QUANTITATIVE MODELLING OF SPATIAL VARIATION IN SWISS GERMAN DIALECTS

Dissertation

zur

Erlangung der naturwissenschaftlichen Doktorwürde
(Dr. sc. nat.)

vorgelegt der

Mathematisch-naturwissenschaftlichen Fakultät

der

Universität Zürich

von

Péter Jeszenszky

aus

Ungarn

Promotionskommission

Prof. Dr. Robert Weibel (Vorsitz)

Prof. Dr. Elvira Glaser

Prof. Dr. Ross Purves

Zürich, 2018

ZUSAMMENFASSUNG

Die Variation, die linguistischen und dialektalen Daten innewohnt, ist ein wichtiges Forschungsfeld, dessen Relevanz sich nicht zuletzt aus den wahrgenommenen sprachlichen Unterschieden und Ähnlichkeiten ergibt. Die ständig zunehmende Anzahl digitaler linguistischer Daten geht mit einer fortschreitenden Entwicklung computergestützter Berechnungsmethoden (zum Beispiel in der Geographischen Informationswissenschaft: GIScience) einher, woraus sich ein großes Potential der Zusammenarbeit zwischen rechnergestützten Wissenschaften und der Linguistik ergibt. Obwohl die Dialektologie die Entstehung von Dialektgebieten gründlich erforscht hat, stellt die quantitative Modellierung der räumlichen Übergänge zwischen Sprachgebieten weiterhin ein Desiderat dar. Der Einbezug räumlicher und quantitativer Verfahren wurde daher in der Sprachforschung zwar häufig gefordert, in der Praxis aber bislang selten durchgeführt.

Das Ziel der vorliegenden Dissertation ist es, zu einem besserem Verständnis der Rolle des geographischen Raumes in der Dialektforschung beizutragen, indem das Potenzial an GIScience-Methoden ausgeschöpft wird. Dabei liegt der Fokus auf der Erarbeitung von Methoden für zwei wichtige Forschungsthemen: der Einfluss geographischer Distanzen auf die dialektale Variation und die quantitative Einschätzung sprachinterner Grenzen beziehungsweise Übergänge. Die im vorliegenden Forschungsvorhaben vorgenommenen Studien basieren auf Daten des Syntaktischen Atlas der deutschen Schweiz (SADS). Der SADS ist ein auf Fragebogenerhebung basierender Dialektatlas mit Fokus auf (morpho-)syntaktischen Phänomenen. Ein besonderes Kennzeichen des SADS besteht darin, dass an jedem Erhebungsort mehrere Antworten erhoben wurden, was ihn von den meisten Dialektumfragen unterscheidet.

Der wesentliche Beitrag dieser Dissertation ist die Verwendung neuer Methoden in der in Dialektologie und Dialektometrie durch die Nutzung etablierter Verfahren aus GIScience und der räumlichen Analyse. Korrelationsanalysen mit verschiedenen räumlichen Granularitäten zwischen linguistischen und geographischen Distanzmatrizen zeigen, dass die geographische Distanz einen Grossteil der Varianz in der schweizerdeutschem Syntax aufzuklären vermag. Die Arbeit zeigt, dass Reisezeiten (für die Jahre 2000, 1950 und 1850) die räumliche Varianz in der Syntax besser wiedergeben als Euklidische Distanz; und dass ältere Reisezeiten zeitgenössische räumliche Varianz in der Syntax besser erklären als aktuellere. Regressions- und Trendoberflächenanalyse wurden sowohl zum quantitativen Vergleich räumlicher Varianz syntaktischer Variablen, als auch zur Prüfung der Übereinstimmung mit theoretischen Modellen angewandt und verdeutlichen die Wichtigkeit der lokalen neben der globalen Analyse. Abgesehen von der Modellierung der Grenzen wird eine Methode vorschlagen, die auf sprachlichen Filtern und räumlichem Clustering für die probabilistische Erfassung sprachinterner Übergänge und Grenzen basiert

und neue Lösungsansätze aufzeigt, die gängige, visuelle Verfahren der Grenzerkennung komplementieren. Während die Fallstudien dieser Dissertation den Nutzen der räumlichen Analyse und Statistik für Fragestellungen der Sprachforschung veranschaulichen, zeigen diese Untersuchungen auch, dass die Linguistik viele interessante und anspruchsvolle Themen für die GIScience Forschung bietet, die zudem Anknüpfungspunkte über sprachwissenschaftliche Fragestellungen hinaus in andere Bereiche der Digital Humanities liefern.

SUMMARY

The variation inherent to linguistic and dialectal data is an important field of research due to the identities created through the perceived linguistic differences and similarities. Digital linguistic data are being produced at an ever-increasing rate, paralleled by advances in computational methods development, including in Geographic Information Science (GIScience), leading to a great potential for collaboration between the computational sciences and linguistics. Although dialectology has thoroughly studied the formation of dialect areas, quantitative modelling of the spatial transition between different linguistic areas stays on the agenda. The inclusion of spatial, quantitative methods has therefore often been advocated in linguistics, yet spatially explicit methodologies have rarely been exploited so far.

The research in this thesis aims to contribute to an improved understanding of the role of geographic space in dialectology by harnessing the potential of methods of GIScience. The focus of this thesis is on establishing methodologies in two key research topics: the role of geographic distances in dialectal variation and the quantitative assessment of language-internal boundaries and transitions. All studies included in the thesis use data from the Syntactic Atlas of German-speaking Switzerland (SADS), a survey-based dialect atlas database focusing on (morpho)-syntactic phenomena. The SADS is characterised by multiple responses at each survey site, distinguishing it from most other dialect surveys.

The main contribution of this thesis lies in introducing new methodologies in dialectology and dialectometry based on methods that already proved their power in GIScience and spatial analysis. Correlation analysis at different spatial granularities between linguistic and geographic distance matrices provides further evidence that geographic distance is responsible for, and thus explains, the majority of the variance found in Swiss German syntax. Furthermore, travel time measures (for years 2000, 1950 and 1850) better reflect syntactic spatial variation than Euclidean distance, and older travel times account for contemporary syntactic spatial variation better than recent ones. Trend surface and regression analysis are used to quantitatively compare spatial variation of syntactic variables and to account for correspondence with theoretical models, concluding the importance of local analysis besides global. Moving beyond the modeling of boundaries, using a combination of linguistic filtering and spatial clustering a methodology is proposed for detecting language-internal transitions and boundaries in a probabilistic way, thus offering a novel solution that may complement conventional, visual methods of boundary delineation. While the case studies of this thesis demonstrate the usefulness of methods from spatial analysis and spatial statistics in addressing linguistic research problems, these same studies also show that linguistics has many interesting and challenging problems to offer for research in GIScience, with the potential of being extended beyond linguistics to other areas of the digital humanities.

Acknowledgements

"Explorers are we, intrepid and bold, Out in the wild, amongst wonders untold. Equipped with our wits, a map, and a snack, We're searching for fun and we're on the right track!"

Bill Watterson

This thesis has been written as part of an interdisciplinary collaborative project between the German Department and the Department of Geography of the University of Zurich. The project SynMod[1], short for *Modellierung morpho-syntaktischer Raumbildung im Schweizerdeutschen (or Modelling morphosyntactic area formation in Swiss German)*, was funded by the Swiss National Science Foundation (SNF) through grants nr. 140716 and nr. 162760. The project was jointly led by Elvira Glaser and Robert Weibel. SynMod has supported the work of Philipp Stöckle as a postdoctoral researcher in the German Department and my work as a PhD candidate in Geography, supervised by Robert Weibel. The SynMod project aimed to develop and apply a series of quantitative methods for the analysis of spatial variation in linguistic data, based on methodological knowledge from GIScience and the semantic knowledge of dialectology, a methodological challenge for both fields.

First of all I would like to thank my family for their ever-present support and love sent from home constantly. I would like to thank Rina for waiting for me patiently and constantlymotivating me from afar.

The biggest volume of thanks, however, should go to my supervisors and colleagues in the SynMod team, Robert Weibel, Elvira Glaser and Philipp Stöckle. I want to thank Röbi for the constant motivation provided, the excellent supervision, his continued support, outside-the-box thinking, for paving the road to success and always trusting me that I can walk it. I want to thank Elvira and Philipp for the help with the linguistic side of the project, the guidance with often confusing terms and concepts in linguistics. I would also like to thank Curdin Derungs and Peter Ranacher for their excellent insights and selfless help with one of the studies in this thesis.

It is very hard to draw boundaries among friends and colleagues at the Department of Geography. As I argue for fuzzy boundaries in this thesis too, I don't want to put you in categories either, only based on where you work(ed). So thanks to all GIS, GeoComputation, GIVA, RSL, 3G, H2K and Human Geography colleagues for having been such a cool and welcoming team where I could really experience collegiality. Thank you Alex, Ale, André, Anita, Annica, Annina, Arzu, Azim, Adrian, Barbara, Barbara, Bea, Beni, Beni, Carla, Christian, Christoph, Curdin, Damien, Dani, Daniel, Daniela, Devis, Diego, Elias, Elisabeth, Elise, Emiliano, Ewa, Fabian, Flurina, Gianluca, Gillian, Halldór, Haosheng, Hendrik, Hoda, Hossein, Huey Shy,

[1] http://www.spur.uzh.ch/en/research/associated/synmod2.html

Ilja, Irene, Ismini, Ivana, Izabela, James, Jing, Julian, Katharina, Katja, Kenan, Kiran, Leonie, Ling, Maitane, Manuela, Marc, Marta, Martin, Max, Maysam, Michele, Michelle, Olga, Oliver, Parviz, Peter, Pia, Qingyang, Raha, Ramya, Reik, Rémy, Rogier, Ronald, Ross, Röbi, Sascha, Sandra, Sanne, Sara, Sara, Shivangi, Simon, Sonja, Stas, Stefan, Tom, Ulrich, Victoria, Walied, Xiu, Yvonne, Zhaoju and everybody that I might have forgotten. I learnt a lot from you in science and in different walks of culture and life.

I'm grateful that I could spend all these years in such a fun and motivating environment, where I not once cursed Monday or uttered 'Thank God it's Friday'. Thank you for all the hikes, sports, skiing days, climbing, running, via ferrata, frisbie games, board game nights, cooking, parties, lindy-hop, swimming, grilling, or just hanging out doing nothing. *'It's not where you are, but who you're with that really matters.'*

Special thanks to the people whom I spent most time with: my office mates. Ali and George, you will always be my BOMFs (Best Office Mates Forever). It was time well wasted :) Also thanks to Julia, Hamouda, Thomas and Cyrill for the good company during the crunch time of my thesis.

