeigt.

Akzeptanzwerte pro Person berechnen

Für jede GP musste pro Antwort ein Akzeptanzwert ausgerechnet werden, welcher den relativen Anteil der spezifischen Antwort an allen akzeptierten Antworten angibt. Akzeptiert eine GP beispielsweise drei Antworten, so erhält jede dieser Antworten einen Akzeptanzwert von einem Drittel.

Zuerst wurde die Attributtabelle geöffnet und als .dbf-Tabelle exportiert. Dieser wurden neue Felder hinzugefügt. In einem ersten Feld wurden die Anzahl Antworten pro Gewährsperson ausgerechnet, indem die Werte aller Antwortfelder zusammengezählt wurden. Für jede Antwort musste zusätzlich ein Akzeptanzfeld geschaffen werden, worin ein Quotient zwischen der spezifischen Antwort und der Summe aller Antworten gebildet wurde. Dabei musste beachtet werden, dass nicht durch 0 geteilt wurde, weshalb dies zuerst mit einer Bedingung getestet wurde. Der nachfolgende VBA-Code zeigt ein Beispiel für das Berechnen des Akzeptanzfeldes für die Antwort 1 der Frage I.1 mittels `Field Calculator` in ArcGIS:

[1] http://notepad-plus-plus.org/, Zugriff: 19.4.2011

```
Dim proz as double
If [S_pro_GP] > 0 Then
proz = [F1_1] / [S_pro_GP]
else
proz = 0
end if
```

Aggregierung nach BFS-Nummer

Bisher wurden nur Tabellen, welche nach einzelnen Gewährspersonen geordnet sind, erstellt. Nun mussten diese nach Herkunftsort aggregiert werden. Die BFS-Nummern vom Bundesamt für Statistik bieten dazu eine geeignete Bezeichnung, da sie von verschiedenen offiziellen Datenkatalogen benutzt werden. Im SADS-Datensatz konnte eine Tabelle exportiert werden, mithilfe welcher den Gewährspersonen eine BFS-Nummer angebunden werden konnte.

Diese Verknüpfungstabelle wurde mit einem Relational Join an die bestehende Attributtabelle angehängt und in einer neuen .dbf-Tabelle abgelegt. Nun konnte nach der BFS-Nummer aggregiert, sprich pro Ort die Akzeptanzen der einzelnen Antworten gemittelt werden. In ArcGIS heisst das Tool dazu `Summary Statistics`. Etwas unglücklich sind die automatisch generierten Feldnamen der daraus resultierenden aggregierten Tabellen, welche mit dem Werkzeug `Table To Table` manuell noch angepasst werden mussten.

Dominante Akzeptanzen

Die bis anhin vorgenommenen Schritte waren für beide Klassierungen identisch. Nun wurde, falls vorhanden, zwischen zwei Klassierungen unterschieden, wofür jeweils separate Tabellen geschaffen wurden. Der nächste Schritt der Tabellenaufbereitung bestand darin, dominante Akzeptanzen für die Antworten, beziehungsweise Klassen, zu finden. Dafür musste die maximale Akzeptanz pro Ort ausgerechnet werden.

Um die Akzeptanzen pro Klasse zu erhalten, sind neue Felder hinzugefügt und darin die Akzeptanzen der zu den Klassen zugehörigen Antworten aufaddiert worden. Als nächstes wurde pro Ort für jede Antwort, sowie für jede Klassierung, die maximale Akzeptanz berechnet, was der an jenem Ort dominanten Antwort oder Klasse entspricht. Dazu wurde nach folgendem Algorithmus vorgegangen:

```
Für jeden Ort
   Speichere alle Felder, die Akzeptanzen enthalten, in einem Array
   Sortiere diesen aufsteigend
   Wenn der letzte Eintrag nicht grösser als 0 ist
      setze maximale Akzeptanz 0 (Klasse kommt nicht vor)
   Wenn der letzte Eintrag grösser als 0 ist
      Wenn der letzte Eintrag dem zweitletzten entspricht
            Setze die maximale Akzeptanzen = 0 (keine dominante
            Akzeptanz)
      Wenn der letzte Eintrag nicht dem zweitletzten entspricht
            Setze die maximale Akzeptanz = dem letzten Eintrag des
            sortierten Arrays
```

Dominante Antworten und Klassen

Insbesondere für die Kalibrierung der Bandbreite (Kapitel 4) ist es interessant, welche Antwort bzw. Klasse pro Ort dominant vorkommt. Folgender Algorithmus speichert mithilfe der dominanten Akzeptanz die dominante Klasse in ein neu zu erstellendes Feld.

```
Speichere die Felder, die Akzeptanzen enthalten, in einem Array [a]

Erstelle einen Klassierungsarray [k] mit den gewünschten
Klassenbezeichnungen

Für jeden Ort

   Wenn das Feld mit der maximalen Akzeptanz den Wert 0 hat

      Setze die dominante Klasse = 0 (keine dominante Klasse)

   Wenn das Feld mit der maximalen Akzeptanz nicht den Wert 0 hat

      Für alle Akzeptanzfelder im Ort

         Wenn der Eintrag an der Stelle i in [a] mit dem Eintrag im
         Feld mit den maximalen Akzeptanzen übereinstimmt

            Setze die dominante Klasse = Eintrag in [k] an i
```

Ein detailliertes, kommentiertes VBA-Script zur Errechnung der dominanten Akzeptanzen und der dominanten Antworten und Klassen ist unter *1_Tabellenaufbereitung/MaxValues.vbs* auf der beigelegten CD auf der Rückseite des Deckblattes vorzufinden.

In dieser Arbeit sind zwei Möglichkeiten für die Bezeichnung der Einträge für die dominanten Felder verwendet worden. Zuerst wurde eine Abkürzung in Anlehnung an die Bezeichnung der Antworten in den SADS-Daten verwendet. Es ergaben sich dadurch aber Probleme, da diese als Text abgespeichert werden müssen, weil sie auch Buchstaben enthalten (z.B. F1_1_SONST). Die Hauptschwierigkeit liegt darin, dass die einzelnen dominanten Antworten keine systematische Reihenfolge haben und daher schwer in automatisierte Prozesse einzugliedern sind. In R ist es später einfacher, mit Integer-Werten zu arbeiten. In einem zweiten Schritt wurde deshalb für jede Antwort ein Integer-Code gewählt, der in den Feldern mit der dominanten Antwort abgespeichert wird. Die genaue Zuordnung ist in der Phänomentabelle in Anhang A nachzuschlagen.

Das Problem ergibt sich nur bei den dominanten Antworten, bei den dominanten Klassen wurde von Beginn weg ein systematisches Vorgehen gewählt.

Wurden auch die dominanten Antworten und Klassen zu den Attributtabellen hinzugefügt, waren diese bereit für die Weiterverarbeitung nach der Methode von Rumpf et al. (2009), welche im Kapitel 3.3 vorgestellt wird. Tabelle 3-2 zeigt ein Beispiel einer fertig aufbereiteten Attributtabelle mit der BFS-Nummer, den Akzeptanzen der Antwortmöglichkeiten und klassierten Varianten (`p_1_1`, `p_1_2`, `p_1_4` & `p_1_5` bzw. `p_K1` & `p_K2`), der Anzahl GP pro Ort (`n_GP`) und den dominanten Akzeptanzen und Antworten bzw. Klassen (`i_dom_A`, `dom_A`, `i_dom_K` & `dom_K`).

Attributes of f1_1K_Attribute_SADS

BFS	p_1_1	p_1_2	p_1_4	p_1_5	n_GP	p_K1	p_K2	i_dom_A	dom_A	i_dom_K	dom_K
1021_1	0.25	0	0	0.125	8	0.25	0.125	0.25	1	0.25	1
1026_1	0.428571	0	0.142857	0.285714	7	0.428571	0.428571	0.428571	1	0	0
1039_1	0.3	0	0	0.5	10	0.3	0.5	0.5	5	0.5	2
1058_1	0.571429	0	0.142857	0.142857	7	0.571429	0.285714	0.571429	1	0.571429	1
1061_1	0	0	0	0.5	6	0	0.5	0.5	5	0.5	2
1062_1	0.333333	0	0	0	6	0.333333	0	0.333333	1	0.333333	1
1069_1	0.3	0	0.2	0.1	5	0.3	0.3	0.3	1	0	0
1086_1	0.461538	0	0	0	13	0.461538	0	0.461538	1	0.461538	1

Record: 1 Show: All Selected Records (0 out of 383 Selected) Options

Tabelle 3-2: Aufbereitete Attributtabelle der Frage I.1K (Finalanschluss)

3.1.4. Erweiterung: Berücksichtigung der Präferenz in der Tabellenaufbereitung

Im bisherigen Arbeitsablauf wurden immer nur akzeptierte Antworten berücksichtigt. Dies führt dazu, dass Aussagen über die Verteilung der dominanten Antworten ein abgeschwächtes, wenn nicht gar ein falsches Bild der Verteilung eines Phänomens abgeben. Dies **betrifft nur die Ankreuzfragen**, bei denen mehrere Antworten zur Verfügung stehen und die GP mehrere von ihnen akzeptierte Antworten und daraus die für sie am natürlichsten erscheinende Variante wählen konnten. Bei den Übersetzungs- und den Ergänzungsfragen ist davon auszugehen, dass die GP bereits die natürliche Variante gewählt haben.

In der Datenbank sind die natürlichen Antworten in einem separaten Feld mit Text-Datentyp abgespeichert, welches auch Mehrfachnennungen enthält. Dieses Feld muss ebenfalls exportiert werden. Die Tabellenaufbereitung ist bis zur Berechnung der Akzeptanzwerte identisch mit dem bis anhin gewählten Verfahren für die akzeptierten Varianten.

Da für die weitere Bearbeitung Mehrfachnennungen in einem Feld nicht berücksichtigt werden können, müssen diese eliminiert werden. Dazu wird eine neue Spalte hinzugefügt, welche den Wert „1“ erhält für Felder mit Mehrfachnennungen und den Wert „0“ für Einfachnennungen. Diese werden in Integer-Werte umgewandelt und in einem neuen Feld übernommen. Bei den übrigbleibenden Feldern mit mehreren Nennungen wird anhand der am selben Ort vorhandenen Einfachnennungen entschieden, welche Antwort dominant ist an diesem Ort und das Resultat wird manuell eingetragen.

Mit der Information, welche Antwort für eine GP natürlich vorkommt, kann eine Tabelle kreiert werden, die als Spalten alle Antwortmöglichkeiten enthält. Pro Person erhält nun das Antwort-Feld, welches der natürlichen Variante entspricht, den Wert „1“, alle anderen den Wert „0“.

Das weitere Vorgehen folgt nun wieder der Methodik der akzeptierten Varianten, beginnend mit der Aggregierung der Varianten.

3.2. Aufbereitung und Abgrenzung des Untersuchungsgebiets

Die bisher aufbereiteten Tabellen haben zwar über die BFS-Nummer einen räumlichen Bezug, sie sind jedoch nicht effektiv verortet. Sie haben noch keine Koordinaten oder eine andere geometrische Entsprechung. Für Analysen in einem GIS und im konkreten Fall für die eingeschlagene Methodik der Kernel Density Estimation sind diese Informationen essentiell, da mit geographischen Lokalitäten und Distanzen gearbeitet wird. In den bisherigen Karten des SDS und auch in ersten erschienenen Karten aus dem SADS ist die Darstellung al Punktkarten gewählt worden. In diversen dialektometrischen Arbeiten (z.B. Goebl 2001; Nerbonne 2009) wurde mit Thiessenpolygonen als Interpolationsform für Flächenkarten gearbeitet. Diese wurden auch in dieser Arbeit verwendet.

Punktdaten

Ausgangspunkt für die Georeferenzierung bildeten zum einen die Angaben im SADS-Datensatz zu den Untersuchungsorten, bei welchen vor allem eine mit einem Index versehene BFS-Nummer und eine Ortsbezeichnung wichtig sind und zum anderen die Ortspunkte im SwissNames-Datensatz der Swisstopo. Ziel war es, für alle 383 BFS-Nummern, die im SADS vorhanden sind, einen entsprechenden Punkt zu haben und in einem Punktdatensatz abzuspeichern.

In den SwissNames-Punkten sind Gemeinden, nach Bevölkerungsgrösse sortiert, in 4 Kategorien abgelegt. Die Verortung der Punkte entspricht den Bevölkerungsschwerpunkten der Gemeinden. Die Punkte sind unter anderem mit BFS-Nummern attribuiert. Wo diese jenen im SADS-Datensatz entsprechen, wurden die Punkte direkt in den Punktdatensatz übernommen. Nicht überall ist dies aber direkt umsetzbar. Im Folgenden sind Schwierigkeiten und Anpassungen aufgelistet.

Gemeindefusionen

Da die aktuellsten verwendeten SwissNames-Daten aus dem Jahr 2007 stammen, sind einige Gemeinden, die in den SADS-Daten noch existierten, durch Gemeindefusionen nicht mehr als Gemeinden mit eigener BFS-Nummer vorhanden. Tabelle 3-3 zeigt, welche Gemeinden davon betroffen sind. In den SwissNames sind diese ehemaligen Gemeinden noch als Weiler auffindbar. Eine Möglichkeit wäre es gewesen, in der Operationalisierung deren Nummern durch die fusionierten Gemeindenummern zu ersetzen. Dadurch wäre aber die Information über die geographische Herkunft der zugehörigen Gewährspersonen verfälscht und die Zahl der Untersuchungsorte reduziert worden. In dieser Arbeit wurden die betroffenen Gemeindenummern mit den Geometrien jener Punkte in SwissNames verknüpft, welche neu als Weiler klassiert sind.

Ehemalige BFS-Nr	Gemeindename vor Fusion	Gemeindenamen durch Fusion
621	Oberwichtrach	Wichtrach
1149	Willisau-Stadt	Willisau
3336	Rapperswil (SG)	Rapperswil-Jona
3692	Medels im Rheinwaldd	Splügen
6067	Reckingen	Reckingen-Gluringen

Tabelle 3-3: Von Gemeindefusionen betroffene Orte im SADS-Datensatz mit den entsprechenden Ortsbezeichnungen.

Mehrere Orte in derselben Gemeinde

Gewisse BFS-Nummern kommen in den SADS-Daten doppelt vor, sprich eine Gemeinde enthält verschiedene Punkte. Dies ist der Fall, wenn verschiedene Dorfkerne in derselben Gemeinde vorhanden sind. Deswegen wurde beim SADS-Projekt mit Indizes gearbeitet. Auch in diesem Fall mussten die Punktdaten über die Ortsnamen in den SwissNames-Punkten manuell ausgewählt werden. Die betroffenen BFS-Nummern sind in Tabelle 3-4 angegeben.

BFS-Nr (SADS)	Ort	BFS-Nr (SADS)	Ort
198_1	Uster	2299_2	Schwarzsee
198_2	Nänikon	3293_1	Mels
567_1	Kiental	3293_2	Weisstannen
567_2	Reichenbach im Kandertal	3294_1	Pfäfers
584_1	Lauterbrunnen	3294_2	Valens
584_2	Mürren	3294_3	Vättis
584_3	Wengen	3377_1	Ricken
768_1	Faulensee	3377_2	Wattwil
768_2	Spiez	3851_1	Davos
908_1	Fankhaus	3851_2	Davos-Monstein
908_2	Trub	4646_1	Ermatingen
2299_1	Plaffeien	4646_2	Triboltingen

Tabelle 3-4: BFS-Nummern von Gemeinden mit mehreren Orten, mit SADS Indizes

Inkonsistente Handhabung von Indizes

Normalerweise werden die BFS-Nummern im SADS abgebildet, indem „_1“ an die BFS-Nummer angehängt wird (Beispiel: Zernez, BFS 3746; im Datensatz 3746_1). Bei mehreren Orten, die zu einer BFS-Nummer gehören, wurde der Index am Ende der Nummer erhöht (siehe Gemeindefusionen). Es gibt aber Ausnahmen, bei denen ein höherer Index angefügt wurde, obwohl nur ein Ort einer BFS-Nummer angehört (Tabelle 3-5).

BFS_Nr	Ort
843_2	Saanen
1614_2	Luchsingen
1707_2	Rotkreuz
2701_2	Basel Stadt
4566_2	Frauenfeld

Tabelle 3-5: Nur einer Gemeinde angehörende Untersuchungsorte mit Indizes grösser als 1

Weitere Anpassungen

- Die BFS-Nummer von Ricken in den SADS-Daten entspricht jener der Gemeinde Wattwil in den SwissNames. Dort gehört Ricken aber gemäss BFS-Nummer zur Gemeinde Ernetschwil. Da in den Daten von Swisstopo der Datenpunkt für Ricken genau auf der Gemeindegrenze von Ernetschwil und Wattwil liegt, ist es aber genauso möglich, die BFS-Nummer von Wattwil zu verwenden. Ricken ist tatsächlich aufgeteilt auf die beiden Gemeinden[2]. Die Lage des Punktes wurde vom SwissNames-Datensatz genommen, die BFS-Nr jedoch dem SADS entsprechend angepasst, um eine Kompatibilität mit den SADS-Daten zu gewährleisten.
- Davos-Monstein ist in SwissNames nur als ‚Monstein' enthalten, dafür aber zweimal, das richtige musste manuell ausgewählt werden.
- Fankhaus ist nicht in den SwissNames enthalten und wurde deshalb manuell auf die Hauptkreuzung (Koordinaten: 635'646, 201'662) im Weiler der Gemeinde Trub gesetzt (Abbildung 3-2).

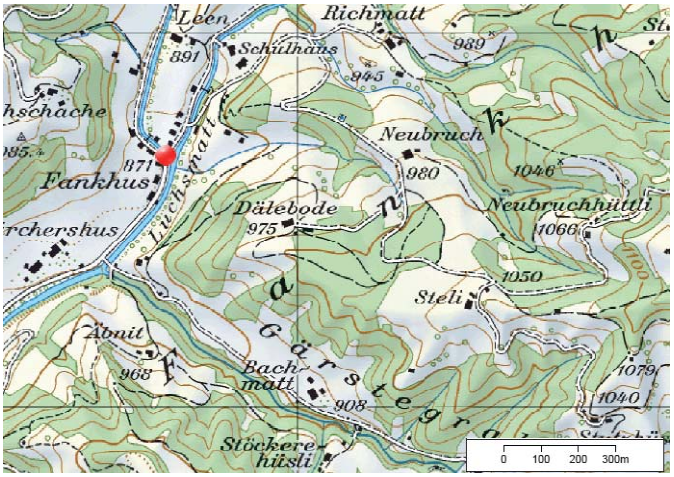


Abbildung 3-2: Georeferenzierung des SADS-Untersuchungsortes „Fankhaus"
Erstellt auf: map.geo.admin.ch, Zugriff: 12.4.2011

Das Endprodukt ist ein Punktdatensatz mit 383 Entitäten, welche jeweils mit einer BFS-Nummer versehen sind, die analog zu den im SADS-Korpus verwendeten Nummern sind.

Thiessenpolygone

Thiessenpolygone bilden ein einfaches Interpolationsverfahren. Bei einer Interpolation werden in einem Gebiet aus einer Menge von Werten an Messpunkten Attributwerte an Orten ohne Messungen abgeleitet (Burrough & McDonnell 1998). Indem um die SADS-Untersuchungspunkte Thiessenpolygone gelegt werden, kann aus den Punkten ein flächenhaftes Bild erzeugt werden. Die Umwandlung in solche Polygone ist eine Standardanwendung in einem GIS und wird hier in ArcGIS durchgeführt. Die Polygone der SADS-Untersuchungsorte wurden mit einer Maske zugeschnitten, die dem nachfolgend erläuterten Deutschschweizer Untersuchungsgebiet entspricht.

[2] Historisches zum Ort Ricken: http://www.ricken.ch/historisches/kapitel8.php, Zugriff: 16.3.2011

Untersuchungsgebiet

Als prinzipielles Untersuchungsgebiet wurde die Deutschschweiz gewählt. Diese Gemeinden wurden mit der aus der Volkszählung 2000 stammenden Arealgliederung nach Sprache vom Bundesamt für Statistik verknüpft. Nun konnte anhand der Information, wie viele Personen in einer Gemeinde welche Sprache sprechen, herausgefunden werden, wie die theoretische Sprachenverteilung in der Schweiz aussieht. Die Hauptsprache wurde dann als solche angenommen, wenn die Mehrheit der Bevölkerung an einem Ort diese Sprache spricht (Abbildung 3-3). Analog der Aufbereitung der Akzeptanzen in Unterabschnitt 3.1.3 wurde ermittelt, welche der vier Landessprachen an einem Ort dominant ist. Diese Information wurde mit dem Ortspunktedatensatz aus SwissNames verknüpft. Die Geometrien mussten nun noch um die manuell hinzugefügten Ortspunkten ergänzt werden.

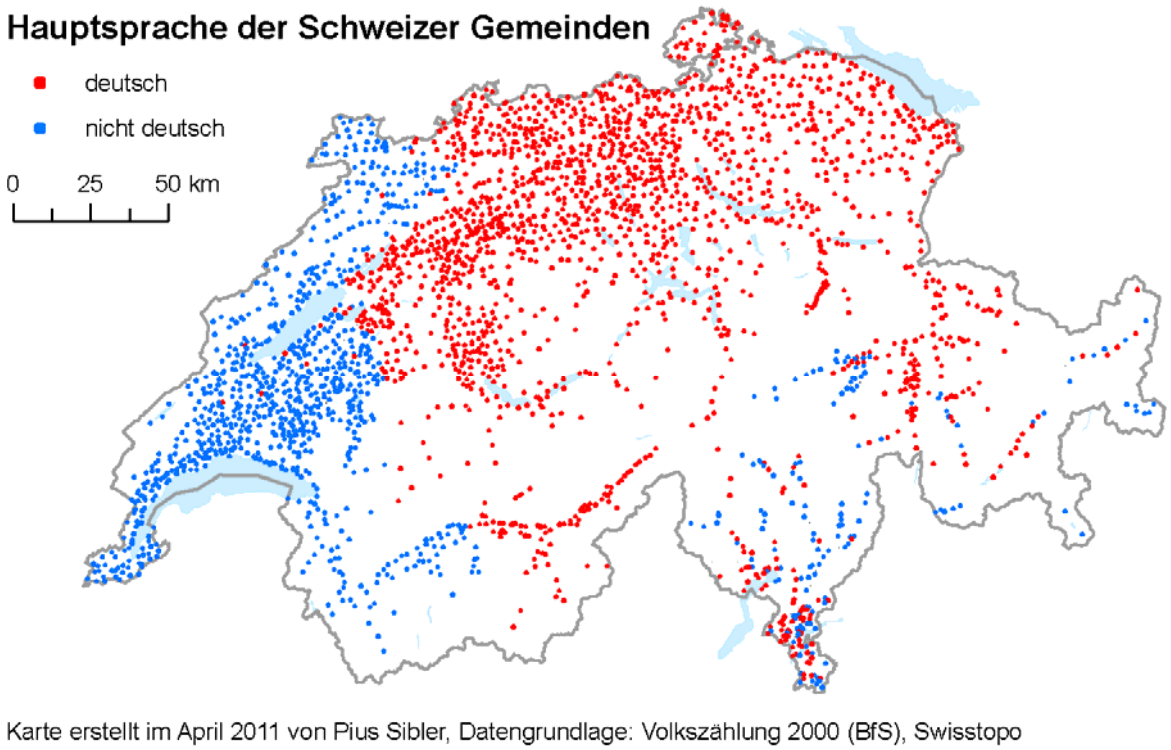


Abbildung 3-3: Schweizer Gemeinden mit aus der Volkszählung 2000 abgeleiteten hypothetischen Hauptsprachen deutsch (rot) und nicht deutsch (blau)

Nun hat man einen Punktdatensatz mit einer Attributtabelle, bei denen die vermuteten Hauptsprachen gemäss Volkszählung 2000 enthalten sind, sowie Informationen aus den SwissNames. Entscheidend ist hier die Information, in welchem Kanton ein Gemeindepunkt zu liegen kommt. So können gewisse Gemeinden als falsch klassiert angenommen werden. Auffällig sind zum Beispiel die Gemeinden im Tessin, bei denen entlang der Regionen um den Lago Maggiore und den Luganersee viele Gemeinden als deutschsprachig bezeichnet werden. Dies macht zwar durchaus Sinn, haben doch viele Deutschsprechende das Bedürfnis nach einem gemütlichen Lebensabend im Tessin. Für eine Sprachkarte, bei welche die Dialektverteilung des Schweizerdeutschen dargestellt werden soll, ist die Berücksichtigung aber nicht geeignet. Es dürfte nur wenige Personen in der Schweiz geben, die das Tessin zur Deutschschweiz zählen. Es befinden auch keine SADS-Datenpunkte in diesem Gebiet. Als deutschsprachig klassierte Gemeinden, die sich im Tessin befinden, wurden aus diesen Überlegungen weggelassen.

Einen Exoten im Swisstopo-Datensatz bildet der Staatswald Galm im Kanton Freiburg, der eine eigene BFS-Nummer (2391) besitzt und in den SwissNames als kleine Gemeinde klassiert ist. Er ist jedoch nicht bewohnt und hat aus diesem Grund auch nicht den administrativen Status einer Gemeinde[3]. Er ist

[3] BFS: Raumgliederungen der Schweiz
http://www.bfs.admin.ch/bfs/portal/de/index/news/publikationen.Document.64476.pdf,
Zugriff: 19.4.2011

ebenfalls nicht für das weitere Vorgehen berücksichtigt worden. Für einige Gemeindepunkte ist keine der vier Landessprachen dominant. Hier wurde geschaut, welche Hauptsprache in den umliegenden Gemeinden gesprochen wird und die Einteilung entsprechend angepasst.

Sind diese Änderungen abgeschlossen, wurden die nun als deutschsprachig eingeteilten Punkte in einem neuen Punktdatensatz abgespeichert. Ein weiterer Spezialfall, die Stadt Fribourg, ist nachträglich noch zu den deutschsprachigen Daten hinzugefügt worden. Die sich auf der deutsch-französischen Sprachgrenze befindende Stadt ist zwar hauptsprachlich französisch sprechend, der SADS enthält dort aber einen Untersuchungspunkt.

Die nun übrig bleibenden Gemeinden wurden in Absprache mit Professorin Elvira Glaser nochmals um Gemeinden vermindert, die als einzelne „Sprachinseln“ auftreten, sowie um die Punkte im Oberengadin, die eine sprachgeschichtlich zu junge Vergangenheit als mehrheitlich deutschsprachige Gemeinden besitzen.

Um alle Gemeindezentroide in der Schweiz wurden als nächstes Thiessenpolygone gelegt und davon die für die weitere Untersuchung nach dem eben vorgestellten Verfahren berücksichtigten deutschsprachigen Orte selektiert. Mit den Schweizer Landesgrenzen zugeschnitten bilden die Polygone schliesslich das Untersuchungsgebiet dieser Arbeit (Abbildung 3-4).

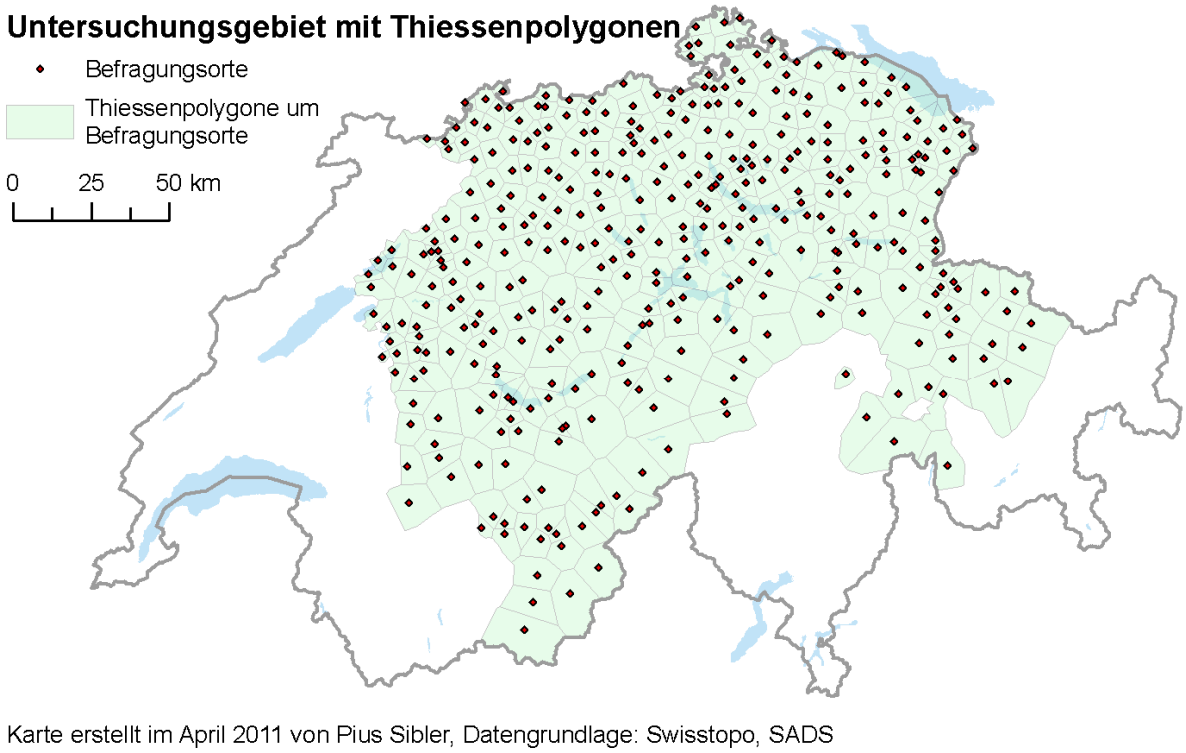


Abbildung 3-4: In Thiessen-Polygone aufgeteiltes Untersuchungsgebiet

Aus den nun vorhandenen Daten lassen sich bereits Flächenkarten erstellen, welche die Originalverteilung der dominanten Akzeptanzen zeigen. Hierzu werden die in Abschnitt 3.1 errechneten dominanten Akzeptanzen und Klassen über die BFS-Nummer mit den Thiessenpolygonen der Untersuchungspunkte verknüpft. Die so erstellten Flächenkarten der „Originalverteilung“ dienen als wichtige Vergleichsgrundlage für spätere geglättete Oberflächen.

3.3. Erstellen von Flächenkarten nach der Methodik von Rumpf et al.

Im Abschnitt 3.1 wurden ausgewählte Fragen des SADS in Tabellen konvertiert, die Akzeptanzen von Klassen und Antworten beinhalten und jeweils pro Untersuchungsort aggregiert sind. Zudem ist die maximale Akzeptanz, wie auch die dazugehörende dominante Antwort oder Klasse, errechnet worden. Darauf aufbauend werden nun Flächenkarten generiert. Der Schwerpunkt liegt auf einer Methodik, die von einer Zusammenarbeit von Augsburgern Linguisten und Ulmern Stochastikern im Rahmen des DFG-Projektes „Neue Dialektometrie mit Methoden der stochastischen Bildanalyse"[4] entworfen wurde (Rumpf et al. 2009). Die methodische Grundlage, die KDE in den Unterabschnitten 3.3.1 und 3.3.2 erklärt. Die Intensitätsschätzung von Sprachdaten von Rumpf et al. (2009) wird in 3.3.3 und die Umsetzung auf die SADS-Daten in 3.3.4 dargelegt. Im restlichen Teil dieses Abschnitts werden Erweiterungen zur Methodik beschrieben.

Von diesem Punkt an wird nur noch mit den zu Klassen zusammengefassten Akzeptanzen und Intensitäten weitergearbeitet (siehe Anhang A). Diese entsprechen den syntaktischen Varianten. Eine Analyse aller Antworten wäre auch möglich gewesen und könnte zu einem späteren Zeitpunkt durchgeführt werden. Diese Arbeit beschränkt sich aus Aufwandsgründen auf die im Vorfeld definierten Varianten der untersuchten Phänomene und macht deshalb nicht noch den Schritt auf die Ebene einzelner Antworten. Vom linguistischen Standpunkt her wäre dies sicherlich interessant, auch eine Variierung der Klassen wäre denkbar. Da diese Arbeit aber hauptsächlich in den Methoden der Geoinformationswissenschaften verankert ist, wird auf eine weitergehende linguistische Untersuchung verzichtet.

3.3.1. Kernel Density Estimation

Die Kernel Density Estimation (KDE) ist ein glättendes Interpolationsverfahren, bei dem aus einer Anzahl gegebener (Mess-)punkte eine Oberfläche erschaffen wird. Um diese Punkte wird dazu eine so genannte Kern-Dichte-Funktion gelegt. Diese kann je nach Bedarf unterschiedlich aussehen. Die Funktion bestimmt, welchen Einfluss ein Punkt, um den ein Kernel gelegt wurde, auf den Wert irgendeines Punktes im Raum hat. Nun kann für jeden Punkt im Raum anhand der Einflüsse der verschiedenen Messpunkte ein eigener Wert geschätzt werden, indem der Einfluss der Messwerte aufsummiert wird. Dadurch wird eine kontinuierliche Oberfläche geschaffen, welche der Punktdichte entspricht. Abbildung 3-5 zeigt eine Dichteschätzung mit dem KDE-Verfahren im eindimensionalen Raum. Es wird schnell klar, dass dort, wo die Punkte nahe beieinander liegen, ein höherer Schätzwert resultiert.

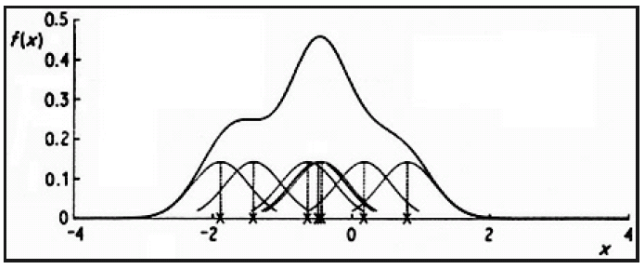


Abbildung 3-5: Kerndichteschätzung basierend auf individuellen Kernels um Untersuchungspunkte (nach Silverman 1986: 14)

Für den zweidimensionalen Raum kann man sich die Kernels als dreidimensionale Glocken vorstellen, das Prinzip bleibt jedoch gleich. Eine Interpolation von Punkten im Raum mittels KDE bildet eine flächenhafte Dichtekarte der Punkte.

[4] Projekt der Deutschen Forschungsgesellschaft (DFG) unter der Leitung von Prof. Dr. Stephan Elspaß & Prof. Dr. Werner König, Universität Augsburg:
www.philhist.uni-augsburg.de/de/lehrstuehle/germanistik/sprachwissenschaft/projekte/dialektometrie/
Zugriff: 19.4.2011

3.3.2. Parameter der KDE

Die beiden zentralen Parameter einer KDE sind die Wahl einer Bandbreite, sowie die Form der Kerndichte-Funktion selbst.

Dichte-Funktion *K*

K bestimmt, welche Kurvenform der auf die Punkte gesetzte Kernel erhält. *K* gilt im Allgemeinen als nicht entscheidend für die Resultate der KDE (Scott 1992). Sie wird deshalb in dieser Arbeit nicht im speziellen untersucht. Es wurde die Standard-Normal-Kerndichtefunktion für den zweidimensionalen Raum verwendet.

Bandbreite *h*

In vielen wissenschaftlichen Behandlungen der KDE (z.B. Silverman 1986, Jones et al. 1996) wird als entscheidende Grösse für das Resultat die Bandbreite *h* der KDE-Funktion angegeben. Sie bestimmt, wie weit sich ein Kernel ausbreitet (Silverman 1986) und dadurch, wie stark die Daten geglättet werden (Sheather & Jones 1991). Sie kann als Massstab für die KDE verstanden werden (Rumpf et al. 2009). Abbildung 3-6 zeigt den Unterschied einer interpolierten Oberfläche bei der Wahl verschiedener Bandbreiten.

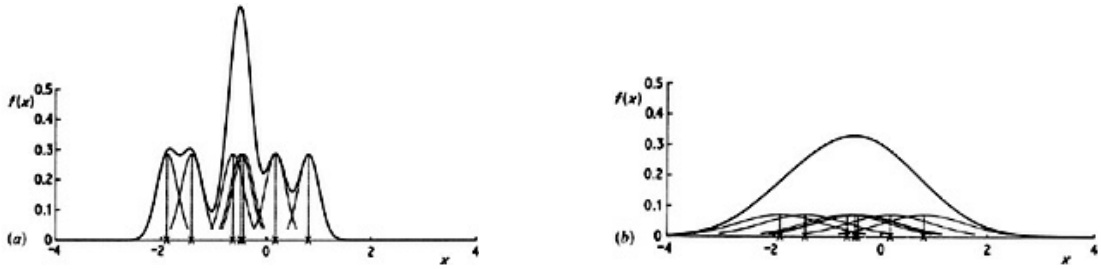


Abbildung 3-6: Unterschiedliche Glättung bei der Wahl einer kleinen Bandbreite (0.2, links) und einer grossen Bandbreite (0.8, rechts) (nach Silverman 1986: 14)

Eine grosse Debatte, wie dieser Glättungsparameter gewählt werden sollte, hat verschiedene Verfahren zu dessen Wahl hervorgebracht. Der einfachste Fall ist die Wahl einer globalen Grösse, die für jeden Punkt über das ganze Gebiet eine identische Bandbreite verwendet. Weiter existieren verschiedene automatisierte Methoden.

Die *Faustregel nach Silverman* (1986) verwendet die Formel 3-1, um eine globale Bandbreite zu definieren, welche in *A* den Interquartilsabstand und die Standartabweichung enthält. In R ist diese unter dem Befehl `bw.nrd0` aufrufbar. Eine angepasste Form nach Scott (1992), die häufiger verwendet wird, ist ebenfalls implementiert unter `bw.nrdx`.

$$h = 0.9An^{-1/5} \tag{3-1}$$

Weiter kann in R zwischen unbiased (`bw.ucv`) und biased (`bw.ucv`) cross-validation gewählt werden. Dies sind Methoden, die versuchen, den Mean Integrated Squared Error MISE durch die Kernelfunktion zu minimieren. Der MISE ist die integrierte mittlere Abweichung der Schätzwerte von den Originalwerten der Messpunkte (Jones et al. 1996).

Diese Verfahren werden von Jones et al. (1996) alle als Verfahren erster Generation beschrieben. Sie sind relativ einfach verständlich, haben aber bezüglich der Performance klare Defizite gegenüber Verfahren zweiter Generation. Ein solches Verfahren ist jenes von Sheather & Jones (1991) und ist ebenfalls in R unter `bw.SJ` implementiert. In der vorliegenden Arbeit wurde mit den verschiedenen Methoden zur Bandbreitenwahl, welche von R offeriert werden, experimentiert. Zusätzlich wurde auch mit verschiedenen manuell gewählten Bandbreiten gearbeitet. Schliesslich ist die Entscheidung auf eine manuell gesetzte Bandbreite von 10‘000 Metern für die Erstellung der Karten gefallen. Mehr dazu ist im Kalibrierungskapitel zu lesen (Kapitel 4).

3.3.3. Intensitätsschätzung von Sprachdaten mithilfe der KDE

Rumpf et al. (2009) verwenden eine angepasste KDE für eine Untersuchung von phonetischen Daten des Sprachatlas von Bayrisch-Schwaben (SBS), um Flächenkarten für die Verteilung einzelner phonetischer Phänomene zu erstellen. Für jede Variante eines Begriffs wird als erstes eine Intensität errechnet, mit welcher sie an den Untersuchungspunkten vorkommen. Anschliessend wird eine Dichteschätzung vorgenommen und zwar nur an den Punkten selbst. Für jeden Punkt wird geschaut, wie stark der Einfluss der anderen Punkte ist. Die Information über die Dichteverteilung wird genutzt, um eine Gewichtung der Intensitäten vorzunehmen. Formel 3-2 zeigt dieses Vorgehen im Detail. Die geschätzte Intensität u einer Variante x am Untersuchungspunkt t_i ist bestimmt durch die Summe aller Kernelfunktionen K für jeden Punkt und der Akzeptanzwerte l_x an den anderen Messpunkten t_j. K ist eine Funktion aus der Distanz $d(t_i, t_j)$ des Untersuchungspunktes zu allen anderen Punkten und der Bandbreite h. Der erste Term $(1/\sum K(d(t_i, t_j)/h)$ dient dabei lediglich der Normalisierung der Intensitätswerte.

$$u_x(t_i) = \frac{1}{\sum_{j=1}^{n} K\left(\frac{d(t_i, t_j)}{h}\right)} \sum_{j=1}^{n} K\left(\frac{d(t_i, t_j)}{h}\right) \cdot l_x(t_j) \tag{3-2}$$

Rumpf et al. (2009) verwenden für ihre Berechnungen eine zweidimensionale Form des weit verbreiteten Standard-Normal-Kernels(Formel 3-3), der als wichtige Eigenschaft besitzt, immer positive Werte zu liefern:

$$K(t) = \frac{1}{2\pi} e^{-\frac{1}{2}t^2} \tag{3-3}$$

Somit entsteht für jede Variante eines Begriffs eine Intensitätskarte. Die verschiedenen Intensitätskarten eines Begriffs werden danach miteinander verschnitten. Dabei wird jeweils die Variante mit der lokal höchsten Intensität für einen Punkt übernommen. Der letzte Schritt zur Erstellung der Flächenkarte ist die Verknüpfung der maximalen Intensitäten pro Ort mit Thiessenpolygonen im Untersuchungsgebiet. Jedem Polygon wird eine der dominanten Variante entsprechende Farbe zugeteilt und die Helligkeit nach der maximalen Intensität gewählt. Als Abgrenzung wählen Rumpf et al. (2009) die konvexe Hülle um die Untersuchungspunkte. Abbildung 3-7 zeigt eine so entstandene Flächenkarte für den Begriff „Kartoffelkraut“. Es sind hier offensichtlich vier Varianten dominant.

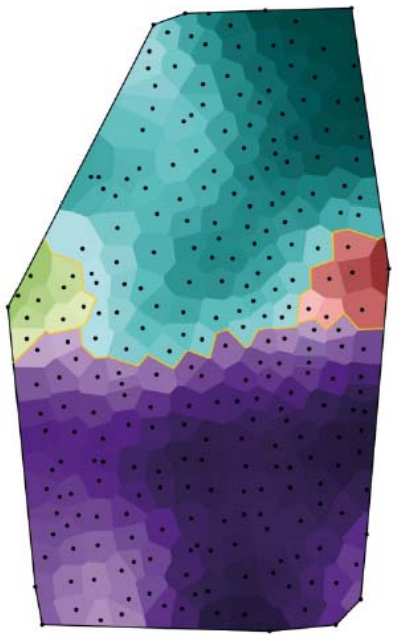


Abbildung 3-7: Flächenkarte der dominanten Intensitäten des Begriffs „Kartoffelkraut“ aus dem SBS (nach Rumpf et al. 2010: 94)

3.3.4. Umsetzung der Methode von Rumpf et al. auf die SADS-Daten

Der grosse Vorteil der Methodik von Rumpf et al. (2009) liegt darin, dass nominale Dialektdaten in eine metrische Skala übersetzt und so zu Flächenkarten interpoliert werden können. Im Abschnitt 3.1 wurden die Daten so aufbereitet, dass diese Methode hier nun auch angewendet werden kann. Es bestehen für jede Frage Varianten, deren Akzeptanzen für jeden Untersuchungsort bekannt sind, sowie die dazu gehörenden Geometrien in Form von Punkten und Thiessenpolygonen. Für die Umsetzung kommt das Open Source Softwarepaket R ins Spiel.[5] Es bietet eine Vielzahl von Möglichkeiten zur Datenanalyse und zur Datenmanipulation. Unzählige Packages machen es zu einem mächtigen Statistiktool. Speziell für die Geoinformationswissenschaft interessant ist die Möglichkeit, Geodaten in R zu importieren und exportieren. Auch können .dbf-Tabellen eingelesen werden. Die eigene Programmiersprache S bietet zudem die Freiheit, eigene Funktionen zu entwerfen, sowie bestehende Funktionen einzubinden und zu erweitern. Mit R wurden alle Berechnungen durchgeführt. Die Darstellung wurde danach in ArcGIS aufbereitet und schliesslich in Adobe Illustrator druckreif gemacht.