I also spent considerable amount of time among the linguists of Zürich, and I find it a pity that I didn't learn more languages while doing so. Agnes, Andreas, Anja, Carlota, Gabi, Hanna, Max, Rik, Sandro, Tanja, Yuta and all others, thanks for the fun times, useful comments, the retreat and workshops together.

My philosophy of what goes around comes around have made a lot of friends for around and beyond the university environment, among them the scores of flatmates (Pierre, Laci, Diego, Ana, Luis and Hyunjin), especially at the dormitory 'Witellika' where I lived for the first 1.5 years. I'm grateful especially for those people who even kept in touch after I moved out or they moved away, such as Carine, Dani, Elena, Katherina, Marco, Tillmann, Rinita, Ronja, Sajjad, Veronika. Also outside the university, like Anita and Federica.

I also want to thank all my friends at home in Hungary for their support and for making me miss home enough to go home whenever I could. Thank you Székesfehérvár, Veszprém, Érd and Budapest.

I had the luck to visit many conferences and workshops during my time as a PhD student, particularly because of the interdisciplinary nature of my research. For some reason linguists love going to lots of conferences, and to find my target audience, I had to 'yield' and join up. It was time well used, as I met dozens of important/cool people that I want to thank for sculpting my views on science, linguistics and geography alike.

I hope to keep in touch with all of you and return to Switzerland as often as I can. But first, I have to try my luck in far away lands yet again, as I am going to Japan.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ALF	**A**tlas **l**inguistique de la **F**rance
DBSCAN	**D**ensity-**B**ased **S**patial **C**lustering of **A**pplications with **N**oise
KDE	**K**ernel **D**ensity **E**stimation
LAJDB	**L**inguistic **A**tlas of **J**apanese **D**ata**b**ase
LAMSAS	**L**inguistic **A**tlas of the **M**iddle and **S**outh **A**tlantic **S**tates
LANE	**L**inguistic **A**tlas of **N**ew **E**ngland
MC	**M**ultiple **C**hoice Question
MDS	**M**ulti**d**imensional **S**caling
MV*X*	**M**ain **V**ariant nr. *X*
OPTICS	**O**rdering **P**oints **t**o **I**dentify the **C**lustering **S**tructure
PCA	**P**rincipal **C**omponent **A**nalysis
SADS	**S**yntaktische **A**tlas der **d**eutschen **S**chweiz (Syntactic Atlas of German-speaking Switzerland)
SBS	**S**prachatlas von **B**ayerisch-**S**chwaben
SDS	**S**prachatlas der **d**eutschen **S**chweiz (Linguistic Atlas of Swiss German)
SAND	**S**yntactische **a**tlas van de **N**ederlandse **d**ialecten (Syntactic Atlas of the Dutch Dialect)
SyHD	**Sy**ntax **H**essischer **D**ialekte (Syntax of Hessian Dialects)
SynMod	SNF-Projekt Modellierung morphosyntaktischer Raumbildung im Schweizerdeutschen (Modelling Morphosyntactic Area Formation in Swiss German)

Cantons of Switzerland (in the German-speaking part)

AG	Canton of Aargau
AI	Canton of Appenzell Inner-Rhodes *(Appenzell Innerrhoden)*
AR	Canton of Appenzell Outer-Rhodes *(Appenzell Ausserrhoden)*
BE	Canton of Berne *(Bern)*
BL	Canton of Basle-Country *(Basel-Landschaft)*
BS	Canton of Basle-City *(Basel-Stadt)*
FR	Canton of Fribourg *(Freiburg)*
GL	Canton of Glarus
GR	Canton of Grisons *(Graubünden)*

LU	Canton of Lucerne *(Luzern)*
NW	Canton of Nidwalden
OW	Canton of Obwalden
SG	Canton of St. Gallen
SH	Canton of Schaffhausen
SZ	Canton of Schwyz
SO	Canton of Solothurn
TG	Canton of Thurgau
UR	Canton of Uri
VS	Canton of Valais *(Wallis)*
ZG	Canton of Zug
ZH	Canton of Zurich *(Zürich)*

Chapter 1

Introduction

1.1 Motivation

Issues of spatial variation in language have not attracted much attention in Geographic Information Science (GIScience) so far. Despite GIScience providing spatial theory as a framework for approaching space-related issues and its large number of potentially valuable methods, the involvement of GIScience in linguistic research has mostly been restricted to visualisation and interpolation (as noted by J. Lee and Kretzschmar, 1993:541 and Hoch and Hayes, 2010).

Since language and the way one speaks it plays such an important role in group formation (Weinreich, Labov, and Herzog, 1968; Labov, 2001; Preston and Robinson, 2005), power relations (Reid and Ng, 1999) and the feeling of belonging, in policy making (Valls, Wieling, and Nerbonne, 2013) and in extra-community phenomena, such as trade (Lameli, Nitsch, et al., 2015), studying the variation of language in space seems to be a worthwhile undertaking.

1.1.1 Dialectology and GIScience

Linguistic geography is an umbrella term combining strands of linguistics that research spatial variation in and among languages, such as linguistic phylogenetics, dialectology and dialectometry. In the context of this thesis, we use the term *dialect geography* as a collective name for dialectology and dialectometry, which are specific subfields of linguistics studying the spatial (and often socio-demographic) distribution of variation *within* languages. *Dialectology* focuses mainly on describing intradialectal variation, primarily examining the description of geographic distribution of said variation; the more quantitative *dialectometry*, on the other hand, deploys various techniques of computational and statistical analysis to identify representative and distinctive features with respect to areal classifications (Wieling and Nerbonne, 2015). The main issues of the subfields include accounting for the robustness and uncertainties of dialect areas, and the causes of dialectal variation and dialect change on different spatial and temporal scales.

Traditional dialect geography dates back to the second half of the 19th century, to Georg Wenker, a German linguist who collected a vast amount of written variations in German dialects, which were then mapped, such as by Wrede et al. (1927). Both

dialectology and geography investigate phenomena that are distributed in space. There have been several hypotheses in dialectology about the spatial distributions and dispersion of linguistic phenomena, but after its popularity in the first half of the 20th century it experienced a dormant period between the 1950s and 1980s (Kortmann, 2002). Since then, however, interest in investigating and explaining the spatial distribution of linguistic phenomena from a more quantitative point of view has risen (Wieling, 2012:4). This is visible not only in dialectology but also in other related fields of linguistics, such as typological research, which focuses on language families (e.g., Nichols, 1992; Bouckaert et al., 2012).

With the recent surge in dialectal data becoming digitally available, coupled with the rapidly increasing computational power and the production of sophisticated analytical tools, a rise in opportunities are visible for answering questions in linguistics as well. At the same time, however, linguistic diversity around the world is decreasing (Harmon and Loh, 2010) and the number of those speaking dialects in the classical sense are dwindling; this is often attributed to different forms of globalisation, such as migration, urbanisation, more liberal world views and decreasing 'mental distances' thanks to the media. Parallel to this, language standardisation policies enforce linguistic convergence more easily.

This race against time puts pressure on both linguistics and computational sciences to develop and implement methodologies towards describing and understanding the processes leading to dialectal variation.

The increasing amounts of linguistic data digitally available come in several formats. The digitalisation of surveys recorded earlier (e.g., LAJDB[1] - *Linguistic Atlas of Japanese Database*; **Kumagai2016**) have added historical importance. Those created digitally owing to new technological capacities (e.g., SyHD - *Syntax of Hessian Dialects*; Fleischer, Kasper, and Lenz, 2012) herald the need for computational sciences to collaborate, especially with regard to the remarkable change in data production thanks to crowdsourcing, which has been successfully implemented for collecting dialectal data using smartphone applications (Leemann, Kolly, Purves, et al., 2016) and webscraping social media (e.g., Eisenstein et al., 2014, 2017; **Nguyen2016**) as well. With the increasing volume of data and computational power available, dialect geography is entering an era where its questions can be answered faster and with greater statistical precision than ever before.

One overarching research topic in dialectology and dialectometry is the classification of observations (or, in a spatial regard, the locations of observations) into categories or, spatially speaking, area formation. One driving force behind linguistic classification and group and area formation appears to be the inherent human need for categorisation, which is broadly discussed in cognitive psychology (Rosch, 1973; E. E. Smith and Medin, 1981; Lakoff, 1987). Identities are generated by noting linguistic differences to distinguish 'our group' from 'the others'. This human psychological need for categorisation also manifests itself most of the time in the

[1] The abbreviations in this listes are listed on page xvi

way we relate to phenomena that essentially vary on continuous basis, such as language. In classical dialectology, this need for categorisation is represented more steadily, based on efforts to delineate certain dialect areas and drawing linguistic *boundaries*. At the level of individual linguistic phenomena, the view of categorisation in space was reinforced through classical dialectal surveys, published in atlases such as the *Sprachatlas des Deutschen Reichs* (1889-1923, posthumus introduction in Wenker, 2013), the *Atlas linguistique de la France (ALF)* (Gilliéron and Edmont, 1902-1910), and the *Linguistic Atlas of New England (LANE)* (Kurath et al., 1939), where only a sole response was recorded at each survey site, assumed to represent the local dialect.

As we have seen above, dialect geography's interests are spatial relations as well as spatial variation and change in language. Similarly to linguistics, GIScience also studies objects and variables in space, along with their representations. Spatial variation, as well as relations and change in spatial phenomena are questions addressed in GIScience, showing overlapping interests of linguistics and GIScience. It was observed already in the early days of dialectology that dialectal differences are sometimes stronger, while weaker in other cases. Experiencing this uncertainty in spatial variation is finally leading to a need of defining boundaries. The origin of this need is very similar in both fields and stems from the human desire for categorisation, in order to deal with the complexity of the phenomena we encounter. Both fields wish to analyse, and for that reason represent, these phenomena – in a GIS, too, in case of GIScience. Therefore they define categories and introduce definitions and delimitations. On the one hand language policies and international linguistic research yielded standardised forms of languages and written standards, such as the International Phonetic Alphabet(IPA) or the Swadesh-list (Swadesh, 1955). On the other hand, GIScience has laid down the ontological status of boundaries (Mark and Csillag, 1989; B. Smith and Varzi, 1997; Galton, 2003) as part of the spatial theoretical framework which could be beneficial in dialect geography as well.

It has often been advocated (e.g., Lameli, Kehrein, and Rabanus, 2010) that linguistics needs quantitative support from other disciplines to deliver more objective results, especially with data (potentially) interesting for linguistics being produced at an ever-increasing rate, parallel to an accelerated increase in algorithm development. Still, despite the number of potentially useful methods in GIScience, however, its application in dialect geography has been rather limited. Lee and Kretzschmar (1993) and Hoch and Hayes (2010) have argued that GIScience should place linguistic research onto its agenda. Such collaboration could also be beneficial for GIScience as an interdisciplinary field: it gains a new area where its conceptual framework and methods are tested and further developed in an environment that is laden with a range of uncertainties, and where the vast variety of linguistic data provides a lot of scope for experiments.

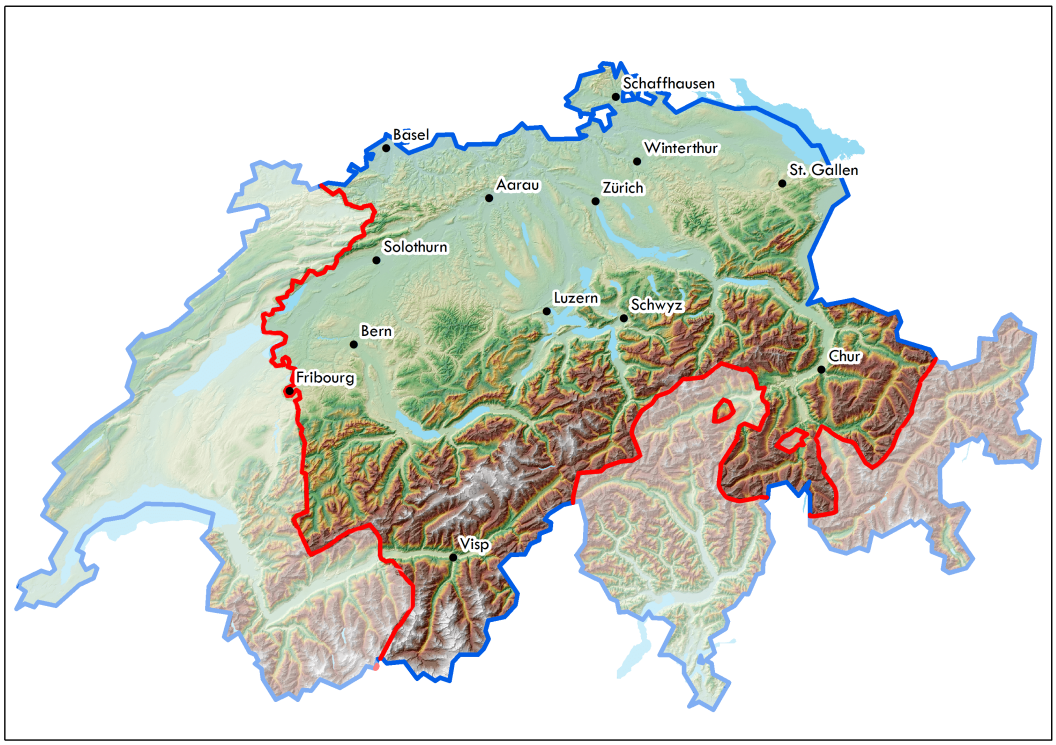

FIGURE 1.1: Topography of the study area of this thesis, the German-speaking municipalities of Switzerland.

1.1.2 Isolation and (dialect) contact, the main drivers of linguistic variation

There is empirical evidence that isolation and, conversely, contact between people speaking certain varieties of a language is crucial from the point of view of linguistic differentiation (Nerbonne and Heeringa, 2007; S. Lee and Hasegawa, 2014). Thus, intuitively, a connection between topography and linguistic diversity may be assumed due to the isolating effect of the former, and the corresponding ease or hardship of transportation (cf. Nichols, 2013). This connection has been discovered early (e.g., Hägerstrand, 1952) and became one of the most important explanations in traditional dialectology (e.g, Wang and Cavalli-Sforza, 1986; Schreier, 2002). However, many other potential factors, such as sociodemographic ones, play a role. It is ultimately the speakers that are in contact (e.g., Bowern, 2013), rather than languages or dialects, so it is crucial that we account for the potential that people can be in touch with each other.

Assessing the effects of geographical factors has been identified as one of the challenges where the spatial theoretical framework as well as the methods of GI-Science and spatial statistics might contribute to research in dialectology. It therefore forms the focus of this thesis. This thesis has concentrated on quantifying the effects of space on the variation in Swiss German syntax, expressed using the database from a survey intended for creating a dialectal atlas. The database used here stems from the Syntactic Atlas of German-speaking Switzerland (*'Syntaktische*

Atlas der deutschen Schweiz' – SADS; Bucheli and Glaser, 2002; Glaser and Bart, 2015). Figure 1.1 presents the study area covered by this survey. The database and the related surveys are presented in detail in Chapter 3. As is visible in Figure 1.1, the topography of Switzerland is very diverse with numerous mountain ranges present, featuring the Alps and the Jura, along with lakes and rivers. On the other hand, the flatter *Swiss Plateau* (German: *'Schweizer Mittelland'*) holds most of the population. These topographic differences and the isolation caused by them are considered to be some of the underlying reasons for the diversity of Swiss German dialects, including its syntax.

1.1.3 Research on Swiss German dialects

The label 'Swiss German' refers to a collection of Alemannic dialects, restricted by the national boundaries of Switzerland as well as the boundaries towards the other languages spoken in Switzerland (French, Italian and Romansh). In Switzerland, dialects enjoy a celebrated status (as opposed to most parts of Germany for example) and are ubiquitously used in everyday life situations, whereas Standard German is mostly confined to writing (Glaser and Bart, 2015). People's fond attachment to their way of speaking, and its power to create cultural identities (attested in e.g., Bartholy, 1992; Esser, 1983; Lameli, 2013) has been drawing interest to researching variation in Swiss German for a long time.

It is a common, naïve surmise that with the advance of increased mobility and forms of new communication (electronic media and modern social media) it is to be expected that dialects will disappear, and the language most used in media will prevail over all variations. Although there is also a trend in Swiss German (Christen, 1998) for such 'levelling', the dialects are still thriving and they remain an important part of the driving forces behind group formation in German-speaking Switzerland. Famous for its diversity and forms deemed peculiar when compared to Standard German, Swiss German varieties have been investigated for a very long time, placing them among the world's best researched dialects. Inspired by the work of Wenker on German dialects (well summarised in Schrambke, 2010), descriptive volumes on dialects (e.g., Stucki, 1917; Henzen, 1927; Hodler, 1969) were a common stream of research. On this basis, the *Language Atlas of German-speaking Switzerland* (*Sprachatlas der deutschen Schweiz* – SDS; Hotzenköcherle et al., 1962-1997) emerged, famed as one of the most comprehensive dialectal research projects of its time, also serving as a role model for the subsequent dialect atlases in the German-speaking areas.

Dialects are an evergreen small-talk topic in German-speaking Switzerland, and corresponding research engages laypeople easily. Thus, projects crowdsourcing dialectal data from lay users of (smartphone) applications are successfully being conducted in Switzerland, based on text and voice recordings (Stark, Ueberwasser, and Göhring, 2014 –; Leemann, Kolly, Grimm, et al., 2015). The *Syntactic Atlas of German-speaking Switzerland* (SADS – Bucheli and Glaser, 2002; Glaser and Bart, 2015), used

as a data source for this thesis, fits into this vibrant assortment of research. Syntax had seen little focus before the SADS in comprehensive dialectological projects, making SADS one of the first survey-based atlas projects to address this field in linguistics.

1.2 Thesis Rationale

Despite the number of potentially useful methods in GIScience, its involvement in dialect geography has been rather limited. As pointed out above, the volume of linguistic data is ever-growing and a number of linguistic hypotheses with a spatial bearing are waiting to be answered. Therefore the overarching motivation of this thesis is to involve GIScience and spatial analysis into developing methodologies to contribute to a deeper understanding of spatial variation in dialectal phenomena and the underlying effects of space on language area formation. The thesis attempts to address the impact of space in an implicit way, through the effects reflected in the resulting spatial variation of the dialectal phenomena in question. This approach is manifested in the **main research objective** of this thesis, concentrating on testing linguistic hypotheses using the methodologies proposed in the thesis:

Developing spatial analysis methodologies to quantify the spatial variation in dialectal phenomena and account for underlying geographic effects.

1.2.1 Research Objectives

In line with important research strands in dialect geography and the potential of the dialect database (SADS) used in this PhD project, three main research objectives have been defined, corresponding to three chapters in this thesis (Chapters 4, 5 and 6).

- ***Research Objective 1***. Develop and experiment with methods to explore the relationship between geographic distances and dialectal variation, in particular Swiss German syntax.
- ***Research Objective 2***. Develop and evaluate methods to quantify the degrees of transition in interdialectal spatial variation.
- ***Research Objective 3***. Automate the detection of interdialectal boundaries and transitions in a data-driven way.

The three chapters corresponding to the above research objectives all culminate in scientific journal papers. They are thus considered building blocks of the thesis.

In the thesis, the main emphasis is placed on *developing methodologies* for testing linguistic hypotheses on the spatial variation in linguistic phenomena. Therefore the

thesis does not aim to provide comprehensive linguistic interpretation on the dialectal phenomena involved. Using the methodologies developed, the thesis intends to stay at a more general, explorative level, in the sense of discovering the potentials of GIScience and exploiting the potentials of the database. Through this process, the proposed methodologies are designed so they can potentially be further developed for other strands of linguistics and beyond, and generalised to be applied in other fields of humanities and social science that concern diverse human-related and statistically referenced data. Potential implementation areas include census-based demographic data, cultural heritage data and digital humanities. Furthermore, the methods elaborated here could potentially be generalised for all kinds of data that have multiple instances pooled into specific geographically anchored areas.