Die Tabelle 3-6 zeigt den Ablauf, wie die Methodik von Rumpf et al. (2009) in R umgesetzt wurde. Der komplette R-Code zu allen untersuchten Fragen ist auf der Software-CD dieser Arbeit beigelegt (*2_KDE*). Basierend auf diesem Ablauf werden die einzelnen Arbeitsschritte erläutert.

Arbeitsschritte		Softwareumgebung
1.	Import der Tabelle mit den Akzeptanzen	R
2.	Import der Punktgeometrien	
3.	Verknüpfung der Akzeptanzwerte mit den Punktdaten	
4.	Durchführung der KDE für alle Klassen	
5.	Bestimmung der maximalen Intensitäten pro Ort	
6.	Zuweisung der dominanten Varianten	
7.	Export der errechneten Intensitäten und der dazugehörigen maximalen Intensitäten als Attribute zu einem Punktdatensatz	
8.	Import der Punktdaten aus R	ArcGIS
9.	Verknüpfung mit den Thiessenpolygonen	
10.	Einfärbung nach dominanter Klasse, Helligkeit nach Intensität	
11.	Kartographische Aufbereitung	Illustrator

Tabelle 3-6: Arbeitsschritte zur Umsetzung der Methode von Rumpf et al. mit der jeweils genutzten Softwareumgebung

1. Import der Tabelle mit den Akzeptanzen der Klassen

In R werden die in Abschnitt 3.1 aufbereiteten Tabellen mit den einzelnen Akzeptanzen importiert und als so genannte DataFrames (df) abgespeichert. Diese bilden die fundamentale Datenstruktur in R (Venables et al. 2010) und können vereinfacht als Matrix mit Objekten und zugehörigen Attributen, in R Variablen genannt, verstanden werden (Bivand et al. 2008). Für den Import von .dbf-Tabellen muss das Package `foreign`[6] aktiviert sein.

2. Import der Punktgeometrien

Um die Punktgeometrien der SADS-Untersuchungs-Lokalitäten importieren zu können, muss das Package `rgdal`[7] geladen sein. Nun muss als erstes ein Koordinatensystem geladen werden. Die Koordinatensysteme werden durch einen Code von der European Petroleum Survey Group Geodesy, dem EPSG-Code bestimmt[8]. Das CH_1903-Koordinatensystem, welches für die verwendeten Geodaten gebraucht wird, hat den Code 4149.

[5] The R Project for Statistical Computing: http://www.r-project.org/, Zugriff: 19.4.2011
[6] Package `foreign`: http://cran.r-project.org/web/packages/foreign/index.html, Zugriff: 19.4.2011
[7] Package `rgdal`: http://cran.r-project.org/web/packages/rgdal/index.html, Zugriff: 19.4.2011
[8] EPSG: http://www.epsg.org/, Zugriff: 19.4.2011

Nun können die Shapefiles mittels *readOGR* eingelesen und in einen SpatialPointsDataFrame (spdf) umgewandelt werden. Dazu ist das Package `maptools`[9] notwendig. Das spdf Datenformat hat den Vorteil, dass damit in R, nebst den üblichen Tabellenkalkulationsmöglichkeiten eines df, zusätzlich auch verschiedene geoanalytische Funktionen durchführbar sind. So können aus spdf direkt wieder Shapefiles exportiert werden. Wichtig ist beispielsweise das Extrahieren von Koordinaten und daraus das Erstellen einer euklidischen Distanzmatrix. Dies wird in Schritt 4 auch gemacht.

3. Verknüpfung der Akzeptanzwerte mit den Punktdaten

Die spdf aus Schritt 2 können nun mit den Attributtabellen aus Schritt 1 über die BFS-Nummer verknüpft werden. So ist sichergestellt, dass sich die Attribute und die Geometrien jeweils auf dieselben Punkte beziehen.

4. Durchführung der KDE für alle Klassen

Der Schlüsselschritt ist die Umsetzung der KDE. Diese wird auf Basis der Methodik von Rumpf et al. (2009) realisiert, indem eine eigene Funktion dafür geschrieben wird. Eine zentrale Rolle bei der Implementierung bildet, wie bereits erläutert, die Wahl der Bandbreite h. Grundsätzlich kann, wie dies von Rumpf et al. (2009) auch verwendet wurde, eine globale Bandbreite für alle Punkte gewählt werden. Es gibt aber auch Vorgehensweisen, in denen h an die Verteilung der Punkte angepasst wird. Die gewählte Implementierung lässt mit wenigen Abänderungen sowohl globale als auch dynamische Bandbreiten zu. Für die dynamischen Bandbreiten ist aus den Koordinaten der Punktgeometrien mithilfe der Funktion `rdist()` eine Distanzmatrix der euklidischen Distanzen zueinander zu erstellen. Basierend darauf kann für jeden Punkt mithilfe von im Package `stats` implementierten Methoden[10] eine individuelle Bandbreite errechnet werden.

Damit für jede Klasse eine separate KDE durchgeführt werden kann, müssen die Klassen in einzelne df gespeichert werden. Für all diese df wird nun in einer Schlaufe jeweils die KDE-Funktion angewendet und deren Resultate werden in einem automatisch generierten df abgelegt.

Dieser df wird danach mit der ursprünglichen Tabelle wieder verbunden, mit Koordinaten verknüpft und anschliessend in einen spdf umgewandelt. Für spätere Analysen und für die Validierung wird nun ein erster Datensatz mit den interpolierten Intensitätswerten für alle Klassen als Punkte-Shapefile exportiert.

5. Bestimmung der maximalen Intensitäten pro Ort

Basierend auf dem df mit allen geschätzten Klassen-Intensitäten wird pro Ort die maximale Intensität bestimmt. Dazu dient die Funktion (*createmax*), die in einem df zeilenweise den Maximalwert bestimmt und in einen df ablegt.

6. Zuweisung der dominanten Varianten

Die dafür implementierte Funktion (*createmaxdf)* erstellt einen neuen df, geht wiederum zeilenweise durch den df mit allen Klassen-Intensitäten und vergleicht diese mit den im vorhergehenden Schritt berechneten Maximalintensitäten. Stimmen die beiden Werte überein, werden die Intensität und die zugehörige Klasse, welche der gerade aktuellen Stelle in der Schlaufe entsprechen, übernommen. Zusätzlich wird auch gleich ein Helligkeitswert aus der Intensität erzeugt, welcher durch die einfache Formel $100 - (Intensität * 100)$ definiert ist. Dieser Wert wird aus kartographischen Überlegungen und aufgrund der Verwendung von ArcGIS für die Kartenaufbereitung gewählt. In ArcGIS werden Werte von 0 (dunkel) bis 100 (hell) und damit Prozentwerte verwendet. Hohe Intensitäten, wie sie bis anhin implementiert wurden entsprechen Werten von 0 bis 1. Aus kartographischer Sicht ist es intuitiver, dunkle Helligkeitswerte für hohe Intensitäten zu wählen, weshalb diese Werte invertiert werden.

[9] Package `maptools`: http://cran.r-project.org/web/packages/maptools/index.html, Zugriff: 19.4.2011

[10] Bandbreitemethoden, im `stats` package installiert: `bw.nrd0, bw.nrdx, bw.bcv, bw.ucv, bw.SJ`: http://stat.ethz.ch/R-manual/R-devel/library/stats/html/bandwidth.html, Zugriff: 19.4.2011

7. *Export der errechneten Intensitäten und der dazugehörigen maximalen Intensitäten als Attribute zu einem Punktdatensatz*

Die in Schritt 6 generierten df werden mit Geometrien und den in Schritt 1 importierten Attributen versehen, in spdf umgewandelt und in ein Shapefile mit Punktegeometrien exportiert.

8. *Verknüpfung mit den Thiessenpolygonen*

In ArcGIS werden die in R aufbereiteten Daten geladen und mit den Thiessenpolygonen über die BFS-Nummer verknüpft.

9. *Einfärbung nach dominanter Klasse, Helligkeit nach Intensität*

Als nächstes werden die Thiessenpolygone nach der dominanten Klasse eingefärbt. Die Farben stammen von der in kartographischen Kreisen anerkannten Color Brewer[11] Homepage. Die gewählte Farbpalette ist für qualitative Daten geeignet und optimiert für den Ausdruck auf Papier. Es wird eine möglichst intensive Palette gewählt, um die Helligkeitsunterschiede durch die Intensitäten besser wahrnehmbar zu machen. Der Helligkeitswert ist direkt durch den in Schritt 6 errechneten prozentualen Intensitätswert definiert.

10. *Kartographische Aufbereitung*

Der letzte Schritt zur Kartengenerierung geschieht im Grafikprogramm Adobe Illustrator. Dort werden die in ArcGIS produzierten Karten mit kartographischen Gestaltungsmitteln, wie der Legende, einem Titel und Quellenangaben versehen.

Eine vereinfachte Darstellung des Prinzips, wie Flächenkarten nach dieser Methode erstellt werden, ist in Abbildung 3-8 am Beispiel der eingeschränkten Klassierung der Frage I.1 zu sehen.

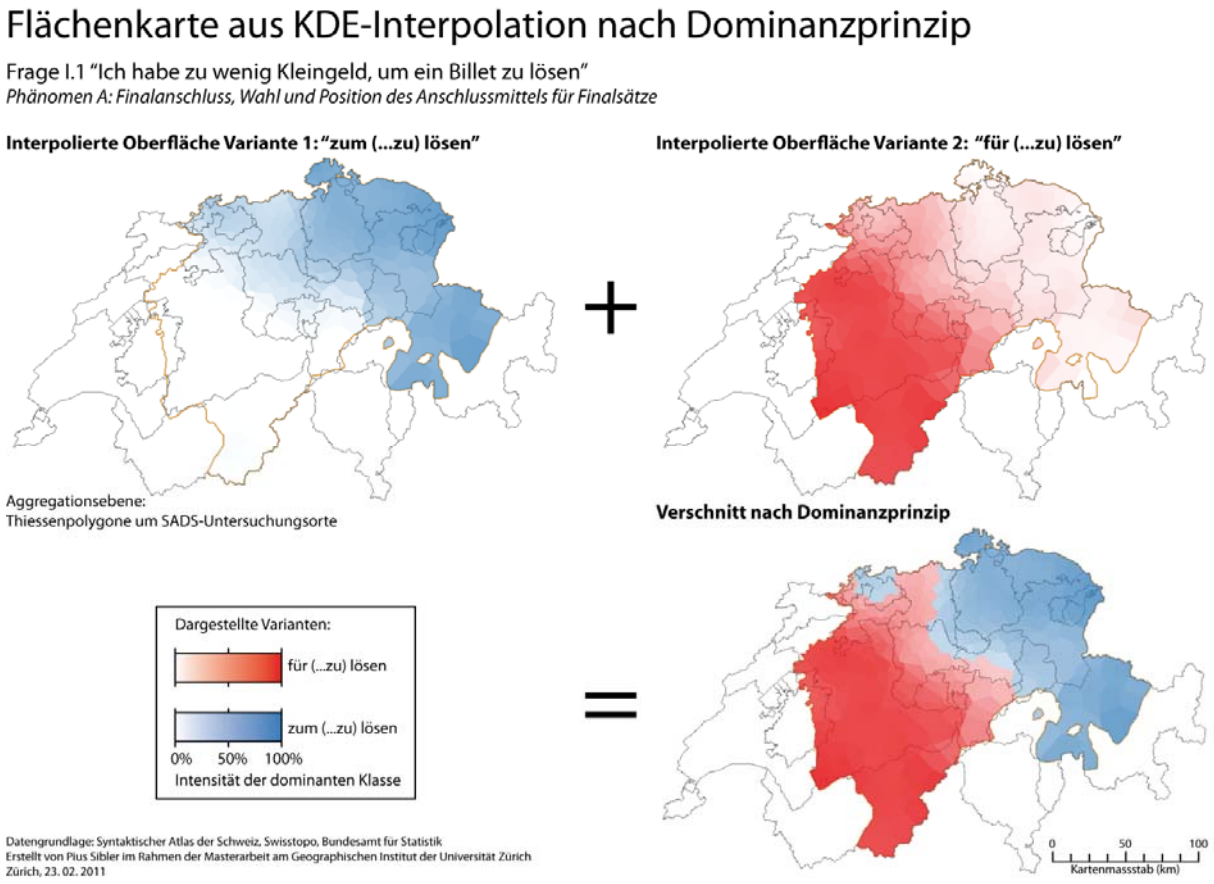


Abbildung 3-8: Die mithilfe von KDE geschätzten Intensitätskarten für jede Klasse werden miteinander verschnitten, indem jeweils die Klasse mit der dominanten Intensität übernommen wird.

[11] Color Brewer 2.0: http://colorbrewer2.org, Zugriff: 19.4.2011

3.3.5. Erweiterung: Aggregierung der Flächenkarten nach Deutschschweizer Gemeinden

Um eine höhere Punktdichte zu erreichen, werden nebst den Intensitäten der SADS-Untersuchungspunkte auch jene für die Gemeindepunkte der Deutschschweizer Gemeinden geschätzt. Dazu bedarf es einer geringfügigen Anpassung bei der Umsetzung der KDE-Formel (3-2). Der Einfluss der Intensitäten der Messpunkte wird nun nicht mehr ausgehend von den Messpunkten mit den Distanzen zu den anderen Untersuchungspunkten gewichtet. Neu wird von allen Gemeindepunkten ausgegangen und deren Distanzen zu den Messpunkten. Die Erweiterung ist bereits in allen R-Skripten im Ordner *2_KDE* auf der CD eingebaut.

3.3.6. Erweiterung: Berücksichtigung der Personenzahl pro Untersuchungsort

In der SADS-Grundlage sind die Anzahl GP je nach Untersuchungspunkt und Frage verschieden (siehe Abbildung 2-5). Dies schlägt sich auch in der Intensitätsschätzung nieder. Personen eines Ortes mit wenig GP haben deshalb einen grösseren Einfluss auf die anderen Messpunkte als solche an einem Ort mit vielen anderen Personen.

Eine Gewichtung nach der Anzahl Gewährspersonen pro Ort ist deshalb eine denkbare Alternative. Es wird dazu die mittlere Personenzahl errechnet und ein Verhältnis g (Formel 3-4) zu den tatsächlich an einem Ort befragten Personen gebildet. g ist grösser als 1 bei Orten mit mehr befragten Personen als dem Durchschnitt und kleiner bei unterdurchschnittlichen Personenzahlen.

$$g_j = \frac{Anz.\ GP\ an\ Ort\ j}{\varnothing\ Anz.\ GP} \tag{3-4}$$

Dieser Quotient wird als Gewichtung in der Formel 3-2 eingesetzt. Damit keine Intensitäten über 1 entstehen, muss zur Normalisierung dieser Quotient dem Nenner des ersten Terms der Originalformal hinzugefügt werden. Die resultierende Formel mit dem Gewichtungsparameter g_j sieht damit wie folgt aus:

$$u_x(t_i) = \frac{1}{\sum_{j=1}^{n} K\left(\frac{d(t_i,t_j)}{h}\right)\cdot g_j} \sum_{j=1}^{n} K\left(\frac{d(t_i,t_j)}{h}\right)\cdot l_x(t_j)\cdot g_j \tag{3-5}$$

Das Skript mit den angepassten Formeln für die Frage I.1K ist der Software-CD beigelegt (*2_KDE/Erweiterungen/*).

3.3.7. Erweiterung: Kartengenerierung für ein ganzes Phänomen

Im bisherigen Vorgehen sind Fragen immer getrennt voneinander betrachtet. Nerbonne (2009) beurteilt Karten, welche aus einem grossen Korpus verschiedener Phänomene aggregiert werden, als geeigneter für linguistische Beurteilungen von Regionen. Davon inspiriert wird prototypisch für das erste Phänomen eine gemeinsame Tabelle erstellt, die aus allen vier untersuchten Fragen zusammen besteht. Da die Klassierung in allen vier Fragen identisch ist, lässt sich eine solche Phänomentabelle mit relativ geringem Zusatzaufwand erstellen. Es müssen lediglich alle Intensitätswerte der vier Fragen klassenweise miteinander verknüpft werden. Dabei wird darauf geachtet, dass die Intensität nach Personenzahl pro Ort gewichtet wird. Auch dieses Skript ist auf der CD zu finden (*2_KDE/Erweiterungen/*).

3.3.8. Erweiterung: Ausweitung auf die dritte Dimension

Bisher ist die Intensität einer Variante immer als Helligkeitswert repräsentiert worden. Dadurch entstehen zweidimensionale Flächenkarten, die jedoch nur die Verteilungen der dominanten Varianten zeigen. Eine ganzheitliche Visualisierung aller Varianten mit allen Intensitätsausprägungen ist nicht möglich, es können auf zwei Dimensionen nicht mehrere Flächen übereinanderer dargestellt werden. Deshalb wird eine dritte Dimension eingeführt. Anstatt mit Helligkeiten werden die Intensitäten nun mit einem Höhenwert versehen. In ESRI ArcScene lassen sich dadurch in kurzer Zeit dreidimensionale Flächenkarten erstellen. Die Methodik zur Schätzung der Intensitäten bleibt gleich, lediglich die Darstellung ändert sich.

Das Koordinatensystem im GIS orientiert sich an der Ausdehnung des Untersuchungsgebiets, welche für die Deutschschweiz in der Grössenordnung von einigen hundert Kilometern liegt. Die Intensitätswerte bewegen sich in einer Spanne von 0 bis 1. Sie müssen deshalb in eine Pseudohöhe umgewandelt werden, damit optisch ein Höhenunterschied erkennbar wird und werden deshalb mit einem Faktor 100‘000 multipliziert.

3.4. Weitere Methoden

Der Schwerpunkt in diesem Teil der Arbeit liegt klar auf der Umsetzung der Methodik von Rumpf et al. auf die Daten des SADS. Es gibt aber auch andere Ideen, wie Dialektkarten aus syntaktischen Daten erstellt werden können. Goebl (z.B. 2007) und Vertreter der holländischen Dialektometrie (z.B. Nerbonne et al. 2008) gehen für die Erstellung von ihren Karten immer von linguistischen Distanzen aus. Diese werden bei Rumpf et al. (2009) mit dem Fokus auf Intensitäten elegant umgangen. Spruit (2008) zeigt, dass sich dialektometrische Methoden von Goebl auf syntaktische Daten umsetzen lassen.

Das Hauptproblem von Syntaxdaten in Bezug auf die Umsetzung von Distanzmatrizen ist deren nominale Skalierung. Es fehlt ein standardisiertes Schema, anhand dessen Syntax beschrieben werden kann. In der Phonetik besteht mit der phonetischen Umschrift ein solches. Nerbonne & Heeringa (1997) nutzen dies, um daraus die Levenshtein-Distanz (LD) zu berechnen (siehe Abbildung 2-3). Dies macht eine objektive Klassierung, beispielsweise über Multidimensional Scaling (MDS), möglich.

Gewisse Distanzmasse sind aber auch für syntaktische Daten möglich. Eine vorgängige Klassierung der Phänomene nach Varianten oder Klassen ist aber nicht zu umgehen und eine absolute Objektivität damit nicht möglich. Gemäss Spruit (2008) sind für syntaktische Daten keine numerischen Distanzen, wie die LD möglich, sondern nur so genannte Ähnlichkeitsmasse. Dies bedeutet, dass zwei Varianten entweder ähnlich oder unähnlich sind. Eine Abstufung, wie nahe beieinander zwei Varianten liegen, wie dies die LD erlaubt, ist hier nicht möglich. Dieses Unterkapitel soll konkret anhand der Ansätze der Hamming-Distanz (HD) und des relativen Intensitätswertes (RIW) zeigen, wie sich Distanzmatrizen auch aus den in dieser Masterarbeit verwendeten Daten erstellen lasen.

3.4.1. Hamming-Distanz

Die HD (Hamming 1950) für syntaktische Daten geht aus von den untersuchten Dialektpunkten. Sie wird jeweils um 1 erhöht, wenn an einem Ort eine Variante auftritt und an einem anderen nicht. Spruit (2008) übersetzt sie in folgenden Algorithmus, der einfach in R umgesetzt werden kann:

```
Für jedes Dialektpaar A und B
    Für jede Variante
        Wenn die Variante in A und nicht in B vorkommt
        Oder wenn sie in B und nicht in A vorkommt
        Erhöhe die Distanz der Dialektpaare A und B um 1
```

Tabelle 3-7 zeigt beispielhaft die Berechnung der HD zwischen drei Orten mit den zwei Finalanschluss Varianten *zum* und *für*, die entweder an einem Ort vorkommen (1) oder nicht (0).

	Ort A	Ort B	Ort C
Variante *zum*	1	1	0
Variante *für*	0	1	1
HD zwischen A&B:	1		
HD zwischen A&C:	2		
HD zwischen B&C:	1		

Tabelle 3-7: Berechnung der Hamming-Distanz zwischen drei Orten mit zwei verschiedenen Varianten

Im auf der Software-CD beiliegenden R-Skript im Ordner *3_syntaktische_Distanzmasse* ist die HD für die Frage I.1E implementiert. Dabei wird in R eine Distanzmatrix erstellt, welche für jede Verbindung der Messpunkte die HD ausrechnet.

3.4.2. Relativer Intensitätswert

Der RIW ist ein Ähnlichkeitsmass, welches von Goebl (1984) erstmals beschrieben wurde. Es setzt die Übereinstimmung von Varianten eines Dialektpaares ins Verhältnis zu allen vorkommenden Varianten der beiden Dialektorte. Dies kann wie folgt in einen Algorithmus übersetzt werden (Spruit 2008):

```
Für jedes Dialektpaar A und B
    Für jede Variante
        Wenn die Variante in A und B vorkommt
        Erhöhe den gemeinsamen Wert um 1
        Wenn die Variante in A und nicht in B vorkommt
        Oder wenn sie in B und nicht in A vorkommt
        Erhöhe den abweichenden Wert um 1
    Teile den gemeinsamen Wert durch die Summe aus gemeinsamem und
    abweichendem Wert
```

Tabelle 3-8 rechnet den RIW für drei Orte mit zwei Varianten aus.

	Ort A	Ort B	Ort C
Variante *zum*	1	1	0
Variante *für*	0	1	1
	gemeinsam (gem)	abweichend (abw)	RIW: gem/(gem+abw)
A&B	1	1	1/(1+1)=0.5
A&C	0	2	0/(1+1)=0
B&C	1	1	1/(1+1)=0.5

Tabelle 3-8: Berechnung des Relativen Identiätswertes (RIW)

Der RIW kann zudem noch über die Auftretenshäufigkeit von ganzen Kombinationen von Varianten gewichtet werden. Dieses Mass wird von Goebl (2006) als Gewichteter Identitätswert (GIW) bezeichnet.

Spruit (2008) und Spruit et al. (2009) empfehlen, den Wert zu invertieren und verweisen auf die Feststellung von Nerbonne & Kleiweg (2007), dass selten vorkommende Varianten nicht linguistisch aussagekräftig sind. Genau diese würden in der Originalberechnung nach Goebl (1984) aber zu stark gewichtet. Auch der RIW ist versuchsweise in R umgesetzt worden und auf dem Skript im Ordner *3_syntaktische_Distanzmasse* auf der CD abgelegt.

1.3.3 Gabmap

Aus den oben erläuterten Ähnlichkeitsmassen lassen sich nun für alle Paarungen zwischen den Untersuchungspunkten linguistische Distanzen bilden und in Matrixform ablegen. Diese Matrix wird als Tabelle exportiert und in einem Texteditor noch weiter aufbereitet, sodass das auf der Basis von Peter Kleiwegs Dialektometrie-Programm RuG/L04[12] entwickelte online-Tool Gabmap[13] sie als Distanztabelle[14] erkennt. Gabmap kann aus solchen Distanzmatrizen und einer Geometrie im kml-Format MDS-Plots und Cluster-Karten erstellen. Das Tool ist jedoch dafür ausgerichtet, linguistische Variation in Datensätzen mit grosser Variantenzahl darzustellen (Nerbonne et al. im Druck). Für die beschränkte Anzahl der Varianten in den SADS-Daten machen die von Gabmap verwendeten Verfahren zur Dimensionsreduktion und Clusterfindung nur beschränkt Sinn. Für eine kombinierte Untersuchung von mehreren Phänomenen, wie sie durchaus umzusetzen wäre mit den vorhandenen Daten, könnte Gabmap aber durchaus eine interessante Option für die Erstellung von Sprachkarten, insbesondere auch für die Analyse der Sprachvariation, sein.

[12] RuG/L04: http://www.let.rug.nl/~kleiweg/L04, Zugriff: 19.4.2011
[13] Gabmap: http://www.gabmap.nl, Zugriff: 19.4.2011
[14] Informationen zu den unterstützten Distanztabellen in Gabmap:
http://www.gabmap.nl/~app/doc/manual/processing.html, Zugriff: 19.4.2011

4. Kalibrierung der Bandbreite

Wie bereits mehrmals erwähnt, beeinflusst die Bandbreite entscheidend die Charakteristik der produzierten Oberfläche. Anhand verschiedener automatisierter und manueller Methoden wird getestet, welche Vorgehensweise für die verwendeten Daten am besten geeignet ist. Der im vorangehenden Kapitel beschriebene Workflow ist deshalb für diverse globale manuelle Bandbreiten und für die fünf bereits in R implementierten automatisierten Methoden durchgespielt worden. Als Testdatensatz wird wieder die Frage I.1 verwendet. Einerseits wird eine quantitative Auswertung (Abschnitt 4.1) über den Vergleich der Intensitäten, der Klasseneinteilung und der Flächenanteile durchgeführt. Andererseits werden die geglätteten Oberflächen qualitativ (Abschnitt 4.2) über den Vergleich mit den Oberflächen der Originalintensitäten kalibriert.

4.1. Quantitative Kalibrierung

Die quantitative Kalibrierung umfasst ein Validierungsmass, welches umschreibt, wie stark sich die Intensitäten mit der Glättung durch die KDE verändern, wie auch zwei Grössen, die sich mit der Veränderung der Klassierung beschäftigen.

Validierung der Intensitäten

Beim Vergleich der Intensitäten wird mithilfe des Root Mean Squared Error (RMSE) überprüft, wie stark sich die durch die KDE geschätzten Intensitäten $\widehat{i_x}$ im Vergleich zu den Originalintensitäten i_x an den Datenpunkten verändern. Damit kann eine Aussage gemacht werden, wie stark geglättet wird:

$$RMSE = \sqrt{\frac{\sum(i_x - \widehat{i_x}}{n-1}} \tag{4-1}$$

Validierung der Klassierung

Weiter wird die mit dem Wert K_{ant} die Verteilung der Klassen überprüft, indem errechnet wird, wie gross der Anteil der gemeinsamen Klassen (K_{gem}) von originaler und interpolierter Oberfläche an allen Messpunkten n ist:

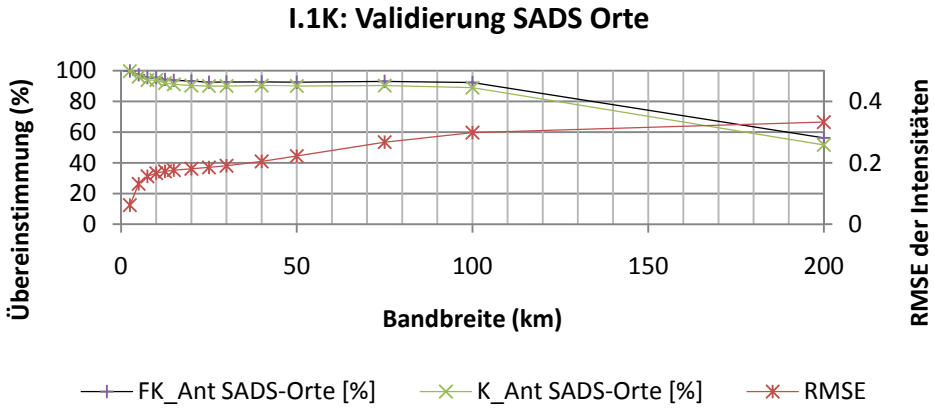
$$K_{Ant} = K_{gem} / n \tag{4-2}$$

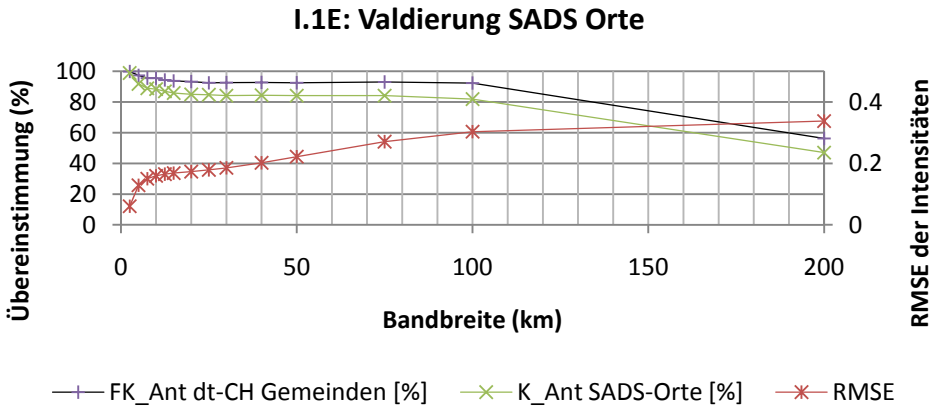
Die beiden Validierungsgrössen sind direkt in den R-Skripten zur KDE (Unterabschnitte 3.3.4 & 3.3.5) umgesetzt und werden automatisch in Tabellen ausgegeben, wenn eine KDE durchgeführt wird. Der RMSE ist sowohl für die dominante Intensität, wie auch für die Intensitäten der einzelnen Klassen implementiert. Die dritte quantitative Grösse FK_{Ant} ist der Grösse K_{Ant} angelehnt und gibt Auskunft darüber, wie gross die gemeinsame Fläche FK_{Gem} der zu den Punkten gehörenden Thiessenpolygone, gemessen an der Gesamtfläche des Untersuchungsgebiets F_{tot} ist:

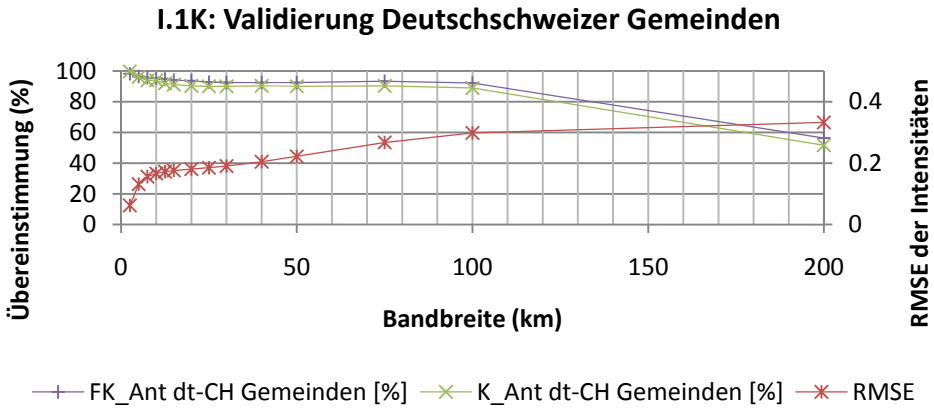
$$FK_{Ant} = FK_{gem} / T_{tot} \tag{4-3}$$

Die Berechnung dieser Grösse wird in ArcGIS mithilfe von Model Builder umgesetzt. Das Modell als ArcGIS Toolbox und der Workflow dazu sind im Ordner *4_Kalibrierung_Bandbreite* auf der beiliegenden Software-CD vorzufinden. Dabei wird unterschieden zwischen den Thiessenpolygonen um die SADS-Untersuchungspunkte und jenen um die Polygone um alle Gemeinden.

Sowohl bei K_{Ant} wie auch bei FK_{Ant} werden die Punkte mit mehreren dominanten Intensitäten nicht berücksichtigt. Diese „0"-Klassen würden das Resultat verfälschen, da der originale Intensitätswert an dieser Stelle nicht 0 ist.







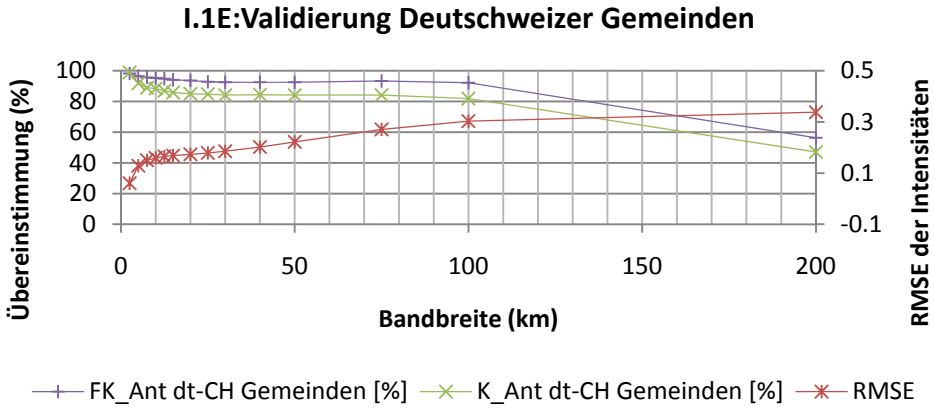


Abbildung 4-1: Quantitative Validierungswerte für die Kalibrierung der Bandbreite der KDE-Interpolation mit manuell gewählten globalen Bandbreiten der Frage I.1K und I.1E für die beiden Aggregierungsebenen (SADS-Untersuchungsorte, 2 Diagramme oben; Deutschschweizer Gemeinden, 2 Diagramme unten)

4.1.1. Manuelle Bandbreitenwahl

Abbildung 4-1 zeigt für verschiedene manuelle Bandbreiten die quantitativen Validierungsresultate für die beiden Klassierungen und beide Aggregierungsebenen der Frage I.1. Die effektiven Werte können in Tabellenform im Anhang B nachgeschlagen werden.

Alle Bandbreiten erzielen einen relativ hohen K_{Ant} und FK_{Ant}, einzig bei der sehr hohen Bandbreite von 200 km nehmen die Übereinstimmungen stark ab. Im Allgemeinen ist der Unterschied in den Flächen kleiner als jener der Punkte. Hierbei handelt es sich wahrscheinlich um ein Skalenphänomen. Bei einem Untersuchungspunkt mit grossflächigem Thiessenpolygon greift die Interpolation weniger stark als bei einem Punkt, der ein kleineres Thiessenpolygon aufweist und damit definitionsgemäss näher bei anderen Punkten liegt. Da die Interpolation über die Distanz zu den umliegenden Punkten gewichtet wird, ändert sich der geschätzte Wert eines Punktes mit nahe liegenden Nachbarn schneller. Dies führt dazu, dass die Übereinstimmung der Zentroide abnimmt und dies stärker als die Übereinstimmung der Flächen.

Die prozentuale Übereinstimmung ist zudem mit Vorsicht zu geniessen. Sie hängt davon ab, welchen Anteil die verschiedenen Klassen im Untersuchungsdatensatz aufweisen. Ist beispielsweise ein Phänomen zu 80 Prozent dominant im Originaldatensatz, so wird eine geglättete KDE-Oberfläche auch hohe Werte der Überdeckung erzielen. Abbildung 4-2 zeigt die originale Verteilung der dominanten Intensitäten für die untersuchte Frage I.1. Im vorliegenden Fall sind beide Varianten der eingeschränkten Klassierung ungefähr in der Hälfte des Gesamtgebiets dominant. Eine maximale Abweichung durch die KDE kann somit einen Grenzwert von ca. 50 Prozent nicht überscheiten. In der erweiterten Klassierung wird noch eine dritte Klasse dominant auf Kosten von Anteilen der ersten beiden Klassen. Hier ist das Potential der Abweichung deshalb etwas höher. Im konkreten Fall wirkt sich dies im Extremfall mit 200‘000 Metern Bandbreite so aus, dass praktisch nur noch eine Klasse bestehen bleibt und die Übereinstimmung beinahe dem Anteil dieser Klasse im ursprünglichen Datensatz entspricht.

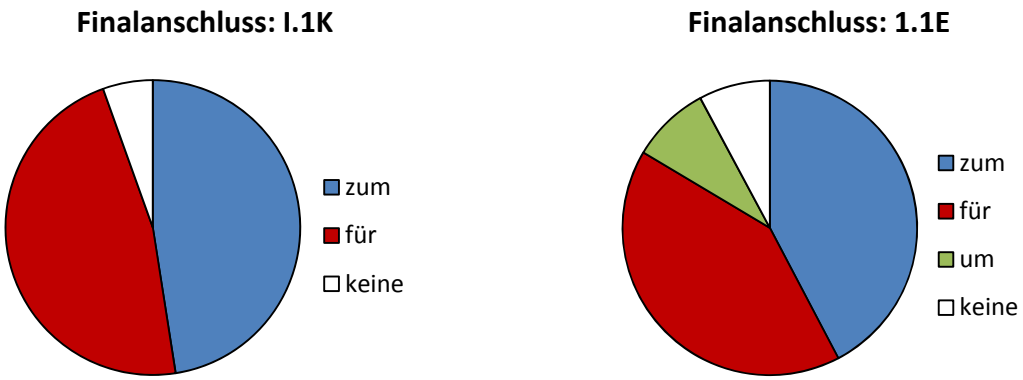


Abbildung 4-2: Verteilung der dominanten Klassen der beiden Aggregierungsebenen im originalen Datensatz der ersten Frage des Finalanschlusses

Mit zunehmender Bandbreite wird die Unsicherheit höher, sprich der RMSE erhöht sich und die prozentuale Übereinstimmung der Klassierungen nimmt ab. Die grössten Unterschiede in der Übereinstimmung werden bei den Bandbreiten bis zu etwa 10‘000 -15‘000 Metern erzeugt. Der Anstieg des RMSE schwächt sich in diesem Bereich ebenfalls ab, nimmt jedoch auch dann noch stärker zu mit zunehmender Distanz als die Übereinstimmung.

Der RMSE steht für die Intensität der Karte, welche sich selbst dann noch ändern kann, wenn die dominante Klasse durch die Glättung bereits flächendeckend identisch ist. Dies erklärt, weshalb der RMSE länger eine Änderung erfährt als die Übereinstimmungen.

Unterschiede zwischen den Klassierungen

Die Validierungswerte der erweiterten Klassierung zeigen eine leicht tiefere Übereinstimmung der Zentroide, jedoch eine etwa gleich grosse Übereinstimmung der Flächen. Die Vermutung liegt nahe, dass durch die Interpolation die erweiterte Klassierung stärker von der Glättung betroffen ist. Der Grund könnte im bereits grösseren Abweichungspotential liegen, welches die erweiterte Klassierung besitzt. Die RMSE-Werte zeigen ein ähnliches Bild.

Unterschiede zwischen den Aggregierungsebenen

K_{Ant} und die RMSE-Werte sind bei beiden Aggregierungsebenen identisch. Dies hängt damit zusammen, dass die Bandbreiten global gleichgesetzt werden und dadurch die Intensitätswerte der umliegenden Punkte gleich stark auf den interpolierten Wert an einem Ort Einfluss nehmen. FK_{Ant} unterscheidet sich ebenfalls nur minim.

Es ist wichtig zu berücksichtigen, dass auf Ebene der SADS-Untersuchungsorte alle Zentroide der interpolierten Oberfläche einen direkten Validierungswert haben. Bei den Zentroiden aller Gemeinden werden nur die Gemeinden validiert, welche einem Untersuchungspunkt entsprechen. Die restlichen Werte können nicht validiert werden, da hierzu kein Referenzdatensatz zur Verfügung steht.

4.1.2. Automatisierte Bandbreitenwahl

Abbildung 4-3 zeigt die Validierungswerte für die mit den automatisierten Methoden zur Bandbreitenwahl erstellten Oberflächen. Es werden bei allen Bandbreite-Auswahlverfahren Übereinstimmungen über 90 % erreicht. Dies gilt für beide Klassen und beide Aggregierungsebenen. Generell überschneiden sich auch hier die Zentroide etwas weniger stark wie die Flächen gemeinsamer Klassierungen, jedoch handelt es sich hierbei nur um wenige Prozente Unterschied. Die für die Übereinstimmung der Zentroide besten Resultate werden bei allen Kombinationen von den beiden Faustregel-Verfahren erzielt. Die Unterschiede betragen aber nur ein bis zwei Prozente, weshalb keinesfalls von einer deutlichen Vormacht gesprochen werden darf.

Der RMSE variiert nur sehr gering zwischen den verschiedenen Bandbreite-Methoden, es macht deshalb keinen Sinn, ein Verfahren als den anderen überlegen zu werten.

In beiden Aggregierungsebenen produzieren die automatisierten Verfahren durchschnittliche Bandbreiten zwischen 11 und 13 Kilometern mit einer Standardabweichung im Bereich von 2 Kilometern (siehe eingefärbte Kästen in Anhang B).

Unterschiede zwischen den Klassierungen

Die eingeschränkte Klassierung erzielt für K_{Ant} und FK_{Ant} die besseren Werte. Diese betragen bei allen Verfahren über 94%. Beim RMSE werden bei der erweiterten Klassierung die besseren Werte erzielt, jedoch nicht in grossem Masse.

Unterschiede zwischen den Aggregierungsebenen

K_{Ant} ist, ausser in einem Fall (nrd0 der eingeschränkten Klassierung), höher auf der Gemeindeebene als auf Ebene der SADS Orte, wenn auch nur minimal. Bei FK_{Ant} ist keine Abhängigkeit von der Aggregierungsebene zu sehen.