1.3 Research Process and the Structure of the Thesis

This thesis investigates spatial variation in dialectal phenomena from two perspectives to achieve the research objectives. On the one hand, dialectal variation is investigated from the point of view of *multiple* dialectal phenomena. On the other hand, dialectology's classical perspective is considered, namely focusing on *individual* linguistic variables, but in a quantitative sense: the boundaries and transitions present between the prevalence areas of corresponding dialectal variants are quantified.[2]

Covariation of geographic and linguistic distances

Dialectometry often focuses its attention on aggregate linguistic variation rather than individual phenomena to better explain the overall variation and represent a more general linguistic difference between localities. The aggregation of different linguistic features is also assumed to reduce the inevitable noisiness of individual features (Nerbonne, 2009). The particular research in the thesis investigates the correlation of a *linguistic distance* defined with different *geographic distance* measures, which operationalise the possibility of language contact between the locations involved in the survey.

The chapter exploring this aspect, Chapter 4, is associated with **Research Objective 1**. Chapter 4 forms part of an article published: **Jeszenszky, Péter**, Philipp Stoeckle, Elvira Glaser, & Robert Weibel (2017). Exploring global and local patterns in the correlation of geographic distances and morphosyntactic variation in Swiss German. *Journal of Linguistic Geography*, 5, 1–23.

Modelling interdialectal boundaries

Dialectology has always been concerned with describing the spatial variation and patterns of spatial distribution in individual linguistic variables. This was, however, rarely accomplished in a quantitative way due to the lack of 'depth' in the available data. Areal studies have mostly concentrated on the internal consistency

[2]Section 2.1 explains terminology such as phenomenon, variables and variants

of dialect areas resulting from classification rather than focusing on their boundaries. Thus, these areas with an intriguing mixture of co-occurring variants were neglected, despite the fact that they pose interesting questions from the point of view of diachronic change in language too. Having data with multiple responses available at each location, however, as is the case in the SADS, allows us to model linguistic concepts and test hypotheses with regard to spatial variation, and give a quantitative and realistic account of them.

Once the appropriate data is available, the boundaries and areas of mixture and transition between two dialectal variants can be perceived as gradients. Based on this gradient concept we investigate the possibilities of modelling and quantifying the patterns of transition between the prevalence areas of variants, for a better understanding of spatial and temporal processes in dialect variation.

The problem of modelling interdialectal boundaries is studied using two different approaches. First, Chapter 5 takes a *hypothesis-driven* approach, focusing on the validation of linguistic concepts (*'isoglosses'* and *'dialect continua'*, to be described in detail in Chapter 2) that address interdialectal boundaries, creating prototype models and testing their correspondence to survey data by means of *trend surface analysis* and *univariate regression analysis*. The chapter exploring this aspect, Chapter 5, is associated with **Research Objective 2**. Chapter 5 forms part of an article under revision: **Jeszenszky, Péter**, Philipp Stoeckle, Elvira Glaser & Robert Weibel (in revision), 'Sharp boundaries vs. gradual transitions: Quantitative models for transitions between areas of dialectal variants', *Journal of Linguistic Geography*, 6(2) special issue.

Another focus is the automated assessment of interdialectal boundaries and transitions, taking space explicitly into account. This is discussed in Chapter 6 with a *data-driven* approach. Instead of fitting models to the variation represented in dialectal data, in this chapter we use *linguistic filtering* and *density-based spatial clustering* to account for boundaries and transitions in a probabilistic manner. The chapter exploring this aspect, Chapter 6, is associated with **Research Objective 3**.

The thesis is structured around the aforementioned three core Chapters 4, 5 and 6, focusing on the three main Research Objectives, respectively.

Leading up to these chapters, Chapter 2 first elaborates on the setting of the research and presents a general literature review. The main role of this chapter is firstly to present general background knowledge relevant for geographers to comprehend some key concepts of linguistics and dialectology, and secondly, to define the main research gaps addressed in the thesis. Importantly, however, specific related work and literature immediately relevant to the subprojects are included in each chapter (Chapters 4, 5 and 6), rather than in the general background chapter. The database used to conduct this research is introduced in Chapter 3. After this, the research specific to the above-stated Research Objectives are presented in Chapters 4, 5 and 6.

Chapters 4, 5 and 6 are seen as separate entities that are functional on their own, thus each of them includes a detailed discussion relevant to the corresponding Research Objective in focus. So there is no overarching discussion chapter included in the thesis. The main contributions, insights and limitations are summarised in a separate, concluding chapter (Chapter 7).

Chapter 2

Background

2.1 Terminology

This first section introduces some key terms and concepts of linguistics that are crucial for the understanding of this thesis. The concepts and terms associated with the methodologies of spatial analysis developed in this thesis will be introduced and explained in the corresponding Chapters 4 to 6.

Dialect

Although the concept of a *'dialect'* is probably known informally to most people, it has to be defined for the sake of a common understanding in this thesis. Here, our notion of dialect is based on Chambers and Trudgill's fundamental book (2004) on dialectology. Every person is a speaker of at least one dialect, as the standard language is just as much a dialect as any other form of language. Dialects can be regarded as subdivisions of a particular language, fulfilling the requirement of *mutual intelligibility*. If two speakers understand each other without having to actively learn each other's language, then their languages are mutually intelligible and thus may be considered dialects of the same language. Note that in this thesis the local variety of language at a certain location will not be called 'the Village B dialect'.

Linguistic levels

Linguistics studies the grammatical fields of languages from several aspects, including language production, meaning, formulation and context of usage. Linguistic research can be subdivided into *linguistic levels*, investigating units of language that establish a functional hierarchy (O'Grady et al., 2016). Starting from the lowest level, *phonetics* is concerned with the physical properties of speech sounds; *phonology* studies the system of sounds and their grammatical roles in a language; *lexicology* studies the function of words, their symbology and meaning; and *morphology* studies the structure and composition of words from morphemes (the smallest meaningful units of a language), such as declination, and their relationship to other words. The next higher linguistic level is *syntax*, which forms the focus of this thesis and which will be explained in more detail in Section 2.2.

Survey

In dialectology and dialectometry the data usually stems from some kind of dialectal survey, often intended for a linguistic atlas which consists of a collection of maps representing the spatial variation in selected dialectal variables. Examples of such atlas projects include the *Sprachatlas der deutschen Schweiz* (SDS, Hotzenköcherle et al., 1962-1997) or the *Linguistic Atlas of the Middle and South Atlantic States* (LAMSAS[1], Kretzschmar, McDavid, et al., 1993). Dialectal surveys may be conducted by means of personal or telephone interviews, online, or through questionnaires sent to people that act as sources of information. These 'informants' are referred to as *respondents* in this thesis. Lately, the field of dialectal data collection has seen the addition of new methodologies. Websites and designated smartphone applications are being developed for collecting dialectal data (e.g., Leemann, Kolly, Grimm, et al., 2015, Leemann, Kolly, and Britain, 2018), which have the potential to yield large data volumes quickly.

Variables and variants

In this thesis, a linguistic or dialectal *'variable'* denotes an attribute of the language which may vary across different language varieties, such as in the language usage of each speaker (cf. Spruit, 2006:494; Glaser, Stoeckle, and Bachmann, accepted). The different ways of expressing a variable are in turn referred to as *'variants'*. In this sense, a linguistic innovation will yield a *new, additional variant* of the linguistic variable affected by the innovation. In this thesis each variable corresponds to one aspect of a survey question (Section 3.1). However, a survey question may cover more than one linguistic variable. In Example 2.1, the same question aims to extract variants for the following different variables: the 'verb particle doubling' in *anfangen* (English: *'to begin'*) and the occurrence of the *infinitive particle 'zu'*. In this example three variants corresponding to the variable 'verb particle doubling' are shown: a variant using the particle *'an'*, another variant with the doubling of the verb *'begin'*, i.e., *'afa'*, and one where the particle and the verb itself do not occur (detailed explanation in Stoeckle, 2016a; 2018).

"Wenn es so warm bleibt, ***fängt*** *das Eis* ***an*** *zu schmelzen!"*

a. *..., **fängt** das Eis **an (zu)** schmelzen.*

b. *..., **fängt** das Eis **afa** schmelzen.*

c. *..., schmilzt das Eis.* (2.1)

"If it stays this warm, the ice will begin to melt."

Phenomenon

In Example 2.1 the survey question extracts information about two different phenomena: the doubling of a verb particle and the occurrence of the *infinitive particle*

[1] http://us.english.uga.edu/cgi-bin/lapsite.fcgi/lamsas/

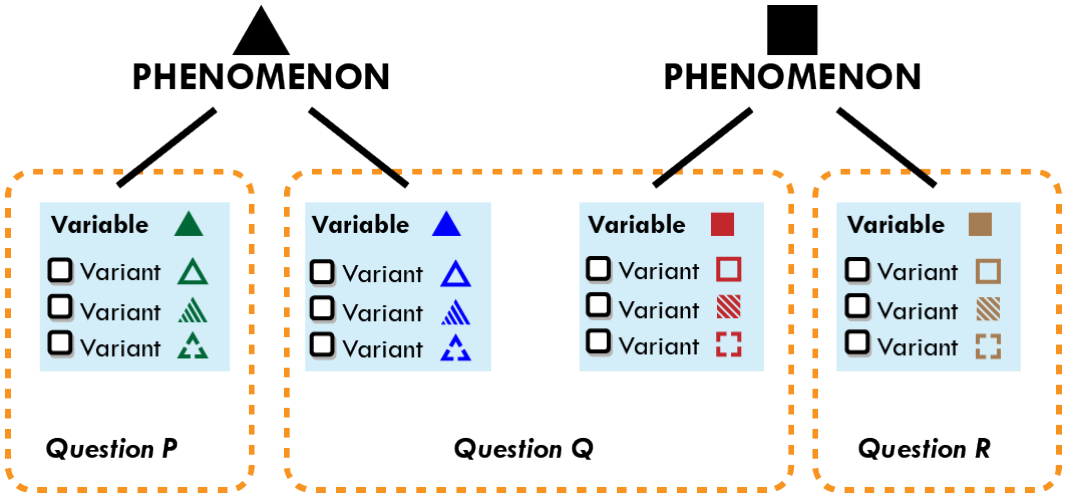

FIGURE 2.1: The hierarchy of terms related to linguistic variables and phenomena.

'zu'. In surveys, a linguistic phenomenon can be investigated using several variables, which means that several survey questions can be directed at extracting information about the above phenomena and that several variables are semantically related as they investigate the same phenomenon. Figure 2.1 illustrates the hierarchy of terms related to linguistic variables and phenomena.

Isogloss

Literally meaning *'equal language'*, as a combination of the Ancient Greek words ισος (*ísos*), meaning 'equal', and γλοσσα (*glossa*), meaning 'tongue' or 'language', the term was first used by Bielenstein (1892) in present Latvia, based on the meteorological term *isotherm*, which denotes a line drawn between locations of same temperature.

The meaning of the isogloss concept can be roughly understood as a boundary line in the geographic landscape, where the occurrence areas of the *variants* corresponding to a linguistic *variable* are expected to be separated. From a gradient point of view these boundaries can be viewed as *'sharp transitions'* between the usage zones of the variants in question. In classical German dialectology, point-symbol maps of variables derived from survey data were often used as a basis for drawing isoglosses, to delineate *'areas'* of different dialectal variants.

A more thorough description of the term is given in Section 2.3.2, as the investigation of this concept plays a crucial role in this thesis.

Prevalence zone and dominance zone

Prevalence zone denotes the usage area of a dialectal variant, where the variant in question occurs often. Thus, the term *prevalence zone* in this thesis is used in a broader sense than *dominance zone*. A *dominance zone* is the area where the variant in question occurs *predominantly*, that is, where it unequivocally builds the majority.

Transition zone

Chiefly used to denominate the areas that exist *between* the *prevalence zones* of different *variants*, *transition zones* are areas of mixture where two, or more, dialectal variants are used in coexistence, and it is often unclear which of these variants are dominant.

Empirically, usage areas of certain variants gradually transition into each other. Where this transition takes place, a wider or narrower *transition zone* is found, marked by the mixing of the co-occurring variants. The extents of a transition zone, similarly to dominance zones, is often fuzzy.

Note that the term transition zone is often used in dialectology to denote areas also at the level of entire dialects, such as a transition zone between Frisian and Low-Saxon dialects (Dros-Hendriks, 2018).

Boundary

The term boundary usually denotes a linear feature delineating areas of two qualities. The expression is used in different ways in the various fields of geography and linguistics. In this thesis *boundary* denotes linear, demarcating features that can be *sharp* or *gradual*. Based on Parker (2006:79), this vernacular word fits well the usage of boundary as an "unspecific divide or separator that indicates limits of various kinds."

Sharp boundaries of a topographic or administrative type will be denoted as *borders* in this thesis, in contrast to the more elusive concept of boundary. Examples include rivers and cantonal borders of Switzerland.

In this thesis, the expression *'interdialectal boundary'* often occurs, denoting either boundaries between 'dialect areas' or, at the level of individual variables, boundaries between prevalence areas of variants. This latter, as discussed later on, comprises the sharp linguistic transitions termed *'isoglosses'* in dialectology.

The ontology of boundaries has been extensively researched in GIScience (Mark and Csillag, 1989; B. Smith and Varzi, 1997; Galton, 2003). The most important implications to this thesis are explained in Section 2.5. In GIScience, the issue sometimes arises as to which of the two concepts – the unary 'boundary of' or the binary 'boundary between' – should have conceptual or logical priority (Galton, 2003). In this thesis, we are interested in modelling the boundaries *between* areas characterised by a certain variant rather than the boundary of a variant (i.e., the extent of the variant's dispersion). Although the spatial distributions of the variants in question often do not cover the whole extent of the study area, our interest is in the spatial transition *between* the usage area of two variants. We have to know what boundaries we want to represent. Even though we might say that the transition (i.e., also transition zone) starts where the homogeneity of one variant ceases, dialectology is interested in finding and describing the ambiguous overlap, the interface of two variant's areas.

2.2 The Linguistic Level of Syntax

The linguistic level of *syntax* focuses on the grammatical structures at the level of sentences. The term is an Ancient Greek compound comprising $\sigma\upsilon\nu$ (*syn*), meaning 'together', and $\tau\alpha\xi\iota\varsigma$ (*táxis*), meaning 'ordering'. Typical examples for syntactic phenomena involve word order and congruence in a sentence or the usage of verb forms in a certain context. Syntax is also commonly discussed together with *morphology*, the study of word forming and relationship to other words in a sentence. *Morphosyntax* (as the combination of the two is often called) is assumed to change at a different rate through language contact than other linguistic levels. Longobardi and Guardiano (2009) argue against common beliefs that it should change more quickly and come to the conclusion that "the diachronic persistence of syntactic properties is sufficiently robust to allow for plausible genealogical reconstructions" (Longobardi and Guardiano, 2009:1695).

For a long time syntactic variation across dialects was assumed to be explained by the principles of oral language production and thus was assumed not to differ from the syntax of the standard language (Löffler, 2003:109, 116) and syntactic variation was claimed to be reducible to other linguistic and non-linguistic levels (Barbiers, 2013). Because of this, studying the geographical structure of syntax was long neglected. Nonetheless, most researchers in the field of dialectology have come to agree that also in syntactic variables structured spatial distribution can be found (e.g., Glaser, 2013; Szmrecsanyi, 2014).

The rising interest in spatial variation of syntax spawned numerous projects, among which some European ones have been organised into the EDISYN[2] (European Dialect Syntax) network. The exciting data sources contributed include the aforementioned SyHD and the *'Syntactische Atlas Van De Nederlandse Dialecten'* (SAND - 'Syntactic Atlas of the Dutch Dialects' - Barbiers et al., 2005). Spruit (2008:24) gives a longer list of recent, successful syntax projects.

It was shown in several studies (Spruit, Heeringa, and Nerbonne, 2009; Uiboaed et al., 2013; Kellerhals, 2014; Scherrer and Stoeckle, 2016) that the spatial distribution patterns in syntax are often different from those seen in lexis, phonology and morphology, and that geographic distances play a different role in the variation at this linguistic level. Based on corpus studies rather than atlas data, where phenomena recorded are actually expected to show spatial variation, Szmrecsanyi (2012:227) finds that "[...] morphosyntax is less amenable to geographic diffusion than e.g., pronunciational variability."

The database used in this thesis, the *'Syntactic Atlas of German-speaking Switzerland'* (SADS - Bucheli and Glaser, 2002; Glaser and Bart, 2015), was also inspired by the working hypothesis that syntactic variation was actually spatially structured. A detailed description of this database follows in Chapter 3.

[2] http://www.dialectsyntax.org/

2.3 Dialectology and Dialectometry

It is natural for languages and dialects to change over time and space. However, the role of different factors and mechanisms responsible for the spatial distribution of linguistic and dialectal phenomena is not yet fully understood (e.g., Lucas, 2015).

The overarching motivation for dialectologic research, phrased by Trudgill (1974:216-217) is still valid: "[...] dialectologists should not be content simply to describe the geographical distribution of linguistic features. They should also be concerned to explain – or perhaps, more accurately, to adduce reasons for – this distribution. Only in this way will we be able to arrive at an understanding of the sociolinguistic mechanisms that lie behind the geographical distribution of linguistic phenomena, the location of isoglosses, and the diffusion of linguistic innovations."

The goal of *dialect geography* (encompassing both dialectology and dialectometry) is to better understand these mechanisms (sociolinguistic, geographic etc.) at the level of dialectal variation (Glaser, 2013). Traditional *dialectology* usually studies the distribution of individual linguistic phenomena, one feature at a time. To characterise the multidimensional nature of dialects, however, linguistic variables need to be investigated at an aggregate level. It is the continuum of variation at the level of dialect areas as well at the level of individual variables which *dialectometry* was meant to explore (Pickl and Rumpf, 2012), with the aim of identifying "general, seemingly hidden structures from a larger amount of features" (Goebl and Schiltz, 1997:13).

The aim of this chapter is not to give an exhaustive overview of dialectology and dialectometry, but to merely identify the research gaps related to the Research Questions given in Section 1.2.1. These research gaps are summarised in Section 2.6 and discussed in more detail in Chapters 4, 5 and 6. The closely relevant background and literature reviews are incorporated within each chapter.

More detailed and comprehensive overview on dialectology is presented by Chambers and Trudgill (2004), and Niebaum and Macha (2006), and on dialectometry by Goebl (2006), Nerbonne (2009) and Wieling and Nerbonne (2015).

2.3.1 Issues in dialect geography

The central issues of dialect geography – description and explanation of dialect area formation and diffusion of variation – according to Glaser (2013) closely resemble the core issue in linguistic typology as stated by Bickel (2007:239): *"What's where why?"*

Szmrecsanyi (2014) notes that the most important issues of concern for contemporary dialectology include the emerging interest in dialect grammar and morphosyntax, and the recent dialectal speech and usage (as opposed to passive dialectal knowledge and investigation of the 'original' local variety). Additionally, there is a focus on aggregational analysis techniques (which roughly correspond to those in dialectometry) and perceptual approaches.

Glaser's (2013) article sums up the questions of concern in dialectology as follows, in many regards setting the theoretical scene for this thesis.

- What is the distribution of the variants, and how can it be described? Are there specific geographical patterns?
- Are we dealing with a continuum, or do we find clear-cut boundaries?
- How can we explain the distribution of the variants, and the location of the isoglosses?

Most quantitative studies on spatial variation in linguistics focus on the internal homogeneity of their groupings and do not explicitly assess the boundaries between them. As Haas points out, "the linguistic coherence of a region is more important than its boundary" (2010:664). However, this thesis argues that since dialect areas are traditionally identified based on the dispersions shared between different individual linguistic variables, spatial characterisation and comparison of the boundaries and transitions of these variables becomes essential.

2.3.2 Quantifying interdialectal boundaries

Dialect area and dialect continuum

Dialect geography addresses the spatial relationship between dialect variants in two different ways.

At the aggregate level, this relationship is approached by the concepts of *dialect area* and *dialect continuum*, one advocating the presence of clear-cut boundaries, and the other assuming gradual transitions, respectively.

Starting out from the inherent, psychologically based (Section 1.1.1) need for categorising and classifying continuous phenomena into areas, traditional dialectology aimed to find delineations of dialect areas, based on perception and by collecting local dialectal data. The boundaries of these areas were sometimes suggested to be sharp, which was partly supported by results of dialectal surveys, which focused on spatial distribution of linguistic phenomena (e.g., Haag, 1898; Wrede, Mitzka, and Martin, 1927; Maurer, 1942).

It has been shown in numerous studies (cf. Heeringa and Nerbonne, 2001; Pickl and Rumpf, 2012), however, that dialect areas rarely have clear-cut boundaries and are mostly characterised by a transition of a certain graduality towards neighbouring varieties.