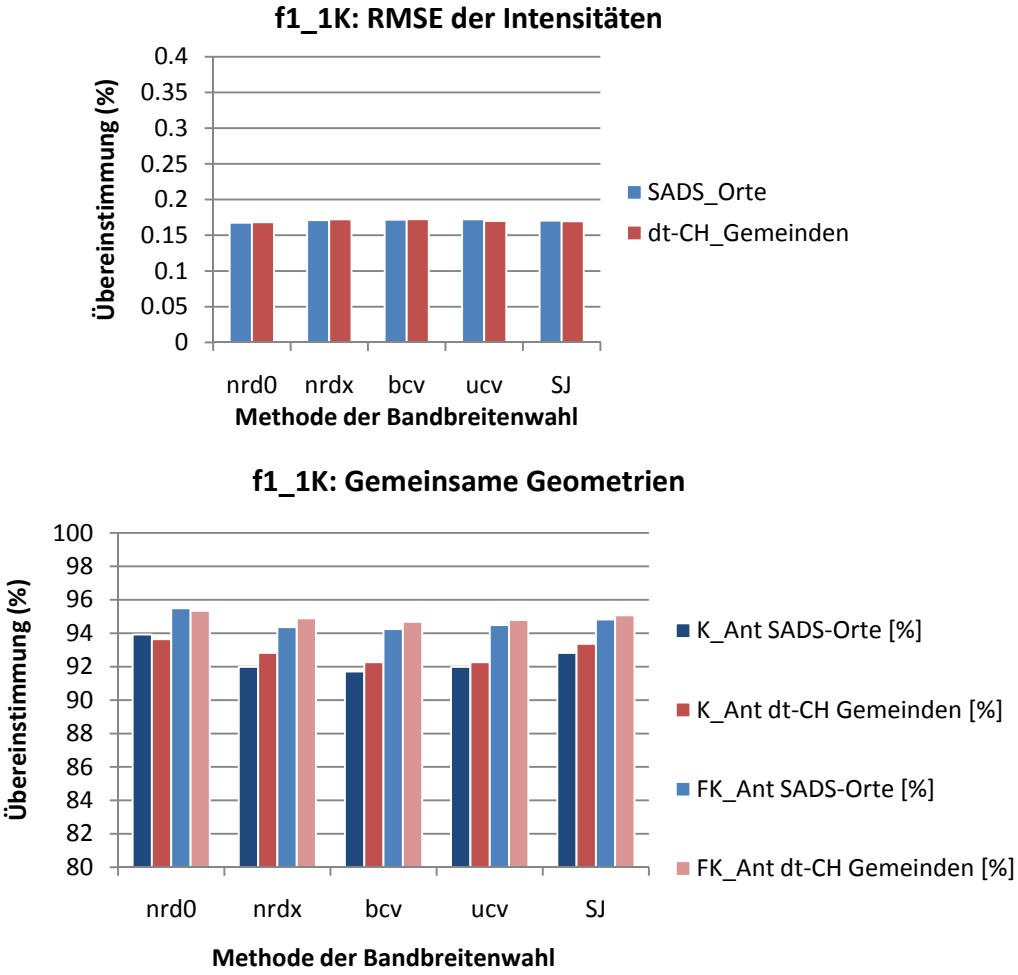


Abbildung 4-3: Validierungsmasse (oben: RMSE; unten: K_{Ant} & FK_{Ant}) für die automatisierten Methoden zur Bandbreitenwahl: `nrd0`: Faustregel nach Silverman, `nrdx`: Faustregel nach Scott, `bcv`: biased cross-validation, `ucv`: unbiased cross-validation, `SJ`: Verfahren nach Sheather & Jones

4.1.3. Fazit

Aus der quantitativen Validierung geht hervor, dass erwartungsgemäss mit zunehmender Bandbreite die interpolierte Oberfläche grössere Fehler der Intensitäten enthält und der Unterschied der Klassierung zur Originalklassierung zunimmt. Es müsste von diesem Standpunkt aus die Empfehlung abgegeben werden, eine kleinstmögliche Bandbreite zu wählen. Dies widerspricht aber dem Gedanken, die Oberfläche zu glätten, um mögliche Areale einfacher zu sehen.

Die manuelle Wahl verschiedener Bandbreiten zeigt, dass ab ungefähr 10 km keine grossen Qualitätseinbussen mit zunehmender Bandbreite entstehen. Erst ab sehr hohen Bandbreiten steigt dann die Unsicherheit wieder an.

In den automatisierten Verfahren resultieren durchschnittliche Bandbreiten, welche in etwa jener der manuell gewählten Bandbreiten von 10, 12.5 und 15 km entsprechen. Dadurch lassen sich die Validierungswerte dieser Bandbreite in Bezug zu jenen der automatisierten setzen (Tabelle 4-1).

Bandbreite-Methode	(Durchschnittliche) Bandbreite (m)	Gemeinsame Fläche SADS-Orte (%)	Gemeinsame Zentroide SADS-Orte (%)	RMSE Zentroide
nrd0	11‘038	90.334	93.923	0.168
nrdx	13‘000	91.374	91.989	0.171
bcv	13‘098	90.221	91.713	0.172
ucv	11‘544	90.446	91.989	0.172
SJ	11‘320	90.755	92.818	0.170
man_10‘000	10‘000	92.608	93.923	0.166
man_12‘500	12‘500	94.513	91.989	0.172
man_15‘000	15‘000	93.937	91.160	0.176

Tabelle 4-1: Quantitative Validierungswerte von automatisierten und 3 manuellen Bandbreitenmethoden. Frage I.1K, Aggregierungsebene: SADS Orte

Die Auswertung der quantitativen Messwerte zeigt kein klares Bild, ob die automatisierte der manuellen Klassifikation überlegen ist. Die manuelle Klassifikation könnte den Vorteil haben, dass sich die Resultate verschiedener Daten besser miteinander vergleichen lassen, da das Verfahren konsistent und deshalb besser kontrollierbar ist. Eine Empfehlung für die Wahl der Bandbreite kann durch eine rein quantitative Analyse der Resultate nicht abgegeben werden, die qualitative Bewertung der aus den interpolierten Werten hervorgehenden Karten ist ebenfalls notwendig. Eine Bandbreite innerhalb von 10‘000 bis 15‘000 Metern scheint aber aus den oben hervorgegangenen Überlegungen sicherlich vernünftig.

4.2. Qualitative Kalibrierung

Für die qualitative Validierung wird anhand des oben bereits verwendeten Testdatensatzes für die zur Verfügung stehenden Methoden jeweils eine Karte erstellt. Diese wird danach nach verschiedenen Gesichtspunkten analysiert. Wie gut ist die Übereinstimmung zur originalen Verteilung? Machen die Klassierung und die Verteilung der Intensitäten Sinn? Wie unterscheiden sich die verschiedenen Resultate voneinander? Zusammen mit den quantitativen Werten des vorherigen Abschnitts wird dann im letzten Unterabschnitt ein Entscheid gefällt, welche Methode für den Rest der Masterarbeit weiterverwendet wird.

4.2.1. Manuelle Bandbreitenwahl

Die Karte C1 im Anhang zeigt die Oberflächen aus der KDE mit verschiedenen manuell gewählten globalen Bandbreiten. Nicht überraschend werden die Konturen mit zunehmender Bandbreite weicher, da immer mehr Untersuchungspunkte auf die geschätzten Werte Einfluss nehmen. Immer mehr „Inseln“ verschwinden. Ab einer Bandbreite von 20 Kilometern verschwindet so die vorher auffällige Insel im Raum Basel im Nordwesten der Karte. Die Linie zwischen den beiden dominanten Varianten verwandelt sich immer stärker von einer komplexen Linie über eine gekrümmte Linie hin zu einer geradenähnlichen Form. Die Intensitätsunterschiede nehmen mit zunehmender Glättung über die Bandbreite ab. Vergleicht man die Karten in den höheren Bandbreiten ab etwa 30 Kilometern mit der originalen Verteilung der Klassen, ist eine klare Zweiteilung zu erkennen, die aber kleinere Variationen nicht mehr berücksichtigt. Die Grenze beginnt ab diesem Wert auch langsam nach Osten zu wandern, der Einfluss der einen Variante (rot) nimmt Überhand und überdeckt im Extremfall bei 200 Kilometern fast das ganze Untersuchungsgebiet.

Vergleicht man die Oberflächen der zwei Aggregierungsebenen ist lediglich ein leichter Unterschied bei der Sanftheit der Karte zu erkennen, welcher auf die höhere Anzahl von Polygonen und damit die höhere Auflösung zurückzuführen ist.

Die manuelle Bandbreitenwahl ist nur für die eingeschränkte Klassierung zu einer Karte aufbereitet, bei der erweiterten Version ist zwar eine weitere Variante dominant, Veränderungen bezüglich der Glättung der Oberfläche sind aber nicht nennenswert anders.

4.2.2. Automatisierte Bandbreitenwahl

Die Karte C2 im Anhang zeigt auf beiden Aggregierungsebenen die Oberflächen der KDE mit den automatisierten Methoden zur Wahl der Bandbreite. Sie geben ein ähnliches Bild ab wie die Karten mit den manuellen Bandbreiten zwischen 10 und 15 Kilometern. Die Glättung ist von Auge nicht klar unterscheidbar. Zwischen den verschiedenen Methoden gibt es ebenfalls keine grossen Unterschiede. Die Faustregel nach Silverman (1986) produziert eine etwas rauere Oberfläche. Hier bleibt auch die Insel um Basel am grössten. Doch auch dieser Unterschied ist nicht bedeutend.

Zwischen den beiden Aggregierungsebenen gibt es eine Nuance. Die höher auflösende Klassierung nach allen Gemeindepunkten der Deutschschweiz erzielt schärfere Grenzlinien, was aber auf die Verteilung der Punkte zurückzuführen ist, die hier im Grenzbereich rasterähnlich ist.

4.2.3. Fazit

Die qualitative Beurteilung der Flächen gestaltet sich als schwierig. Keine klaren Unterschiede sind bei den automatisierten Methoden erkennbar. Im manuellen Fall ist eine erwartete Glättung mit zunehmender Bandbreite ersichtlich. Wo genau eine Grenze gezogen werden soll, ist nicht klar. Ebenso wenig kann ein Favorit zwischen manueller und automatisierter Klassierung ausgemacht werden. Hier wäre eine Expertenmeinung von Linguisten, die sich mit dem Deutschschweizer Sprachraum befassen, hilfreich, um abzugrenzen, ab wann eine Glättung zu stark ist, um die Variation der Intensitäten und der Klassen zu zeigen.

Aus der durchgeführten Kalibrierung scheint eine manuelle Bandbreite von 10’000 Metern sinnvoll. Dies hat mehrere Vorteile. Die Wahl einer globalen Bandbreite ermöglicht ein konsistentes Vorgehen für alle Datensätze und ist deshalb vergleichbar. Zudem ist sie einfach nachvollziehbar. Die Resultate, die mit dieser Bandbreite für die Frage I.1K erzielt werden, unterscheiden sich nicht stark von jenen der automatisierten lokalen Bandbreitemethoden. Die mittlere Bandbreite dieser Methoden befindet sich in der gleichen Grössenordnung. Das Bild, welches sich durch die KDE mit dieser Bandbreite zeigt, ist jenem der Originalverteilung der Intensitäten ähnlich.

5. Resultate: Flächenkarten nach Rumpf et al.

Einen Hauptteil dieser Arbeit bildet die Umsetzung der Methodik von Rumpf et al. (2009) (Abschnitt 3.3). Diese wird, sofern zwei Klassierungen vorhanden sind, auf die eingeschränkte Klassierung der Fragen der drei Phänomene angewendet. Einerseits hätte eine Ausweitung auf die erweiterte Klassierung den Umfang der Arbeit gesprengt, andererseits sind die Resultate für die als Prototyp dienende Frage I.1 nicht gross abweichend. Mit der eingeschränkten Klassierung sind zudem meist schon die klar dominanten Varianten abgedeckt. Die aufbereiteten Karten beinhalten jeweils eine Flächenkarte mit den originalen Intensitäts- und Klassenverteilungen und pro Aggregierungsebene je eine Karte mit der interpolierten Oberfläche.

Allen Resultaten gemein ist eine Glättung durch die KDE. Die unruhige Erscheinung der nicht interpolierten Originaloberflächen wird umgewandelt in ein kontinuierlich wirkendes, geglättetes Bild. Zwischen den beiden Aggregierungsebenen, den Thiessenpolygonen um die Untersuchungspunkte und jenen um alle Deutschschweizer Gemeinden, sind überall nur Unterschiede in der Glättung zu sehen. Die Oberfläche wirkt bei letzterer etwas weniger rau und deshalb sanfter. Die Verteilungen der Intensitäten und der Klassen ändert sich zwischen den beiden Aggregierungsebenen nicht.

In den folgenden Abschnitten werden nun die Karten der drei behandelten Phänomene (5.1 bis 5.3) und die implementierten Erweiterungen (5.4) einzeln präsentiert und beschrieben.

5.1. A: Finalanschluss

Es resultieren in allen vier Fragen zwei dominante Varianten, wobei die Variante *zum* einen grösseren Anteil ausmacht. Das erste Phänomen produziert Karten, auf denen eine Ost-West-, sowie eine Nord-Süd-Verteilung der dominanten Varianten zu erkennen ist. Die Ausprägung dieser Feststellung unterscheidet sich je nach Frage aber deutlich. Verschiedene Fragen zum selben Phänomen ergeben somit verschiedene Verteilungen der dominanten Varianten.

Die Intensitäten nehmen einerseits zur Variantengrenze hin ab, andererseits ist in Richtung der Grenzen des Untersuchungsgebietes eine Zunahme erkennbar, insbesondere an exponierten Lagen wie in Graubünden oder im Oberwallis.

Frage I.1: „für/zum ein Billet (zu) lösen“

Die erste Frage des Finalabschluss-Phänomens hat als einzige der vier untersuchten Fragen eine Dominanz der ersten Variante *für (zu)*. Sie erstreckt sich über ein grosses Gebiet im Westen und Süden der Deutschschweiz. Die zweite dominante Variante *zum (zu)* ist vor allem in der östlichen Hälfte des Gebietes dominant und bildet dort im Raum Hinterrhein und im Appenzell besonders intensive Areale. Zudem gibt es noch eine Insel im Raum Basel, in welcher diese Variante dominant auftritt. Die Grenzlinie zwischen den beiden dargestellten Varianten macht eine Linkskurve von Norden nach Süden hin. In dem Gebiet um die Grenze ist im originalen Bild teilweise keine Variante dominant, was sich in den weissen Flächen widerspiegelt. Im gleichen Gebiet sind auch klar tiefere Intensitäten und somit hellere Farbtöne zu erkennen.

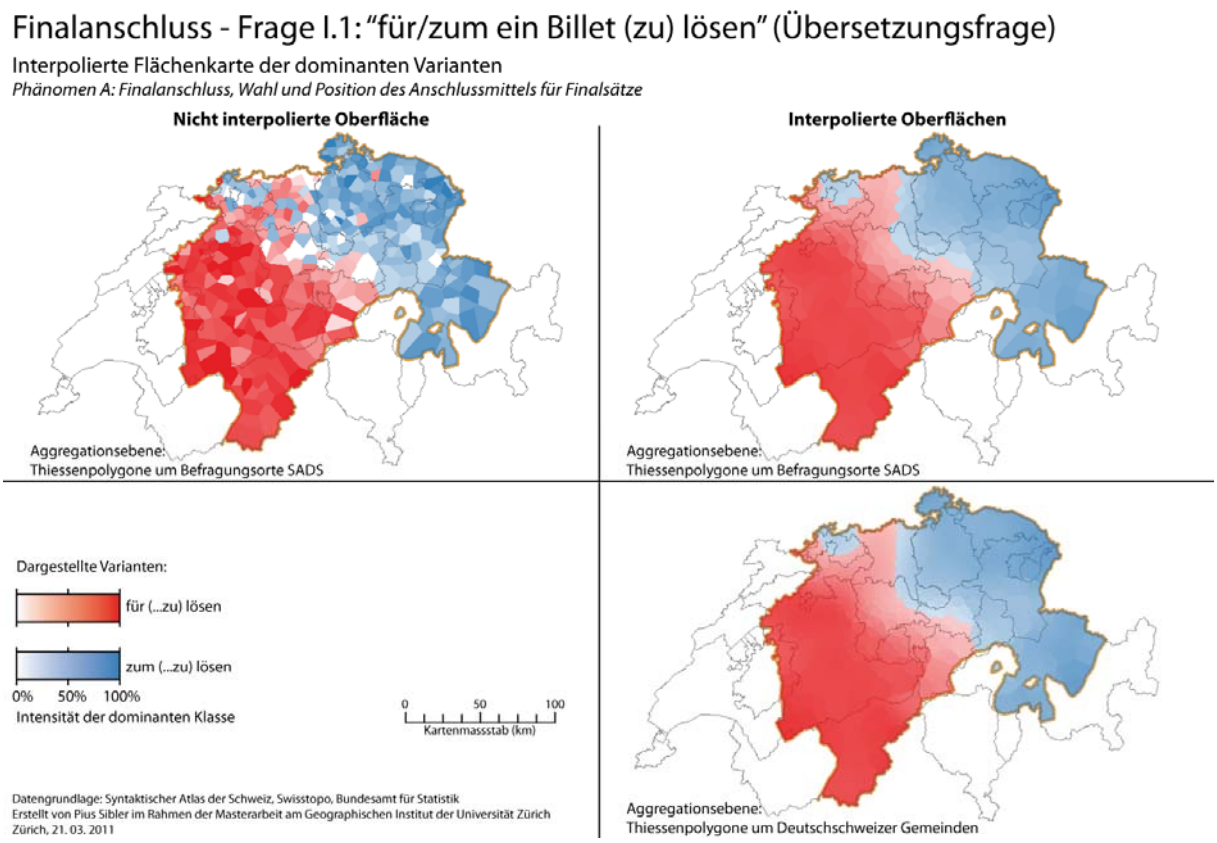


Abbildung 5-1: Finalanschluss: Flächenkarten der Frage I.1

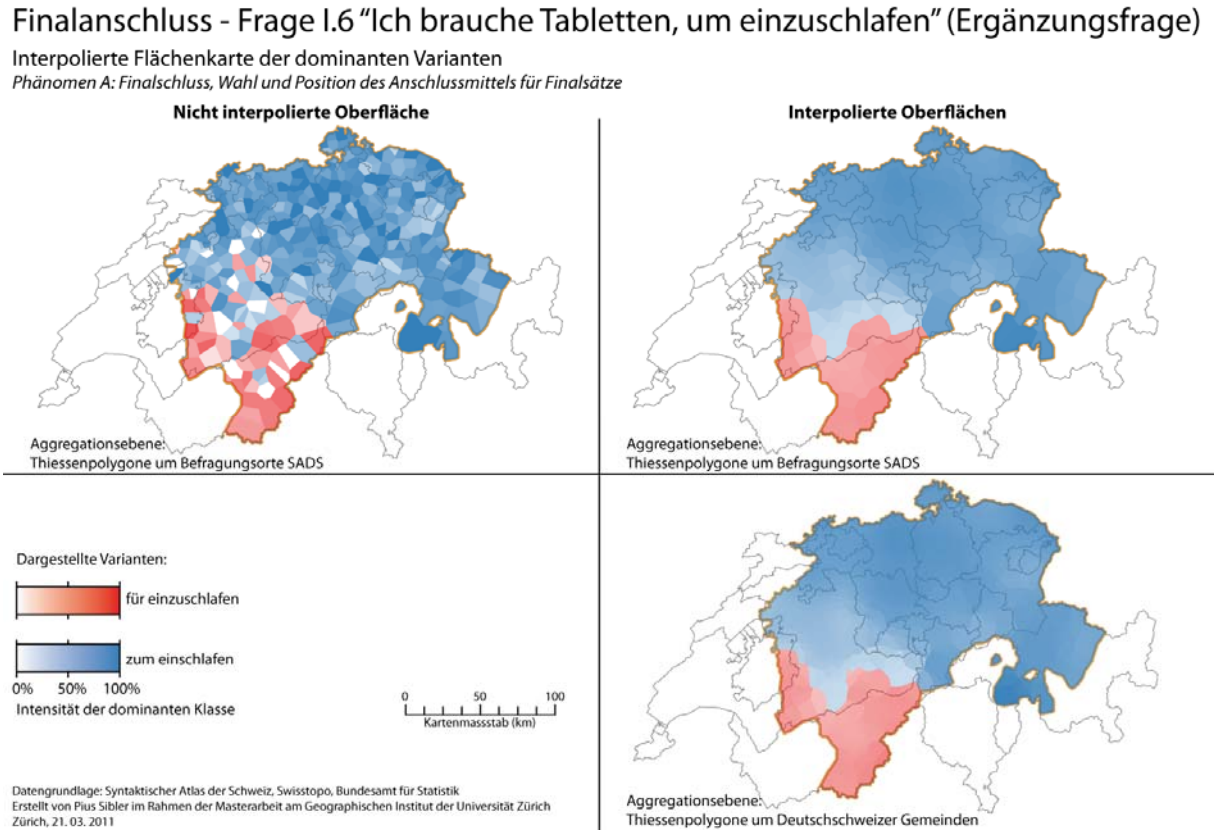


Abbildung 5-2: Finalanschluss: Flächenkarten der Frage I.6

Frage I.6: „Ich brauche Tabletten, um einzuschlafen“

Von allen Fragen zum Finalanschluss hat diese Frage die deutlichste Dominanz einer Variante. Die *zum* Variante überdeckt bis auf wenige Flächen im Oberwallis und im südöstlichen Bern und Freiburg die ganze Deutschschweiz. Die Grenze verläuft hier nicht von Osten nach Westen, sondern von Norden nach Süden. Im Gebiet Bern sind niedrigere Intensitäten festzustellen, ebenso im Bereich Appenzell im Nordosten des Untersuchungsgebiets. Im Hinterrheintal im Südosten der Deutschschweiz sind dagegen stärkere Intensitäten als sonst auf der Karte zu erkennen.

Frage I.11: „... um ein Buch zu lesen“

In diesen Karten machen die dominanten Varianten etwa gleich grosse Anteile des Bildes aus. Es ist eine fast geradenförmige Grenzlinie zwischen den beiden Varianten zu erkennen, zu welcher hin die Intensitäten gleichmässig abnehmen. Wieder sind die Grenzgebiete von einer stärkeren Intensität geprägt. Auffällig ist auch, dass im Originaldatensatz keine Gebiete ohne klare dominante Variante zu erkennen sind.

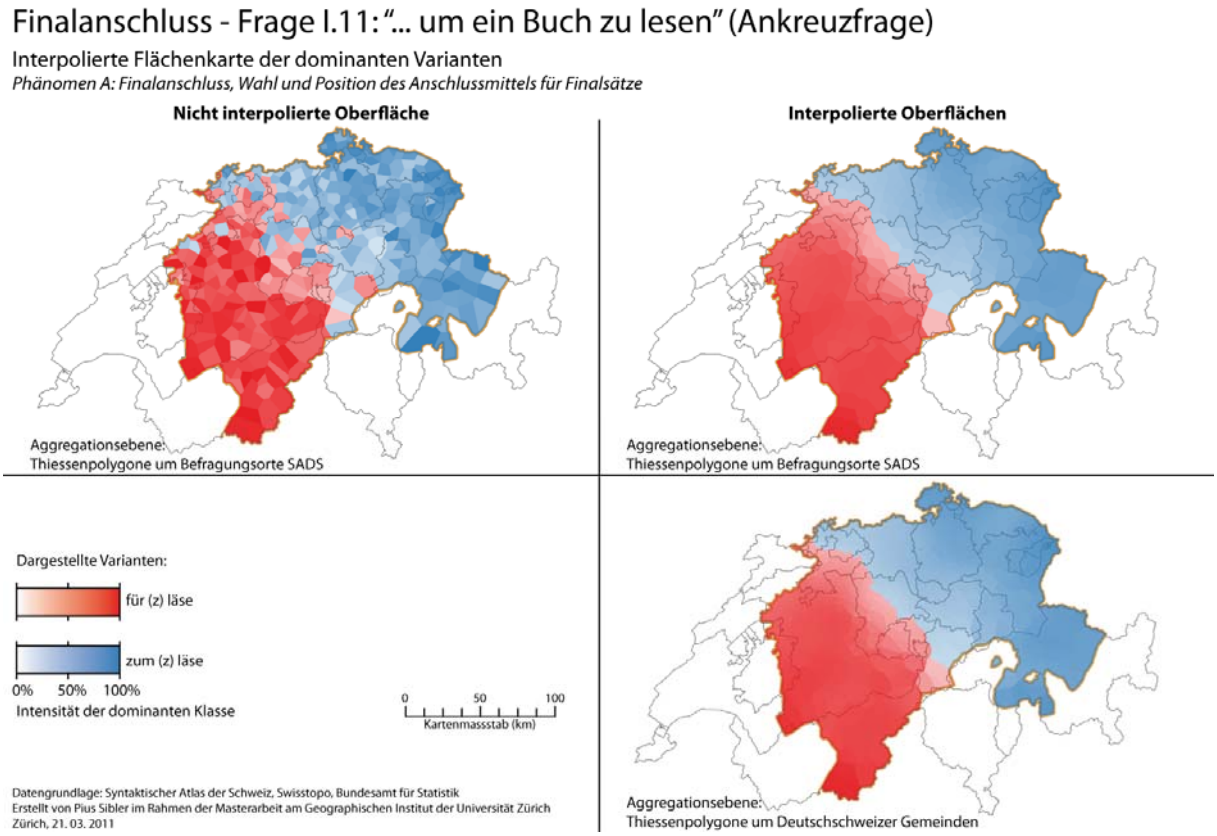


Abbildung 5-3: Finalanschluss: Flächenkarten der Frage I.11

Frage IV.14: „Du musst das Licht anzünden, um zu lesen“

Hier sind die Intensitäten im mittleren Bereich der Karte weniger intensiv als bei den vorangegangenen Fragen. Dies ist auch an den vielen weissen Flächen im Original zu erkennen. Die Grenze zwischen den dominanten Varianten verläuft hier von Nordwesten nach Südosten. Hohe Intensitäten sind wieder an den Grenzen des Gesamtgebietes zu erkennen.

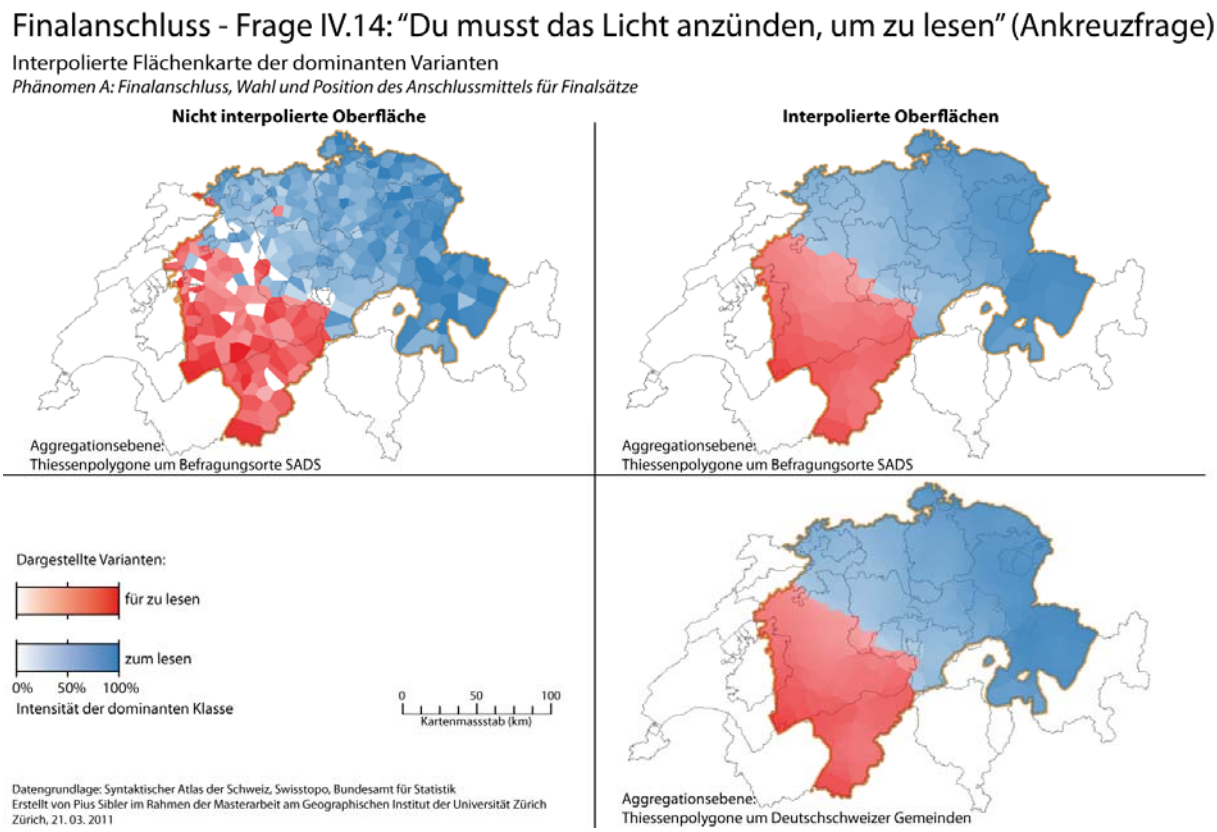


Abbildung 5-4: Finalanschluss: Flächenkarten der Frage IV.14

5.2. B: Komparativ

Die Oberflächen zum zweiten Phänomen ergeben alle ein sehr ähnliches Bild, weshalb hier nicht einzeln auf die Karten eingegangen wird. Überall ist *als* gegenüber den anderen drei möglichen Varianten in den interpolierten Karten klar dominant. Die restlichen Varianten sind nur vereinzelt im Originalbild zu sehen, werden mit der Glättung aber eliminiert. Gebiete mit erhöhter Intensität befinden sich wieder an exponierten Stellen, so etwa in Basel, im Wallis und am deutlichsten in Graubünden. Die letzte Frage gibt insgesamt ein etwas blasseres Bild ab; sonst sind nur minime Unterschiede zwischen den verschiedenen Fragen zu erkennen.

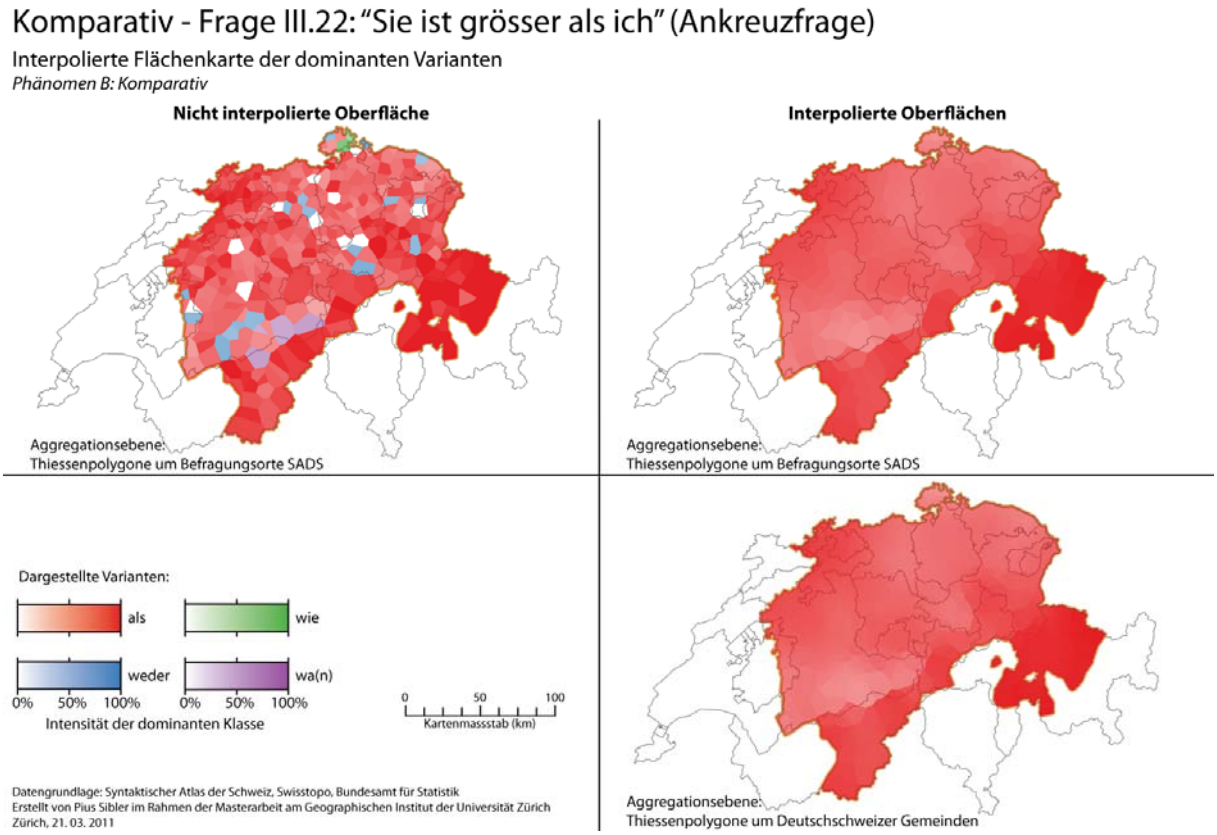


Abbildung 5-5: Komparativ: Flächenkarten der Frage III.22

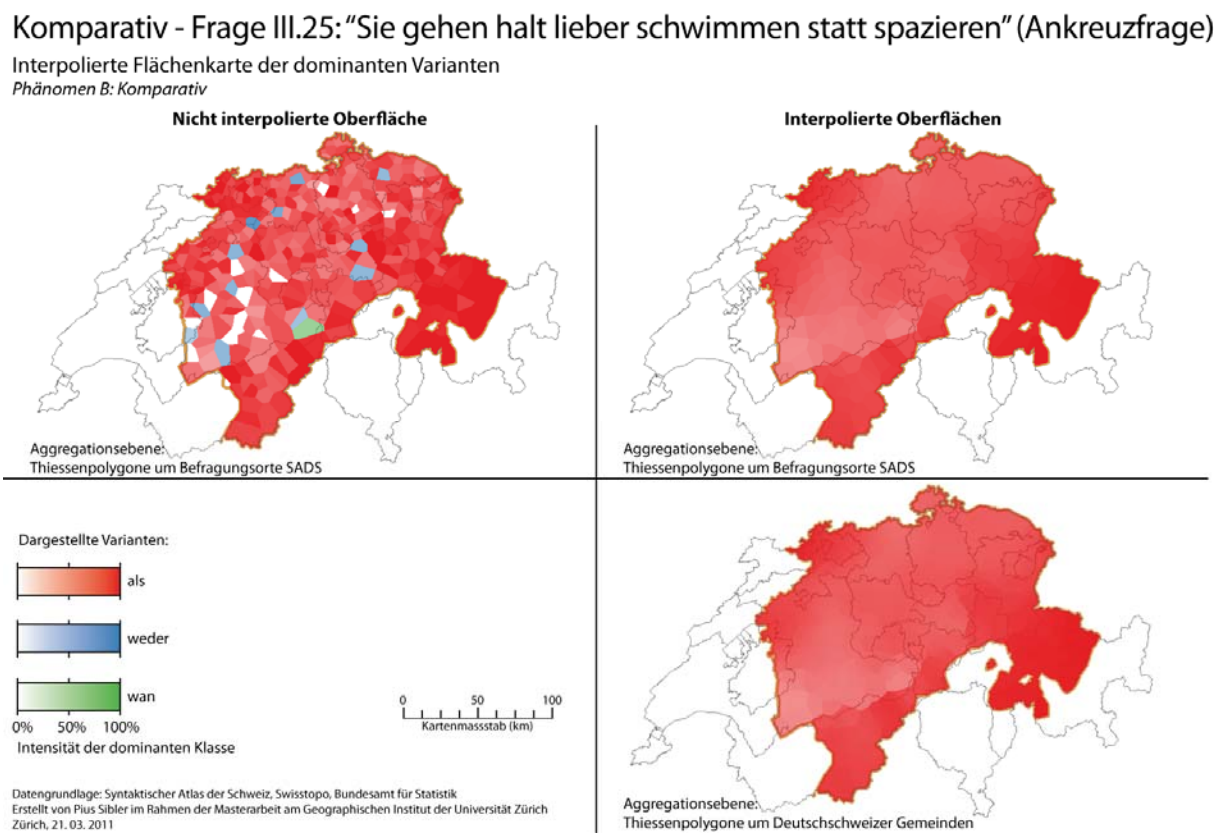


Abbildung 5-6 Komparativ: Flächenkarten der Frage III.25

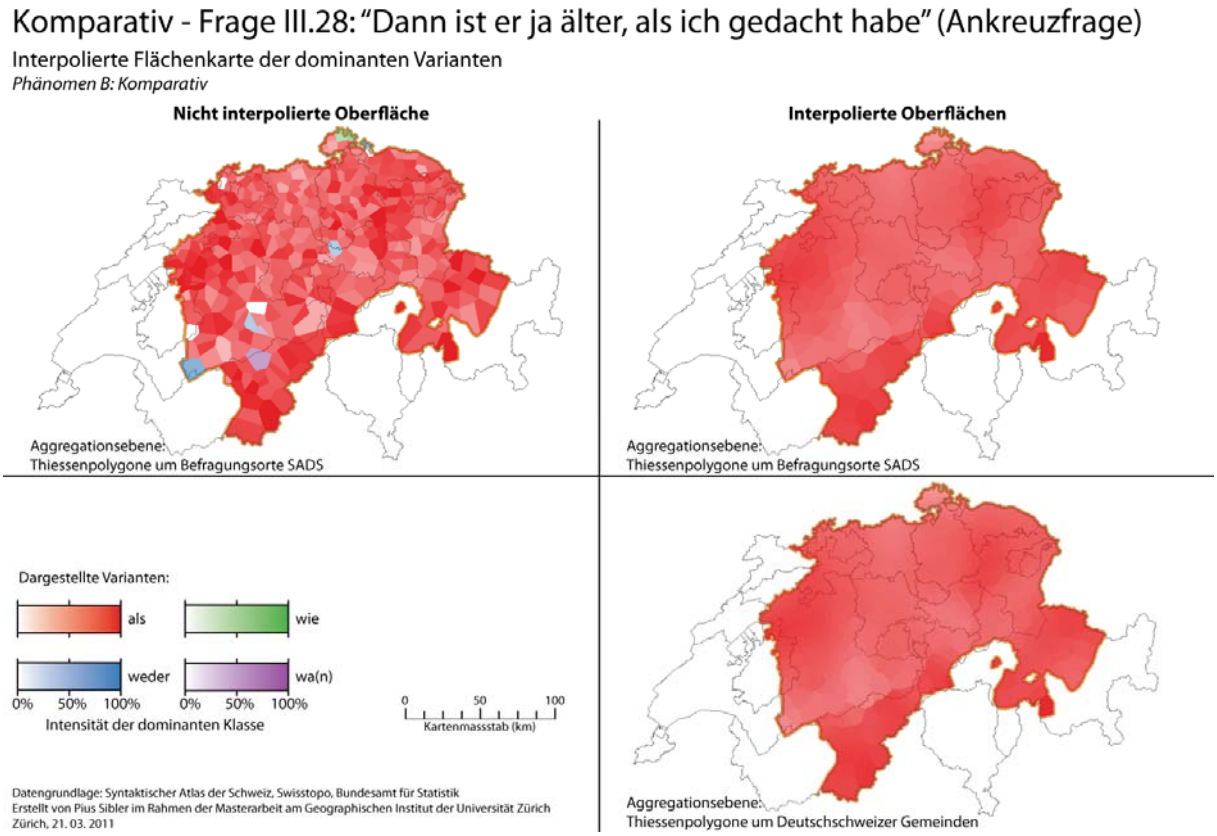


Abbildung 5-7: Komparativ: Flächenkarten der Frage III.28

5.3. C: Artikelverdoppelung

Auf den Bildern des dritten Phänomens sind wenige Gemeinsamkeiten festzustellen. Es lässt sich lediglich sagen, dass die Intensitäten im Allgemeinen eher schwächer sind als bei den anderen beiden Phänomenen. Die Verteilungen wirken einmal zufällig, ein andermal dominiert eine Variante. Auch in den Originaldatensätzen sind keine Ähnlichkeiten feststellbar.

Frage I.10: „Also d Susi wär e ganz e liebi Frau für de Markus“

Die originale Karte wirkt sehr uneinheitlich. Eine Ausnahme bildet der Südwesten der Deutschschweiz, wo die nachgestellte Variante „ganz e liebi“ wenig in Konkurrenz zu den anderen Varianten steht. In den interpolierten Oberflächen ist dies auch gut zu erkennen. Hier werden auch Tendenzen zur Clusterbildung der Artikelverdoppelung erkennbar. Die erste Variante zeigt im Bereich Schwyz/Glarus wie auch im Gebiet um Basel eine Häufung. Eine kleine Insel dieser Variante ist zudem im nördlichen Thurgau zu sehen. Die zweite Variante mit dem vorangestellten Artikel hat erhöhte Intensitäten im Furkagebiet, im Südwallis und etwas abgeschwächt im Raum Appenzell. Die beiden nach der KDE übrigbleibenden Varianten durchdringen sich in der östlichen Hälfte der Deutschschweiz. Ein klarer Trend über die gesamte Fläche ist nicht erkennbar. Die ursprünglich drei dominanten Varianten werden durch die Interpolation auf zwei reduziert, der vorangestellte Artikel kommt nicht mehr vor.

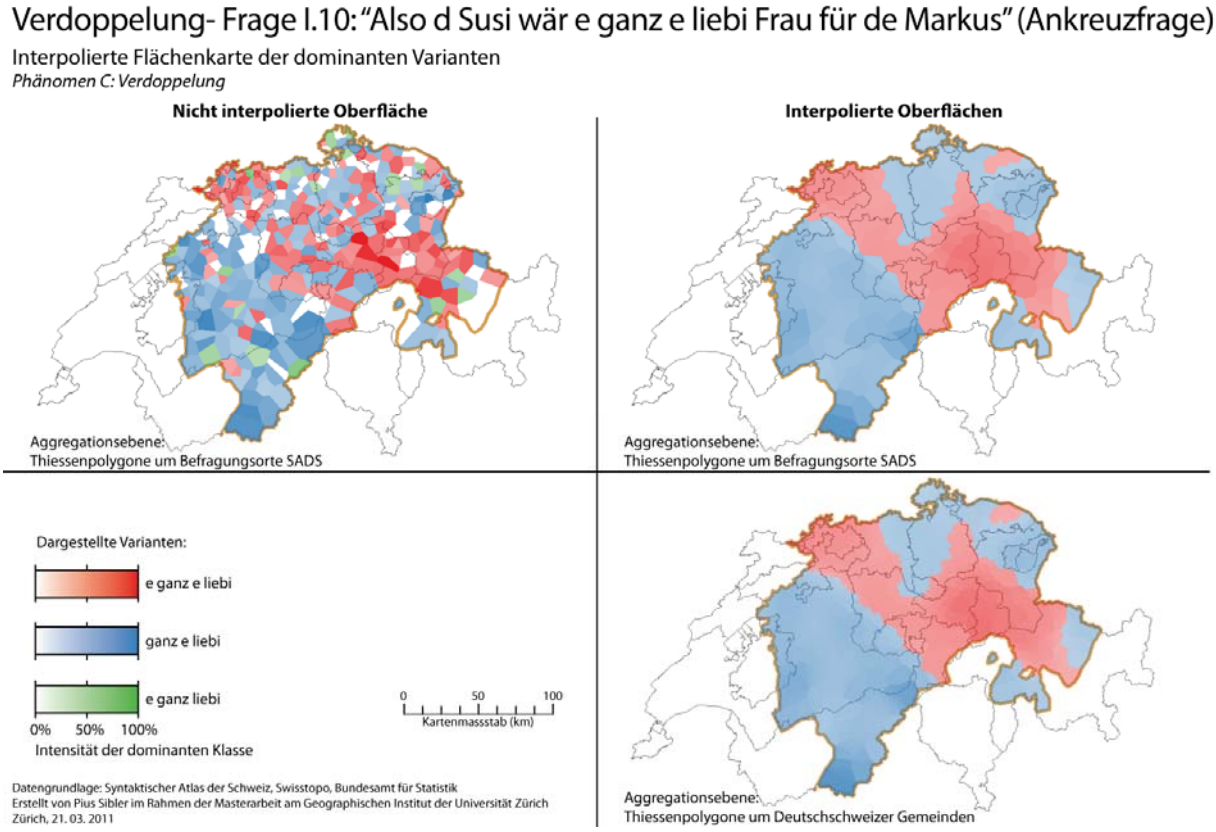


Abbildung 5-8: Artikelverdoppelung: Flächenkarten der Frage I.10

Frage II.10: „Aber du häsch (de) vil (de) schöner Garte“

Ein gänzlich anderes Bild als die erste Karte gibt diese Frage wieder. Hier dominiert die nachgestellte Variante (*vil de schöner*), bei welcher die KDE die wenigen Gebiete mit der vorangestellten Variante *de vil schöner* im originalen Bild eliminiert. Es entsteht eine praktisch homogene Oberfläche. Die Variante mit der Verdoppelung kommt erst gar nicht vor. Die Karte zeigt im Original nur wenige weisse Flecken, welche sich lose über das Gebiet verteilen.