In the same way that dialectal variation manifests itself in different ways at the aggregate level, the clear-cut and the gradual view of dialectal variation are also present at the level of individual linguistic variables. On the one hand, the dialect area concept becomes apparent as sharp boundaries between *prevalence areas* of variants. These boundaries are the *isoglosses*, where the occurrence areas of variants are

expected to be separated. On the other hand, the dialect continuum concept manifests itself as *gradual transitions* between the *dominance areas* of variants.

In general, the linguistic community started to view dialects increasingly as a spatial continuum of gradually changing language, with some parts showing only very subtle differences between locations, and some parts exhibiting sharper transitions (Pickl and Rumpf, 2012). There is an agreement in variationist linguistics today that spatial distributions are usually not as crisply delimited as isoglosses drawn on maps would suggest; inherent to the transitions in single phenomena, continua are present. Chambers and Trudgill (2004) claim that the reality of linguistic variability is literally hidden by isoglosses, thus calling for a revision of the two concepts. Still, dialectology continues to use both concepts of dialect boundary in parallel, without resolving their incompatibility (Chambers and Trudgill, 2004:105). It has to be noted that Chambers and Trudgill came to this observation based on phonological studies as spatial variation of syntax has been under-researched until recently; the spatial distribution in syntactic variables and its differences therein from other linguistic levels are still not fully understood.

With the recent advent of richer data at an appropriate spatial granularity, however, models can be more finely tuned and it is possible to test methodologies that overcome the opposition of concepts. The issue of spatial transition at the level of dialect areas and, more specifically, at the level of individual dialectal variables has been handled in numerous ways. However, to date, no quantitative method has been proposed that focuses on boundaries, taking space explicitly into account.

Isoglosses and sharp boundaries

After their conception at the end of the 19th century (Section 2.1), *isoglosses* have become a central element of interest in dialectology. Isoglosses were traditionally drawn on point-symbol maps derived from dialectal surveys to determine spatial variation in the usage of individual variables. Coinciding isoglosses, so-called *'isogloss bundles'*, were used to quantitatively account for dialect areas at different scales (e.g., Haag, 1898; Maurer, 1942; Kurath, 1972) by highlighting boundaries between (mostly) homogenous dialect areas (cf. Bloomfield, 1933; Händler and Wiegand, 1982). Early on, however, researchers of dialect geography pointed out that isogloss bundles do not fulfil all expectations as a means of delimiting dialect areas (Bloomfield, 1933; Freudenberg, 1966). It has also been noted that different variables tend to show different patterns of regional variation (Chambers and Trudgill, 2004). As seen in earlier in this section, the issue of delimitations in the dialect landscape has not ceased to be a key question in dialectology, and isoglosses are still a common feature in classical dialectological researches for representing interdialectal changes in space. Thus, for linguistics the quality and effect of these boundaries, their gradual or sharp nature is relevant (cf. Wattel and Reenen, 2010).

Isoglosses are usually represented as sharp lines on maps and thus, importantly, they leave room for misinterpretations about the possible graduality of the transition. Interestingly, isoglosses indicate a major focus on class-boundaries, while in general through classification most attention is paid to class-affiliations, with boundaries themselves often being an implicit side product, as is also the case in most dialectal research (e.g., Daan and Blok, 1969; Heeringa and Nerbonne, 2001; Heeringa, 2004; Rumpf et al., 2009). In practice, drawing isoglosses usually entails ignoring sites that would, corresponding to the opposing variant, render the areas to be delimited less homogeneous, because they are located on the 'wrong side' of the theoretical line. Thus, technically speaking, these sites are considered outliers. It is understood that such smoothing is in the interest of meaningful interpretation, needed to identify underlying regional linguistic patterns, and is employed by almost every research in dialectology (Grieve, 2014).

The intuitive use of smoothing is also partly rooted in its visual advantage. Even if gradual transition is present in the given dialectal data, correctly representing it on a map using linear features is difficult, as the transition may exhibit different patterns between different dominance zones. Thus, drawing a line at the estimated middle of the transition, or two lines where the mixing of variants appears to cease (Chambers and Trudgill, 2004:90; Glaser, 2013:214) is still considered the best interpretable approach for visualisation, when lacking data with higher local resolution. Consequently, Labov, Ash and Boberg (2006) express the need in dialect geography for an automatic method for drawing such boundaries, without the preconceived notions of an analyst.

Francis (1983:5) states that such boundaries do "[...] not mark a sharp switch from one word to the other, but the centre of a transitional area where one comes to be somewhat favoured over the other." The vague definitions of boundaries in areal linguistics thus point toward a research gap worth investigating from the perspective of GIScience and implementing methods of the spatial sciences.

Dialect continua and gradual transitions

Alternatives to categorical, clear-cut boundaries were proposed early on (e.g., Bloomfield, 1933), with the suggestion that dialect areas are organised in a continuum without sharp boundaries. Accounting for the spatial relationship of dialect areas in a quantitative manner, however, was only initiated after the 'quantitative turn' seen in the 1960s in social sciences (Pickl, 2013). Séguy's work (Séguy, 1971) is, thus, regarded as the start of dialectometry as a subdiscipline. Séguy's approach, establishing a measure of linguistic distance between locations based on multiple linguistic features, marked the change of concept in dialectal areality. The dialect landscape was thus no longer seen as a mosaic put together from distinct dialect areas, but as a spatial continuum of gradually changing language, with some parts showing only

very subtle differences between locations, and other parts exhibiting sharper transitions (Pickl and Rumpf, 2012:202). Since then, researchers have strived to quantitatively account for the similarity of dialects by grouping language locations along multiple dimensions (e.g., Goebl, 1982; Kessler, 1995; Heeringa, 2004).

As mentioned above, the common recognition in modern dialectology is that geographic distributions may involve continuous transitions. The notion of gradual transitions has also been discussed in the more qualitative paradigms of dialectology. Importantly, Seiler (2005), using examples from Swiss German dialectal survey data (SADS), proposed a theoretical model for a particular gradual transition, called *'inclined plane'*, instead of the classical isogloss. Investigating the phenomenon of the *infinitival purposive clause* in Swiss German, he noted that different variables tend to follow different patterns of regional variation, even when the variables correspond to the same linguistic phenomenon. Based on his observations, Glaser (2013:214) calls for "systematically comparing the areal distribution of the variants and the (non-)correspondence of the respective geographic areas in a comprehensive manner." She also suggests that "dialectometrical analysis techniques may help us to spot similar patterns, while the analysis and interpretation of commonalities is reserved for qualitative dialectological analysis." The work by Seiler (2005) also gave rise to the study of transitions in Chapter 5.

Of course, the notion of gradual boundaries is not only a linguistic issue. Mark and Csillag (1989) propose that boundary lines situated between categories or classes which occur over contiguous regions of geographic space are far more similar (mathematically and geographically) to elevation contours than they are to linear features such as coastlines and rivers. Parker (2006) offers a conceptual framework – the *'continuum of boundary dynamics'* – for the 'fluidity' of boundaries in space, which could be used in numerous disciplines. Based on the approaches in his article (Parker, 2006:82), boundaries can be described and placed along a spectrum of increasing graduality. In spatial analysis, Leung (1987) quantifies regions in between dominance zones ('cores') with gradual transitions, with the decrease of an attribute of Region A and the increase of another attribute of Region B. His example of climatic zones can be set in parallel with dialectal areal patterns structuring the dialect continuum.

Numerous studies have been conducted with the aim of quantitatively accounting for variant areas or, in an aggregate manner, for dialect areas. Research groups in Augsburg and Ulm, Germany (Rumpf et al., 2009, 2010; Pickl and Rumpf, 2012) have quantitatively analysed spatial structures of variant areas in the SBS (*'Sprachatlas von Bayerisch-Schwaben'*, König (ed.), 1996-2009), a lexical dialect atlas that relies on multiple respondents (normally 1-2, up to 6) per survey site. They used spatial interpolation methods (termed 'dialectometric intensity estimation') to detect prevalent structures, devised measures to estimate the 'homogeneity', 'complexity' and 'compactness' of dominance areas and accounted for the similarity across maps of different variables (Rumpf et al., 2010).

Concurrently, Grieve et al. (2011) used a large corpus of American English lexical data. As summarised by Wieling and Nerbonne (2015), they aimed to obtain a quantitative counterpart to the traditional analysis of regional linguistic variation, which consists of identifying isoglosses, bundles of isoglosses, and finally dialect regions. They identified patterns of regional linguistic variation (through local variation statistics) taking into account the variant used in surrounding locations and therefore smoothing the geographical pattern. The second step taken by Grieve et al. was to apply factor analysis to identify groups of linguistic variables showing a similar regional pattern (i.e., determining the linguistic basis). The third and final step was to apply hierarchical cluster analysis to the factor scores per location to determine the resulting geographical clusters. Nerbonne et al. (1999) used multidimensional scaling (MDS), a dimension reduction technique, to reduce large dialectal matrices to a three-dimensional space, associated with the three components of the RGB colour space, to show the graduality of the transitions between dialect areas. Since then, MDS has become a common tool for dialectometric visualisations (Heeringa, 2004; Nerbonne, 2010a; Kellerhals, 2014), showing differences between survey sites with regards to a multitude of phenomena.

As mentioned in Section 2.3.1, most research has so far concentrated on the internal homogeneity of the areal features found. In contrast, interdialectal transitions themselves, however, have rarely been investigated quantitatively or placed along such a scale of graduality. Informally though, transition zones are often described and used in dialectology (Scherrer, 2012; Pickl, 2013; Scholz et al., 2016), but the concept itself lacks a clear definition. Using quantitative methods, however, it should be possible to propose models to define dominance and transition zones. Further, this thesis deems it feasible to quantify the gradual transition present between dominance areas of variants directly from the original data (i.e., without smoothing), in order to overcome the conflict between the concepts of the isogloss and the dialect continuum, respectively (Chapters 5 and 6).

The fact that spatial patterns in individual variables usually do not show sharp boundaries becomes even more visible when the spatial and attribute granularity of a survey is dense enough. Dialectal surveys relying on multiple informants per survey site, such as the ALT (Lexical Atlas of Tuscany – Giacomelli et al., 2000), the SyHD (Fleischer, Kasper, and Lenz, 2012), the SBS (König (ed.), 1996-2009) and more recently, popular online and smartphone-based dialect surveys providing massive data with geolocations (e.g., **Leemann2016a**), facilitate the discovery of local dialectal variation. Because of the potential of having multiple variants present at each survey location, such as in the database used in this thesis (SADS - Chapter 3), the spatial distribution of variants can be assessed with an increased spatial granularity and precision.