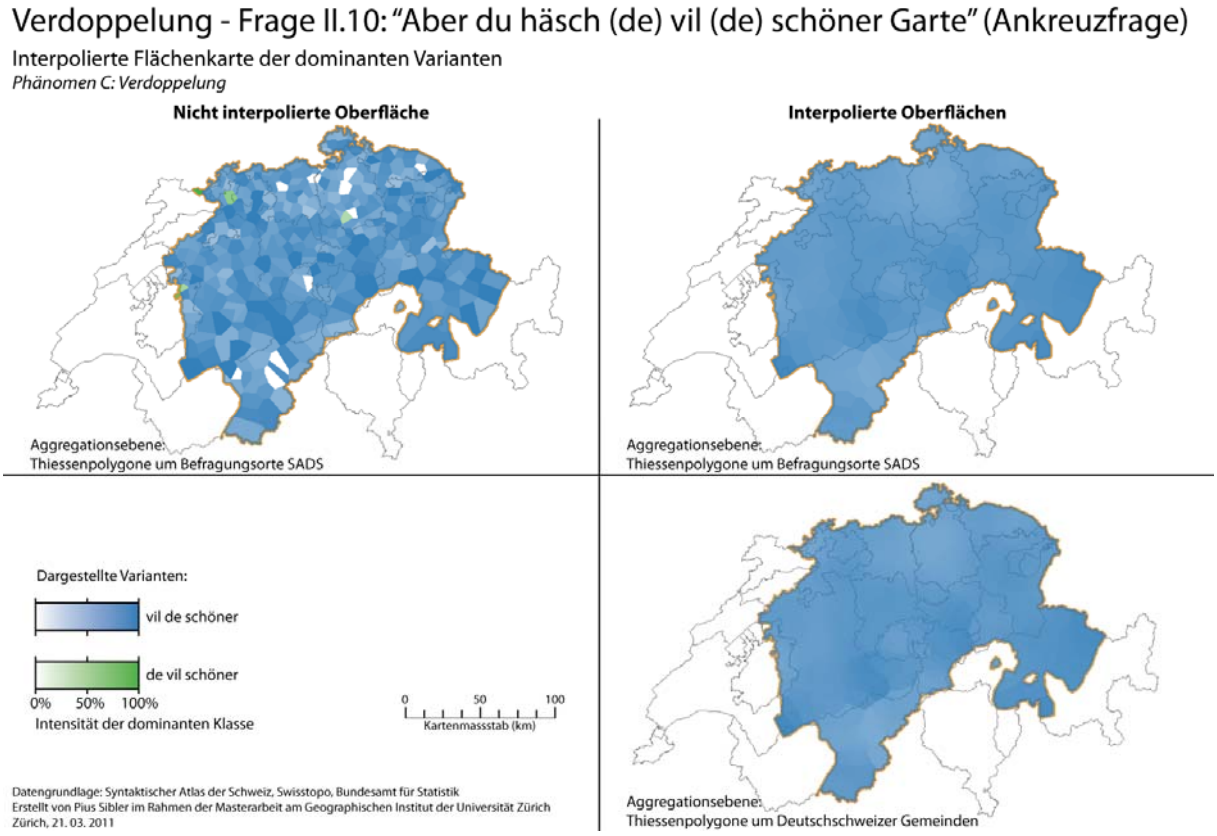


Abbildung 5-9: Artikelverdoppelung: Flächenkarten der Frage II.10

Frage IV.1: „Martina wäre eine ganz gute Gemeindepräsidentin“

Auch bei der dritten Frage ist eine Variante klar dominant, jedoch im Gegensatz zu II.10 der vorangestellte und nicht der nachgestellte Artikel. In der Originalkarte tauchen alle drei Varianten auf. Auffällig sind dort auch die vielen Polygone, in welchen keine Variante dominiert. Im interpolierten Bild bleiben nur noch zwei Varianten übrig, wobei die Variante mit dem vorangestellten Artikel fast das gesamte Gebiet abdeckt und die Artikelverdoppelung nur zwei kleine Inseln in der Innerschweiz und südlich von Chur bildet. Bei dieser Frage ist wieder eine kleine Intensitätszunahme zu den Grenzen des Untersuchungsgebiets hin erkennbar. Die Flächen wirken aber insgesamt blasser im Vergleich zu allen anderen erstellten Karten.

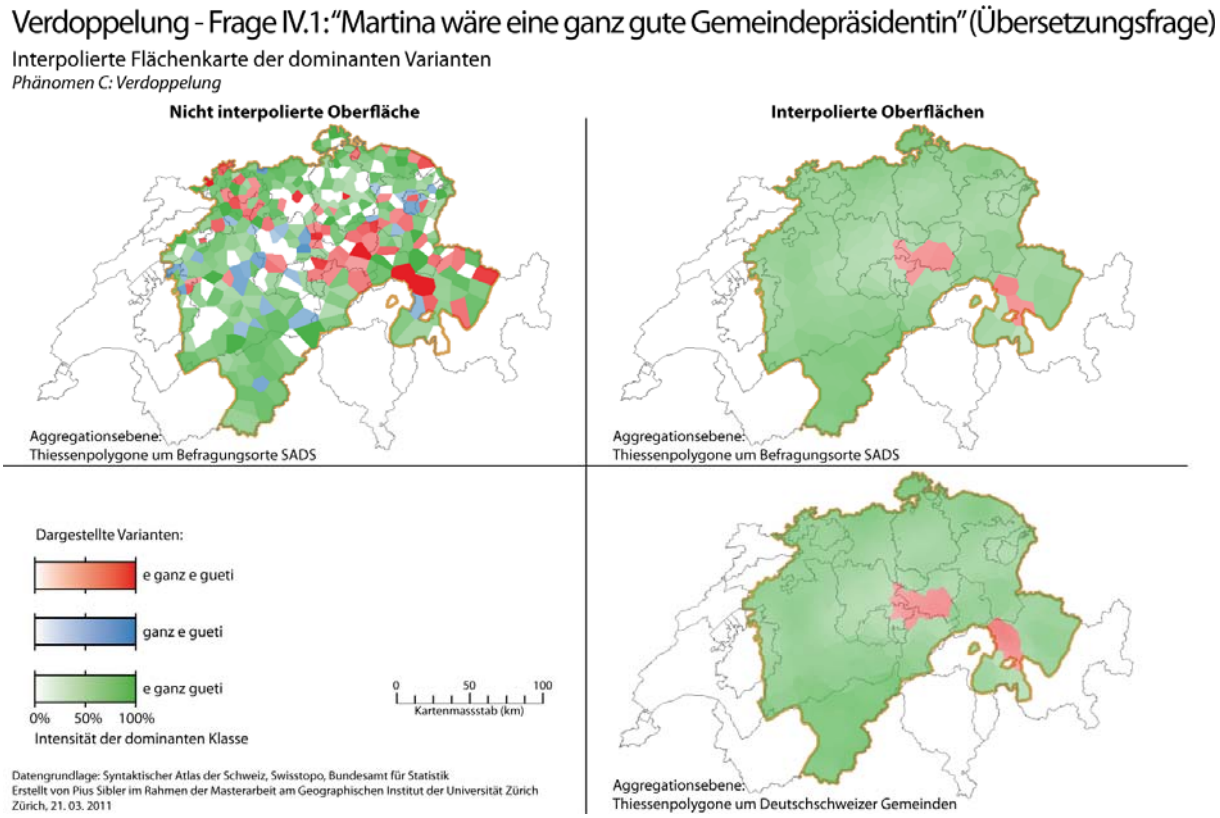


Abbildung 5-10: Artikelverdoppelung: Flächenkarten der Frage IV.1

5.4. Erweiterungen

Kombinierte Karte Finalanschluss

Die Finalanschluss-Karte, welche aus der Kombination der vier zu diesem Phänomen untersuchten Fragen besteht, macht bereits in der Originalverteilung einen geglätteten Eindruck mit zwei praktisch homogenen Gebieten im Nordosten und im Südwesten. Das nordöstliche Gebiet, welches die *zum* Variante abdeckt, umfasst etwa 2/3 des Untersuchungsgebiets. Interpoliert kommt diese Unterteilung noch deutlicher zum Vorschein. Verglichen mit den Einzelkarten zu den Fragen ist die Intensität kleiner. Zwischen den beiden Aggregierungsebenen zeigt sich kein erkennbarer Unterschied.

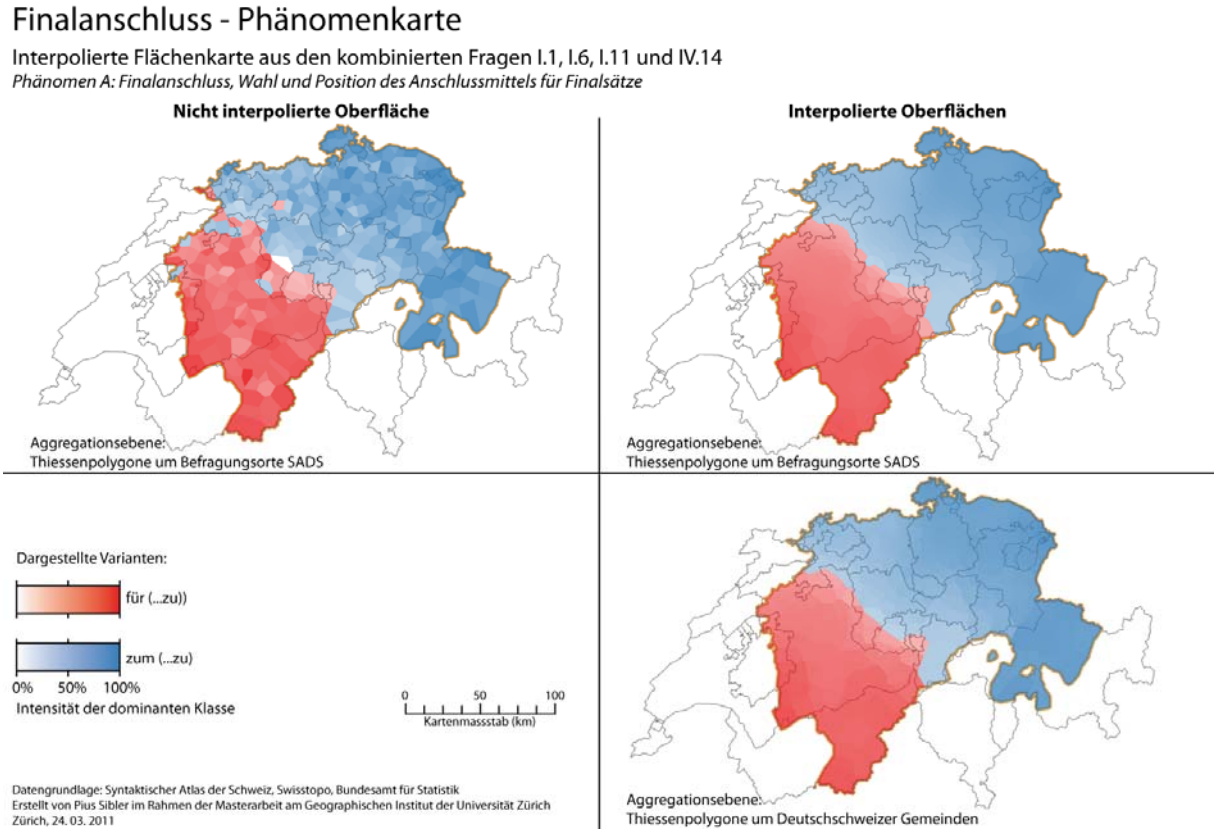


Abbildung 5-11: Kombinierte Karte des Finalanschlusses aus den Intensitäten der vier behandelten Fragen

Unterscheidung von Präferenz und Akzeptanz

Die Verteilung der Klassen zwischen den präferierten und den dominanten Varianten unterschiedet sich je nach Frage. Exemplarisch zeigen dies die Frage IV.14 des Finalanschlusses und die Frage I.10 des Artikelverdoppelungsphänomens. Während beim ersten Fall ein praktisch identisches Muster resultiert, ist im zweiten Beispiel eine klare Veränderung zusehen, die Verdoppelungsvariante nimmt insgesamt zu. Allen Resultaten gemeinsam ist eine Intensivierung der dominanten Varianten.

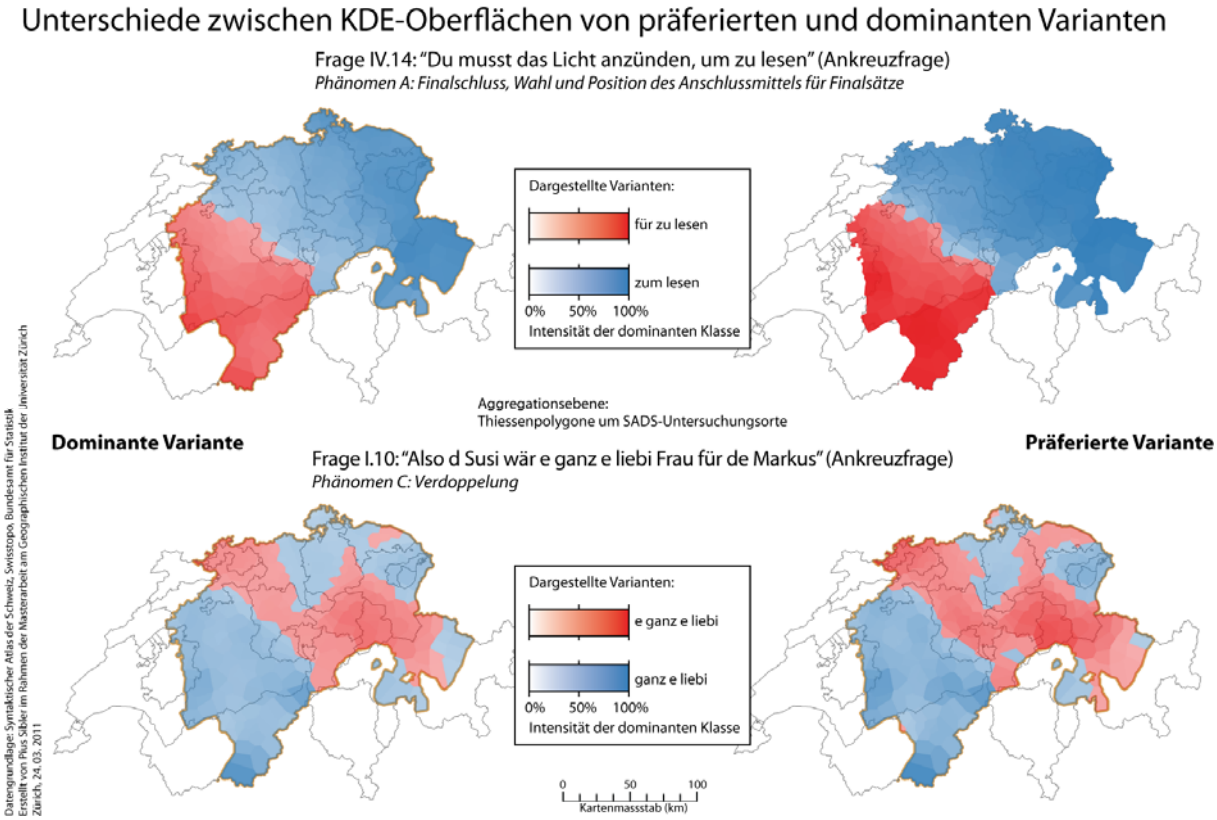


Abbildung 5-12: Flächenkarten mit den Intensitäten der dominanten Varianten (links) und der präferierten Varianten (rechts) für die Fragen IV.14 (Finalanschluss) und I.10 (Artikelverdoppelung)

Gewichtung nach Personenzahl

Bei der Gewichtung nach Personenzahl an den Untersuchungspunkten ergibt sich von Auge aus kein erkennbarer Unterschied zu den nicht gewichteten Oberflächen. Nimmt man die mittleren Intensitäten der verschiedenen Karten, so ist eine, wenn auch nur minime, Zunahme bei beiden Aggregierungsebenen erkennbar (Tabelle 5-1):

Aggregierungsebene	Durchschnittliche Intensität	
	ungewichtet	*gewichtet*
Untersuchungsorte	40.78	41.15
Gemeindepunkte	42.26	42.87

Tabelle 5-1: Mittlere Intensitäten der dominanten Varianten der Frage I.10 (Artikelverdoppelung)

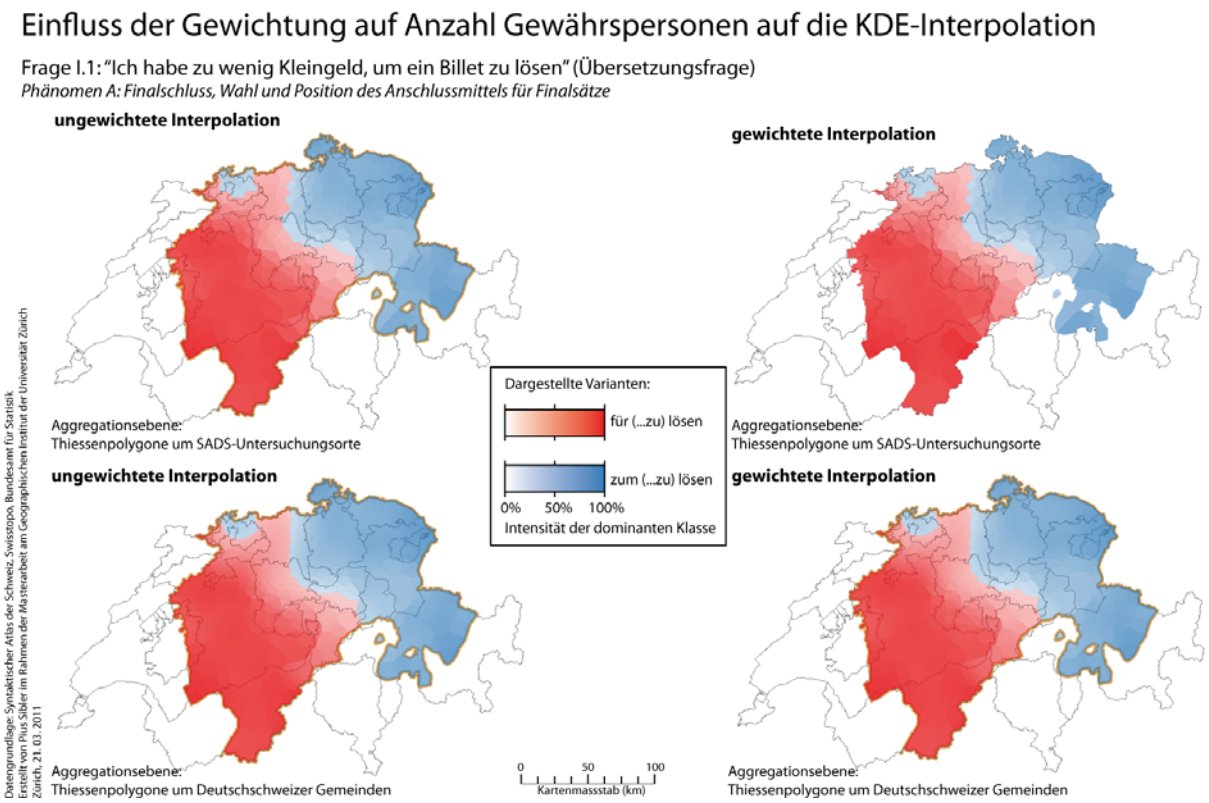


Abbildung 5-13: Flächenkarten mit den ungewichteten (links) und nach GP gewichteten (rechts) Intensitäten der Varianten für die Frage I.1 (Finalanschluss)

3D-Darstellung

Eine dreidimensionale Darstellung auf einen Papierausdruck abzubilden ist schwierig. Als Beispiel ist das Resultat für I.1K als .wrl-File auf der Daten-Disc im Ordner *5_3D* beigelegt. Dieses kann mit jedem vrml-Viewer[1] angeschaut werden. Man sieht darauf wie je eine Variante zum Osten und zum Westen deutlich dominiert und zur Mitte hin ab einem gewissen Punkt beginnt, abzunehmen. Dabei ist der Abnahmetrend bei der *für* Variante zu Beginn steiler als bei der *zum* Variante. Dort setzt der Trend jedoch früher ein und flacht ungefähr von der Mitte des Gebiets an vollständig ab. Die Variante kommt ab dieser Grenze praktisch nicht mehr vor. Die *für* Variante kommt dagegen auch in jenem Gebiet noch vor, in welchem die *zum* Variante zu beobachten ist.

Als Vorschau ist in Abbildung 5-14 ein statischer Screenshot der 3D-Ansicht zu sehen:

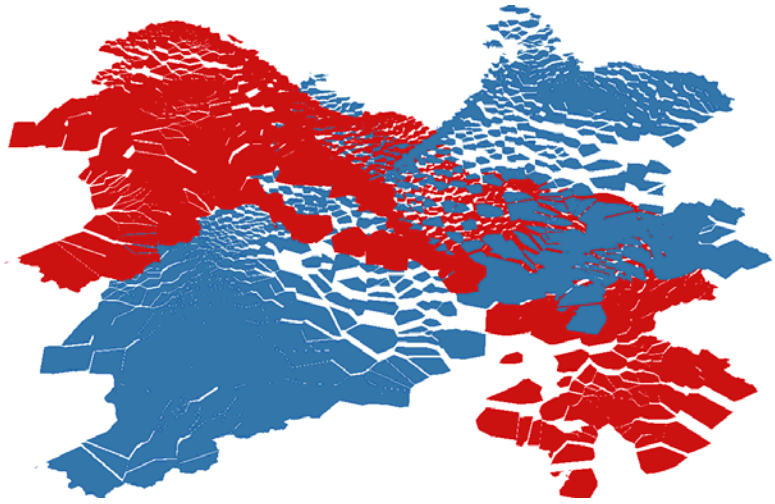


Abbildung 5-14: Screenshot der 3D-Repräsentation der Frage I.1K (Finalanschluss)

[1] Beispielsweise mit dem Free WRL Viewer: Beigelegt auf CD (Ordner *5_3D*) oder zum Download verfügbar unter: http://freewrl.sourceforge.net/, Zugriff: 3.4.2011

6. Diskussion der Flächenkarten

Diese Arbeit bezieht sich bei der Generierung von Flächenkarten lediglich auf eine Methode (Rumpf et al. 2009). Dies ist von einem ganzheitlichen Standpunkt her gesehen nicht genügend, um die Möglichkeiten und Grenzen dieser Visualisierungs- und Analyseform für syntaktische Phänomene abschliessend zu diskutieren. Das muss beim Lesen dieses Kapitels unbedingt beachtet werden. In der Folge werden zuerst einige methodische Überlegungen zum gewählten Vorgehen (Abschnitt 6.1) und im Anschluss die Resultate selbst diskutiert (Abschnitt 6.2).

6.1. Methodik

Im ersten Hauptteil der Arbeit wird die Methodik nach Rumpf et al. (2009) zur Erstellung von Flächenkarten auf syntaktische Daten des Syntaktischen Atlas der Deutschen Schweiz angewandt (Teil II). Die beiden Hauptschritte, die Tabellenaufbereitung und die Transformation in Flächenkarten, sollen hier kritisch gewürdigt und Optimierungsvorschläge erwähnt werden. Zudem soll die Methodik im Allgemeinen diskutiert und auf Mängel und Erweiterungsmöglichkeiten eingegangen werden.

Tabellenaufbereitung

Da diese Daten nicht immer einheitlich gegliedert vorliegen, bedarf es einer ziemlich aufwändigen Aufbereitung der Untersuchungsdaten, um Tabellen mit Intensitäten akzeptierter linguistischer Varianten pro Untersuchungsort aufzubereiten (vgl. Abschnitt 3.1). Um dieses suboptimale manuelle Verfahren zu umgehen, wäre eine Neukonzipierung der Datenbank in ein konsistentes Schema mit Feldern gleichen Datentyps und konsistenter Bezeichnung nötig. Erschwerend dürfte sich dabei die Vielseitigkeit der erhobenen Daten auswirken. Je nach Frageart und Phänomen wurden für das SADS-Projekt ganz unterschiedliche Mengen an verschiedenen Varianten und Zusatzinformationen gesammelt. Eine Vereinheitlichung der Datengrundlage innert nützlicher Frist ist nicht möglich gewesen. Weil es sich bei dieser Arbeit um einen explorativen Denkanstoss dafür handeln soll, was im Bereich syntaktischer Flächenkarten möglich ist, ist der eingeschlagene Weg aber vertretbar.

Erstellen von Flächenkarten mithilfe der KDE

Ist der erste, aufwändige Schritt der Tabellenaufbereitung (3.1) abgeschlossen, können mit dem ausgearbeiteten R-Skript in relativ kurzer Zeit Intensitäts-Punktdatensätze generiert werden (3.3.4). Es sind lediglich einige Änderungen bei den Pfadangaben, Benennungen und Variantenzuordnungen zu machen. Der Aufbau des Skripts mit eigenen Funktionen für den Kernel und die Bandbreitenwahl machen es sehr anpassungsfähig und für die explorative Datenanalyse geeignet.

Eher umständlich und zeitintensiv ist die Weiterverarbeitung der Punktdaten in Flächenform in ArcGIS, worin für jede Karte einzeln die Intensitätswerte zu den Thiessenpolygonen angefügt, nach dominanter Variante eingefärbt und die Helligkeit nach Intensität gewählt werden muss. Eine automatisierte Erstellung solcher Karten mittels der Plotmöglichkeiten, die R bietet, beispielsweise in den Erweiterungen `maptools` oder der mächtigen Toolbox `ggplot`[1] würde den Arbeitsprozess erheblich beschleunigen und wahrscheinlich auch qualitativ verbessern. So erstellte Karten kommen zwar nicht an die Visualisierungsqualitäten von ArcGIS in Kombination mit Illustrator heran, wären aber ohne zusätzlichen Aufwand erstellbar und somit eine gute Entscheidungsbasis für die explorative Datenanalyse. Ist eine optimale Kombination von Parametern gefunden, kann für den kartographischen Feinschliff immer noch die herkömmliche Vorgehensweise gewählt werden.

Vorbehalte müssen bei der Wahl der Parameter geäussert werden. Alle präsentierten Karten resultieren aus der Auswertung von Resultaten einer Frage bei sich ändernden Bandbreitewahl-Methoden (vgl. Kapitel 4). Konsequenterweise müsste für jede Frage eine individuelle Analyse gemacht werden, welche Bandbreite die vernünftigsten Resultate liefert. Es könnte sein, dass sich die optimale Bandbreite je nach Phänomen oder Frage unterscheidet. Ein Einbezug der Resultate aus der geostatistischen Analyse oder

[1] ggplot2 Plotting System: http://had.co.nz/ggplot2/, Zugriff: 9.4.2011

zumindest der Strukturdaten von Rumpf et al. (2010) (siehe Abschnitt 8.1), welche im zweiten Hauptteil der Arbeit behandelt werden, könnte verlässlichere Beurteilungsgrundlagen liefern.

Weiter wird immer derselbe Kernel verwendet, unter der Annahme, K hätte keinen grossen Einfluss auf die Resultate. Dies könnte man auch noch verifizieren, indem verschiedene Formeln für K eingesetzt würden.

Untersuchungssetting

Die Wahl einer globalen Bandbreite berücksichtigt die lokale Verteilung der Punkte nicht. Gebiete mit geringen Punktdichten erhalten so Intensitäten, die von wenigen umliegenden Punkten bestimmt werden und die Anfälligkeit auf Ausreisser erhöht sich. Die verwendeten Validierungsgrundlagen mit dem RMSE und den Klassenanteilen (Abschnitt 4.1) sind globale Messgrössen und können diese Ausreisser nicht aufdecken. Besonders betroffen davon sind Regionen an den Grenzen des Untersuchungsgebiets, bei denen automatisch weniger Messpunkte im Einzugsradius vorzufinden sind. Dadurch können so genannte Randeffekte entstehen. Möglichkeiten zu deren Korrektur existieren (Diggle et al. 1994). Die Normalisierung nach Anzahl Punkte der geschätzten Oberfläche sollte diese etwas eindämmen (Rumpf et al. 2010). Randeffekte müssen bei der Beurteilung der Resultate aber bedacht werden.

Sprache ist ein kontinuierliches Phänomen, das nicht an der Landesgrenze endet[2]. Eine Ausweitung des Untersuchungssettings auf den gesamten deutschsprachigen Raum wäre interessant.

Kritik an der Distanzmessung

Die Repräsentation von Punkten als Thiessenpolygone (Abschnitt 3.2) bietet eine einfache, wenn auch gleichwohl begrenzte Methodik. Es ist nicht der Realität entsprechend, anzunehmen, dass jeder Untersuchungsort von den, für ihn gemäss euklidischer Distanz am nächsten gelegenen, Messpunkten am stärksten beeinflusst wird. Die geographische Distanz als einzige Einflussgrösse für die Verteilung von sprachlichen Phänomenen zu wählen greift wohl zu kurz. Geht man davon aus, dass der Kontakt und damit die Interaktion die entscheidende Grösse für die Verteilung des Dialekts ist (Gooskens 2004; Seiler 2008; Wiersma et al. 2011), so wird in der Deutschschweiz schnell klar, dass eine Repräsentation über die Luftdistanz nicht sonderlich geeignet ist. Topographische Hürden bilden beispielsweise entscheidende Sprachbarrieren, die nicht berücksichtigt werden in dieser Methode. Ebenso werden historisch gewachsene Kulturgrenzen vernachlässigt.

Um die Topographie zu berücksichtigen wäre beispielsweise die effektive Wegdistanz eine Möglichkeit der Distanzoperationalisierung. Ebenfalls ist die Reisezeit zwischen den Punkten eine denkbare Alternative, wie dies Gooskens (2004) schon für linguistische Daten in Norwegen erfolgreich gezeigt hat, wo die Interaktion ebenfalls durch topographische Barrieren gehemmt wird.

Vor- und Nachteile der Flächenkarten

Der direkte Vergleich mit bisher erstellten Punktkarten zeigt Chancen, aber auch Einschränkungen, die interpolierte Oberflächen gegenüber einer Punktdarstellung haben. In Flächenform kann nur eine Variante pro Ort dargestellt werden, Punktkarten mit verschiedenen Symbolen vermögen dagegen, mehrere Möglichkeiten am selben Ort parallel darzustellen. Dafür kann mit der Intensität vermittelt werden, wie stark die dominante an einem Ort vertreten ist und mit den Helligkeitsunterschieden kann ein Eindruck von räumlicher Kontinuität entstehen. Für Experten in Deutschschweizer Dialektologie bietet die neue Darstellungsform von syntaktischen Phänomenen, die sich auch auf andere grammatische Bereiche ausweiten lässt, eine ungewohnte Form und ist wahrscheinlich auf den ersten Blick weniger intuitiv als die Punktkarten. Diese haben durch den SDS bereits eine grosse Tradition und sind vertrauter.

Die Punktkarten sind im Vergleich zu den interpolierten Oberflächen ehrlicher, was die Verteilung der Varianten betrifft. In Gebieten mit niedriger Anzahl Messpunkten resultieren automatisch weniger Symbole. Dagegen werden mit den bisher im SADS verwendeten Punktgrössen, welche die Messpunkte

[2] Elvira Glaser im Interview im Uni Magazin 4/2008:
http://www.kommunikation.uzh.ch/publications/magazin/unimagazin-08-4/unimagazin-08-4-31.pdf,
Zugriff: 19.4.2011

lediglich in zwei absolute Klassen einteilen, Untersuchungsorte mit vielen befragten Personen automatisch überrepräsentiert, was mit der Intensitätslösung der Flächenkarten nicht geschieht.

Andererseits ist die flächendeckende Repräsentation trügerisch, da sie impliziert, dass an jeder Lokalität in dem Gebiet irgendeine Sprache gesprochen wird, was für nicht bewohnte Gebiete nicht stimmt. Durch die Thiessenpolygone werden zudem Orte mit wenig Nachbarn auf den Resultaten flächenmässig überrepräsentiert gegenüber solchen in dicht befragten Regionen. Hier könnte eine dasymetrische Karte (Eicher & Brewer 2001) Abhilfe schaffen, die nur Gebiete berücksichtigt, die bewohnt sind. Dasymetrische Karten werden beispielsweise standardmässig in Kartenpublikationen des BFS verwendet.

Als Option, um Verzerrungen der Flächenkarten entgegenzuwirken, könnte eine kombinierte Lösung angestrebt werden, welche die interpolierten Oberflächen um die Messorte erweitert.

Effizienz der Algorithmen

Die geschriebenen Skripte sind nicht sonderlich performant und sind auch nicht mit Augenmerk darauf entstanden. Sollte die Methodik auf eine grössere Auswahl von Fragen angewandt werden, wären Überlegungen, wie die Effizienz zu erhöhen ist, angebracht. Der grösste Arbeitsaufwand betrifft aber die einmalige Aufbereitung der Tabellen, weshalb eine Effizienzsteigerung nicht prioritär behandelt werden muss.

6.2. Diskussion der Resultate

Die drei untersuchten Phänomene deuten darauf hin, dass es keine einheitliche syntaktische Verteilung in der Deutschschweiz gibt. Vielmehr muss von einer sehr unterschiedlichen Verteilung je nach Phänomen ausgegangen werden, wie dies bereits von Bucheli & Glaser (2002) bemerkt wurde. Um dies zu belegen, bräuchte es allerdings weiterführende Untersuchungen mit einem grösseren Volumen an Phänomenen und Fragen.

Es kann nur teilweise eine deutlich räumlich beeinflusste Struktur erkannt werden. Beim Finalanschluss sind klar Areale mit den beiden Varianten *für* und *zum* zu erkennen, der Komparativ bildet keine sichtbaren Areale und wird von der *als* Variante klar dominiert und bei der Artikelverdoppelung ist das Bild je nach Frage unterschiedlich. Gemeinsamkeiten zwischen den verschiedenen Phänomenen zu finden, gestaltet sich als sehr schwierig. Innerhalb der Phänomene sind aber, zumindest bei den ersten beiden, durchaus Ähnlichkeiten sichtbar. Die Beurteilung, weshalb sich die erhaltenen Bilder ergeben, wird deshalb zunächst für die einzelnen Phänomene vorgenommen. Anschliessend wird auf weitere Feststellungen eingegangen.

Finalanschluss

Frühere Dokumente belegen die Existenz einer Isoglosse zwischen *für* und *zum* beim Finalanschluss (Weber 1987; Hodler 1969). Seiler (2005) stellt die Frage, wo sich diese syntaktische Grenze befindet. Er erkennt sie richtig als Übergangsregion und nicht als scharfe Grenzlinie. Dies lässt sich sehr schön in den Resultaten (Abschnitt 5.1) nachvollziehen. Mit Ausnahme eines Ausreissers (I.6) zeichnen die vier untersuchten Fragen dieses Grenzgebiet gut nach. Noch deutlicher wird dies auf der kombinierten Karte zum Phänomen A (Abbildung 5-11).

Seiler (2005) stellt fest, dass die *für* Variante insgesamt am häufigsten vorkommt. Auf den Resultaten dieser Arbeit entsteht eher der Eindruck, die *zum* Variante sei verbreiteter. Der Grund liegt darin, dass die *für* Variante seltener dominant vorkommt als die *zum* Variante, jedoch im Gesamtgebiet auftritt. Dies ist erst in der 3D-Darstellung (Abbildung 5-14) nachvollziehbar. Zudem könnten grosse Thiessenpolygone im Gebiet der *für* Variante das Bild verfälschen. Eine Schwäche der Methodik scheint zu sein, dass sie durch das Dominanzprinzip der Einfärbung und der Einteilung in Thiessenpolygone nur eine eingeschränkte Aussagekraft über die Gesamtverteilung besitzt.

Die erwartete West-Ost-Verteilung kann durch die Resultate in eine Nordost-Südwest-Verteilung verfeinert werden.

Zuletzt gilt es noch, die Idee der „schiefen Ebene“, die Seiler hatte, zu überprüfen. Sie beinhaltet gegen Osten hin eine Abnahme der Dichte der Orte, an welchen die *für* Variante vorkommt und eine Abnahme der relativen Häufigkeit gegenüber anderen Varianten. Beide Hypothesen werden durch die Beobachtungen mit der Abnahme der Intensitäten zur Grenzregion hin erhärtet. Somit verläuft die syntaktische Grenze anscheinend tatsächlich nicht nur entlang einer Linie, sondern entlang von kontinuierlich zur Übergangsregion hin abnehmenden Ebenen.

Insgesamt scheint für die Repräsentation des Anschlussmittels in Finalsätzen die flächige Darstellung nach der Methode von Rumpf et al. (2009) geeignet zu sein.

Komparativ

Die Auswahl der Fragen zum Komparativ entstammt dem Text von Friedli (2005), der diese zur Untersuchung des Komparativanschlusses verwendete. Er befand das Phänomen für arealbildend. Genauer wird die *als*-Variante als flächendeckend und die anderen drei Varianten *weder, wie* und *wan* als örtlich begrenzt beschrieben. Die Flächenkarten dieser Arbeit vermögen diese Feststellungen nicht zu widerspiegeln. Die dominante Variante *als* überdeckt die gesamte Deutschschweiz bei allen Fragen.

Die Methodik darf auf ihre Aussagekraft hin kritisch hinterfragt werden. Sie eignet sich nicht zur Entdeckung lokaler Varianten. Allenfalls hätte die dominante Variante von den restlichen getrennt werden sollen. Die Interpolation der übrigen drei Varianten hätte danach vielleicht aufschlussreichere Information über die Arealbildung des Komparativs ergeben.

Deutlich manifestieren sich bei den Karten dieses Phänomens die Randeffekte gegen die Grenzen der Deutschschweiz hin, an denen die Intensitäten stärker geschätzt werden als im Innern des Untersuchungsgebiets.

Artikelverdoppelung

Glaser & Frey (2007) stellen keine regionale Präferenz bei Verdoppelungsphänomenen im Allgemeinen fest. Sie unterschieden sich je nach Phänomen. In dieser Arbeit kann dies nicht nachgewiesen werden, da lediglich ein Phänomen, die Verdoppelung des indefiniten Artikels, untersucht wird und andere Verdoppelungsphänomene wie jene des W-Wortes keine Beachtung finden. Wohl aber kann ein Unterschied innerhalb des Phänomens festgestellt werden. Drei Fragen ergeben drei völlig unterschiedliche Karten, welche keine Gemeinsamkeiten besitzen.

Steiner (2005) untersucht lediglich die Frage I.10 „(e) ganz (e) liebi Frau“ unter anderem auf die geographische Verteilung hin. Sie konstatiert eine Ablehnung der Artikelverdoppelung im Wallis und im Kanton Bern. Das Kerngebiet erstreckt sich laut Steiner entlang eines Bandes von Baselland über die Zentralschweiz hin zu Nordgraubünden. Sie weist darauf hin, dass es sich dabei nicht um die einzige vorkommende Variante im Untersuchungsgebiet handelt. Das erwähnte Band ist auf der interpolierten Oberfläche ebenfalls erkennbar, wenn auch mit einer deutlich komplexeren Struktur (vgl. Abbildung 5-8). Es wird teilweise von der nachgestellten Variante durchdrungen und weist keine kontinuierlichen Intensitäten auf. Wo genau die Artikelverdoppelung vorkommt und wo nicht, vermag die Oberfläche nicht zu zeigen. In der nicht interpolierten Oberfläche ist eine solche Tendenz jedoch zu sehen.

Zu den Fragen II.10 und IV.14 existiert keine Literatur, was die Diskussion erschwert. Die beiden Karten zeigen ein sehr gegensätzliches Bild, in einer wird die nachgestellte, in der anderen die vorangestellte Variante dominant akzeptiert. Dies unterstreicht die Unabhängigkeit verschiedener Fragen bei der Artikelverdoppelung.

Schliesslich hilft die Umwandlung in Flächenkarten begrenzt bei der Frage I.10, deren stark segregierte Originaloberfläche in eine räumlich konsistentere Bänderung umgewandelt werden kann. Bei den beiden anderen Fragen ergibt sich, ähnlich dem Komparativ, ein homogenes Bild einer dominanten Variante über das gesamte Gebiet.

Randeffekte

In den isolierten Gebieten in Graubünden und in den Regionen nahe der Grenze sind erhöhte Intensitäten erkennbar. Dabei handelt es sich eventuell um bereits angesprochene Randeffekte. Da die Schätzung der Intensitäten einen einheitlichen Radius um alle Punkte legt, haben nur wenige Untersuchungspunkte einen Einfluss auf die Einfärbung dieser Regionen. Die Wahrscheinlichkeit bei einer geringen Anzahl von Nachbarn eine einheitliche Klassierung zu erhalten ist grösser als bei vielen einfliessenden Messwerten.

Erweiterungen

Von den verschiedenen Variationen ist wahrscheinlich die kombinierte Finalanschlusskarte (Abbildung 5-11) die spannendste. Sie zeigt eine Möglichkeit zur Aggregierung einzelner Fragen zu einer integrierten Phänomenkarte, die den Gesamteindruck des zu Grunde liegenden Phänomens widergibt.

Die Berücksichtigung der Gewährspersonenzahl, der Einbezug von Präferenzen und auch die Änderung der Aggregierungsebene haben keine grossen Unterschiede in den resultierenden Karten bewirkt. Dies könnte ein Hinweis auf die Robustheit der Methodik sein.

Interessant könnte in Zukunft die dreidimensionale Darstellung sein, da sie eine integrierte Betrachtungsweise verschiedener Intensitäten in einer Visualisierung ermöglicht. Die zweidimensionalen Flächenkarten können dies nicht.

Aussagekraft der Resultate

Die linguistische Aussagekraft der Resultate kann nicht beurteilt werden, da der Autor keinen sprachwissenschaftlichen Hintergrund aufweist. Deshalb würde sich eine Beurteilung durch Experten anbieten. Eventuell ist aber auch das Stadium noch zu früh, jetzt schon umfassende linguistische Beurteilungen aus den erhaltenen Karten vorzunehmen. Die erstellten Karten haben aber gezeigt, dass grundsätzlich die flächenhafte Darstellung von syntaktischen Daten möglich ist, wenn auch nicht für alle Phänomene gleich gut.