2.3.3 The effect of geographic factors in dialect geography

As stated in Section 1.1.2, isolation and, conversely, contact between people speaking certain varieties of a language, is crucial from the point of view of linguistic differentiation (Schreier, 2009; S. Lee and Hasegawa, 2014). It is, ultimately, the speakers themselves that are in contact (Bowern, 2013). Logically, it follows that in order to explain linguistic differences between regions or survey sites, it is necessary to account for people's potential to meet. This potential is, in turn, highly dependent on the geographic (especially topographic) settings in the study area. Holman et al. (2007) explicitly associate the concept of 'isolation by distance' among the world's languages, which is inspired by population genetics, with the situation faced by dialectology. The axiomatic role of geography structuring language (Nerbonne and Kleiweg, 2007:154) has also been pointed out in Section 1.1.2 and has been tested in numerous studies.

In this thesis the issue of contact potential is especially interesting, due to the varied topography present in the study area of the German-speaking part of Switzerland.

There are many potential factors which might affect how languages vary, including, for instance, settlement size, social class, sex, and educational level (Spruit, 2008). Of these, however, in large-scale, quantitative studies, geographical factors have shown a predominant influence (Nerbonne and Heeringa, 2007).

Trudgill (1974) confirms that it is not unreasonable to consider social barriers as a part of the conceptual model either. This thesis, however, does not focus on influencing features that lack spatial variation, nor on the sociodemographic factors of language contact, which are undoubtedly included in the scope of the language contact paradigm, yielding phenomena such as urban sociolects, internet-slang or standardisation spreading through media (Christen, 1998; Goel et al., 2016; Dürscheid, in press). These would all, doubtless, pose interesting challenges for GIScience, but are beyond the scope of this thesis.

Language diffusion models

Different theoretical models have been developed to examine the diffusion of linguistic innovations, partly within the diachronic linguistics paradigm, as change can only truly be captured in the temporal dimension. The 'wave model', sometimes called 'contagion diffusion' (**Bailey1994** Britain, 2002, 2010) assumes that innovations spread like ripples on a pond when a pebble is cast into it, radiating over time from a central area. 'Urban hierarchical' or 'cascade diffusion' (Trudgill, 1974; **Bailey1994** Chambers and Trudgill, 2004; Britain, 2002, 2010) – also known as Trudgill's *linguistic gravity* – assumes that innovations spread from larger populations towards smaller ones: first between cities, then to smaller towns and finally to the countryside, corresponding to the mobility patterns of the population. 'Cultural hearth diffusion' is described by Horvath and Horvath (2001, 2002), as being where

innovations get accepted in different sizes of population within the same region before moving on to other parts of the country.

The type of diffusion found appears to be heavily context-dependent, however, and there are examples of the same variant diffusing differently in different places (Britain, 2010). Naturally, innovations do not only originate from cities. Glaser (2013:198), however, claims that "[...] diffusion models suggest that we should be dealing with a more or less homogeneous distribution of the new feature within a larger area." In the spatial dimension of linguistic innovation diffusion, the effects of historical isolation in dialectal variation are assumed to be slowly overwritten by the effects of increased mobility and better transport, and overwhelmed by mass media often spreading a standard(ised) language variation.

Functional models in the temporal domain, however, seem to uniformly adhere to the S-curve model (Yokoyama and Sanada, 2009; Blythe and Croft, 2012), independent of the spatial diffusion model. The S-curve model assumes an innovation spreading similarly to the shape of a logistic function, between the proportion of usage and age (not unlike the theory about adaption of innovations in economics; Rogers, 1995). Of course the diffusion of language features, too, can be better investigated using data with better spatial (and temporal) granularity, especially with *'apparent temporal depth'* (Cukor-Avila and Bailey, 2013; Stoeckle, 2016a), that is, incorporating respondents of different ages.

Three observations can be made regarding language contact and diffusion research. First, better accounting for the contact patterns between language varieties (i.e., speakers or survey sites) will help explain dialectal differences. Second, if spatial variation is quantitatively assessed at a better spatial and attribute granularity, especially in individual phenomena, the models of diffusion can also be validated. And third, based on better modelled interdialectal boundaries we can infer future scenarios of the dialectal landscape with regards to the variables in question.

Influence of geographical factors

As geographic factors have the potential to crucially influence language contact and thus impact on linguistic variation (e.g., Wang and Cavalli-Sforza, 1986), there has been a stream of research in modern dialectology and dialectometry investigating the relationship of geographic distances with linguistic variation (e.g., Séguy, 1971; Goebl, 1982; Heeringa and Nerbonne, 2001; Gooskens, 2004; Heeringa, 2004; Nerbonne and Kleiweg, 2007; Shackleton, 2007; Szmrecsanyi, 2012; Lameli, 2013; Pickl, Spettl, et al., 2014; Scherrer and Stoeckle, 2016). In most cases, Euclidean distance was used to represent the *geographic distance* between the survey sites of the dialectal data. Gooskens (2004) was the first to operationalise the possibility of contact using travel times. Since then, rather few studies (e.g., Lameli, Nitsch, et al., 2015) have attempted to explain dialectal variation using geographic distance measures deemed to be more realistic than Euclidean distance for expressing a possibility for dialectal

contact. Stanford (2012) tested the correlation with 'rice-paddy distances' in the clan-based society of the indigenous Sui people in southwest China, while Szmrecsanyi (2012) used travel times and Trudgill's linguistic gravity index on British syntax from corpus data. However, multiple studies for different languages (van Gemert 2002, cited in Spruit, Heeringa, and Nerbonne, 2009:1638-1639; Stanford, 2012 and Szmrecsanyi, 2012) have found that travel times are not a better predictor for dialectal variation than Euclidean distances. Yet, it is believed in this thesis, in countries dominated by rugged topography – such as Switzerland – travel times, which inherently incorporate effects of isolation caused by topography, *should* be a better predictor of linguistic variation, as exemplified by the study of Gooskens (2004) on Norwegian dialects.

The *linguistic distance*, providing the linguistic basis of aggregate dialect variation, has been identified in numerous ways depending on the linguistic level in question. Grouping linguistic variables with regards to spatial distribution is another essential research approach. Most contemporary dialectometric studies use principal components analysis (PCA) (e.g., Shackleton, 2005) or factor analysis (e.g., Nerbonne, 2006; Pröll, 2013a, 2014) to detect linguistic items showing similar geographical patterns. A detailed overview on the methods applied in finding the linguistic basis of aggregate dialect variation is delivered by Wieling and Nerbonne (2015).

Nerbonne and Kleiweg judge the influence of geography as predominant (Nerbonne and Kleiweg, 2007), leading them to formulate the 'Fundamental Dialectological Postulate' (FDP), which posits that "geographically proximate varieties tend to be more similar than distant ones". As this observation also holds more generally for most other geographically distributed phenomena, the discipline of geography knows a very similar postulate, first formulated by Tobler (1970:236): "Everything is related to everything else, but near things are more related than distant things." The universality of this observation later led others to call it 'Tobler's First Law of Geography' (Sui, 2004). Both postulates describe an effect that is commonly known as spatial autocorrelation (Griffith, 1987).

On the other hand, corpus linguistics claims that the effect is considerably smaller than research based on atlas data shows, and the explanation power depends on the linguistic level too. For instance, Szmrecsanyi (2012) found only a very weak effect in his corpus-based study of English dialects, causing him to conclude that the impact of geography *per se* is overrated.

Most dialectometric investigations not directed at syntax (e.g., Heeringa and Nerbonne, 2001; Nerbonne, 2009; Nerbonne, 2010b; Pickl, Spettl, et al., 2014) found that a sublinear, logarithmic model better describes the relationship between the linguistic and geographic distance than a linear model.

All studies mentioned above restricted their attention to the 'global' level, i.e., the entire dialect area in question, not investigating correlation in smaller geographic subsets of these areas. Hence, they miss out on discovering regional differences in

correlation structures, and on delivering possible explanations of regionally different linguistic variation patterns stemming from local effects such as topographic barriers and, conversely, interconnectedness.

Similarly, a number of studies have assessed the effects of single geographic features, some of which form fiat boundaries (B. Smith and Varzi, 2000; Vogt et al., 2012) in (contemporary) perception. As an early example, Haag (1898) claimed that linguistic boundaries often coincide with political borders. This relationship has recently been investigated by Kürschner and Gooskens (2011), and by Pickl (2013), who also examined the effects of rivers. Sieber (2017), in his master's thesis, analysed the effects of cantonal borders, (historical) boundaries of Christian denominations and other administrative borders on Swiss German syntax, using the SADS database. He has shown how often neighbouring regions share features, with regards to a known external boundary placed between them. This approach is easily extensible to a larger geographical scale. However, Sieber's study did not investigate the separating effects of these sharp boundaries at the local level, i.e., in smaller spatial subsets.

2.4 Recent Research on Swiss German Dialectal Data

As showed earlier (Section 1.1.3), Swiss German is among the top researched dialects in Europe, with a plethora of linguistic interests assessed in the last more than 100 years (e.g., Sonderegger, 1962; Lötscher, 1983; Börlin, 1987; Schrambke, 2010; Sonderegger, 2013; Bucheli Berger and Landolt, 2014)[3].

Besides physical barriers inducing isolation, other, (socio)cultural factors are considered to have influenced the evolution of Swiss German dialects (e.g. Hotzenköcherle, 1961). These include historical isolation in the times before the modern Swiss Confederation was formed, more *'sharp'* boundaries, like those of administrative subdivision, or boundaries of confession (Roman catholic vs. protestant). In the last two decades, several dialectometric studies focused on Swiss German, for example Kelle (2001) and Scherrer (2012), who worked with the SDS data. Later, Kellerhals and Scherrer (2014), and Scherrer and Stoeckle (2016) incorporated SADS's syntactic variables into a similar analysis, comparing the spatial variation across different linguistic levels. They did not, however, account for the effect in the linguistic variation caused by spatial variation. In other words, they did not explicitly quantify the impact of geographic distances. An exception is the recent master's thesis by Sieber (2017), which demonstrated the influence of administrative, and partly also cultural, borders on spatial variation in Swiss German dialects, as represented in the SADS.