Vergleich mit den Resultaten von Pickl & Rumpf:

Der Sprachatlas von Bayrisch-Schwaben hat ein praktisch rasterförmig angeordnetes Untersuchungssetting, was bei der Anwendung von Thiessenpolygonen ein bienenwabenartiges Bild produziert. Zudem befinden sich die Messpunkte in einem mehrheitlich flachen Gebiet. Diese beiden Punkte, in denen sich der SBS klar vom SADS unterscheidet, machen die euklidische Distanz für die Operationalisierung sinnvoller als im vom Alpenkamm durchzogenen Deutschschweizer Gebiet.

Randeffekte lassen sich auch bei Rumpf et al. (2009; 2010) erkennen (vgl. Abbildung 3-7), obwohl diese durch die Normalisierung in der KDE-Formel reduziert werden sollten.

Daneben zeigen ihre Resultate viele, relativ kompakte Regionen. Es sind dort aber bei der Erstellung dieser Karten auch mehr Varianten eingegangen. Zudem wurden mehrere hundert Fragen untersucht, weshalb nicht klar wird, ob für die Publikation nur möglichst arealbildende Resultate ausgewählt wurden, oder ob sich diese Feststellung generell machen lässt. Weiter bilden phonetische Daten die Untersuchungsgrundlage. Diese sind wie lexikalische lokal stärker segregiert als syntaktische Phänomene, denen laut Christen et al. (2010) früher oft gar keine räumliche Gliederung zugesprochen wurde. Vergleicht man die Flächenkarten mit jenen aus dem Kleinen Sprachatlas der deutschen Schweiz, so kommt man intuitiv zu einem ähnlichen Schluss. Es existieren mehr Varianten für phonetische Phänomene als hier syntaktische Varianten verwendet worden sind. Die vermutete räumliche Unabhängigkeit syntaktischer Daten lässt sich aber nicht bestätigen. Weitere Untersuchungen dazu sind im zweiten Hauptteil nachzulesen.

Teil III: Geostatistische Analyse dialektsyntaktischer Phänomene

Im zweiten Hauptteil markieren in der Sprachforschung geäusserte raumbezogene Hypothesen zu den untersuchten Phänomenen den Startpunkt (Kapitel 7). Um sie zu überprüfen, werden passende geostatistische Verfahren ausgewählt. Nach einer Erläuterung der methodologischen Grundlagen werden sie auf die SADS-Daten umgesetzt (Kapitel 8). Nach der Präsentation (Kapitel 9) bildet die Diskussion der Resultate der Geostatistik in Kapitel 10 den Abschluss dieses Teils.

7. Raumbezogene Hypothesen

Für alle untersuchten Phänomene wurden im Vorfeld Vermutungen geäussert, welche räumlichen Verteilungen und Abhängigkeiten vorhanden sein könnten. Diese Hypothesen wurden vor allem auf Basis der im SADS-Kontext erstellten Punktkarten aufgestellt und in Absprache mit Gabriela Bart und Prof. Elvira Glaser verfeinert. Tabelle 7-1 zeigt diese Hypothesen für die untersuchten Phänomene. Darin sind ebenfalls die geostatistischen Verfahren aufgelistet, mit welchen sie überprüft werden sollen. Eine spezielle Rolle nehmen die Strukturkenngrössen ein, die von Rumpf et al. (2009) beschrieben werden. Sie sollen ebenfalls zur Beurteilung der Hypothesen herangezogen werden.

Phänomen	Vermutung über die räumliche Verteilung	Räumliche Hypothesen	Gewählte Verfahren zur Überprüfung
A: Finalanschluss	markante Nordost-Südwest-Verteilung	Räumliche Autokorrelation der beiden dominanten Varianten entlang der NO-SW-Achse, keine Autokorrelation entlang NW-SO-Achse	Moran's I, Semivariogramme von 2 Untersuchungsbändern auf verschiedenen Achsen
	Verteilung der Varianten entlang von schiefen Ebenen	Trend nachweisbar	„schiefe Ebenen": Trendoberflächenanalyse
B: Komparativ	Vorherrschende dominante Variante	keine Autokorrelation der dominanten Variante	Moran's I
	einzelne Varianten bilden zusätzlich kleinere Areale	lokale räumliche Autokorrelation hoher Intensitäten bei mindestens einer nicht dominanten Variante	Getis-Ord G_i^*
C: Artikelverdoppelung	Phänomen ist zufällig verteilt	Keine räumliche Autokorrelation	Moran's I

Tabelle 7-1: Vermutungen zur räumlichen Verteilung der untersuchten Phänomene mit zugehörigen Hypothesen und Verfahren der Geostatistik

Finalanschluss

Die erste Vermutung, die auf Basis von Punktkarten aus den SADS-Daten entstand, ging von einer Ost-West-Verteilung der zwei dominanten Varianten aus. Aus den Resultaten des ersten Hauptteils geht aber hervor, dass eher eine Nordost-Südwest-Verteilung vorliegt. Insgesamt muss folglich erstens eine globale Autokorrelation vorhanden sein, was mit dem Moran's Index bestätigt werden soll. Zweitens soll gezeigt werden, dass diese Abhängigkeit grösser entlang der NW-SO-Achse ist als von NO nach SW hin. Hierzu werden zwei Bänder entlang den Hauptrichtungen auf die Punktdaten gelegt und je ein Semivariogramm aus den Intensitäten der beiden dominanten Varianten erstellt.

Seiler (2005) hatte die Vermutung, dass sich für den Finalanschluss die zwei dominanten Varianten entlang von „schiefen Ebenen" bewegen. So sollen die Varianten zunächst klar vorherrschend sein ohne erkennbare Abnahme der Akzeptanz zum Gebiet hin, an welchem die andere Variante dominiert. Ab einer gewissen Grenze nimmt dann die Akzeptanz gleichmässig ab, bis die Variante schliesslich nicht mehr

vorkommt. Was bereits mit der dreidimensionalen Darstellung der ersten Frage gezeigt werden konnte, soll nun mittels Trendoberflächen, die an die Akzeptanzwerte angepasst werden, erhärtet werden.

Komparativ

Die erstellten Flächenkarten zu diesem Phänomen machen bereits klar, dass eine Variante klar dominant vorkommt im gesamten Untersuchungsgebiet. Diese Feststellung lässt sich mit Moran's I statistisch überprüfen, indem von keiner globalen Autokorrelation der dominanten Variante im Gebiet ausgegangen wird. Die Variante ist unabhängig von der geographischen Lage dominant, sie bildet keine Areale. Karten aus Friedli (2005) zeigen zusätzlich eine räumliche Konzentration der anderen Varianten in Gebieten mit relativ hohen Intensitäten. Es kann daraus geschlossen werden, dass kleine Cluster entstehen. Die lokale geostatistische Messgrösse G_i^* hilft bei der Detektion von solchen Hot Spots.

Artikelverdoppelung

Die Hypothese zu diesem Phänomen ist einfach. Es wird davon ausgegangen, dass keine räumliche Autokorrelation vorhanden ist. Diese Vermutung scheint von den Eindrücken der erstellten Flächenkarten her vernünftig. Die Vermutung soll mit Moran's I nachgewiesen werden.

8. Methodik

Strukturkenngrössen nach Rumpf et al. (2009) sind ein einfaches Werkzeug, um, basierend auf aufbereiteten Sprachdaten, einen groben Überblick über deren Verteilung zu erreichen. Sie werden deshalb gleich als erstes behandelt (Abschnitt 8.1).

Danach wird ein Methodenmix von geostatistischen Verfahren zur weiteren räumlichen Analyse der behandelten Sprachdaten vorgestellt. Der Geostatistik inhärent ist der räumliche Bezug. Es wird überprüft, ob eine Verteilung von Punkten durch die geographische Distanz beeinflusst wird. Bevor auf die einzelnen geostatistischen Verfahren eingegangen werden kann, muss deshalb als erstes das Grundkonzept, die räumliche Autokorrelation, erläutert werden (Abschnitt 8.2).

Es gibt verschiedene Messgrössen zur Diagnose, ob ein Phänomen räumlich abhängig verteilt ist, wobei jene, die in dieser Arbeit zum Zug kommen, in den Unterabschnitten 8.3.1 und 8.3.2 kurz beschrieben werden. Unterabschnitt 8.3.3 befasst sich mit dem Semivariogramm, einer mächtigen Visualisierungs- und Analysemöglichkeit für räumliche Abhängigkeit. Im letzten Unterabschnitt (8.3.4) wird die Trendoberflächenanalyse vorgestellt, ein an die Varianzanalyse angelehntes Verfahren für räumliche Daten.

8.1. Strukturkenngrössen

Zur Charakterisierung der Flächenkarten schlagen Pickl & Rumpf (unveröffentlicht) drei so genannte Strukturkenngrössen vor: Die Gesamtgrenzlänge, das mittlere Auftretensgewicht, sowie zwei Grössen zur Bestimmung der Homogenität der Karte.

8.1.1. Komplexität *C*

$$C = \frac{\sum GL_{spez}}{\sum GL_{\max}} \tag{8-1}$$

Hier werden alle Gebiete, in denen eine Variante dominant vorkommt, zusammengenommen und daraus errechnet, wie lang die Grenze des entstehenden Flächenmusters ist. Diese Summe GL_{spez} wird anschliessend ins Verhältnis zur summierten Länge aller Grenzen GL_{max} gesetzt. Dieser relative Wert C erlaubt Aussagen über die **Komplexität der Karte**. Ist C klein, so ist mit wenigen, kompakten Gebieten zu rechnen. Die Karte ist nur wenig komplex. Je grösser aber C und damit die Gesamtgrenzlängen der verschiedenen Varianten, relativ gesehen zur maximal möglichen Grenzlänge, desto unruhiger und zerstückelter sind die einzelnen Varianten über den Raum verteilt. Eine hohe Komplexität der resultierenden Karte ist die Konsequenz.

Umsetzung auf die SADS-Daten

Da in den zugrundeliegenden Arbeiten (Rumpf et al. 2009; Rumpf et al. 2010; Pickl & Rumpf unveröffentlicht) nicht im Detail steht, welche Linien als Grenzen verwendet werden sollen, ist im Folgenden die Aussengrenze des Gesamtgebiets aPG nicht zu den Grenzlängen hinzugezählt worden, da sie keine Grenze zwischen zwei Gebieten darstellt. Ebenfalls nicht erwähnt ist der hier auftretende Fall der „weissen Flecken“, an denen in einem Gebiet keine Variante dominant ist. Da diese zum Gesamtgebiet gehören, werden Grenzen zu diesen Gebieten in den hier vorgenommenen Berechnungen einbezogen. Ein Weglassen hätte einen positiven Einfluss auf C und damit eine Verfremdung zur Folge.

Das Konzept der Gesamtgrenzlänge wird in ArcGIS umgesetzt. Dazu werden die auf die Grenzen der Deutschschweiz zugeschnittenen Thiessenpolygone der Untersuchungsgebiete verwendet, in Verbindung mit den aus dem SADS-Datensatz entnommenen, nach Varianten gegliederten Klassierungsattributen. Anschliessend wird ein Zusammenschluss (`Dissolve`) der einzelnen Gebiete nach dominanter Variante vorgenommen. Die Grenzen zwischen benachbarten Gebieten, welche dieselbe dominante Variante besitzen, werden dadurch eliminiert.

Als nächstes werden in der Attributtabelle mittels `Calculate Geometry` die Grenzlängen der resultierenden Polygone aufsummiert. Von dieser Summe wird aPG abgezogen, wodurch GL_{spez} resultiert.

Die maximale Gesamtgrenzlänge GL_{max} kann wieder mit `Calculate Geometry` berechnet werden. Als erstes wird die Länge aller vorhandenen Polygongrenzen PG ermittelt. Da in ArcGIS in Shapefiles Polygone gemäss dem so genannten „Spaghettimodell" dargestellt werden, existieren für jede Grenze zwischen zwei Polygonen A und B zwei Linien, die Aussenlinie von Polygon A und jene von Polygon B. Die Gebietsgrenzen kommen nur einfach vor, da sie jeweils nur von einem Polygon benutzt werden. Aus diesen Überlegungen folgt, dass zuerst aPG abgezogen wird und anschliessend die übriggebliebene Summe halbiert wird (Formel 8-2):

$$GL_{\max} = \frac{\sum PG - \sum aPG}{2} \tag{8-2}$$

Schliesslich wird GL_{spez} durch GL_{max} dividiert, um die angestrebte Strukturkenngrösse zu erhalten.

8.1.2. Gebietskompaktheit der Fläche einer Variante $\bar{l}_x$ bzw. einer Karte $\bar{L}$

Die **Gebietskompaktheit** $\bar{l}_x$ einer Variante x in dem ihr zugeordneten Gebiet $T(x)$ ist definiert durch die Formel:

$$\bar{l}_x = \frac{1}{|T(x)|} \sum_{t_j \in T(x)} l_x(t_j) \tag{8-2}$$

Diese Kenngrösse entspricht dem gewichteten Mittel der Akzeptanz l_x einer einzelnen Variante an allen Betrachtungspunkten $T(x)$, an denen sie dominant vorkommt. Hohe Werte dieses mittleren Auftretensgewichts bedeuten, dass die Dominanz einer Variante in deren Vorkommensgebiet insgesamt hoch ist. Weiter lässt sich aus der Gebietskompaktheit der einzelnen Varianten eine **Gebietskompaktheit der Karte** $\bar{L}$ (Formel 8-3) errechnen. Diese entspricht dem nach $|T(x)|$ gewichteten Mittel der Gebietskompaktheiten aller Varianten.

$$\bar{L} = \sum_x \frac{|T(x)|}{n} \cdot \bar{l}_x = \frac{1}{n} \sum_x \sum_{t_j \in T(x)} l_x(t_j) \tag{8-3}$$

Rumpf et al. (2009) sehen die Bedeutung von $\bar{L}$ vor allem darin, dass daraus geschlossen werden kann, wie sinnvoll die Umwandlung einer zugrundeliegenden Punktkarte in eine Flächenkarte ist. Ist die Gebietskompaktheit einer Karte hoch, so kann mit geringer Veränderung der tatsächlichen Werte eine Flächenkarte erstellt werden, welche die Punktkarte in wenige, zusammenhängende Gebiete einteilt. Ist $\bar{L}$ dagegen klein, so entstehen bei geringen Veränderungen der Originalwerte unruhige und heterogene Karten. Kompakte Karten zu erstellen verlangt in diesem Fall, dass die Werte der interpolierten Karte stark vom Original abgeändert werden müssen.

Umsetzung auf die SADS-Daten

Das mittlere Auftretensgewicht entspricht den bereits für die Attributtabellen der einzelnen Fragen berechneten Akzeptanzen einer Variante an einem Ort. Da diese Tabellen in R für die KDE bereits importiert wurden, lässt sich die Gebietskompaktheit in einer einfachen Funktion umsetzen. Hier der Algorithmus dazu:

Umsetzung Formel 8-2 - Gebietskompaktheit $\overline{l_x}$ einer Variante x in dem ihr zugeordneten Gebiet $T(x)$:

```
Für alle Gebiete
   Wenn Variante x dominant vorkommt
      Addiere die Akzeptanz von x auf (aufsummierte Akzeptanzen)
      Addiere die Anzahl Gebiete auf (Zahl der Vorkommensgebiete)
Teile die aufsummierten Akzeptanzen durch die Zahl der Vorkommensgebiete
```

Umsetzung Formel 8-3 für die Gebietskompaktheit $\bar{L}$ der Karte:

```
Für alle Gebiete
      Für alle Varianten
            Wenn Variante x dominant vorkommt
            Summiere die Akzeptanzen auf
            Bilde die Summe der aufsummierten Akzeptanzen
      Berechne Gesamtzahl der Gebiete
      Teile die aufsummierten Summen aller Akzeptanzen durch die Gesamtzahl
      der Gebiete
```

8.1.3. Homogenität eines Gebiets $\overline{b_x}$ bzw. einer Karte $\bar{B}$

Das Vorgehen zur Berechnung von $\overline{b_x}$ (Formel 8-4) und $\bar{B}$ (Formel 8-5) ist analog zur Berechnung der Kompaktheitsmasse, nur werden hier, anstatt der gemessenen Akzeptanzwerte, die durch die Interpolation geschätzten Intensitätswerte $b(t_j)$ der Varianten an einem Ort verwendet. Die Autoren verwenden diese Kennzahlen, um einen Eindruck über die **Homogenität** der Verteilung **einer einzelnen Variante**, bzw. **der gesamten Karte**, zu erhalten. Geringe Werte von $\overline{b_x}$ bedeuten, dass die Variante x in jenen Gebieten, in denen sie vorkommt, stark in Konkurrenz zu weiteren dort vorkommenden Varianten steht, wogegen hohe Werte für eine starke Dominanz der einzelnen Variante in diesen Gebieten sprechen. Ist $\bar{B}$ hoch, so sind die Gebiete einer Karte stark segregiert, bei tiefen Werten ist eine grössere Durchmischung der einzelnen Varianten zu erwarten.

$$\overline{b_x} = \frac{1}{|T(x)|} \sum_{t_j \in T(x)} b(t_j) \tag{8-4}$$

$$\overline{B} = \sum_x \frac{|T(x)|}{n} * \overline{b_x} \tag{8-5}$$

Umsetzung auf die SADS-Daten

Auch diese Kennzahlen sind am einfachsten in R umzusetzen. Da diese Kennzahlen direkt verwandt sind mit den Berechnungen zur Gebietskompaktheit, müssen im Code lediglich die tatsächlichen Intensitäten der Varianten an einem Gebiet durch die mittels KDE geschätzten Intensitäten ersetzt werden.

Die Skripte zur Berechnung der Kompaktheiten und Homogenitäten sind auf der Software-CD im Ordner *6_Strukturkenngrössen* beigelegt.

8.2. Räumliche Autokorrelation:

Sucht man in der Literatur nach einer Definition von räumlicher Autokorrelation, so wird man mit grosser Wahrscheinlichkeit dem berühmten „ersten Gesetz der Geographie“ von Waldo R. Tobler begegnen:

> „Everything is related to everything else, but near things are more related than distant things.“
>
> (Tobler 1970: 236)

Es gibt wohl keine Definition, die einleuchtender das Thema der räumlichen Abhängigkeit umschreibt. Es kann eine Einteilung in negative und positive Autokorrelation gemacht werden. Positive Autokorrelation bedeutet eine Bildung von Clustern, sprich eine Klumpung eines Phänomens. Untersuchungsobjekte, die nahe beieinander liegen, erzielen dabei ähnliche Werte. Negative Autokorrelation bedeutet eine regelmässige Abwechslung verschiedener Werte, was im Extremfall bei zwei Klassen einem Schachbrettmuster entspricht. Keine Autokorrelation herrscht vor, wenn ein Phänomen zufällig über den Raum verteilt ist (O’Sullivan & Unwin 2010). Abbildung 8-1 zeigt schematisch für 2 Klassen Punktverteilungen der drei möglichen Ausprägungen von räumlicher Autokorrelation.

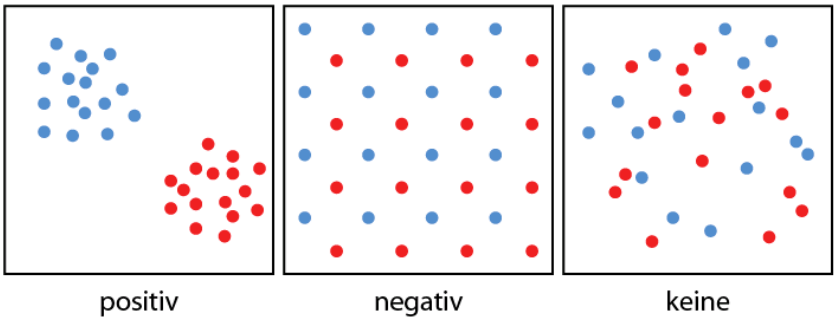


Abbildung 8-1: Die drei Formen von räumlicher Autokorrelation am Beispiel einer Punktverteilung mit 2 Klassen

Das Konzept der räumlichen Autokorrelation ist die Basis für geostatistische Verfahren. Verschiedene Methoden für deren Messung, die für die Grundhypothesen zur Verteilung der Sprachphänomene gebraucht werden, sind im folgenden Abschnitt erläutert.

Test auf räumliche Autokorrelation

Mithilfe eines Z-Tests kann auf räumliche Autokorrelation getestet werden. Ausgegangen wird von der „Complete Spatial Randomness“ (CSR)-Hypothese. Sie besagt, dass ein Punktdatensatz zufällig verteilt ist, sodass keine räumlichen Muster zu erkennen sind (O’Sullivan & Unwin 2010).

Die Nullhypothese heisst demnach:

H_0: „Die räumliche Verteilung ist zufällig.“

Zwei Masse sind wichtig für die Beurteilung des Testergebnisses, der p-Wert und der Z-Wert.

***p*-Wert**

Der p-Wert sagt aus, mit welcher Wahrscheinlichkeit eine statistische Grösse einer bestimmten Verteilung folgt. Unter der CSR-Annahme bedeutet dies hier, dass mit dieser Wahrscheinlichkeit eine zufällige räumliche Verteilung vorliegt. Wird ein gewisser p-Wert unterschritten, wird die Nullhypothese verworfen und von einer statistisch signifikanten Verteilung gesprochen.

***Z*-Wert**

Der Z-Wert besagt, wie viele Standardabweichungen eine Verteilung von der zufälligen Verteilung entfernt ist und weist darauf hin, wie stark ein räumlicher Zusammenhang ist.

Wie verschiedene Z- und p-Werte eingeordnet werden können, wird in Abbildung 8-2 gezeigt.

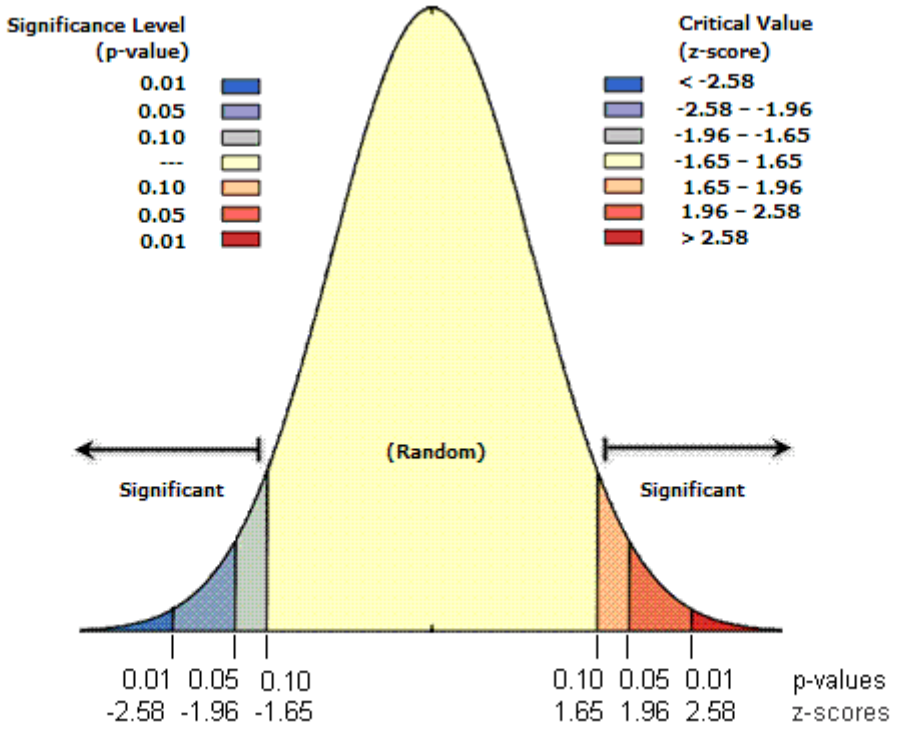


Abbildung 8-2: Normalverteilungskurve mit kritischen ***p***-Werten und ***Z***-scores verschiedener Signifikanz-Levels (nach ArcGIS Desktop Help 10.0[1])

8.3. Verwendete geostatistische Methoden

Je nach Hypothese sind andere geostatistische Verfahren sinnvoll. Die ersten beiden folgenden Unterabschnitte befassen sich mit Grössen, die Auskunft über eine bestimmte räumliche Charakteristik geben. Sie alle gehen von einer Gewichtung aus, welche in Matrix-Form abgespeichert wird. In dieser Gewichtungsmatrix W werden alle Punkte miteinander verglichen und daraus ein spezifischer Wert abgeleitet. Das Gewicht einer Punktkombination wird mit w_{ij} abgekürzt. Verschiedene Möglichkeiten der Gewichtung können zur Bestimmung dieses Wertes angewendet werden, beispielsweise Distanz oder die Bikonnektivität, sprich, ob zwei Punkte benachbart sind oder nicht (Rogerson 2010). Bei allen Verfahren hat die Gewichtung einen entscheidenden Einfluss auf das Resultat. (O'Sullivan & Unwin 2010).

8.3.1. Moran's I

Moran (1950) beschrieb einen Index, um zu bestimmen, ob ein Phänomen eine zufällige oder eine abhängige räumliche Struktur besitzt. Moran's I (Formel 8-6) ist eine Übertragung von einem nicht-räumlichen Korrelationsmass und verlangt standardmässig räumliche Einheiten, die mindestens intervallskaliert sind (Moran 1950).

$$I = \left[\frac{n}{\sum_{i=1}^{n} \left(y_i - \bar{y} \right)^2} \right] \times \left[\frac{\sum_{i=1}^{n} \sum_{i=1}^{n} w_{ij} \left(y_i - \bar{y} \right) \left(y_j - \bar{y} \right)}{\sum_{i=1}^{n} \sum_{i=1}^{n} w_{ij}} \right] \qquad (8\text{-}6)$$

Die zentrale Grösse ist die **Kovarianzmatrix** (Formel 8-7), wobei i und j verschiedene räumliche Einheiten bzw. Zonen beschreiben und y_i bzw. y_j die Datenwerte an diesen Stellen. Die Kovarianz zwischen den zwei Einheiten ergibt sich durch die Multiplikation der Unterschiede von zwei Zonen vom Mittelwert $\bar{y}$ aller Datenwerte. Diese Kovarianzwerte werden über die Distanzbeziehung, welche in der

[1] What is a z-score? What is a p-value?
http://help.arcgis.com/en/arcgisdesktop/10.0/help/index.html#/What_is_a_z_score_What_is_a_p_value/005p00000006000000/, Zugriff: 20.4.2011

Gewichtungsmatrix enthalten ist, gewichtet. Die restlichen Terme in Moran's Index dienen lediglich dessen Normalisierung, damit die Anzahl und Grössen der Untersuchungswerte keinen Einfluss auf das Resultat nehmen (O'Sullivan & Unwin 2010).

$$Kovarianz = \sum_{i=1}^{n}\sum_{i=1}^{n} w_{ij}\left(y_i - \bar{y}\right)\left(y_j - \bar{y}\right) \tag{8-7}$$

Der Index hat einen Wertebereich von -1 bis 1. Ein **positives *I*** bedeutet, dass Objekte, die nahe beieinander liegen, Werte auf der gleichen Seite des Mittelwertes haben, sprich dass eine **positive Autokorrelation** im untersuchten Gebiet vorliegt. Ist ***I* negativ**, so sind die Werte der naheliegenden Objektpaarungen vornehmlich nicht auf derselben Seite des Mittelwertes, wodurch eine **negative Autokorrelation** abgeleitet werden kann. **Werte um 0** entsprechen einer zufälligen Verteilung der Punktwerte. In diesem Fall gibt es **keine Autokorrelation**.

Test auf Autokorrelation

Um mithilfe von Moran's *I* zu einer statistisch signifikanten Aussage zu gelangen, kann ein *Z*-Score gemäss der Formel 8-8 errechnet werden. $E[I]$ ist dabei der Erwartungswert, welcher im Falle von *I* dem theoretischen Mittelwert entspricht und $V[I]$ der Varianz von *I* (Rogerson 2010).

$$Z_I = \frac{I - E[I]}{\sqrt{V[I]}} \tag{8-8}$$

Wird bei einem Test auf räumliche Autokorrelation die Nullhypothese verworfen, so kann mithilfe von *I* nicht nur ausgesagt werden, ob räumliche Autokorrelation im untersuchten Gebiet existiert, sondern auch in welchem Masse und ob diese positiv oder negativ ist (Getis 2010).

Umsetzung auf die SADS-Daten

Moran's *I* kann nur auf die Intensitätswerte einer Klasse auf einmal angewendet werden, da er kategoriale Unterschiede nicht berücksichtigt. Somit werden für alle Klassen der eingeschränkten Klassierung die Werte für *I*, der *p*-Wert und der *Z*-Score separat errechnet.

Der `Geostatistical Analyst` in ArcGIS bietet ein Tool (`Analyzing Patterns -> Spatial Autocorrelation (Moran's I)`), um Moran's *I* und gleichzeitig die statistischen Masse für die Autokorrelation für einen Punktdatensatz mit Attributwerten zu berechnen. Die Resultate werden dabei einerseits in der Konsole ausgegeben, andererseits kann auf Wunsch eine kleine Grafik erstellt werden, die darüber informiert, ob eine räumliche Abhängigkeit statistisch gesehen existiert (Abbildung 8-3).

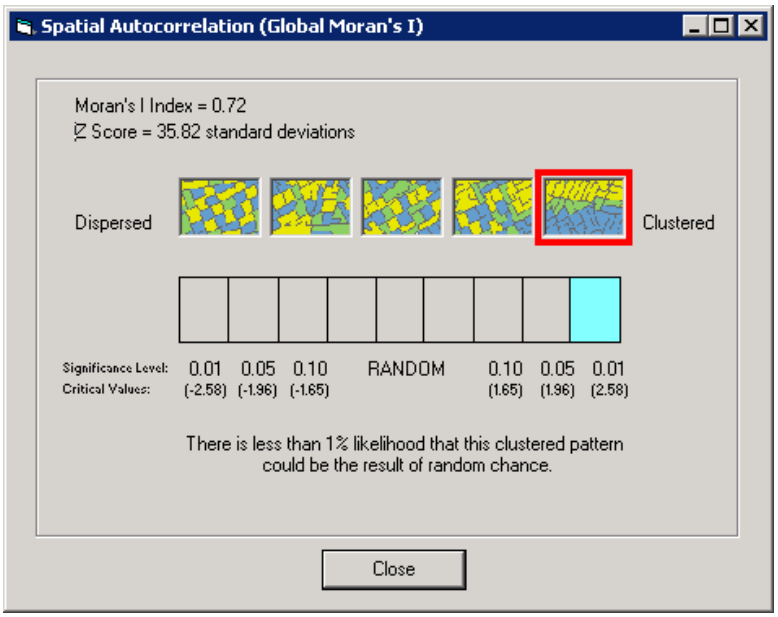


Abbildung 8-3: Beispiel einer grafischen Ausgabe in ArcGIS für die Berechnung von Moran's *I*

Es gilt zu beachten, wie die Distanzen gewichtet werden sollen. Um eine bessere Vergleichbarkeit zu erzielen, wird bei Moran's I, wie auch beim nachfolgend erläuterten Getis-Ord G_i^*, eine feste Distanz bestimmt, innerhalb der die Werte der umliegenden Punkte berücksichtigt werden sollen. Getis & Ord (1992) empfehlen als Faustregel, eine Distanz d zu wählen, die gewährleistet, dass jeder Punkt im Umreis dieser etwa 8 Nachbarn besitzt. w_{ij} ist dadurch eine 1/0 Matrix, welche allen Beziehungen zwischen Punkten einer räumlichen Verteilung den Wert „0" zuordnet, wenn diese ausserhalb d liegen

Diese Distanz, über `Calculate Distance Band` in ArcGIS ermittelt, beträgt für die SADS-Punkte im Durchschnitt 13'585 Meter. In der Analyse wird der Wert auf 15'000 Meter aufgerundet, da die Punkte etwas ungleichmässig verteilt sind und deshalb auch solche mit wenigen Nachbarn, etwa in Berggebieten, von genügend Punkten beeinflusst werden sollen.

8.3.2. Getis-Ord G_i

Das zur G statistics Famile gehörende Mass G_i dient als Ergänzung zu Moran's I. Damit lassen sich kleinräumige Verteilungen von Punkten, an welchen sich Phänomene häufen (Hot Spots) oder besonders selten vorkommen (Cold Spots), beschreiben (Getis & Ord 1992).

G_i (Formel 8-9) ist ein Mass für die räumliche Konzentration eines Gebietes (Getis & Ord 1992). Im Gegensatz zu Moran's I ist es keine globale, sondern eine lokale statistische Grösse (Rogerson 2010). Es wird für jeden Untersuchungspunkt einzeln aus dem Messwert des Punktes x_i und der Werte x_j der innerhalb d liegenden benachbarten Punkte berechnet, womit lokal bestimmt wird, ob ein Phänomen im kleinen Raum über- oder untervertreten ist.

$$G_i(d) = \frac{\sum_{i=1}^{n} w_{ij}(d) x_i x_j}{\sum_{i=1}^{n} (d) x_i x_j} \tag{8-9}$$

Aus dem Produkt der Attributwerte aller Punkte könnte wieder eine globale Statistik, *General G* errechnet werden. Da mit Moran's I bereits eine solche vorliegt, wird hier aber darauf verzichtet.

Eine kleine Erweiterung zu G_i bietet G_i^*. Hier werden nicht nur die umliegenden Punkte berücksichtigt, um G_i eines Punktes zu berechnen. Der Messpunkt selbst wird auch betrachtet, was geringfügige Anpassungen an der Berechnung zur Folge hat. Die Resultate ändern sich dadurch nicht gross, gerade bei grösseren Datensätzen wie beim SADS.

Test auf Autokorrelation

Die Nullhypothese lautet:

H_0: „Ein Set von x Werten innerhalb d an Ort i ist zufällig verteilt".

Ord & Getis (1995) stellen eine standardisierte G_i^*-Statistik vor, die so angepasst wird, dass sie der Standardabweichung der Normalverteilung entspricht und somit gleich als Z-Wert verwendet werden kann. Daraus kann wieder geschlossen werden, ob eine Verteilung signifikant ist oder nicht. Bei hohen positiven Z-Werten liegt eine Häufung von Werten vor, die grösser als der Durchschnittswert sind und damit einen Hot Spot bilden. Im umgekehrten Fall, sprich bei hohen negativen Z-Werten, ist es ein Cold Spot.

Umsetzung auf die SADS Daten:

Die Berechnungen der lokalen Verteilungen werden wieder einzeln auf die Varianten der eingeschränkten Klassierungen angewendet.

In ArcGIS lässt sich die standardisierte G_i^* Statistik über das Werkzeug „High/Low Clustering" in der „Spatial Statistics" Toolbox für eine Punktmenge berechnen. Als Gewichtungsmethode wird dasselbe fixe Distanzband wie bei der Berechnung von Moran's I gewählt, um eine ungefähre Zahl von Nachbarn zu berücksichtigen.

Die dadurch generierten Shapefiles erhalten als Attribute die Grösse G_i^*, sowie p. Die Punkte werden automatisch nach dem Z-Wert eingefärbt. Rote Farbtöne entsprechen signifikant positiven, blaue signifikant negativen und gelbe nicht signifikanten Häufungen.

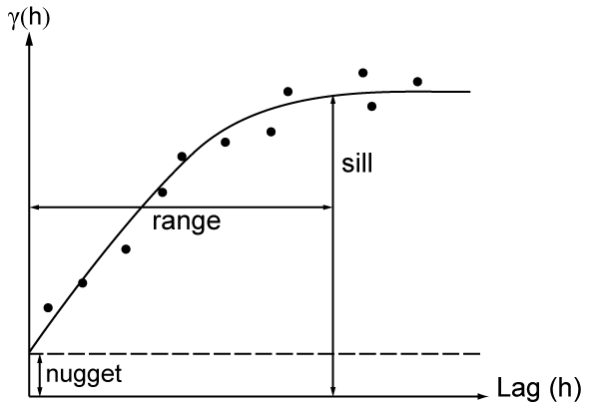
8.3.3. Semivariogramm

Bevor auf das Semivariogramm eingegangen werden kann, sollte die „regionalized variable theory" erwähnt werden. Diese unterteilt die räumliche Variation von Messwerten in drei Komponenten. Eine *strukturelle Komponente* mit konstantem Mittelwert oder einem Trend; eine zufällige, aber räumlich korrelierende Komponente, *regionalisierte Variable* genannt und eine räumlich nicht korrelierende, zufällige Komponente, die als *Rauschen* bezeichnet wird (Burrough & McDonnell 1998).

Sind diese Bedingungen für eine Reihe von Messpunkten erfüllt, so kann die Semivarianz $\hat{\gamma}(h)$ (Formel 8-10) errechnet werden. Sie beschreibt die Hälfte der Varianz der Messwerte in Abhängigkeit der Distanz h zwischen den Punkten. n entspricht der Anzahl Messpunkte und $z(x_i)$ den Messwerten an der Stelle x_i.

$$\hat{\gamma}(h) = \frac{1}{2n}\sum_{i=1}^{n}\left[z\left(x_i\right) - z\left(x_i + h\right)\right]^2 \qquad (8\text{-}10)$$

Geplottet gegen die Distanz kann aus den Semivarianzen ein experimentelles Variogramm erstellt werden. h wird dabei als *lag* bezeichnet. Um die Aussage eines Variogramms, welches oft eine schwer lesbare Punktwolke erzeugt, zu vereinfachen, werden die Semivarianzwerte erstens pro lag-Klasse gemittelt und es wird versucht, eine mathematische Funktion, das so genannte Variogrammmodell, zu fitten (O'Sullivan & Unwin 2010). Der daraus resultierende Graph (Abbildung 8-4) besteht aus dem *lag*, einem *nugget*, einer *range* und einer *sill*. Das *nugget* entspricht dem Rauschen der „regionalized variable theory" und die *range* zeigt, in welchem Distanzspektrum eine räumliche Abhängigkeit der Messwerte vermutet werden kann. Schliesslich beschreibt die *sill* das Höchstmass der Semivarianz[2].



8-4: Semivariogramm mit nugget, range, sill und lag (nach Burrough & McDonnell 1998: 135)

Umsetzung auf die SADS-Daten

Beim Finalanschluss werden bei allen Fragen Semivariogramme für die Varianten-Intensitäten der eingeschränkten Klassierung gefittet. Als erstes werden zwei Untersuchungsbänder über die Punktdaten gelegt. Ausgehend von der Phänomenkarte A des ersten Hauptteils sind sie so angeordnet, dass sie einmal parallel zum optisch vermuteten Trend laufen, also von Südwesten nach Nordosten und einmal vertikal dazu von Nordwesten nach Südwesten (Abbildung 8-5). Die Idee ist, dass im SW-NO-Untersuchungsfenster eine räumliche Abhängigkeit über ein grösseres Gebiet zu sehen ist.

[2] Geoinformatik-Lexikon Uni Rostock:
http://www.geoinformatik.uni-rostock.de/lexikon.asp, Zugriff: 20.4.2011

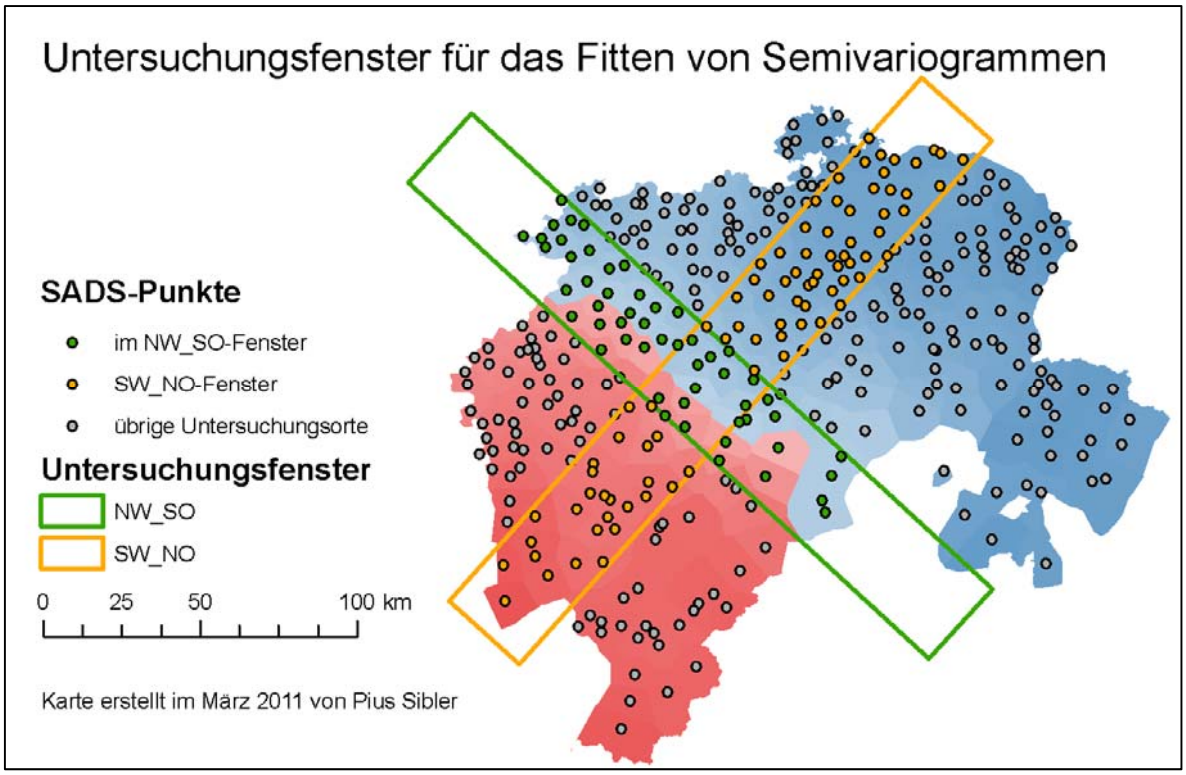


Abbildung 8-5: Untersuchungsfenster entlang (orange) und vertikal (grün) zum vermuteteten SW-NO-Trend

Mithilfe des `Geostatistical Wizards` in der `Geostatistical Analyst` - Erweiterung von ArcGIS wird mit Ordinary Kriging ein experimentelles Variogramm der Intensitäten für jede Klasse der vier Fragen erstellt. Im Dialog sind die lag-Grösse, welche mehrere Punkte innerhalb eines bestimmten Distanzradius zusammenfasst, sowie die Anzahl dieser lags, einstellbar. Sie werden gemäss der Faustregel gewählt, ungefähr die Hälfte der maximalen Distanz der gewählten Punktwolke zu betragen, wenn man die lag-Grösse mit der Anzahl lags multipliziert (Isaacs & Srivastava 1989). Als mathematisches Modell dient das häufig eingesetzte sphärische Modell (O'Sullivan & Unwin 2010).

8.3.4. Trendoberflächenanalyse

Die Trendoberflächenanalyse (TA) ist eine globale Interpolationsmethode. Global heisst, dass die Informationen aller Messpunkte Einfluss auf die Interpolation haben (Burrough & McDonnell 1998).

Bei der Bildung von Trendoberflächen wird versucht, eine Ebene n-ter Ordnung so in eine Punktwolke von verhältnisskalierten und georeferenzierten Messpunkten zu legen, dass die Summe der quadrierten Abstände zu den Punkten minimiert wird. Eine Trendoberfläche entspricht also einer räumlichen Regression (Burrough & McDonnell 1998). Die einfachste Möglichkeit ist eine lineare Trendoberfläche, sprich eine Ebene erster Ordnung, welche durch das Polynom

$$z_i = \beta_0 + \beta_1 x_i + \beta_2 y_i \qquad (8\text{-}11)$$

definiert ist, wobei z_i der Messwert, x_i und y_i die Koordinaten des Punktes i und β_0, β_1 und β_2 Koeffizienten sind. Ebenen höherer Ordnung gleichen sich immer stärker der tatsächlichen Verteilung der Punkte an, Aussagen über allgemeine Trends machen aber zunehmend weniger Sinn, da sie immer komplexer werden.

Hat man für eine Punktwolke eine Trendoberfläche gerechnet, so können die Abstände der Punktwerte zur „gefitteten" Ebene ausgerechnet werden. Diese Abstände werden „Residuen" genannt. Sie sind der Ausgangspunkt für die eigentliche Trendoberflächenanalyse, bei der versucht wird, eine statistische Aussage über eine Punktverteilung zu erhalten. Dazu wird die Technik der Varianzanalyse, auch als *F*-Test bekannt, verwendet (Burrough & McDonnell 1998). Als „Goodness of Fit"-Wert dient R^2. Es wird auch Bestimmungsmass oder Determinationskoeffizient genannt (Pearson & Lee 1897). R^2 ergibt sich gemäss

Formel 8-12 aus der Summe aller quadrierten Residuen ε_i (SSE) und der Summe der quadrierten Unterschiede der Messwerte z_i (SS_z) zu ihrem Mittelwert $\bar{z}$.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \varepsilon_i^{\,2}}{\sum_{i=1}^{n} (z_i - \bar{z})^2} = 1 - \frac{SSE}{SS_z} \tag{8-12}$$

Nun kann noch getestet werden, ob R^2 signifikant ist oder nicht. Dazu dient der F-Test (O' Sullivan & Unwin 2010):

$$F = \frac{R^2 \,/\, df_{Oberfläche}}{(1 - R^2) \,/\, df_{Residuen}} \tag{8-13}$$

$df_{Oberfläche}$ steht dabei für die Freiheitsgrade der gefitteten Oberfläche, welche der Anzahl Koeffizienten der Ebenengleichung minus 1 entspricht. $df_{Residuen}$ steht für die Anzahl Freiheitsgrade der Residuen, welche durch die Anzahl Messpunkte minus $df_{Oberfläche}$ minus 1 gegeben ist.

Die Nullhypothese lautet dabei:

H_0: „Es gibt keine Unterschiede zwischen den Varianzen."

Sie wird dann verworfen, wenn der errechnete F-Wert grösser als der entsprechende Grenzwert in der F-Verteilung ist (Wonnacott & Wonnacott 1972). Dies bedeutet, dass es einen signifikanten Unterschied der Varianzen gibt und die Verteilung der Messwerte durch den gefitteten Trend beeinflusst ist.

Umsetzung auf die SADS Daten

Für die vier Fragen des Finalanschlusses werden in ArcGIS über `Spatial Analyst-> Trend` jeweils für die beiden Varianten separat Trendoberflächen erster bis vierter Ordnung errechnet. Danach werden die Trendwerte der Punkte über `Extract Values To Points` als Attribute den Datensätzen mit den Untersuchungspunkten angefügt. Daraus können durch Vergleich mit den Originalintensitäten an derselben Stelle die Residuen berechnet werden. Die weiteren Schritte der Trendoberflächenanalyse, die Berechnung des Determinationskoeffizienten die Bestimmung der Signifikanz durch den F-Test werden in einer Excel-Tabelle vorgenommen.

9. Resultate

Abschnitt 9.1 präsentiert die Resultate zu den Strukturkenngrössen, die direkt aus den im ersten Hauptteil aufbereiteten Daten zu den Flächenkarten extrahiert werden können. Im Abschnitt 9.2 werden anschliessend die Resultate der geostatistischen Verfahren vorgestellt, um die in Kapitel 7 aufgestellten Hypothesen zu den drei syntaktischen Phänomenen zu prüfen. Um die Resultate einfacher zu verstehen, sind in Tabelle 9-1 zur Erinnerung nochmals die Fragen mit den zugehörigen Abkürzungen aufgelistet.

Finalanschluss	
I.1K	„für/zum ein Billet (zu) lösen“
I.6K	„Ich brauche Tabletten, um einzuschlafen“
I.11K	„...um ein Buch (zu) lesen“
IV.14K	„Du musst das Licht anzünden um zu lesen“
Komparativ	
III.22	„Sie ist grösser als ich“
III.25K	„Sie gehen halt lieber schwimmen statt spazieren“
III.28K	„Dann ist er ja älter, als ich gedacht habe“
Artikelverdoppelung	
I.10	„Also s’Susi wär e ganz e liebi Frau für de Markus“
II.10	„Aber du häsch (de) vil (de) schöner Garte“
IV.1	„Martina wäre eine ganz gute Gemeindepräsidentin“

Tabelle 9-1: Untersuchte Phänomene mit den zugehörigen SADS-Fragen und Abkürzungen

9.1. Strukturkenngrössen

Die Resultate der in Abschnitt 8.1 beschriebenen von Rumpf et al. (2009) vorgeschlagenen drei Strukturkenngrössen, die Komplexität $\bar{C}$, die Kompaktheit (Kartenkompaktheit $\bar{L}$ und Gebietskompaktheit $\overline{l_x}$) und die Homogenität ($\bar{B}$ bzw. $\overline{b_x}$) werden im Folgenden nach den drei untersuchten syntaktischen Phänomenen gegliedert.

9.1.1. A: Finalanschluss

Tabelle 9-2 zeigt die Strukturkenngrössen zum Finalanschluss. C, welches sich aus den Grenzlängen der durch die dominanten resultierenden Gebiete errechnet, ist für alle vier Fragen relativ niedrig. Das heisst, die Fragen zum Finalanschluss scheinen eine eher niedrige Komplexität zu haben und zu zusammenhängenden Arealen zu tendieren.

Die Kartenkompaktheit $\bar{L}$ ist durchgehend hoch, wie auch die dazugehörigen Gebietskompaktheiten der zwei dominanten Klassen. Die Werte dazu bewegen sich alle zwischen etwa 0.6 bis 0.7, wobei in der zweiten Frage (I.6K) die Gebietskompaktheit sich etwas zugunsten der Variante *zum* verschiebt. Eine hohe Gebietskompaktheit bedeutet, dass die Original-Intensitäten einer Variante innerhalb ihres dominanten Gebiets, verglichen mit den Intensitäten der übrigen Varianten, hoch ist.

Die Homogenitätswerte zeigen ein sehr ähnliches Bild. Ausgangspunkt bilden nicht mehr die gemessenen, sondern die geschätzten Intensitäten. $\bar{B}$ und $\overline{b_x}$ werden durch die KDE Glättung überall minim kleiner, wobei dieser Effekt am stärksten bei der zweiten Frage zu beobachten ist.

	I.1K (Abb. 5-1)	I.6K (Abb. 5-2)	I.11K (Abb. 5-3)	IV.14K (Abb. 5-4)
C	0.16	0.16	0.11	0.13
$\overline{L}$	0.60	0.67	0.65	0.64
$\overline{l_{für}}$	0.68	0.47	0.68	0.63
$\overline{l_{zum}}$	0.59	0.72	0.63	0.69
$\overline{B}$	0.59	0.66	0.64	0.65
$\overline{b_{für}}$	0.64	0.42	0.67	0.58
$\overline{b_{zum}}$	0.55	0.69	0.61	0.68

Tabelle 9-2: Strukturkenngrössen zum Finalanschluss

9.1.2. B: Komparativ

Die Strukturwerte für den Komparativ (Tabelle 9-3) unterscheiden sich stark von jenen des Finalanschlusses. Einzig die Komplexität C erzielt für die ersten beiden Fragen (III.22 und III.25K) vergleichbare Werte. Die dritte Frage besitzt eine sehr niedrige Komplexität.

Die Kartenkompaktheiten und –homogenitäten sind bei allen Fragen sehr hoch. Ein Blick auf die Gebietsgrössen zeigt jedoch, dass sich die hohen Werte lediglich auf die dominierende Variante mit dem Vergleichswort *als* beschränken. Die Kompaktheitsmasse erreichen für die drei Varianten *weder, wie* und *wan* noch knapp 50 Prozent Anteil in den von ihnen dominierten Gebieten. In den Homogenitätswerten sind sie nicht mehr vertreten, da diese Varianten von der dominanten *als*-Variante verdrängt werden, die als einzige bei der Kartenhomogenität mit einbezogen wird.

	III.22 (Abb. 5-5)	III.25K (Abb. 5-5)	III.28K (Abb. 5-6)
C	0.17	0.15	0.05
$\overline{L}$	0.69	0.74	0.70
$\overline{l_{als}}$	0.72	0.77	0.71
$\overline{l_{weder}}$	0.56	0.59	0.44
$\overline{l_{wıe}}$	0.59	0	0.45
$\overline{l_{wan}}$	0.51	0.56	0.47
$\overline{B}$	0.70	0.75	0.70
$\overline{b_{als}}$	0.70	0.75	0.70
$\overline{b_{weder}}$	0	0	0
$\overline{b_{wıe}}$	0	0	0
$\overline{b_{wan}}$	0	0	0

Tabelle 9-3: Strukturkenngrössen zum Komparativ

9.1.3. C: Artikelverdoppelung

Das letzte Phänomen zeigt im Gegensatz zu den anderen Phänomenen kein konsistentes Bild über alle Fragen hinweg. Die erste (I.10) und die dritte Frage (IV.1) gleichen sich etwas stärker als die zweite Frage (II.10), die völlig aus dem Rahmen fällt. Bei I.10 und IV.1 fallen die hohen Kartenkomplexitäten auf. Die Kompaktheiten und die Homogenitäten erhalten durchschnittliche Werte, wobei bei IV.1 die Variante mit vorgestelltem Artikel etwas heraussticht. Die zweite Frage erzielt sehr hohe Kompaktheits- und Homogenitätswerte und dazu eine sehr niedrige Kartenkomplexität.

	I.10 (Abb. 5-8)	II.10 (Abb. 5-9)	IV.1 (Abb. 5-10)
C	0.56	0.07	0.55
$\bar{L}$	0.48	0.76	0.55
$\overline{l_{dop}}$	0.56	0	0.62
$\overline{l_{nach}}$	0.56	0.79	0.58
$\overline{l_{vor}}$	0.49	0.62	0.68
$\bar{B}$	0.47	0.77	0.54
$\overline{b_{dop}}$	0.47	0	0.44
$\overline{b_{nach}}$	0.47	0.77	0
$\overline{b_{vor}}$	0	0	0.54

Tabelle 9-4: Strukturkenngrössen zur Artikelverdoppelung

9.2. Geostatistische Methoden

Sind die Strukturmasse aus dem vorangegangenen Abschnitt noch direkt mit der Methodik zur Erstellung der Flächenkarten in Verbindung zu setzen, so sind die folgenden Resultate separat dazu zu betrachten. Die Basis zur Anwendung verschiedener geostatistischer Methoden bilden nur noch die gemessenen Akzeptanzen der einzelnen Varianten. Die durch KDE geschätzten Intensitäten werden nicht mehr berücksichtigt. Für alle Phänomene wird der Moran's Index errechnet, der im vorliegenden Fall besagt, wie stark die Verteilung der Akzeptanzwerte einer Variante räumlich beeinflusst ist. Zusätzliche Methoden werden je nach Hypothese abwechselnd eingebunden.

9.2.1. A: Finalanschluss

Moran's *I*

Tabelle 9-5 beinhaltet für alle Varianten der vier Fragen zum Finalanschluss Moran's Index, sowie die dazugehörigen *Z*- und *p*-Werte. Der Index zeigt für alle Varianten und Fragen positive Werte, wobei die *für* Variante durchgehend ein höheres Niveau als die *zum* Variante erzielt. Somit weisen alle Varianten eine positive Autokorrelation auf und bilden zusammenhängende Gebiete.

Der sehr niedrige *p*-Wert zeigt für alle Varianten eine signifikante räumliche Abhängigkeit. Er lässt sich direkt aus den hohen *Z*-Werten ableiten. Die Resultate zu I.6K fallen im Vergleich zu den anderen Varianten etwas schwächer aus, sind aber immer noch stark signifikant.

Frage	I.1K		I.6K		I.11K		I.14K	
Variante	für	zum	für	zum	für	zum	für	zum
Moran's *I*	0.73	0.66	0.53	0.42	0.75	0.73	0.69	0.60
Z-Wert	33.13	30.25	24.36	19.04	34.21	33.47	31.47	27.32
p-Wert	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**

Tabelle 9-5: Moran's *I* Werte für den Finalanschluss

Semivariogramme:

Für die beiden Untersuchungsbänder sind die Semivarianzen gegenüber der Distanz in Semivariogrammen geplottet worden (Abbildungen 9-1 & 9-2). Die Werte für die gewählten lag-Gruppen und die erhaltenen Resultate zu range, sill und nugget dazu sind in Tabelle 9-6 und aufgelistet. Die Maximaldistanz zwischen den Punkten des NW-SO-Untersuchungsbandes beziffert sich auf 130'552 Meter, wodurch der Grenzwert des Produktes aus Grösse und Anzahl lags bei etwa 65 Kilometern liegt. Die untersuchten Gebiete werden in zehn lag-Gruppen mit 6'500 Metern lag-Grösse unterteilt. Die Punkte im zweiten Untersuchungsstreifen weisen eine maximale Distanz von 202'916 Metern auf. Damit sollte gemäss Faustregel die Multiplikation von Grösse und Anzahl der lags etwa 100 Kilometer betragen. Wieder wurden zehn Klassen gewählt, diesmal aber mit einer Spannbreite von zehn Kilometern.

Die Semivariogramme des NW-SO-Bandes zeigen alle ein stärker gestreutes Bild als jene des SW_NO-Bandes, in welchem die Punkte optisch stärker korrelieren. Die range, in der eine Zunahme der Semivarianzen zu beobachten ist, deckt einen deutlich kleineren Radius ab. Das gefittete sphärische Modell (violett-blaue Linie) flacht bei Radien zwischen zehn und zwanzig Kilometern ab. Beim SW-NO-Band erstreckt sich die range über das ganze Gebiet. Zudem ist ein Unterschied zwischen den Varianten zu beobachten. Bei der *zum* Variante wird die sill schneller erreicht als bei der *für* Variante. Die Ausnahme bildet das Diagramm der Frage IV.14K, bei der die *zum* Variante ein sehr schwach ansteigendes, kaum abflachendes Bild abgibt.

Daraus kann geschlossen werden, dass für die NW-SO Richtung nur eine lokale räumliche Abhängigkeit besteht, während diese sich in SW-NO über die gesamte Länge des Untersuchungsbandes erstreckt und damit auf eine grösserskalige räumliche Abhängigkeit hinweist.

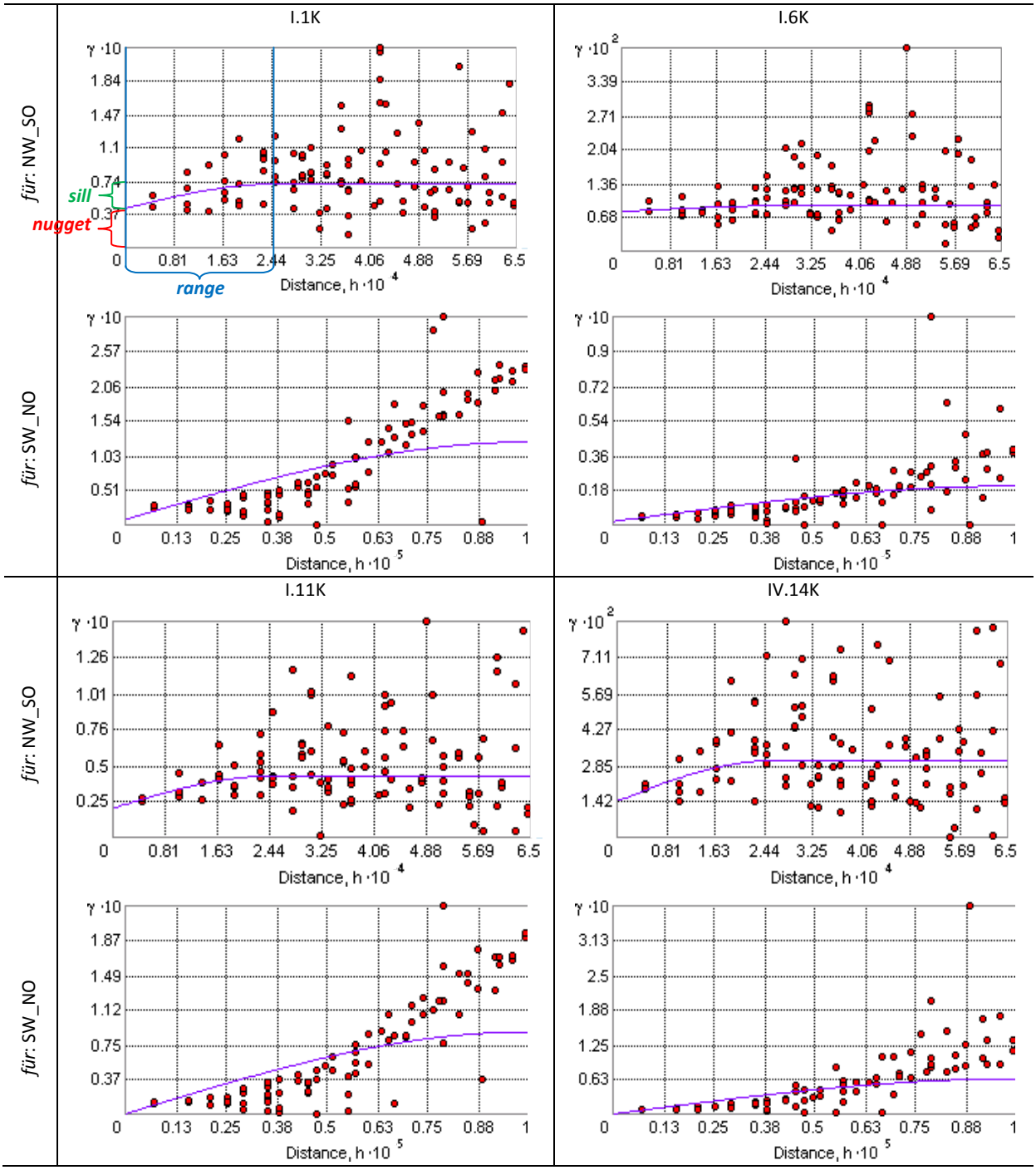


Abbildung 9-1: Experimentelle Semivariogramme der *für* Variante der vier Fragen zum Finalanschluss für die Punkte innerhalb der beiden Untersuchungsbänder mit hervorgehobenen Grössen nugget (rot), sill (grün) und range (blau) im ersten Diagramm (oben links)

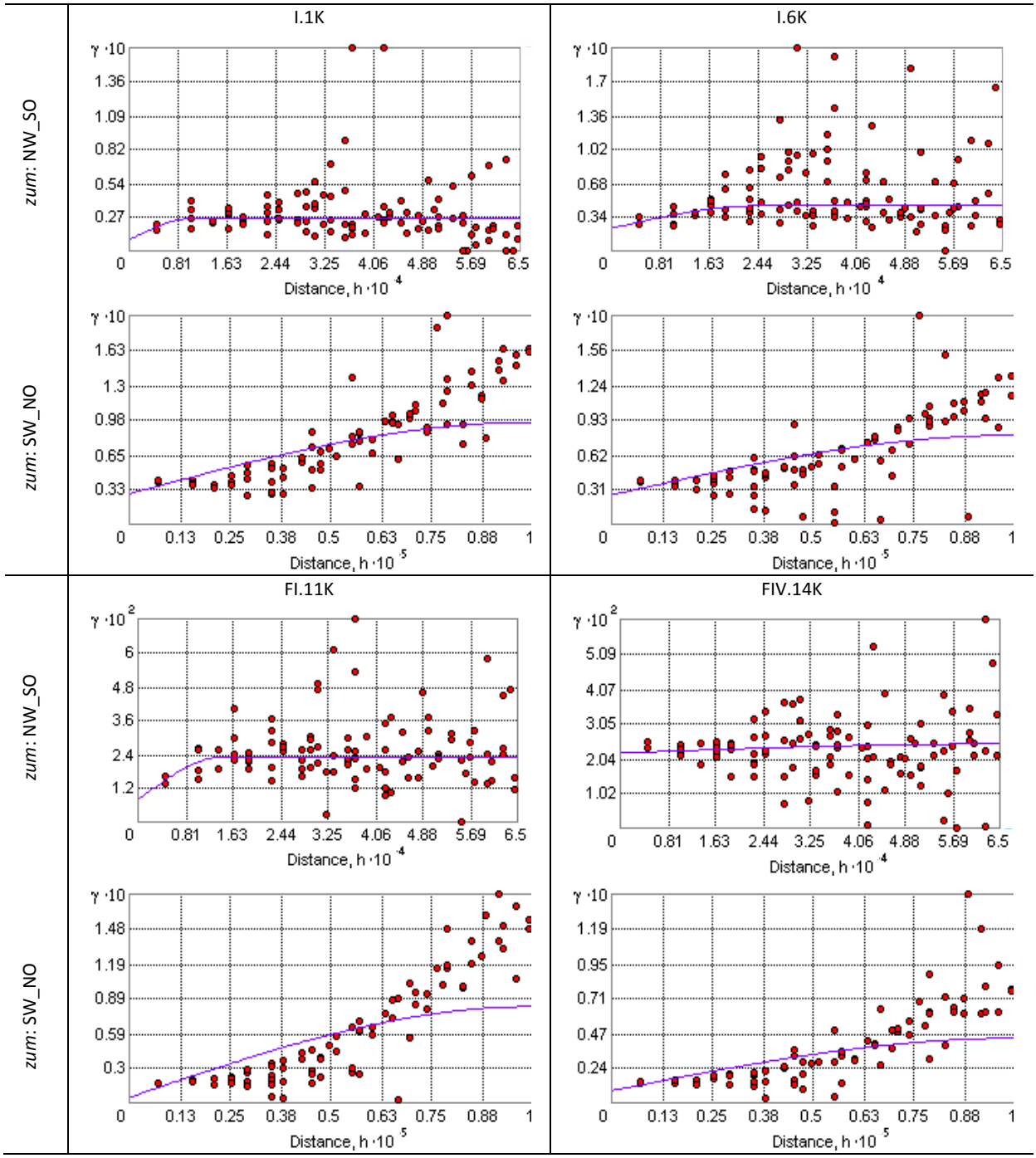


Abbildung 9-2: Experimentelle Semivariogramme der *zum* Variante der vier Fragen zum Finalanschluss für die Punkte innerhalb der beiden Untersuchungsbänder

Variante *für*	lag [km]	range [km]	sill	Nugget
I.1K_NW_SO	6.5	24.7	0.0270	0.0430
I.1K_SW_NO	10	99.2	0.1145	0.0082
I.6K_NW_SO	6.5	31.7	0.0012	0.0079
I.6K_SW_NO	10	99.2	0.0183	0.0018
I.11K_NW_SO	6.5	24.4	0.0225	0.0203
I.11K_SW_NO	10	99.2	0.0183	0.0018
IV.14K_NW_SO	6.5	99.2	0.0883	0
IV.14K_SW_NO	10	99.2	0.0635	0
Variante *zum*	**lag [km]**	**range [km]**	**sill**	**Nugget**
I.1K_NW_SO	6.5	10.6	0.0173	0.0087
I.1K_SW_NO	10	99.2	0.0669	0.0285
I.6K_NW_SO	6.5	24.7	0.0226	0.0228
I.6K_SW_NO	10	99.2	0.0534	0.0262
I.11K_NW_SO	6.5	14.7	0.0154	0.0077
I.11K_SW_NO	10	99.2	0.0775	0.0046
IV.14K_NW_SO	6.5	64.5	0.0266	0.0220
IV.14K_SW_NO	10	99.2	0.0358	0.0083

Tabelle 9-6: Lag, range, sill und nugget für die experimentellen Semivariogramme der *für* und *zum* Varianten

Trendoberflächenanalyse:

Die letzte Methode, welche für den Finalanschluss angewendet wird, ist die Trendoberflächenanalyse. In die Punktdaten wurden Ebenen erster bis vierter Ordnung gefittet. Abbildung 9-3 bildet die erzielten Korrelationskoeffizienten und die F-Werte für die Residuen zwischen den Trendoberflächen und den Messwerten ab. Damit ein Trend mit einer Zuverlässigkeit von 99% signifikant ist, müssen die F-Werte die Grenzwerte, welche durch die F-Verteilung gegeben sind, überschreiten. Signifikante Werte sind in Abbildung 9-3 mit einem Stern markiert. Die drei Fragen I.1K, I.11K und IV.14K markieren eine Gruppe mit sehr ähnlichen Ergebnissen mit einem stark signifikanten Trend (hohe F-Werte), der einen Grossteil der Varianz erklärt (hohes R^2). Die Frage I.11 erzielt die höchsten Resultate. R^2 nimmt zu mit der Erhöhung der Ordnung, der F-Wert hingegen nimmt etwas ab, was jedoch beim entsprechenden Grenzwert ebenfalls der Fall ist. Die Variante *zum* erzielt bei der ersten Frage leicht schwächere Werte als die *für* Variante. Bei den letzten zwei Fragen ist dies nicht zu beobachten.

Es ist wieder die Frage I.6, die ein anderes Bild zeigt. Hier sind die Werte der Variante *für* zwar signifikant, jedoch etwas weniger hoch. Bei dieser Frage ist eine starke Zunahme von R^2 bei der Erhöhung der Trendfläche von der ersten zur zweiten Ordnung sichtbar, auch die Signifikanz nimmt hier deutlich zu. Die Variante *zum* ist für die ersten drei Ordnungen nicht signifikant mit einem Trend erklärbar, erst die vierte Ordnung ergibt eine schwache Überschreitung des Grenzwertes, erklärt aber immer noch weniger als 10% der Varianz. Die effektiven F-Werte und Bestimmungsmasse (R^2) und die durch die verschiedenen Ordnungen vorgegebenen Grenzwerte können in der Tabelle in Anhang D nachgeschlagen werden.

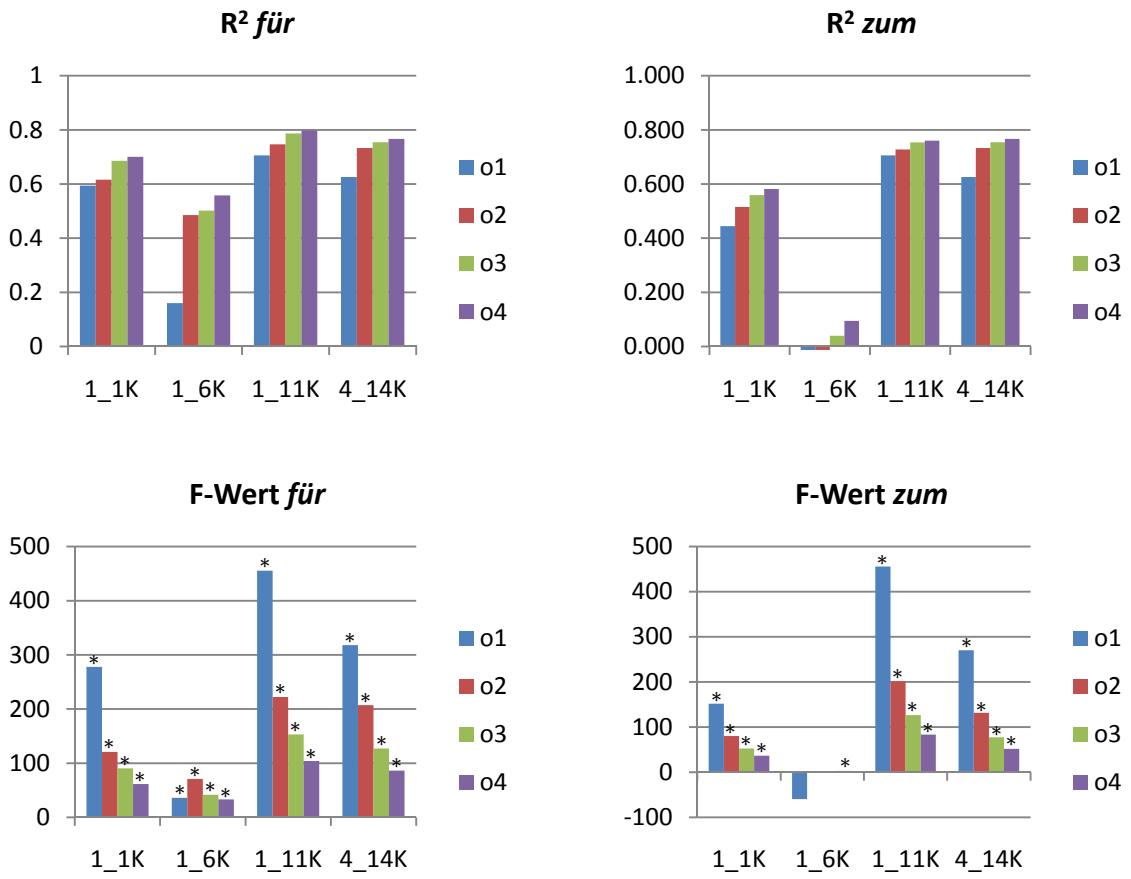


Abbildung 9-3: Bestimmungsmasse (R^2) und F-Werte (p=0.01) der TA für die dominanten Finalanschlussvarianten

9.2.2. B: Komparativ

Moran's *I*

Der Komparativ zeigt mit einem eher tiefen Moran-Index eine kleine positive räumliche Autokorrelation, die jedoch signifikant auf dem 99%-Niveau ist. Die Z-Werte sind immer noch hoch, wenn auch klar tiefer als beim Finalanschluss. Die dritte Frage zeigt, verglichen mit den anderen Fragen, etwas tiefere Werte für die Varianten *wan* und *wieder*, ansonsten ist die Ausprägung der Varianten-Werte innerhalb und zwischen den Fragen ähnlich. Global gesehen formen die Varianten zwar Areale, diese sind aber deutlich weniger ausgeprägt als beim Phänomen A.

Frage	III.22				III.25K				III.28K			
Variante	als	weder	wie	wan	als	weder	wie	wan	als	weder	wie	wan
Moran's I	0.29	0.27	0.27	0.28	0.22	0.24	0.29	0.26	0.19	0.10	0.32	0.18
Z-Wert	13.29	12.35	12.46	13.36	10.34	11.22	13.56	12.31	8.53	4.82	14.59	8.92
p-Wert	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**

Tabelle 9-7: Moran's ***I*** Werte für den Komparativ

Getis-Ord G_i^*

Die lokale Messgrösse G_i^* misst, wie stark die lokale räumliche Dominanz einer Variante ist. Sie kann direkt als Z-Wert verwendet werden (Getis 2010). Da sie für jeden Untersuchungspunkt einzeln errechnet wird, ist eine Darstellung als Punktkarte sinnvoll.

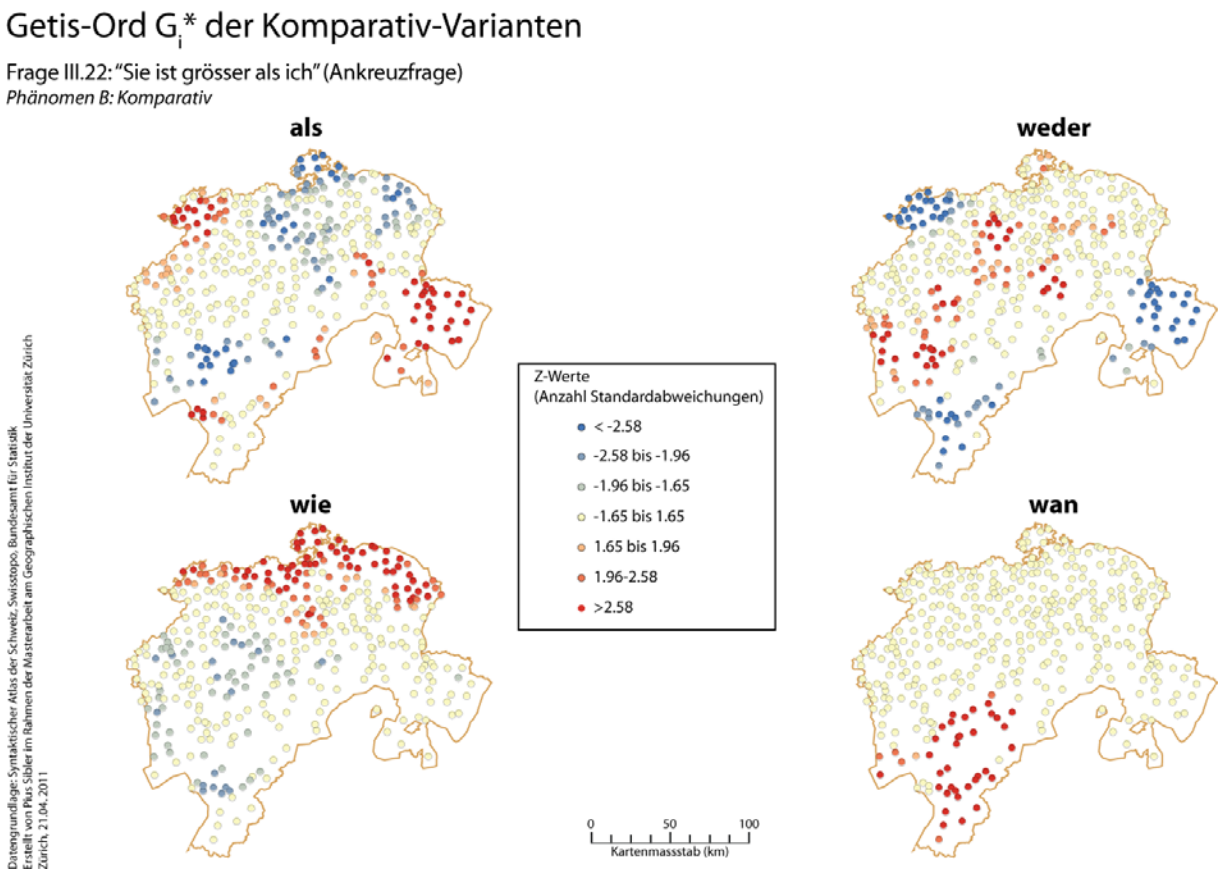


Abbildung 9-4: Getis-Ord G_i^* der vier Varianten der Frage III.22 (Komparativ)

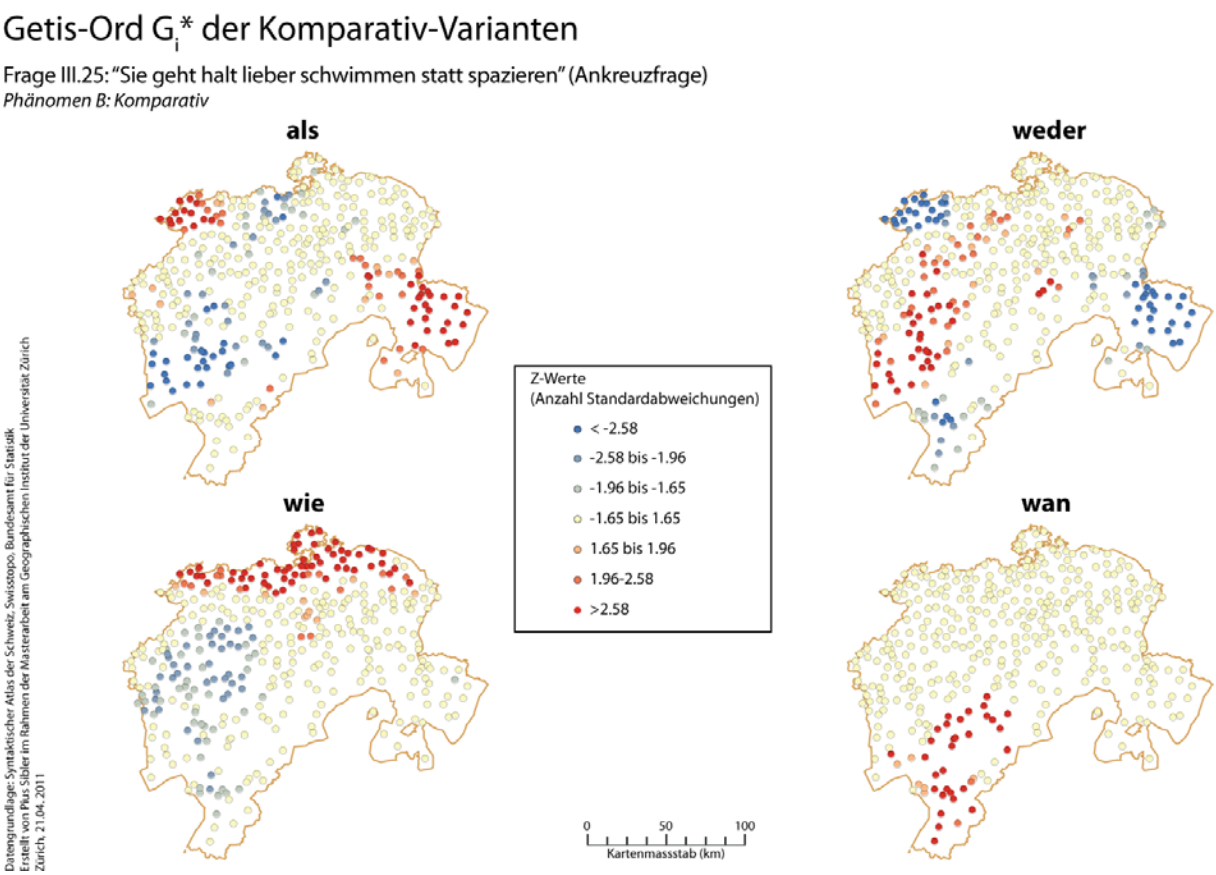


Abbildung 9-5: Getis-Ord G_i^* der vier Varianten der Frage III.25 (Komparativ)

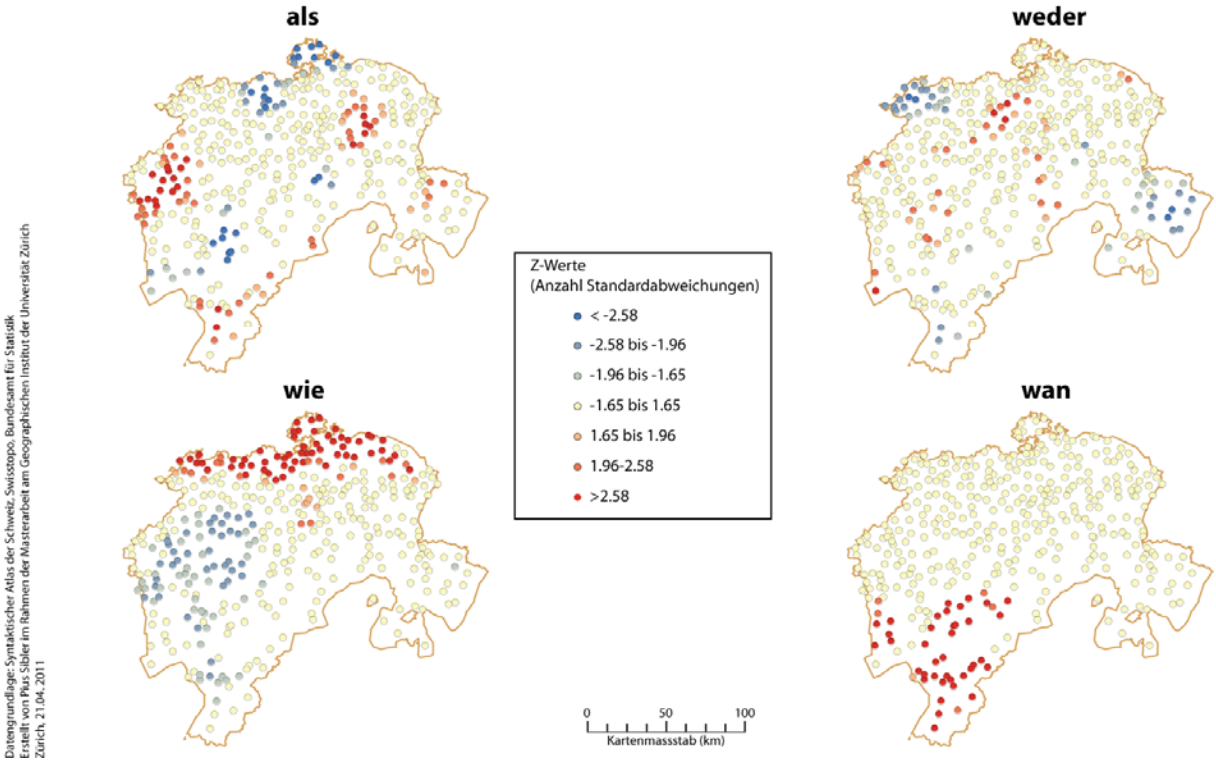


Abbildung 9-6: Getis-Ord G_i^* der vier Varianten der Frage III.28 (Komparativ)

Insgesamt zeigen die drei Fragen zum Komparativ übereinstimmende Resultate, die Verteilungen der Cluster mit überhöhten und tiefen Werten sind sehr ähnlich.

Die Resultate präsentieren für die dominante Variante *als* Hot-Spots in Graubünden und in der Region Basel, jedoch nur in den Fragen III.22 (Abbildung 9-4) und III.25 (Abbildung 9-5). In der letzten Frage III.28 verschieben sich die Hot Spots etwas in westliche Richtung (Abbildung 9-6). Cold Spots sind für alle Fragen in der Südwestschweiz zu erkennen, in der ersten und dritten Frage zudem im Gebiet um Schaffhausen.

Bei der Variante *wie* befinden sich Cold Spots tendenziell in der Westschweiz, währendem entlang der nördlichen Landesgrenze bei allen Varianten sehr deutliche Häufungen auftreten.

Weder kommt in Graubünden signifikant unterdurchschnittlich vor, ebenso im Raum Basel, wogegen in den ersten beiden Fragen Mehrungen im Mittelland zu sehen sind. Die letzte Frage bildet keine deutlichen Häufungsgebiete.

Wan tritt in der Südwestschweiz im Gebiet Wallis und Berner Oberland deutlich gehäuft in Erscheinung, in der restlichen Schweiz formt die Variante keine Clusters.

9.2.3. C: Artikelverdoppelung

Moran's *I*

Moran's Index ist bei allen Fragen zur Artikelverdoppelung leicht positiv und für alle Varianten signifikant. Die Nähe zu null lässt aber darauf schliessen, dass das Phänomen eine beinahe zufällige Verteilung aufweist. Zwischen den Fragen lässt sich keine Gemeinsamkeit bei den Ausprägungen der Indizes erkennen.

Frage		I.10			II.10			IV.1	
Variante	Dopp	nach	vor	Dopp	nach	vor	Dopp	nach	vor
Moran's *I*	0.22	0.16	0.06	0.07	0.08	0.11	0.11	0.15	0.06
Z-Wert	10.15	7.45	2.88	3.32	3.94	4.95	5.31	6.84	2.83
p-Wert	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.01**

Tabelle 9-8: Moran's ***I*** Werte für die Artikelverdoppelung

10. Diskussion der geostatistischen Untersuchungen

Bisher sind nur sehr wenige Arbeiten erschienen, welche die Methoden der Geographischen Informationswissenschaft für sprachwissenschaftliche Untersuchungen verwendet haben. Getis-Ord G_i^*und Moran's I wurden allerdings bereits von Grieve (2009) für seine Doktorarbeit für die Untersuchung der regionalen Variation von sprachlichen Phänomenen in den USA angewandt. Er berechnet die Intensitäten für die Orte durch die relativen Anteile von linguistischen Varianten pro Ort an der Gesamtzahl der Varianten und folgt damit der Idee des RIW (Goebl 1982).

Hoch & Hayes (2010) heben die Vorteile eines Einbezugs von GIS für die sprachgeographische Forschung hervor und erwähnen ebenfalls Verfahren zur Beschreibung räumlicher Autokorrelation, wie Point Pattern Analysis-Masse oder das Semivariogramm als mögliche Mittel, um die geographische Verteilungen von Dialekten zu untersuchen. Sie führen aber selbst keine Untersuchung durch, sondern ermutigen lediglich dazu.

Schliesslich bietet die den ersten Hauptteil der Arbeit dominierende Umsetzung der KDE auf Dialektdaten von Rumpf et al. (2009) eine Verknüpfung geoinformationswissenschaftlicher Grundlagen mit der Linguistik.

Diese Arbeit zählt damit zu den ersten wissenschaftlichen Auseinandersetzungen in der Verschmelzung von GIS und Linguistik. Entsprechend ist sie als explorativer Vorstoss zu betrachten. Die Resultate sind nur schwer in den nur spärlich vorhandenen Kontext einzuordnen und müssen mit Vorsicht genossen werden. Es können Aussagen über die Verteilung von syntaktischen Phänomenen gemacht werden, aber nicht, wie diese Muster entstanden sind.

Für die Strukturkenngrössen wurden zwar einige R-Skripte geschrieben, das Gros der verwendeten Methoden im zweiten Hauptteil ist aber in ArcGIS umgesetzt worden. Dies darf kritisch angefügt werden, da mit der Statistikumgebung R ein viel mächtigeres Werkzeug für die (Geo-)Statistik zur Verfügung stünde. Die enorme Fülle an möglichen Analysemöglichkeiten birgt aber auch die Gefahr, sich darin zu verlieren. Die für diese Arbeit ausreichende Toolbox an statistischen Verfahren in ArcGIS ist in dieser Beziehung etwas weniger problematisch.

Die methodische Herangehensweise wird im folgenden Abschnitt 10.1. diskutiert, die Resultate daraus im darauffolgenden Abschnitt 10.2. Im letzten Abschnitt 10.3. werden die Grundhypothesen aus Kapitel 7 aus den gefundenen Schlüssen bewertet.

10.1. Methodik

In diesem Abschnitt werden konkret die Stärken und Schwächen der verwendeten Methoden behandelt. Dabei wird der Bezug zum sprachwissenschaftlichen Kontext hergestellt.

Strukturkenngrössen

Die Strukturkenngrössen von Rumpf et al. (2009) sind einfach zu berechnen und geben bereits Hinweise darauf, wie ein sprachliches Phänomen räumlich verteilt ist. Sie sind keinesfalls unabhängig voneinander. Eine hohe Komplexität bedeutet gleichzeitig eine niedrige Kompaktheit von Varianten. $\bar{C}$ und $\bar{B}$ korrelieren also negativ miteinander (Rumpf et al. 2010). Sie sind darauf ausgelegt, möglichst schnell einen Überblick über die Variation verschiedener Fragen eines grossen Korpus von Karten zu geben. Sie sind jedoch auf globale Einschätzungen beschränkt, lokale Feinheiten sind nicht erkennbar aus den Kenngrössen. Zudem lassen sie keine statistisch signifikanten Aussagen zu. Da sie aber für jede Karte separat berechnet werden, können sie deren strukturelle Charakteristiken einzeln wiedergeben, was beim integralen Ansatz der Groninger Dialektometrie unberücksichtigt bleibt (Pickl & Rumpf unveröffentlicht).

Mit den Strukturkenngrössen können integrierte Aussagen über die Verteilung *aller* Varianten eines Phänomens gemacht werden, was sie in dieser Beziehung allen anderen verwendeten Verfahren aus der Geostatistik überlegen macht.

Moran's I und Getis-Ord G_i^*

Die beiden Masse sind zwei Beispiele aus einer ganzen Reihe von statistischen Massen zur Beschreibung von Punktmustern. Anstelle von I hätten Geary's C (Geary 1954) oder Ripley's K (Ripley 1977) verwendet werden können. Die getesteten Hypothesen ändern sich dadurch nicht. Die beiden Masse reichen damit aus für diese Arbeit, allenfalls wäre ein Vergleich von Resultaten verschiedener Kenngrössen interessant gewesen.

Moran's I, wie er hier in der Arbeit verwendet wird, ist eine globale Kenngrösse, welche keine Aussagen über die regionalen Charakteristiken zulässt. Es gibt auch einen Local Moran's I, der einzelne Werte für Messpunkte zulässt (O'Sullivan & Unwin 2010). Wie bei vielen statistischen Verfahren ist bei Moran's I Vorsicht bei der Signifikanz der statistischen Aussage geboten. Statistische Verfahren, wie beispielsweise die Monte Carlo Simulation, helfen hier die Sicherheit zu erhöhen. (O'Sullivan & Unwin 2010).

G_i^* dient als lokale Grösse und kann, in Kartenform präsentiert, Hinweise auf lokal über- oder unterrepräsentierte Variablen geben. Sie ist somit die einzige verwendete Methode dieser Arbeit, welche regionale Dialektunterschiede aufzudecken vermag. Weitere Methoden, bespielsweise in dem Boots & Okabe (2007) integrativen local spatial statistical analysis (LoSSA) – Ansatz, wären ebenfalls interessant gewesen.

In der Untersuchung wird mit einem fixen Distanzband für die Bestimmung der Nachbarschaftsbeziehung gearbeitet. Alternative Möglichkeiten für eine stärkere Aussagekraft von I und G_i^* wären die Gewichtung über die inverse Distanz oder der Miteinbezug von lokalen Begebenheiten gewesen. Allerdings ist die Vergleichbarkeit der Masse eingeschränkt, wenn unterschiedliche Gewichtungsmethoden zur Anwendung kommen.

Da beide Verfahren die Normalverteilung in der Nullhypothese verwenden, lassen sie sich gut kombinieren. Ein weiterer Vorteil ist die einfache Verständlichkeit.

Semivariogramm

Hoch & Hayes (2010) sehen das Semivariogramm als nützliches Mittel, um sprachliche Variation zu untersuchen und besser zu verstehen. Da die Parameter, sprich die Grösse und Anzahl der lags für das Semivariogramm manuell gewählt werden können und in dieser Arbeit auch werden, sind die Resultate jedoch statistisch schwierig zu beurteilen. Als Hinweis auf eine mögliche Verteilung kann das Semivariogramm aber durchaus Berechtigung haben. Kritisch muss die Handhabbarkeit für Linguisten beurteilt werden. Um ein Semivariogramm zu erstellen und daraus sinnvolle Informationen ableiten zu können, muss ein vertieftes geostatistisches Grundwissen vorhanden sein (Burrough & McDonnell 1998).

Trendoberflächenanalyse

Die TA kann theoretisch eine Ebene beliebiger Ordnung verwenden. Im Extremfall könnte eine Oberfläche gefittet werden, die jeden Messpunkt berücksichtigt. Die Aussagekraft einer TA nimmt aber mit jeder zusätzlichen Ebenenordnung ab und das Resultat büsst an Verständlichkeit ein (Burrough & McDonnell 1998). Bei niedrigen Ordnungen bietet die TA jedoch eine intuitive Möglichkeit zur Erkennung von globalen räumlichen Abhängigkeiten. Auf lokale Besonderheiten kann die TA definitionsgemäss nicht eingehen. Für weiterführende Untersuchungen würde sich eventuell auch die logistische Regressionsanalyse anbieten, die Übergänge zwischen dominanten Varianten in anderer Form modellieren kann (Wattel & van Reenen 2010).

10.2. Resultate

Der folgende Abschnitt soll erörtern, ob die verschiedenen umgesetzten Verfahren geeignet sind für syntaktische Daten sowie Einschränkungen bezüglich deren Aussagekraft aufzeigen. Zudem soll diskutiert werden, wie die erzielten Resultate zu interpretieren sind.

Strukturkenngrössen

Der Vergleich mit den Karten aus Teil 2 zeigt bei allen Phänomenen eine grosse Übereinstimmung mit den entsprechenden Strukturkenngrössen. Da sowohl die Karten wie auch die Strukturkenngrössen dieselbe

methodische Basis aufweisen, ist dies zu erwarten. Der Finalanschluss erzielt niedrige Komplexitätswerte für die gesamte Karte und gleichzeitig hohe Homogenitäts- und Kompaktheitswerte für beide Gebiete der Varianten, ebenso für die Karte. Damit kann bereits eine Arealbildung vermutet werden, die das Gebiet in zwei kompakte Teilgebiete aufteilt. Für den Komparativ gilt ähnliches, wobei hier auf die Übermacht der *als* Variante hingewiesen werden muss, welche nur bei der kombinierten Verwendung der Gebiets- und Kartenkompaktheiten bzw. –homogenitäten ersichtlich wird. Die Artikelverdoppelung bildet zwei heterogene Karten (I.10 und IV.1), was sich ebenfalls in den hohen Komplexitätsgrössen manifestiert. Gleichzeitig sind bei diesen Karten die Homogenitäts- und Kompaktheiten mässig ausgeprägt. Die eher dem Gesamteindruck des Komparativs gleichende Frage I.10 erzielt eine entsprechend niedrige Komplexität und hohe Werte für die anderen beiden Kenngrössen.

Die Strukturwerte, die Rumpf et al. (2009) für ihre Kartoffelkrautkarte erhalten, (Abbildung 3-7) gleichen am ehesten den Werten des Finalanschlusses. Dies ist dadurch zu erklären, dass sie auch eine Karte mit geschlossenen, klar getrennten Gebieten zeigt, wenn auch zwei Varianten mehr vorhanden sind.

Man kann festhalten, dass $\bar{C}$ ein zuverlässiges Mass für die Komplexität der untersuchten Fragen abgibt, $\bar{B}$ und $\bar{L}$ aber nur in Kombination mit den entsprechenden Gebietsgrössen aussagekräftig sind.

Die Richtung, beziehungsweise die Lage der Gebiete können mit den Strukturkenngrössen nicht erklärt werden.

Moran's *I* und Getis-Ord G_i^*

Moran's *I* ist für alle untersuchten Fragen eingesetzt worden, da für alle Hypothesen geklärt werden sollte, ob benachbarte Messwerte ähnlich sind, womit eine räumliche Abhängigkeit besteht und konkret auf die SADS-Daten bezogen, ob diese Muster generieren. Wichtiger als die daraus gewonnene Einsicht, dass sämtliche Fragen eine signifikante globale räumliche Abhängigkeit besitzen, ist die Ausprägung dieser in Form von verschieden hohen Index-Werten.

Der Finalanschluss erzielt hier bei allen Fragen hohe Werte für beide Varianten. Dies bedeutet eine hohe positive Autokorrelation und damit eine eindeutige Arealbildung innerhalb der Varianten. Für den Komparativ sind diese Werte kleiner, deuten aber immer noch auf eine leicht positive Autokorrelation hin. Es erstaunt, dass hier keine Unterschiede zwischen der über das ganze Gebiet dominierenden *als* Variante und den anderen drei Varianten zu erkennen sind. Die Artikelverdoppelung erhält einen Index nahe null. Folglich bilden die Varianten keine Areale.

G_i^*-Berechnungen sind lediglich für den Komparativ durchgeführt worden. Für die anderen Phänomene sind keine Hypothesen aufgestellt worden, welche einen Gebrauch gerechtfertigt hätten. Erwähnenswert sind die Clusterbildungen für die *wan* Variante im Gebiet Wallis und Berner Oberland, sowie die Häufung der *wie* Variante entlang der Nordgrenze der Deutschschweiz (vgl. Abbildungen in Unterabschnitt 9.2.2). Sie entsprechen damit den Folgerungen von Friedli (2005). *Als* und *weder* lassen ein weniger klares Muster von Häufungen erkennen.

Semivariogramm

Das Semivariogramm, wie es in dieser Arbeit eingebunden ist, bietet die einzige Methode, welche Richtungsabhängigkeiten aufzeigen kann. Die aufgrund der Resultate des ersten Hauptteils gelegten Untersuchungsfenster ermöglichen es, für den Finalanschluss die räumliche Abhängigkeit entlang der vermuteten Trends zu untersuchen. Die Semivariogramme in den Abbildungen 9-1 und 9-2 zeigen deutlich, dass die räumliche Abhängigkeit viel kleinräumiger von Nordwesten nach Südosten ist als senkrecht dazu. Grund dafür könnte die Querung des Alpenraums sein, währendem das andere Untersuchungsband topographisch weniger überprägtes Gebiet einschliesst. Die Semivariogramm-Methode unterstützt jedenfalls die Hypothese eines Übergangs von Nordost nach Südwest. Statistisch signifikante Aussagen können durch die Semivariogramme aber nicht gemacht werden.

Die Untersuchungsfenster sind manuell gelegt, eine Verschiebung der Lage und Grösse der Bänder könnte andere Resultate liefern. Die Sensitivität bezüglich Richtung und Grösse der Bänder wurde hier nicht untersucht, muss jedoch kritisch angefügt werden.

Trendoberflächenanalyse

Diese Methode wurde mit konkretem Bezug auf die Theorie der schiefen Ebene von Seiler (2005) für den Finalanschluss implementiert. Für drei der vier Fragen dazu kann ein signifikanter und stark ausgeprägter Zusammenhang bereits mit einer simplen Ebene erster Ordnung für beide Varianten nachgewiesen werden (vgl. Abbildung 9-3). Frage I.6 zeigt lediglich einen schwachen Zusammenhang, der zudem nur noch teilweise signifikant ist.

Die Hypothese, wonach der Finalanschluss einem linearen Trend folgt, kann somit für drei der vier Fragen klar bestätigt werden, für eine jedoch nicht. Ein Grund dafür könnte die Form der Befragung sein. I.6 ist die einzige behandelte Ergänzungsfrage. Weiter könnte das konkrete Verb, auf welches sich der Finalanschluss bezieht, einen Einfluss haben. Das sind jedoch nur Vermutungen, die ohne linguistisches Expertenwissen angestellt wurden. Ein Expertengespräch könnte hier mehr Klarheit schaffen.

10.3. Beurteilung der Grundhypothesen

Die in Kapitel 7 aufgestellten Vermutungen werden hier aufgrund der geostatistischen Resultate beurteilt.

Finalanschluss: markante Nordost-Südwest-Verteilung und Verteilung der Varianten entlang von 2 schiefen Ebenen

Der Anschluss des Finalmittels ist das am gründlichsten untersuchte Phänomen. Für die geprüften Hypothesen sind vor allem die Trendoberflächenanalyse, welche die schiefen Ebenen bestätigt und die Semivariogramme der Untersuchungsfenster, welche den Verdacht der Nordost-Südwest-Verteilung erhärten, hilfreich. Die Strukturmasse nach Rumpf et al. (2009) und Moran's I nehmen eine unterstützende Funktion ein, indem sie eine starke Arealbildung implizieren. Die Grundhypothesen können damit bestätigt werden.

Komparativ: Dominanz einer Variante über das ganze Gebiet und Bildung von einzelnen Clustern für die übrigen Varianten

Moran's I unterstützt die Hypothese keiner Autokorrelation der dominanten *als* Variante nicht. Es ist damit eine signifikante räumliche Abhängigkeit zu erwarten. Der mässig positive Wert kann dabei aber weder eindeutig sagen, ob die Variante positiv korreliert, noch ob es sich um eine zufällige Verteilung handelt. Ähnliche Werte werden für den Index auch bei den anderen Varianten gemessen.

Besser geeignet für die Behandlung dieser Hypothese scheint G_i^*. Diese Grösse lässt für die zwei Varianten *wan* und *wie* klare Hot Spots erkennen, die in der Literatur, welche dieser Hypothese zu Grunde liegt (Friedli 2005), Übereinstimmung finden.

Die Strukturkenngrössen unterstützen die Analyse des Komparativs wenig. Sie unterstreichen lediglich die Dominanz der *als* Variante.

Artikelverdoppelung: Zufällige Verteilung des Phänomens

Moran's I konnte zeigen, dass eine zufällige Verteilung der einzelnen Varianten eine realistische Hypothese ist, wobei klar tiefere Z-Werte als bei den anderen Phänomenen erreicht werden und nach wie vor überall eine leichte Tendenz zur Arealbildung zu erkennen ist. Die Strukturkenngrössen vermögen dies nicht zu untermauern, sie sind je nach Frage sehr unterschiedlich. Eine Ergänzung um weitere Messgrössen würde eventuell eine gehaltvollere Aussage zulassen.

Teil IV: Fazit

11.Schlussfolgerungen und Ausblick

In diesem letzten Teil wird zunächst rekapituliert, was in dieser Masterarbeit erreicht wurde (Abschnitt 11.1). Die Forschungsfragen werden beantwortet (11.2). Weiter werden die Grenzen der Aussagekraft der Resultate aufgezeigt (11.3) und schliesslich wird in Abschnitt 11.4 ein Blick in die Zukunft des behandelten Themas gewagt. Offene Baustellen und mögliche Erweiterungen in der jungen Forschungszusammenarbeit bilden dabei den Mittelpunkt.

11.1. Erreichtes

Die Arbeit hat gezeigt, dass es möglich ist für nominalskalierte syntaktische Daten Flächenkarten zu generieren. Die Methodik von Rumpf et al. (2009) konnte erfolgreich umgesetzt werden. Sie generiert mithilfe der Kernel Density Estimation (KDE) Thiessenpolygonkarten, welche die Verteilung der dominanten Varianten zeigt und über die Helligkeitsvariation deren Intensität angibt.

Eine räumliche Analyse der Verteilung von Sprachvarianten konnte mit herkömmlichen Methoden der Geoinformationswissenschaft und der Geostatistik durchgeführt werden. Diese linguistischen Varianten sind aber getrennt voneinander betrachtet worden. Statistische Aussagen über die generelle Verteilung eines Phänomens konnten so nicht erzielt werden. Diese können aber teilweise durch die Strukturkenngrössen von Rumpf et al. (2009) beschrieben werden, welche eine Schnittstelle zwischen der Geostatistik und der Flächenkartengenerierung bilden.

Die Hauptziele der Arbeit wurden erreicht. Es konnten syntaktische Flächenkarten aus dem SADS-Datensatz erstellt werden. Ebenfalls konnte mit den dafür errechneten Intensitäten räumliche Grundhypothesen mit geostatistischen Methoden beurteilt werden, wobei diese nicht überall gleich starke Aussagen zulassen.

11.2. Forschungsfragen und Antworten

Die einleitend aufgestellten Forschungsfragen werden in diesem Abschnitt, soweit dies möglich ist, aus den Erkenntnissen des Arbeitsprozesses beantwortet.

Hauptziel 1: Erstellen von syntaktischen Flächenkarten

- Wie lassen sich syntaktische Daten in Flächenkarten umwandeln und welche Methoden der Geoinformationswissenschaft sind dazu geeignet?

Die Kernpunkte zur Erstellung von Flächenkarten aus den Punktdaten bilden einerseits die Interpolation der Punkte zu Thiessenpolygonen, welche eine distanzabhängige Umwandlung von Punkten in Flächenobjekte bietet und andererseits die Umwandlung der SADS-Daten von einer Nominal- in eine Verhältnisskala über die Bildung von Variantenintensitäten. So muss keine ratioskalierte linguistische Distanz errechnet werden, was für syntaktische Phänomene ein bisher ungelöstes Problem bleibt. Aus den Intensitäten kann pro Ort die dominante Variante gewählt werden, das entsprechende Thiessenpolygon mit einer bestimmten Farbe eingefärbt und dessen Helligkeit nach Intensität gewählt werden. Die Intensitäten der Varianten können durch die Kernel Density Estimation geglättet werden.

- Welche Vor- und Nachteile bilden syntaktische Flächenkarten gegenüber herkömmlichen Punktkarten?

Von einem linguistischen Standpunkt aus konnte diese Frage nicht beantwortet werden, Expertenmeinungen wären dazu von Nöten. Rein technisch gesehen ist der Aufwand für die Erstellung von Flächenkarten erheblich grösser als für Punktkarten. Dafür entsteht ein gesamträumliches Bild, welches gerade mit Blick auf mögliche Arealbildungen einer Punktdarstellung überlegen ist. Allerdings erlaubt die flächenhafte Darstellung lediglich eine Variante pro Polygon, die Punktekarten des SADS können dagegen mehrere Varianten in Form verschiedener Symbole verwenden. Schliesslich kann durch

die Helligkeitswahl nach Intensität bei den Flächenkarten stufenlos eine Information zur Dominanz der Varianten abgebildet werden, was sich bei den Punktekarten als schwierig gestaltet.

Hauptziel 2: Beurteilung von räumlichen Zusammenhängen in der Deutschschweizer Dialektsyntax mit geostatistischen Methoden

- Welche geostatistischen Methoden helfen, Aussagen über die räumliche Verteilung von syntaktischen Phänomenen zu machen?

Strukturkenngrössen zur Komplexität, Kompaktheit und Homogenität können einen ersten Anstoss geben für die Beschreibung eines Phänomens. Hinzu kommen geostatistische Verfahren wie Moran's I, welcher positive und negative räumliche Autokorrelation quantifiziert, oder Getis-Ord G_i^* für die Erkennung von Häufungen und Cold Spots. Weiter können mit der Trendoberflächenanalyse Hypothesen zu räumlichen Trends getestet werden und mithilfe von Semivarogrammen die Richtungsabhängigkeit räumlicher Abhängigkeiten überprüft werden. Nicht jedes Verfahren eignet sich gleich gut für die Überprüfung linguistischer Hypothesen. Es muss auf die spezifischen Anforderungen jeder Grundhypothese eingegangen werden. Eine Kombination der Methoden steigert die Aussagekraft von erkannten räumlichen Charakteristiken.

- Sind in den untersuchten Daten räumliche Abhängigkeiten erkennbar?

Der ***Finalanschluss*** hat zwei dominante Areale, welche einer Nordost-Südwest-Verteilung folgen. Dies bestätigen die dafür generierten Flächenkarten, wie auch die Trendoberflächenanalyse und die Semivarogramme mit Untersuchungsbändern entlang und quer zur vermuteten Trendrichtung. Die Strukturkenngrössen in Kombination mit Moran's I weisen zudem auf eine starke Arealbildung hin.

Der ***Komparativ*** bildet lokale Hot Spots einzelner Varianten im nördlichen Grenzgebiet und im Berner Oberland und Wallis, was mit G_i^*-Karten gut zu erkennen ist. Eine Variante ist dominant über das gesamte Deutschschweizer Untersuchungsgebiet, was auf den Flächenkarten eindeutig abzulesen ist. Moran's I hilft hier nicht, um die deutliche Übermacht dieser Variante hervorzuheben. Diese Dominanz ist aus den Strukturkenngrössen zur Gebietshomogenität und –kompaktheit deutlich zu sehen aber nicht mittels der gewählten statistischen Tests zu belegen.

Bei der ***Artikelverdoppelung*** sind je nach Karte sehr unterschiedliche Verteilungen zu sehen, weshalb eine allgemeine Hypothese hier schon Schwierigkeiten verursacht. Bei der Karte I.10 und IV.1 sind auf den nicht interpolierten Flächenkarten keine klaren Muster zu erkennen, was auf eine Zufallsverteilung der Varianten hinweist. Es wurde lediglich Moran's I als statistische Grösse verwendet. Die Vermutung einer zufälligen Verteilung muss bei allen Karten signifikant verworfen werden, die tiefen Werte von I weisen aber auf eine niedrige Autokorrelation hin.

11.3. Grenzen

Trotz interessanter Resultate muss die Aussagekraft dieser Arbeit kritisch bewertet werden. Sie enthält sprachwissenschaftlich (noch) nicht abgestützte Resultate und wandelt auf wenig erforschtem Gebiet. Einige wichtige Einschränkungen sollen hier nochmals unterstrichen werden.

Explorativer Charakter

Die Resultate dieser Arbeit bilden einen ersten Schritt in Richtung Generierung von Flächenkarten und geostatistische Analyse syntaktischer Phänomene in der Deutschschweiz. Es muss nochmals klargemacht werden, dass es sich dabei um eine explorative Arbeit handelt, welche versucht, bestehende Methoden an neu erhobenen Daten zu testen. Das verwendete Datenvolumen von 11 Fragen zu drei Phänomenen lässt keine allgemeingültigen Schlüsse zu.

Begrenzte Variantenzahl

Es wurden zwischen zwei und vier Varianten pro Phänomen einbezogen, zum Teil sind aber deutlich mehr Varianten im SADS-Programm erhoben worden. Vor allem für die geostatistische Analyse könnte eine Ausdehnung auf mehr Varianten die Resultate noch entscheidend beeinflussen.

Getrennte Betrachtung einzelner Varianten

Der Geostatistik-Teil bezieht sich grösstenteils nur auf die Verteilungen einzelner Varianten. Die Interaktion verschiedener Varianten eines Phänomens kann so nur erahnt, jedoch nicht direkt beschrieben werden.

Unbekannter Einfluss der Fragetypen

Der Einfluss von verschiedenen Untersuchungsmethoden wurde nicht berücksichtigt. Es ist aber gut möglich, dass die verschiedenen Fragearten (Ergänzung, Übersetzung und Ankreuzen) das Resultat verändern.

Subjektivität

Die Unterteilungen in verschiedene Varianten aus den verschiedenen Antworten im SADS ist manuell vorgenommen worden. Sie ist mithilfe von Experten zusammengestellt worden und hat daher eine wissenschaftliche Grundlage. Trotzdem ist dadurch mit einem subjektiven Einfluss auf die resultierenden Karten zu rechnen.

11.4. Ausblick

Der nächste Schritt in der Zusammenarbeit von Linguistik und GIScience wäre die sprachwissenschaftliche Validierung der hier vorgestellten Produkte. Eine Beurteilung der Flächenkarten durch Experten kann darüber entscheiden, ob der von Rumpf et al. (2009) vorgestellte Ansatz für syntaktische Daten einen Gewinn darstellt, oder ob diese Methode nicht weiterverfolgt werden soll. Ebenfalls könnte dadurch eine Einordnung der geostatistischen Ergebnisse ermöglicht werden. Widerspiegeln die statistischen Kenngrössen erwartete räumliche Verteilungen? Welche der Kenngrössen sind besonders gut für die Beschreibung von syntaktischen Mustern geeignet?

Eine Erweiterung könnte, wie auch von Spruit (2008) vorgeschlagen, die Untersuchung von demographischen Charakteristiken und Grenzen auf die Raumbildung sein. Ebenso könnten spezifische geographische Abhängigkeiten, wie der Einfluss von Gebirgs- oder Wasserbarrieren auf die Deutschschweizer Sprachlandschaft, getestet werden.

Die Arbeit hat sich auf wenige Fragen und Phänomene der Dialektsyntax der Deutschschweiz beschränkt. Würde die Methodik für alle Fragen durchgeführt, liessen sich durch die Strukturkenngrössen Karten mit ähnlicher Struktur zusammenfassen. Dies geht in Richtung Aggregierung, wie sie die Groninger Dialektometrie (Heeringa 2004, Nerbonne & Wiersma 2006) anstrebt. Eine Optimierung der Datenbankstruktur hinsichtlich konsistenter Verwendung von Feldnamen und Datentypen wäre dazu allerdings notwendig, um automatisierte Abläufe für die Aufbereitung erstellen zu können.

Vertiefte Untersuchungen könnten auch im Bereich der Ähnlichkeitsmasse, wie der Hamming-Distanz (HD) oder dem relativen und gewichteten Identitätswert (RIW bzw. GIW) in Kombination mit gabmap aussichtsreich sein. Komplexere statistische Verfahren wie die Clusteranalyse oder Multidimensional Scaling könnten vertieftere Erkenntnisse über die Syntax bringen.

Vorstellbar ist eine Ausweitung der Methode von Rumpf et al. (2009) auf andere grammatische Bereiche, um auch dort Flächenkarten zu erstellen. Eine umfassende Datengrundlage ist bereits mit dem SDS vorhanden. Es ist sogar eine Ausdehnung auf andere Sprachen in Erwägung zu ziehen.

Die Darstellung von dreidimensionalen Verteilungen der Phänomene könnte die Begrenztheit auf eine Variante pro Gebiet aufheben. Prototypisch wurde eine solche 3D-Karte erstellt (siehe Daten-CD: 5_3D). Sie scheint vielversprechend zu sein. Die dritte Dimension könnte eine Chance für künftige qualitative Untersuchungen und sollte weiter untersucht werden.

Die Trendoberflächenanalyse könnte um geländeanalytische Komponenten erweitert werden. Die mittlere Exposition und Steigung könnten als Indizien dafür genommen werden, in welche Richtung Trends hauptsächlich laufen und ob eine globale Tendenz besteht bzw. wie stark diese ist.

Schliesslich wäre eine Untersuchung der Zeitkomponente sehr interessant, um die Verschiebung von Sprachareale oder die Durchmischung verschiedener Varianten zu visualisieren. Dazu fehlt zurzeit aber schlichtweg die Datengrundlage.

Das syntaktische Stiefkind hat noch einen langen Weg hin zur Lieblingsschwiegertochter der Dialektologie vor sich.

Literatur

Auer, P. & Schmidt, J. E. (Hrsg.) (2010): Language and Space. An International Handbook of Linguistic Variation. Vol. 1 Theories and Methods. Mouton De Gruyter, Berlin.

Barbiers, S., Cornips, L. & van der Kleij, S. (2002): Syntactic Microvariation. Meertens Institute Electronic Publications in Linguistics, Amsterdam.

Bivand, R. S., Pebesma, E. J. & Gómez-Rubio, V. (2008): Applied Spatial Data Analysis with R. Springer, New York.

Boots, B. (1999): Spatial tessellations. In: Longley, P. A., Goodchild, M. F., Maguire D. J. & Rhind, D. W. (Hrsg.): Geographical Information Systems. Principles and Technical Issues. Zweite Ausgabe. John Wiley & Sons, New York, S. 503-526.

Boots, B. & Okabe, A. (2007): Local statistical spatial analysis: Inventory and prospect. International Journal of Geographical Information Science, Vol. 21, Nr. 4, S. 355–375.

Bucheli, C. & Glaser, E. (2002): The Syntactic Atlas of Swiss German Dialects: Empirical and Methodological Problems. In: Barbiers, S., Cornips, L. & van der Kleij, S. (Hrsg.): Syntactic Microvariation, Vol. 2. Meertens Institute Electronic Publications in Linguistics, Amsterdam, S. 41-73.

Bucheli Berger, C. (2008): Neue Technik, alte Probleme: auf dem Weg zum Syntaktischen Atlas der Deutschen Schweiz (SADS). In: Elspass, S. & König, W. (Hrsg.): Sprachgeographie digital – die neue Generation der Sprachatlanten. Olms, Hildesheim, S. 29–44.

Burrough, P. A. & McDonnell, R. A. (1998): Principles of Geographical Information Systems. Oxford University Press, Oxford.

Christen, H., Glaser, E. & Friedli, M. (Hrsg.) (2010): Kleiner Sprachatlas der deutschen Schweiz. Huber, Frauenfeld.

Diggle, P. J., Fiksel, T., Grabarnik, P., Ogata, Y., Stoyan, D. & Tanemura, M. (1994): On parameter estimation for pairwise interaction point processes. International Statistical Revue, Vol. 62, S. 99-117.

Eicher, C. L. & Brewer, C. A. (2001): Dasymetric Mapping and Areal Interpolation. Implementation and Evaluation. Cartography and Geographic Information Science, Vol. 28, Nr. 2, S. 125-138.

Friedli, M. (2005): Si isch grösser weder ig! Zum Komparativanschluss im Schweizerdeutschen. In: Christen, H. (Hrsg.): Dialektologie an der Jahrtausendwende. Linguistik Online, Vol. 24, Nr. 3. http://www.linguistik-online.de/24_05/friedli.pdf, Zugriff: 19.4.2011

Geary, R. C. (1954): The Contiguity Ratio and Statistical Mapping. The Incorporated Statistician, Vol. 5, Nr. 3, S. 115–145.

Getis, A. & Ord, J. K. (1992): The analysis of spatial association by use of distance statistics. Geographical Analysis, Vol. 24, Nr. 3, S. 189-207.

Getis, A. (2010): Spatial Autocorrelation. In: Fischer, M. M. & Getis, A. (Hrsg.): Handbook of Applied Spatial Analysis. Springer, Berlin/Heidelberg/New York.

Glaser, E (1997): Dialektsyntax: eine Forschungsaufgabe. In: Ott, P et. al. (Hrsg.): Bericht über das Jahr 1996. Schweizerisches Wörterbuch. Schweizerdeutsches Idiotikon. Rotkreuz, S. 11-32.

Glaser, E. & Frey, N. (2007): Doubling Phenomena in Swiss German Dialects. Meertens Institute, Amsterdam.

Glaser, E. (2008): Syntaktische Raumbilder. In: Ernst, P. & Patocka, F. (Hrsg.): Dialektgeographie der Zukunft. Akten des 2. Kongresses der Internationalen Gesellschaft für Dialektologie des Deutschen (IGDD), Stuttgart, S. 85-111.

Goebl, H. (1982): Dialektometrie. Prinzipien und Methoden des Einsatzes der Numerischen Taxonomie im Bereich der Dialektgeographie. Denkschriften der Österreichischen Akademie der Wissenschaften, phil.-hist. Klasse, Band 157, Wien.

Goebl, H. (1984): Dialektometrische Studien. Anhand italoromanischer, rätoromanischer und galloromanischer Sprachmaterialien aus AIS und ALF. 3 Bände. Niemeyer, Tübingen.

Goebl, H. (2001): Arealtypologie und Dialektologie. In: Haspelmath, M., König, E., Österreicher, W. & Raible, W. (Hrsg.): Language Typology and Language Universals. An International Handbook, Vol. 2., de Gruyter, Berlin/New York, S. 1471-1491.

Goebl, H. (2006): Recent Advances in Salzburg Dialectometry. Literary and Linguistic Computing, Vol. 21, Nr. 4, S. 411–435.

Goebl, H. (2007): Dialektometrische Streifzüge durch das Netz des Sprachatlasses AIS. Ladinia, Vol. 31, S. 187-271.

Gooskens, C., (2004): Norwegian dialect distances geographically explained. In: Gunnarson, B. L., Bergström, L., Eklund, G., Fridella, S., Hansen, L. H., Karstadt, A., Nordberg, B., Sundgren, E. & Thelander, M. (Hrsg.): Language Variation in Europe. Papers from the Second International Conference on Language Variation in Europe ICLAVE 2, Uppsala, S. 195-206.

Grieve, J. (2009): A Corpus-Based Regional Dialect Survey of Grammatical Variation in Written Standard American English. Dissertation (Manuskript). University of Northern Arizona. https://perswww.kuleuven.be/~u0064311/GrieveDissertationFinal.pdf, Zugriff: 20.4.2011

Haag, K. (1898): Die Mundarten des oberen Neckar- und Donaulandes. Buchdruckerei Egon Hutzler, Reutlingen.

Hamming, R W. (1950): Error detecting and error correcting codes. Bell System Technical Journal, Vol. 29, Nr. 2, S. 147–160.

Heeringa, W. (2004): Measuring Dialect Pronunciation Difference Using Levenshtein Distance. PhD thesis, University of Groningen.

Hoch S. & Hayes J. J. (2010): Geolinguistics: The Incorporation of Geographic Information Systems and Science. The Geographical Bulletin, Vol. 51, Nr. 1, S. 23-36.

Hodler, W. (1969): Berndeutsche Syntax. Francke, Bern.

Isaacs, E. H. & Srivastava, R. M. (1989): An Introduction to Applied Geostatistics. Oxford University Press, New York.

Jones, M. C., Marron, J. S. & Sheather, S. J. (1996): A Brief Survey of Bandwith Selection for Density Estimation. Journal of the American Statistical Association, Vol. 91, S. 401-407.

Keller, A. v. (1855): Anleitung zur Sammlung des schwäbischen Sprachschatzes. In: Einladungsschrift der Universität Tübingen zum 27. September 1855. Fues, Tübingen.

Kessler, B. (1995): Computational dialectology in Irish Gaelic. Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics, EACL, Dublin, S. 60-67.

Kortmann, B. (2010): Areal variation in syntax. In: Auer, P. & Schmidt, J. E. (Hrsg.): Language and Space. An International Handbook of Linguistic Variation. Vol. 1 Theories and Methods. Mouton De Gruyter, Berlin, S. 837-864.

Lameli, I., Kehrein, R. & Rabanus, S. (Hrsg.) (2010): Language and Space. An international Handbook of Linguistic Variation. Vol. 2 Language Mapping. Mouton De Gruyter, Berlin/New York.

Lee, J. & Kretzschmar, Jr., W. (1993): Spatial analysis of linguistic data with GIS functions. International Journal of Geographical Information Systems, Vol. 7, Nr. 6, S. 541-560.

Löffler, H. (2003): Dialektologie. Eine Einführung. Narr Studienbücher, Tübingen.

Moran, P. A. P. (1950): Notes on continuous stochastic phenomena. Biometrika, Nr. 37, S. 17-33.

Nerbonne, J. & Heeringa, W. (1997): Measuring Dialect Distance Phonetically In: Coleman, J. (Hrsg.): Workshop on Computational Phonology. Special Interest Group of the Association for Computational Linguistics. Madrid, S.11-18.

Nerbonne, J. & Kleiweg, P. (2003): Lexical Distance in LAMSAS. In: Nerbonne J. & Kretzschmar, Jr. W. (Hrsg.): Computational Methods in Dialectometry. Special issue of Computers and the Humanities, Vol. 37, Nr. 3, S. 339-357.

Nerbonne, J. & Kretzschmar, Jr., W. (2003): Introducing Computational Methods in Dialectometry In: Nerbonne J. & Kretzschmar, Jr. W. (Hrsg.): Computational Methods in Dialectometry. Special issue of Computers and the Humanities, Vol. 37, Nr. 3, S. 245-255.

Nerbonne, J. & Wiersma, W. (2006): A Measure of Aggregate Syntactic Distance. In: Nerbonne, J. & Hinrichs, E. (Hrsg.): Linguistic Distances Workshop at the joint conference of International Committee on Computational Linguistics and the Association for Computational Linguistics, Sydney, S. 82-90.

Nerbonne, J. & Kleiweg, P. (2007): Toward a Dialectological Yardstick. Journal of Quantitative Linguistics, Vol. 14, Nr. 2, S. 148–167.

Nerbonne, J., Kleiweg, P., Heeringa, W. & Manni, F. (2008): Projecting Dialect Distances to Geography Bootstrap Clustering vs. Noisy Clustering. In: Christine Preisach, Lars Schmidt-Thieme, Hans Burkhardt & Reinhold Decker (Hrsg.): Data Analysis, Machine Learning, and Applications. Proceedings of the 31st Annual Meeting of the German Classification Society, Springer, Berlin, S. 647-654. (Studies in Classification, Data Analysis, and Knowledge Organization)

Nerbonne, J. (2009): Data-Driven Dialectology. Language and Linguistics Compass, Vol. 3, Nr. 1, S. 175-198.

Nerbonne, J. & Heeringa, W. (2009): Measuring Dialect Differences In: Auer, P. & Schmidt, J. E. (Hrsg.): Language and Space. An International Handbook of Linguistic Variation. Vol. 1 Theories and Methods. Mouton De Gruyter, Berlin, S. 550-567.

Nerbonne J. (**2010):** Mapping aggregate variation. In: Lameli, A., Kehrein, R. & Rabanus, S. (Hrsg.): An international Handbook of Linguistic Variation. Vol. 2 Language Mapping. Mouton De Gruyter, Berlin/New York, S. 476-495.

Nerbonne, J., Colen, R., Gooskens, C., Kleiweg, P. & Leinonen, T. (im Druck): Gabmap - A Web Application for Dialectology.
http://www.let.rug.nl/nerbonne/papers/Gabmap-long-2011.pdf, Zugriff: 19.4.2011

Ord, J. K. & Getis, A. (1995): Local spatial autocorrelation statistics. Distributional Issues and an application. Geographical Analysis, Vol. 27, Nr. 4, S. 286-306.

O'Sullivan D. & Unwin, D. J. (2010): Geographic Information Analysis. Zweite Ausgabe. John Wiley & Sons, Hoboken.

Pearson, K. & Lee, A. (1897): On the Distribution of Frequency (Variation and Correlation) of the Barometric Height at Divers Stations, Philosophical Transactions of the Royal Society of London. Series A, Vol. 190, S. 423-469.

Pickl, S. & Rumpf, J. (unveröffentlicht): Dialectometric Concepts of Space. Towards a Variant-Based Dialectometry.

Ripley, B. D. (1977): Modelling Spatial Patterns. Journal of the Royal Statistical Society, Series B., Vol. 39, S. 172-212.

Rogerson, P. A. (2010): Statistical methods for geography. A student's guide. Dritte Ausgabe. SAGE, London.

Rumpf, J., Pickl, S., Elspass, S., König, W. & Schmidt, V. (2009): Structural analysis of dialect maps using methods from spatial statistics. Zeitschrift fur Dialektologie und Linguistik, Vol. 76, Nr. 3, S. 280–308.

Rumpf, J., Pickl, S., Elspass, S., König, W. & Schmidt, V. (2010): Quantification and Statistical Analysis of Structural Similarities in Dialectological Area-class Maps. Dialectologia et Geolinguistica, Vol. 18, S. 73-98.

Hotzenköcherle, R., Trüb, R. et al. (1962–2003): Sprachatlas der deutschen Schweiz., Bände I–IX, Bern/Tübingen.

Schrambke, R. (2010): Language and Space. Traditional dialect geography. In: Auer, P. & Schmidt, J. E. (Hrsg.): Language and Space. An International Handbook of Linguistic Variation. Vol. 1 Theories and Methods. Mouton De Gruyter, Berlin, S. 107-125.

Schwarz, E. (1950): Die Deutschen Mundarten. Vandenhoeck & Ruprecht, Göttingen.

Scott, D. W. (1992): Multivariate Density Estimation. Theory, Practice, and Visualization. John Wiley & Sons, New York.

Séguy, J. (1973): La dialectométrie dans l'Atlas linguistique de la Gascogne. Revue de linguistique romane, Vol. 37, S. 1–24.

Seiler, G. (2005): Wie verlaufen syntaktische Isoglossen, und welche Konsequenzen sind daraus zu ziehen? In: Eggers, E., Schmidt, J. E. & Stellmacher, D. (Hrsg.): Moderne Dialekte – neue Dialektologie. Steiner (ZDL Beiheft 130), Stuttgart, S. 313–341.

Seiler, G. (2008): Syntaxgeographie und Plastizität der Grammatik. In: Valentin, J.-M., Vinckel, H. (Hrsg.): Akten des XI. Internationalen Germanistenkongresses in Paris 2005 »Germanistik im Konflikt der Kulturen«. Bd. 4, Bern, S. 49–58.

Sheather, S. J. & Jones, M. C. (1991): A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation. Journal of the Royal Statistical Society. Series B, Vol. 53, Nr. 3, S. 683-690.

Silverman, B. W. (1986): Density Estimation for Statistics and Data Analysis. Chapman and Hall, London.

Spruit, M. (2006): Measuring syntactic variation in Dutch dialects. In: Nerbonne, J. & Kretzschmar, Jr., W. (Hrsg.): Literary and Linguistic Computing, special issue on Progress in Dialectometry: Toward Explanation, Vol. 21, Nr. 4, Oxford University Press, Oxford, S. 493–506.

Spruit, M. R. (2008): Quantitative perspectives on syntactic variation in Dutch dialects. PhD thesis, University of Amsterdam LOT Dissertation Series, LOT, Utrecht, 2008.

Spruit, M. R., Heeringa, W. & Nerbonne, J. (2009): Associations among linguistic levels. Lingua, Vol. 119, Nr. 11, The forests behind the trees, Elsevier, S. 1624-1642.

Steiner, J. (2005): Also d' Susi wär e ganz e liebi Frau für de Markus! Zur Verdoppelung des indefiniten Artikels in der adverbiell erweiterten Nominalphrase im Schweizerdeutschen. In: Christen, H. (Hrsg.): Dialektologie an der Jahrtausendwende. Linguistik Online, Vol. 24, Nr. 3. http://www.linguistik-online.de/24_05/steiner.pdf, Zugriff: 19.4.2011

Steiner, J. (2006): Syntaktische Variation in der Nominalphrase – ein Fall für die Dialektgeographin oder den Soziolinguisten In: Klausmann, H. (Hrsg.): Beiträge der 15. Arbeitstagung zur alemannischen Dialektologie. Schloss Hofen, Vorarlberg [Schriften der VLB, Band 15], Graz-Feldkirch, S. 109–113.

Steiner, J. (im Druck): ›E ganz e liebi Frau‹ – Zu den Stellungsvarianten in der adverbiell erweiterten Nominalphrase im Schweizerdeutschen. Dissertation, Universität Zürich.

Tobler, W. (1970): A computer movie simulating urban growth in the Detroit region. Economic Geography, Nr. 46, Vol. 2, S. 234-240.

Venables, W. N., Smith, D. M. & The R Development Core Team (2010): An Introduction to R. http://cran.r-project.org/doc/manuals/R-intro.pdf, Zugriff: 20.4.2011

Wattel E. & van Reenen, P. (2010): Probabilistic Maps. In: In: Lameli, A., Kehrein, R. & Rabanus, S. (Hrsg.): An international Handbook of Linguistic Variation. Vol. 2 Language Mapping. Mouton De Gruyter, Berlin/New York, S. 495-505.

Weber, A. (1987): Zürichdeutsche Grammatik. Ein Wegweiser zur guten Mundart, Schweizer Spiegel Verlag, Zürich.

Wenker, G. (1877): Das rheinische Platt. Den Lehrern des Rheinlandes gewidmet. Düsseldorf (Selbstverlag).

Wenker, G. (1881): Sprachatlas von Nord- und Mitteldeutschland. Auf Grund von systematisch mit Hilfe der Volksschullehrer gesammeltem Material aus circa 30.000 Orten. Einleitung. Straßburg.

Wiersma, W., Nerbonne, J. & Lauttamus (2011): Automatically Extracting Typical Syntactic Differences from Corpora. Literary and Linguistic Computing, Vol. 26, Nr. 1, S. 107-124.

Wonnacott, T. H. & Wonnacott, R. J. (1972): Introductory Statistics. Zweite Ausgabe. John Wiley & Sons, New York.

Anhang

A: Klassierung der behandelten Phänomene

A: Finalanschluss, Wahl und Position des Anschlussmittels für Finalsätze

Grundhypothese: Phänomene mit einer markanten Ost-West-Verteilung

Fett gedruckt: eingeschränkte Klassierung (nach Seiler 2005)

Kursiv: Erweiterte Klassierung

I.1 ... für/zum ein Billet (zu) lösen (Übersetzungsfrage)

Klasse	Variante	Antwort-Nr.	Code	Antwort
1	**FÜR (... ZU)**	**1.1** **1.2**	**1** **2**	**Ich ha z wenig Münz, für es Billet z lööse** **Ich ha z wenig Münz, für es Billet lööse**
2	**ZUM (...ZU)**	**1.4** **1.5**	**3** **4**	**Ich ha z wenig Münz, zum es Billet z lööse** **Ich ha z wenig Münz, zum es Billet lööse**
3	*UM... (ZU)*	*1.6* *1umlösen*	*5* *6*	*Ich ha z wenig Münz, um es Billet z lööse (chönne z lööse)* *Ich h a z wenig Münz, um es Billet lööse*
4	*FÜR ZUM (ZU...)*	*1.7* *1.8*	*7* *8*	*Ich ha z wenig Münz, für zum es Billet z lööse* *Ich ha z wenig Münz, für zum es Billet lööse*
5	*ZU-Infinitvanschluss ohne explizite Nebensatzeinleitung*	*1.3*	*9*	*Ich ha z wenig Münz, es Billett z lööse*
6	*FINITER NEBENSATZ MIT DASS BZW. DAMIT*	*1dass*	*10*	*Ich ha z wenig Münz, dass/damit (i cha es Billet lööse)*

I.6 ... für/zum ein(zu)schlafen (Ergänzungsfrage)

Klasse	Variante	Antwort-Nr.	Code	Antwort
1	**FÜR (... ZU)**	**6.1** **6.2**	**1** **2**	**... für iizschlafe** **... für iischlaafe**
2	**ZUM (...ZU)**	**6.4** **6.5**	**3** **4**	**... zum iischlaafe** **... zum iizschlafe**
3	*UM (... ZU)*	*6.6* *Sonst-Feld 37*	*5* *6*	*... um iizschlaafe* *... um iischlaafe*
4	*FÜR ZUM (ZU...)*	*6.7* *6.8*	*7* *8*	*... für zum iizschlaafe* *... für zum iischlaafe*
5	*FINITER NEBENSATZ MIT DASS BZW. DAMIT*	*6.9*	*9*	*... dass/damit i cha iischlafe*
6	*NORMALISIERUNG DES TYPS FÜR DAS (EIN-)SCHLAFEN*	*6.10* *6.14*	*10* *11*	*... für s Schlaafe* *... für s iischlafe*

I.11 … um ein Buch zu lesen (Ankreuzfrage)

Klasse	Variante	Antwort-Nr.	Code	Antwort
1	**FÜR (… ZU)**	**11.1** **11.2**	**1** **2**	**… für es Buech z läse** **… für es Buech läse**
2	**ZUM (…ZU)**	**11.4** **11.5**	**3** **4**	**… zum es Buech läse** **… zum es Buech z läse**
3	*UM… ZU*	*11.6*	*5*	*… um es Buech z läse*
4	*FÜR ZUM (ZU…)*	*11.7* *11.8*	*6* *7*	*… für zum es Buech z läse* *… für zum es Buech läse*
5	*ZU-Infinitvanschluss ohne explizite Nebensatzeinleitung*	*11.3*	*8*	*… es Buech z läse*
6	*FINITER NEBENSATZ MIT DASS BZW. DAMIT*	*11.9*	*9*	*… dass (da/as) i cha es Buech läse*

IV.14 … für/zum (zu) lesen (Ankreuzfrage)

Klasse	Variante	Antwort-Nr.	Code	Antwort
1	**FÜR (… ZU)**	**14.1** **14.2**	**1** **2**	**Muesch s Liecht aazünde für z läse** **Muesch s Liecht aazünde für läse**
2	**ZUM (…ZU)**	**14.3** **14.5**	**3** **4**	**Muesch s Liecht aazünde zum läse** **Muesch s Liecht aazünde zum z läse**
3	*UM… ZU*	*14.6*	*5*	*Muesch s Liecht aazünde um z läse*
4	*FÜR ZUM (ZU…)*	*14.7* *14.8*	*6* *7*	*Muesch s Liecht aazünde für zum z läse* *Muesch s Liecht aazünde für zum läse*
5	*NORMALISIERUNG DES TYPS FÜR DAS LESEN*	*14.9*	*8*	*… fürs läsä*
6	*FINITER NEBENSATZ MIT DASS BZW. DAMIT*	*14.11*	*9*	*… dass / as chasch läse*

B: Komparativ

Grundhypothese: Phänomene mit einer dominanten Variante, die überall vorkommt und einzelner Varianten, die kleinere Areale bilden

Fett gedruckt: eingeschränkte Klassierung (nach Friedli 2005)

Kursiv: weitere Varianten

III.22 Sie ist grösser als ich (Ankreuzfrage)

Klasse	Variante	Antwort-Nr.	Code	Antwort
1	**ALS**	**22.1**	**1**	**Si isch grösser als ich**
2	**WEDER**	**22.2**	**2**	**Si isch grösser weder ich**
3	**WIE**	**22.3**	**3**	**Si isch grösser wie ich**
4	**WA(N)**	**22.4**	**4**	**Si isch grösser wan ich**

III.25 Sie gehen halt lieber schwimmen statt spazieren (Ankreuzfrage)

Klasse	Variante	Antwort-Nr.	Code	Antwort
1	**ALS**	**25.1**	**1**	**Si gönd halt lieber go schwimme als go lauffe**
2	**WEDER**	**25.3**	**2**	**Si gönd halt lieber go schwimme weder go lauffe**
3	**WIE**	**25.2**	**3**	**Si gönd halt lieber go schwimme wie go lauffe**
4	**WA(N)**	**25.5**	**4**	**Si gönd halt lieber go schwimme wan go lauffe**
5	*(AN)STATT*	*25.4*	*5*	*Si gönd halt lieber go schwimme statt go lauffe*

III.28 Dann ist er ja älter, als ich gedacht habe (Ankreuzfrage)

Klasse	Variante	Antwort-Nr.	Code	Antwort
1	**ALS**	**28.2**	**1**	**Dänn isch er ja älter als ich gmeint han**
2	**WEDER**	**28.1**	**2**	**Dänn isch er ja älter weder ich gmeint han**
3	**WIE**	**28.3**	**3**	**Dänn isch er ja älter wie ich gmeint han**
4	**WA(N)**	**28.9**	**4**	**Dänn isch er ja älter wan ich gmeint han**
5	*WEDER + ZUSATZ*	*28.4* *28.5* *28.6* *28.18*	*5* *6* *7* *8*	*weder as* *weder als* *weder dass* *weder was*
6	*ALS + ZUSATZ*	*28.7* *28.8* *28.13* *28.17*	*9* *10* *11* *12*	*als as* *als dass* *als was* *als wie*
7	*WIE + ZUSATZ*	*28.12*	*13*	*wie dass*
8	*WAN + ZUSATZ*	*28.10*	*14*	*wan dass*
9	*OHNE VERGLEICHSWORT + ZUSATZ*	*28.14* *28.15*	*15* *16*	*dass* *as*

C: Artikelverdoppelung

Grundhypothese: keine Areale werden gebildet

I.10 Also d Susi wär e ganz e liebi Frau für de Markus (Ankreuzfrage)

Klasse	Variante	Antwort-Nr.	Code	Antwort
1	DOPPELUNG DES INDEFINITEN ARTIKELS	10.1	1	e ganz e liebi Frau
2	EINFACHES AUFTRETEN NACH ADVERB	10.2	2	ganz e liebi frau
3	EINFACHES AUFTRETEN VOR ADVERB	10.3	3	e ganz liebi frau

II.10 Aber du häsch de vil de schöner Garte (Ankreuzfrage)

Klasse	Variante	Antwort-Nr.	Code	Antwort
1	DOPPELUNG DES INDEFINITEN ARTIKELS	10.1	1	de vil de schöner Garte
2	EINFACHES AUFTRETEN NACH ADVERB	10.2	2	vil de schöner Garte
3	EINFACHES AUFTRETEN VOR ADVERB	10.3	3	de vil schöner Garte

IV.1 Martina wäre eine ganz gute Gemeindepräsidentin (Übersetzungsfrage)

Klasse	Variante	Antwort-Nr.	Code	Antwort
1	DOPPELUNG DES INDEFINITEN ARTIKELS	1.1	1	e ganz e gueti Gmeindspräsidentin
2	EINFACHES AUFTRETEN NACH ADVERB	1.3	2	ganz e gueti Gmeindspräsidentin
3	EINFACHES AUFTRETEN VOR ADVERB	1.2	3	e ganz gueti Gmeindspräsidentin

B: Tabelle mit den Resultaten der Bandbreitenkalibrierung

VALIDIERUNGSTABELLEN FÜR DIE WAHL DER BANDBREITE DER KDE

I.1K: manuelle, globale Bandbreitenwahl

Aggregierungsebene: SADS-Untersuchungsorte

bw [m]	Fläche gesamt [m^2]		Überschneidungsfläche [m^2]	FK_Ant SADS-Orte [%]		K_Ant SADS-Orte [%]	RMSE Zentroide
	mit 0-Fl.	*ohne 0-Fl.*		*mit 0-Fl.*	*ohne 0-Fl.*		
2500	24'753'430'507	23'608'415'058	23'572'382'908	95.229	99.847	99.724	0.062
5000	24'753'430'507	23'608'415'058	22'982'909'324	92.847	97.350	96.133	0.132
7500	24'753'430'507	23'608'415'058	22'558'847'747	91.134	95.554	93.923	0.156
10000	24'753'430'507	23'608'415'058	22'544'904'308	91.078	95.495	93.923	0.166
12500	24'753'430'507	23'608'415'058	22'266'225'439	89.952	94.315	91.989	0.172
15000	24'753'430'507	23'608'415'058	22'125'513'213	89.384	93.719	91.160	0.176
20000	24'753'430'507	23'608'415'058	22'021'901'027	88.965	93.280	90.331	0.181
25000	24'753'430'507	23'608'415'058	21'829'453'210	88.188	92.465	90.055	0.185
30000	24'753'430'507	23'608'415'058	21'870'333'359	88.353	92.638	90.055	0.191
40000	24'753'430'507	23'608'415'058	21'901'360'673	88.478	92.769	90.331	0.204
50000	24'753'430'507	23'608'415'058	21'839'323'209	88.227	92.507	90.055	0.222
75000	24'753'430'507	23'608'415'058	21'963'918'517	88.731	93.034	90.331	0.267
100000	24'753'430'507	23'608'415'058	21'800'366'035	88.070	92.342	88.950	0.298
200000	24'753'430'507	23'608'415'058	13'277'329'269	53.638	56.240	51.657	0.333

Validierung	2500	5000	7500	10000	12500	15000	20000
RMSE_K_1	0.045	0.111	0.134	0.144	0.149	0.153	0.158
RMSE_K_2	0.053	0.116	0.138	0.147	0.153	0.157	0.164
RMSE_idomK	0.062	0.132	0.156	0.166	0.172	0.176	0.181
Kla_vali	0.997	0.961	0.939	0.939	0.920	0.912	0.903

Validierung	25000	30000	40000	50000	75000	100000	200000
RMSE_K_1	0.163	0.169	0.181	0.197	0.243	0.275	0.320
RMSE_K_2	0.169	0.174	0.185	0.197	0.231	0.256	0.292
RMSE_idomK	0.185	0.191	0.204	0.222	0.267	0.298	0.333
Kla_vali	0.901	0.901	0.903	0.901	0.903	0.890	0.517

Aggregierungsebene: Deutschschweizer Gemeinden

bw [m]	Fläche gesamt [m²]		Überschneidungsfläche [m²]	FK_Ant dt-CH Gemeinden [%]		K_Ant dt-CH Gemeinden [%]	RMSE Zentroide
	mit 0-Fl.	*ohne 0-Fl.*		*mit 0-Fl.*	*ohne 0-Fl.*		
2500	24'753'430'507	23'608'415'058	23'185'635'524	93.666	98.209	99.724	0.062
5000	25'604'094'414	23'608'415'058	22'778'584'672	88.965	96.485	96.133	0.132
7500	25'604'094'414	23'608'415'058	22'577'884'166	88.181	95.635	93.923	0.156
10000	25'604'094'414	23'608'415'058	22'497'526'437	87.867	95.295	93.923	0.166
12500	25'604'094'414	23'608'415'058	22'386'262'467	87.432	94.823	91.989	0.172
15000	25'604'094'414	23'608'415'058	22'213'978'776	86.759	94.093	91.160	0.176
20000	25'604'094'414	23'608'415'058	22'117'474'884	86.383	93.685	90.331	0.181
25000	25'604'094'414	23'608'415'058	21'908'855'783	85.568	92.801	90.055	0.185
30000	25'604'094'414	23'608'415'058	21'845'340'773	85.320	92.532	90.055	0.191
40000	25'604'094'414	23'608'415'058	21'838'574'097	85.293	92.503	90.331	0.204
50000	25'604'094'414	23'608'415'058	21'833'395'399	85.273	92.481	90.055	0.222
75000	25'604'094'414	23'608'415'058	22'036'337'905	86.066	93.341	90.331	0.267
100000	25'604'094'414	23'608'415'058	21'770'005'635	85.025	92.213	88.950	0.298
200000	25'604'094'414	23'608'415'058	13'311'232'363	51.989	56.383	51.657	0.333

Validierung	2500	5000	7500	10000	12500	15000	20000
RMSE_K_1	0.045	0.111	0.134	0.144	0.149	0.153	0.158
RMSE_K_2	0.053	0.116	0.138	0.147	0.153	0.157	0.164
RMSE_idomK	0.062	0.132	0.156	0.166	0.172	0.176	0.181
Kla_vali	0.997	0.961	0.939	0.939	0.920	0.912	0.903

Validierung	25000	30000	40000	50000	75000	100000	200000
RMSE_K_1	0.163	0.169	0.181	0.197	0.243	0.275	0.320
RMSE_K_2	0.169	0.174	0.185	0.197	0.231	0.256	0.292
RMSE_idomK	0.185	0.191	0.204	0.222	0.267	0.298	0.333
Kla_vali	0.901	0.901	0.903	0.901	0.903	0.890	0.517

I.1E: manuelle, globale Bandbreitenwahl

Aggregierungsebene: SADS-Untersuchungsorte

bw [m]	Fläche gesamt [m^2]		Überschneidungsfläche [m^2]	FK_Ant SADS-Orte [%]		K_Ant SADS-Orte [%]	RMSE Zentroide
	mit 0-Fl.	*ohne 0-Fl.*		*mit 0-Fl.*	*ohne 0-Fl.*		
2500	24'753'430'507	23'095'954'464	22'918'451'872	92.587	99.231	98.867	0.060
5000	24'753'430'507	23'095'954'464	21'460'121'131	86.696	92.917	91.785	0.129
7500	24'753'430'507	23'095'954'464	20'981'663'500	84.763	90.846	88.952	0.150
10000	24'753'430'507	23'095'954'464	20'863'773'945	84.286	90.335	88.385	0.160
12500	24'753'430'507	23'095'954'464	20'650'500'927	83.425	89.412	86.969	0.165
15000	24'753'430'507	23'095'954'464	20'477'731'481	82.727	88.664	85.836	0.168
20000	24'753'430'507	23'095'954'464	20'267'513'292	81.878	87.754	84.986	0.173
25000	24'753'430'507	23'095'954'464	20'925'729'384	84.537	90.603	84.703	0.179
30000	24'753'430'507	23'095'954'464	20'075'065'476	81.100	86.920	84.136	0.185
40000	24'753'430'507	23'095'954'464	20'011'805'278	80.845	86.646	84.419	0.202
50000	24'753'430'507	23'095'954'464	19'961'506'155	80.641	86.429	84.136	0.222
75000	24'753'430'507	23'095'954'464	20'035'783'914	80.941	86.750	84.136	0.271
100000	24'753'430'507	23'095'954'464	19'690'089'499	79.545	85.253	81.870	0.303
200000	24'753'430'507	23'095'954'464	11'984'850'653	48.417	51.892	47.025	0.338

Validierung	2500	5000	7500	10000	12500	15000	20000
RMSE_K_1	0.045	0.111	0.134	0.144	0.149	0.153	0.158
RMSE_K_2	0.053	0.116	0.138	0.147	0.153	0.157	0.164
RMSE_K_3	0.044	0.101	0.124	0.133	0.138	0.140	0.144
RMSE_K_4	0.003	0.010	0.012	0.013	0.014	0.014	0.014
RMSE_K_5	0.002	0.008	0.011	0.012	0.012	0.013	0.013
RMSE_K_6	0.015	0.033	0.041	0.043	0.044	0.045	0.045
RMSE_idomK	0.060	0.129	0.150	0.160	0.165	0.168	0.173
Kla_vali	0.989	0.918	0.890	0.884	0.870	0.858	0.850

Validierung	25000	30000	40000	50000	75000	100000	200000
RMSE_K_1	0.163	0.169	0.181	0.197	0.243	0.275	0.320
RMSE_K_2	0.169	0.174	0.185	0.197	0.231	0.256	0.292
RMSE_K_3	0.146	0.147	0.149	0.151	0.153	0.154	0.154
RMSE_K_4	0.014	0.014	0.014	0.014	0.014	0.014	0.014
RMSE_K_5	0.014	0.014	0.014	0.014	0.014	0.014	0.014
RMSE_K_6	0.046	0.046	0.046	0.046	0.046	0.046	0.046
RMSE_idomK	0.179	0.185	0.202	0.222	0.271	0.303	0.338
Kla_vali	0.847	0.841	0.844	0.841	0.841	0.819	0.470

Aggregierungsebene: Deutschschweizer Gemeinden

bw [m]	Fläche gesamt [m^2]		Überschneidungsfläche [m^2]	FK_Ant dt-CH Gemeinden [%]		K_Ant dt-CH Gemeinden [%]	RMSE
	mit 0-Fl.	*ohne 0-Fl.*		*mit 0-Fl.*	*ohne 0-Fl.*		
2500	24'753'430'507	23'095'954'464	22'280'186'192	90.008	96.468	98.867	0.060
5000	24'753'430'507	23'095'954'464	21'242'920'988	85.818	91.977	91.785	0.129
7500	24'753'430'507	23'095'954'464	20'886'738'734	84.379	90.435	88.952	0.150
10000	24'753'430'507	23'095'954'464	20'771'772'798	83.915	89.937	88.385	0.160
12500	24'753'430'507	23'095'954'464	20'616'629'768	83.288	89.265	86.969	0.165
15000	24'753'430'507	23'095'954'464	20'482'763'898	82.747	88.686	85.836	0.168
20000	24'753'430'507	23'095'954'464	20'239'209'887	81.763	87.631	84.986	0.173
25000	24'753'430'507	23'095'954'464	20'048'063'366	80.991	86.803	84.703	0.179
30000	24'753'430'507	23'095'954'464	19'988'739'938	80.751	86.546	84.136	0.185
40000	24'753'430'507	23'095'954'464	19'965'193'767	80.656	86.445	84.419	0.202
50000	24'753'430'507	23'095'954'464	19'937'269'841	80.543	86.324	84.136	0.222
75000	24'753'430'507	23'095'954'464	20'054'544'360	81.017	86.831	84.136	0.271
100000	24'753'430'507	23'095'954'464	19'663'398'706	79.437	85.138	81.870	0.303
200000	24'753'430'507	23'095'954'464	12'018'753'747	48.554	52.038	47.025	0.338

Validierung	2500	5000	7500	10000	12500	15000	20000
RMSE_K_1	0.045	0.111	0.134	0.144	0.149	0.153	0.158
RMSE_K_2	0.053	0.116	0.138	0.147	0.153	0.157	0.164
RMSE_K_3	0.044	0.101	0.124	0.133	0.138	0.140	0.144
RMSE_K_4	0.003	0.010	0.012	0.013	0.014	0.014	0.014
RMSE_K_5	0.002	0.008	0.011	0.012	0.012	0.013	0.013
RMSE_K_6	0.015	0.033	0.041	0.043	0.044	0.045	0.045
RMSE_idomK	0.060	0.129	0.150	0.160	0.165	0.168	0.173
Kla_vali	0.989	0.918	0.890	0.884	0.870	0.858	0.850

Validierung	25000	30000	40000	50000	75000	100000	200000
RMSE_K_1	0.163	0.169	0.181	0.197	0.243	0.275	0.320
RMSE_K_2	0.169	0.174	0.185	0.197	0.231	0.256	0.292
RMSE_K_3	0.146	0.147	0.149	0.151	0.153	0.154	0.154
RMSE_K_4	0.014	0.014	0.014	0.014	0.014	0.014	0.014
RMSE_K_5	0.014	0.014	0.014	0.014	0.014	0.014	0.014
RMSE_K_6	0.046	0.046	0.046	0.046	0.046	0.046	0.046
RMSE_idomK	0.179	0.185	0.202	0.222	0.271	0.303	0.338
Kla_vali	0.847	0.841	0.844	0.841	0.841	0.819	0.470

I.1K: automatisierte Bandbreitenwahl

Aggregierungsebene: SADS-Untersuchungsorte

	Fläche gesamt [m^2]		Überschneidungsfläche [m^2]	FK_Ant SADS-Orte [%]		K_Ant SADS-Orte [%]	RMSE
bw-Methode	*mit 0-Fl.*	*ohne 0-Fl.*		*mit 0-Fl.*	*ohne 0-Fl.*		
nrd0	24'753'430'507	23'608'415'058	22'544'904'308	91.078	95.495	93.923	0.168
nrdx	24'753'430'507	23'608'415'058	22'278'424'355	90.001	94.366	91.989	0.171
bcv	24'753'430'507	23'608'415'058	22'249'646'241	89.885	94.245	91.713	0.172
ucv	24'753'430'507	23'608'415'058	22'307'302'773	90.118	94.489	91.989	0.172
SJ	24'753'430'507	23'608'415'058	22'386'437'416	90.438	94.824	92.818	0.170

Validierung	nrd0	nrdx	bcv	ucv	SJ
RMSE_K_1	0.144	0.148	0.149	0.147	0.146
RMSE_K_2	0.149	0.153	0.153	0.152	0.151
RMSE_idomK	0.168	0.171	0.172	0.172	0.170
Kla_vali	0.939	0.920	0.917	0.920	0.928

Globale Statistik automatisierte bw SADS Orte		
bw-Methode	*avg bw*	*stdev bw*
nrd0	11037.771	2033.643
nrdx	13000.041	2395.179
bcv	13097.597	2465.349
ucv	11543.938	2482.268
SJ	11319.927	1872.575

Aggregierungsebene: Deutschschweizer Gemeinden

	Fläche gesamt [m^2]		Überschneidungsfläche [m^2]	FK_Ant dt-CH Gemeinden [%]		K_Ant dt-CH Gemeinden [%]	RMSE
bw-Methode	*mit 0-Fl.*	*ohne 0-Fl.*		*mit 0-Fl.*	*ohne 0-Fl.*		
nrd0	24'753'430'507	23'608'415'058	22'507'437'520	90.927	95.337	93.646	0.168
nrdx	24'753'430'507	23'608'415'058	22'403'029'111	90.505	94.894	92.818	0.172
bcv	24'753'430'507	23'608'415'058	22'353'193'180	90.303	94.683	92.265	0.172
ucv	24'753'430'507	23'608'415'058	22'380'162'026	90.412	94.797	92.265	0.170
SJ	24'753'430'507	23'608'415'058	22'445'152'128	90.675	95.073	93.370	0.170

Validierung	nrd0	nrdx	bcv	ucv	SJ
RMSE_K_1	0.144	0.148	0.149	0.146	0.146
RMSE_K_2	0.148	0.152	0.153	0.152	0.150
RMSE_idomK	0.168	0.172	0.172	0.170	0.170
Kla_vali	0.936	0.928	0.923	0.923	0.934

Globale Statistik automatisierte bw dt-CH Gemeinden		
bw-Methode	*avg bw*	*stdev bw*
nrd0	10989.106	1850.936
nrdx	12942.725	2179.991
bcv	13259.919	2276.375
ucv	11494.126	2337.143
SJ	11405.990	1712.937

I.1E: automatisierte Bandbreitenwahl

Aggregierungsebene: SADS-Untersuchungsorte

bw-Methode	Fläche gesamt [m^2]		Überschneidungsfläche [m^2]	FK_Ant SADS-Orte [%]		K_Ant SADS-Orte [%]	RMSE
	mit 0-Fl.	*ohne 0-Fl.*		*mit 0-Fl.*	*ohne 0-Fl.*		
nrd0	24'753'430'507	23'608'415'058	21'372'613'266	86.342	90.530	88.669	0.134
nrdx	24'753'430'507	23'608'415'058	21'285'908'883	85.992	90.162	87.252	0.138
bcv	24'753'430'507	23'608'415'058	21'257'130'769	85.875	90.040	86.686	0.138
ucv	24'753'430'507	23'608'415'058	21'257'130'769	85.875	90.040	86.686	0.136
SJ	24'753'430'507	23'608'415'058	21'304'120'180	86.065	90.240	87.535	0.135

Bandbreite	nrd0	nrdx	bcv	ucv	SJ
RMSE_K_1	0.144	0.148	0.149	0.147	0.146
RMSE_K_2	0.149	0.153	0.153	0.152	0.151
RMSE_K_3	0.134	0.138	0.138	0.136	0.135
RMSE_K_4	0.013	0.013	0.013	0.013	0.013
RMSE_K_5	0.011	0.011	0.012	0.011	0.011
RMSE_K_6	0.043	0.044	0.044	0.044	0.044
RMSE_idomK	0.159	0.162	0.163	0.165	0.163
Kla_vali	0.887	0.873	0.867	0.867	0.875

Globale Statistik automatisierte bw SADS Orte		
bw-Methode	*avg bw*	*stdev bw*
nrd0	11037.771	2033.643
nrdx	13000.041	2395.179
bcv	13097.597	2465.349
ucv	11543.938	2482.268
SJ	11319.927	1872.575

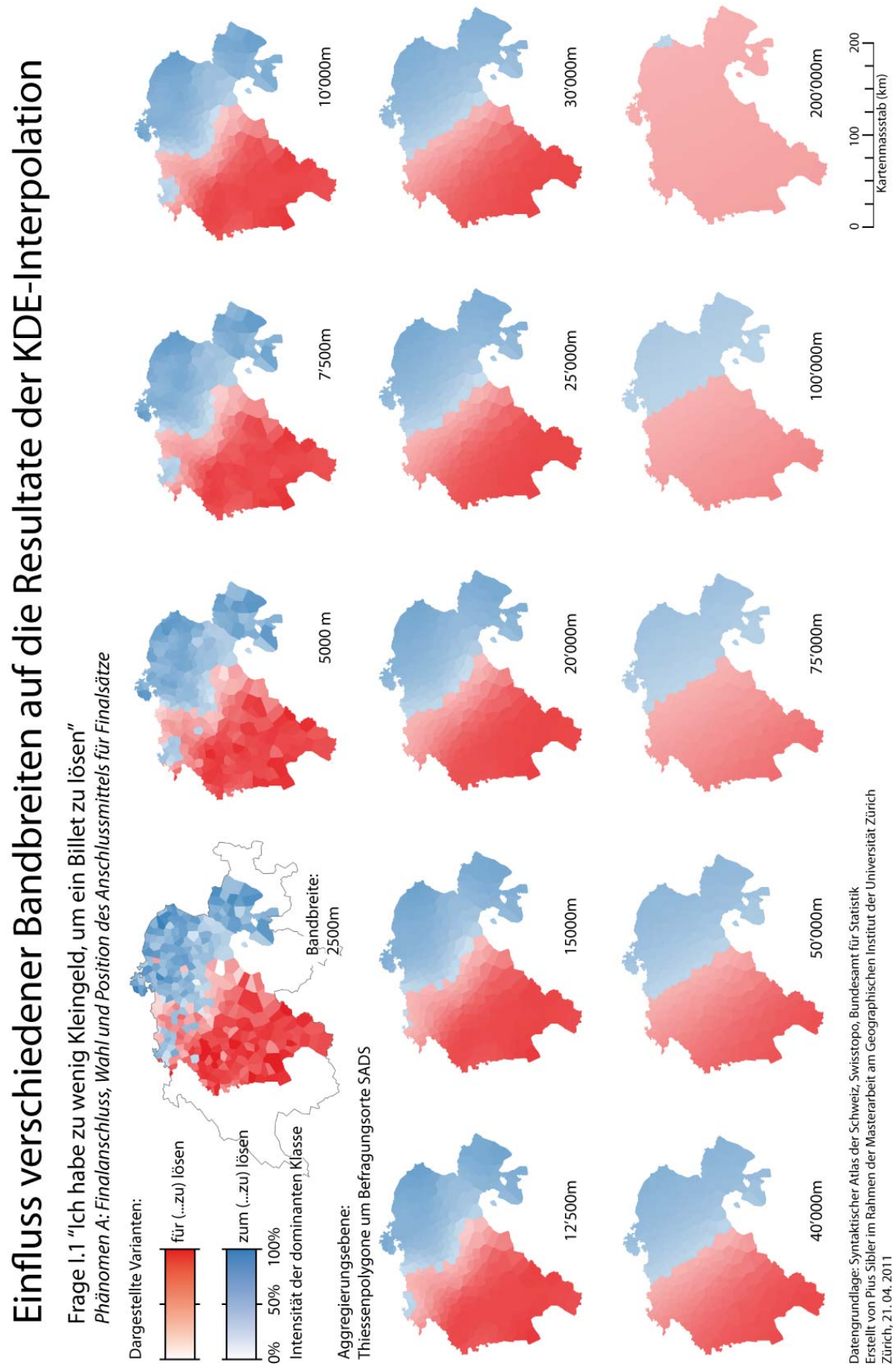
Aggregierungsebene: Deutschschweizer Gemeinden

bw-Methode	Fläche gesamt [m^2]		Überschneidungsfläche [m^2]	FK_Ant dt-CH Gemeinden [%]		K_Ant dt-CH Gemeinden [%]	RMSE
	mit 0-Fl.	*ohne 0-Fl.*		*mit 0-Fl.*	*ohne 0-Fl.*		
nrd0	24'753'430'507	23'608'415'058	21'350'703'072	86.254	90.437	88.669	0.134
nrdx	24'753'430'507	23'608'415'058	21'276'149'818	85.952	90.121	87.819	0.137
bcv	24'753'430'507	23'608'415'058	21'238'263'679	85.799	89.961	87.252	0.137
ucv	24'753'430'507	23'608'415'058	21'258'986'907	85.883	90.048	87.535	0.135
SJ	24'753'430'507	23'608'415'058	21'300'404'458	86.050	90.224	88.102	0.135

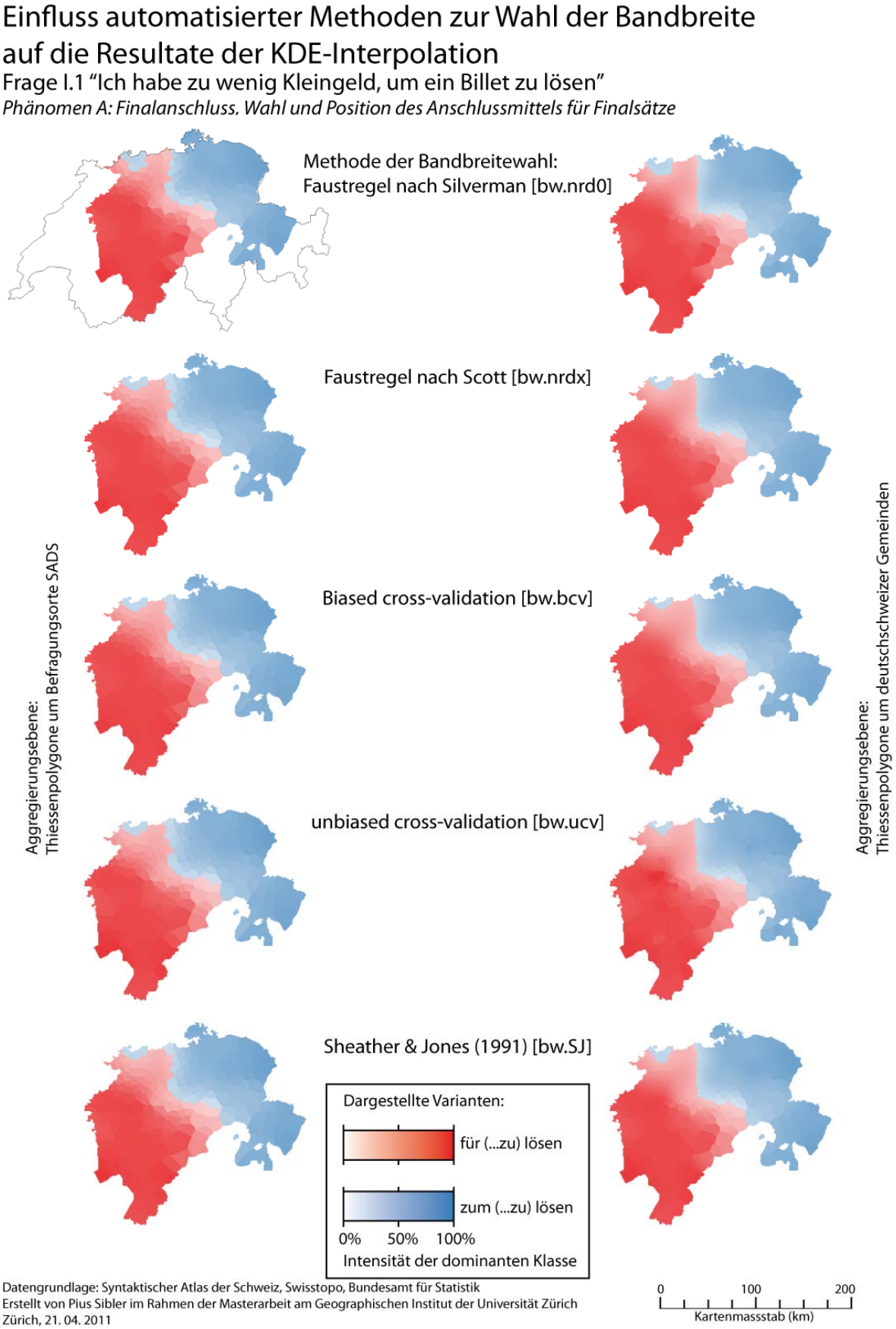
Bandbreite	nrd0	nrdx	bcv	ucv	SJ
RMSE_K_1	0.144	0.148	0.149	0.146	0.146
RMSE_K_2	0.148	0.152	0.153	0.152	0.150
RMSE_K_3	0.134	0.137	0.137	0.135	0.135
RMSE_K_4	0.013	0.013	0.013	0.013	0.013
RMSE_K_5	0.011	0.011	0.011	0.011	0.011
RMSE_K_6	0.043	0.044	0.044	0.043	0.043
RMSE_idomK	0.162	0.165	0.165	0.163	0.163
Kla_vali	0.887	0.878	0.873	0.875	0.881

Globale Statistik automatisierte bw dt-CH Gemeinden		
bw-Methode	*avg bw*	*stdev bw*
nrd0	10989.10646	1850.935888
nrdx	12942.72538	2179.991157
bcv	13259.91935	2276.375323
ucv	11494.12644	2337.142523
SJ	11405.99012	1712.937314

C1: Interpolierte Oberflächen mit manuell gewählten globalen Bandbreiten



C2: Interpolierte Oberflächen mit automatisierten Methoden zur Bandbreitenwahl



D: Tabelle mit den Resultaten der Trendoberflächenanalyse

R^2				
Variante *für*	I.1K	I.6K	I.11K	IV.14K
o1	0.594	0.160	0.706	0.626
o2	0.616	0.485	0.747	0.733
o3	0.686	0.502	0.787	0.754
o4	0.700	0.558	0.798	0.767
Variante *zum*				
o1	0.444	-0.456	0.706	0.626
o2	0.516	-0.020	0.728	0.733
o3	0.559	0.040	0.753	0.754
o4	0.581	0.094	0.760	0.767
***F*-Wert**				
Variante *für*	I.1K	I.6K	I.11K	IV.14K
o1	277.959**	36.192**	455.404**	318.137**
o2	120.926**	71.069**	222.302**	207.108**
o3	90.371**	41.768**	153.175**	127.238**
o4	61.473**	33.177**	103.963**	86.364**
Variante *zum*				
o1	151.891**	-59.513	455.590**	270.245**
o2	80.232**	-1.509	201.401**	131.612**
o3	52.597**	1.713	126.648**	77.513**
o4	36.512**	2.741**	83.285**	51.932**
F-Grenzwerte (p= 0.01)				
	Freiheitsgrade	Grenzwerte		
o1	F(2:380)	4.66		
o2	F(5:377)	3.07		
o3	F(9:373)	2.46		
o4	F(14:368)	2.13		

Bestimmtheitsmasse und *F*-Werte der dominanten Finalanschlussvarianten und Grenzwerte der ***F***-Funktion (p=0.01) mit den entsprechenden Freiheitsgraden für die verschiedenen Ebenenordnungen (o1-o4)

E: Inhalt der Software-CD

- A_digitale Arbeit in PDF-Form

- B_digitale Anhänge:

Ordner	Inhalt	Entsprechende Abschnitte
1_Tabellenaufbereitung	• VBA-Skript zur Bestimmung der maximalen Akzeptanzen und dominanten Klassen (`maxValues.vbs`) • Detaillierter Ablauf für die SADS-Tabellenaufbereitung am Beispiel I.1 (`Tabellenaufbereitung_GIS.pdf`)	3.1.4
2_KDE	• /automatisierte Bandbreiten: R-Skripte für die Intensitätsschätzung mit KDE mit automatisierten Bandbreite Methoden für beide Klassierungen der Frage I.1 (`KDE_A1_K_automatisierte_Bandbreiten.r`) • /manuelle Bandbreite 10000: R-Skripte für die Intensitätsschätzung mit KDE mit einer globalen Bandbreite von 10‘000 Metern für beide Klassierungen aller behandelten Fragen (`KDE_A1_K_10000.r`)	3.3.4 & 3.3.5
	• /Erweiterungen: R-Skript mit nach GP gewichteter KDE am Beispiel I.1K (`KDE_A1_K_10000_gew_GP.r`) R-Skript für die kombinierte KDE der eingeschränkten Klassierung des Finalanschlusses (`Phaenomen_A_K_10000_kombi.r`)	3.3.7
3_syntaktische_Distanzmasse	• R-Skript zur Bestimmung der Hamming-Distanz und des Relativen Identitätswertes am Beispiel I.1E (`syntaktische_Distanzmasse_1_1E.r`)	3.4.1 & 3.4.2
4_Kalibrierung_Bandbreite	• ArcToolbox zur Berechnung der Grösse FK_{Ant} (`Flächenüberschneidung.tbx`) • Workflow zur Berechnung von FK_{Ant} (`Flächenüberschneidung.pdf`)	4.1
5_3D	• Vrml-Datei mit der 3D-Darstellung der Frage I.1K der auf Aggregierungsebene der Gemeinden (`f1_1K_Punkte_3D.wrl`) • Setup (Windows) für den FreeWRL-Viewer (`SetupFreeWRL_1.22.10.msi`)	5.4
6_Strukturkenngrössen	• R-Skripte zur Berechnung der Homogenitäts- & Kompaktheits-Strukturgrössen für die eingeschränkte Klassierung aller behandelten Fragen (`Bsp. A1_K_Strukturkenngrössen.r`)	8.1.2 & 8.1.3

Persönliche Erklärung

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbständig verfasst und die den verwendeten Quellen wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Zürich, 29. April

Pius Sibler