The hypothesis of Seiler (2005), mentioned in Section 2.3.2, which assumes a largely west-east variation in some phenomena (e.g., the *'infinitival purposive clause'*) with gradual transition forming *inclined planes*, was tested by Sibler (2011) in his

[3] A searchable database of studies on Swiss German dialects is available at www.ds.uzh.ch/dialbib/index.html

classification study. For each variant, a three dimensional surface is built using the proportion of respondents using the given variant at each survey location (termed *'intensity surface'*). Sibler was testing trend surfaces of different order on the intensity surface of different variants. He did not, however, take the relationship of corresponding variants into account and also did not explicitly account for the gradual nature of boundaries.

Stoeckle (2016), using variables from the SADS database, dealt with inter-personal variation at the global and local levels, determining the syntactic variability for each location using hot-spot analysis (the same method as used by Grieve et al., 2011). In another study (Stoeckle, 2018), he conducted an *'apparent time analysis'*, exploiting the fact that at each survey site respondents of different ages provided information, to determine dynamic and stable regions from the point of view of dialect change. This study involved extralinguistic factors, both geographic and sociodemographic. He did not, however, take an aggregate of variables into account for the investigation of the correlation, nor did he quantitatively account for the boundaries found in the variation.

2.5 GIScience and Dialectology

2.5.1 The potential of GIScience in linguistics

Understanding the role of space, distances and their relationships to other variables is a key element to understanding any phenomenon that is distributed over a geographic area (cf. Hoch and Hayes, 2010).

There are two main strands of GIScience-related research in linguistics. One is interested in the *language **of** space*, i.e., how people talk about space and how, for instance, they navigate through space and how they describe places (Egorova, 2018). The other one is dealing with *language **in** space*, interested in understanding the mechanisms of language through its spatial variation. The latter forms the wider research strand of this thesis.

The three most important ways in which GIScience and spatial theory have been affecting linguistics are the following. First, GIScience contributes to highlighting and analysing patterns in linguistic data and to showing the variation in language (popular methods include boundary definition, point pattern analysis, kriging, semivariogram, Cronbach's α, geographically weighted regression and spatial autocorrelation measures such as Moran's I, Getis-Ord G_i*). Second, methods used in GIScience allow showing the relationship of linguistic variables and other spatially distributed phenomena using correlation analyses with measures such as Pearson's r and Kendall's τ, the Mantel-test, regression models such as the generalized additive model (Wolk, 2014; Wieling, Montemagni, et al., 2014), and geographically weighted regression (Fotheringham, Brunsdon, and Charlton, 2002), barrier detection such as the Monmonier-analysis (Manni, Guérard, and Heyer, 2004), and least-cost models

and route planning. And third, GIScience methods allow testing for the effects of spatial dependence in data and correcting for them, as exemplified by the works of Grieve (2011) and Nguyen and Eisenstein (Nguyen and Eisenstein, 2017).

Ultimately, GIScience provides methods and theories that may help the spatial understanding of language, offering "...an articulation of spatial theory as a framework for approaching hypotheses in linguistics research" (Hoch and Hayes, 2010:28). Besides, GIS offers geolinguistics a range of possibilities for visualisation of geographic relationships and enables the creation of multiple alternative maps for comparison.

2.5.2 Concepts of GIScience relevant for dialectology

Several models from spatial information theory are relevant for the distribution of dialectal phenomena. Regarding their demarcation, linguistic boundaries can be put in parallel with boundaries of climatic regions or forest types (Leung, 1987; Mark and Csillag, 1989), that is, with uncertain and fuzzy boundaries. Probability surfaces of class membership (Mark and Csillag, 1989; Pröll, 2013a) and fuzzy set membership functions (Girard and Larmouth, 1993; Scholz et al., 2016) have been used to describe and locate boundaries. GIScience has researched the ontological status of boundaries (e.g., B. Smith and Varzi, 1997; Galton, 2001, 2003) and the framework they contributed would be beneficial in relation to boundaries in dialect geography as well.

Dialectology approaches the spatial distribution of features or dialectal variants as areas that are either sharply separated or gradually transition into each other. These approaches correspond to GIScience's boundary concepts of *objects* and *fields*, respectively (e.g., Galton, 2004). "In a field-based model, a transition zone can be represented analogically *as* a transition zone, with the intermediate field values directly representing the gradation in the underlying reality. In an object-based model, on the other hand, everything is biased towards a crisp all-or-nothing style of representation..." (Galton, 2003:169), and in order to represent indeterminate boundaries one has to define transition zones with sharp outer boundaries. This duality described by Galton can be discovered in a very similar way in dialectology's two approaches to describing dialectal variation in space: isoglosses and dialect continua.

With the presence of uncertainty, however, abstractions may be necessary. Dialectology's contemporary approach (e.g., Sibler, 2011; Pickl and Rumpf, 2012) is reflected in Mark & Smith's article: "Appropriate generalization methods may involve construction of surfaces representing probability of class membership, generalization or smoothing of such surfaces, and 'contouring' the probabilities to find boundaries" (Mark and Csillag, 1989:2).

Explicit research on linguistic boundaries from the side of GIScience include Manni et al.'s (2004) Monmonier-analysis for finding boundaries and Scholz et al.'s (2016) robustness test on boundaries of different porosities using fuzzy logic.

Two other relevant spatial properties are spatial dependence and spatial heterogeneity (e.g., Anselin et al., 1996) that are in connection with the First Law of Geography (Tobler, 1970:236), that is, "Everything is related to everything else, but near things are more related than distant things", and includes spatial autocorrelation, to some degree always present in linguistic data. Goodchild (2004) proposed a Second Law, about spatial heterogeneity, whose corollary is that global standards and local standards will always differ. Therefore Goodchild's proposition is that 'average' places do not exist and it is always worth investigating local relations and variation besides global.

Viewing distributions in linguistic data considering these two approaches is central in this thesis.

2.5.3 Experiments with GIScience methods in Dialectology

As noted before, researching spatial variation in dialectal data from the point of view of GIScience is still a little-trodden area, despite the number of potentially available tools in GIS, which have already proven their value in other interdisciplinary studies.

Linguistics, nevertheless, naturally gravitates towards using methods used in GIScience, however, mainly for visual representation. Namely, Voronoi diagrams (Voronoi, 1908) have often been used for representing probable answers in unsurveyed areas, connecting each point in the study area to the closest survey site. Besides map production and this salient example, linguistic researchers have indeed adopted other theories and methods often used in GIScience, such as spatial autocorrelation (e.g., Thomas, 1980; Grieve, Speelman, and Geeraerts, 2011), interpolation by kernel density estimation (Rumpf et al., 2009) and kriging (Grieve, 2013). Willis (2017) used geographically weighted regression to investigate language change by plotting S-curves across a region for a linguistic feature hypothesised to be undergoing change.

Lee and Kretzschmar (1993) analysed the *'Linguistic Atlas of the Middle and South Atlantic States'* (LAMSAS) using methods of spatial autocorrelation and point pattern analysis, advocating more collaboration with GIScience and geolinguistics. Hoch and Hayes (2010) neatly summarise what methodology areas in GIScience could be useful for geolinguistics and thus how it is connected to spatial theory. They discuss GIS tools and techniques, such as point pattern analysis, semivariance and kriging interpolation, deemed potentially beneficial for geolinguistics. They criticise traditional choropleth and isoline mapping techniques – and thus the visualisation of variation by isoglosses in dialect geography – for ignoring the possibly gradual nature of the boundaries and the "the complex multi-attribute nature of dialect space" (Hoch and Hayes, 2010:29). They do, however, admit that cartographic generalisation and smoothing are necessary to create an interpretable map. Hoch and Hayes (2010) emphasise the potential of an interface between geography and linguistics:

"GIS and the geographical approach offer researchers of geolinguistics many possibilities for advancing and reexamining theory, hypotheses, and data visualisation. GIS and GIScience can offer an articulation of spatial theory as a framework for approaching hypotheses in linguistics research. In addition, GIS can simply make much research in geolinguistics faster and easier" (2010:28). As a methodological example, they propose a *'gradation zone'*, resembling the *transition zones* discussed before (Section 2.1). They also assume that Kronenfeld's (2007) TIN-based models, representing transition between areas of more certain class memberships, should be useful in linguistics. Finally, the authors suggest using probability surfaces of class membership (such as Mark and Csillag, 1989) and fuzzy set membership functions (such as Girard and Larmouth, 1993) to better describe and locate class and spatial boundaries by noting variations in the rate of dialect change across space.

Lee and Kretzschmar (1993) and Hoch and Hayes (2010) have called for an action from the side of GIScience, but, ever since, contributions from GIScience in dialect geography have not been frequent. As an example, Sibler et al. (2011, 2012) used point pattern analysis, trend surface analysis, interpolation methods (kernel density estimation; KDE) and investigated spatial autocorrelation (using semivariograms and Moran's Index) in the same Swiss German dialect data as used in this thesis (SADS).

Lameli (2013) and Pickl (2013) both express their surprise over the paucity of research applying spatial statistical methods to dialectal data analysis, and Pickl further deplores that the main research interest in dialectology has been to determine dialect areas and boundaries. Pickl's linguistic thesis itself (2013) tests hypotheses using spatial interpolation, measurements of regional complexity and factor analysis, beside visual analysis. He also summarises some of the achievements (Pickl, 2013:14) of the Ausgburg/Ulm dialectometric research project (Rumpf et al., 2009, 2010; Pickl and Rumpf, 2012; Pröll, 2013b), focusing on the application of spatial statistical methods in dialecology. The interest of the corresponding project was on the 'variation of the variation', thus the differences and peculiarities in the spatial distribution of individual variables and variants. Pröll's (Pröll, 2013b) thesis offered methods of grouping dialectal variables based on the similarities in their spatial distributions. The explicit focus, however, was not on quantifying the graduality of boundaries. Chagnaud et al.'s (2017) cartographic tool is, contrasting the aims of Chapter 6 in this thesis, mimicking isogloss-drawing.

Hoch and Hayes (2010) claim that quantitative analyses of linguistic data were careful to include sampling bias and statistical independence (Guy, 1993; Kretzschmar and Schneider, 1996); however spatial dependence is not consistently considered, which might constrain unbiased and independent sampling in space.

As an example of analysing aggregate linguistic variation, Sieber's (2017) master thesis in GIS specifically analysed the effects of *fiat*, administrative boundaries to assess their effect on syntactic differences using multiple regression models, correlation analysis and least cost paths.

Scholz et al. (2016) investigated the diffusion of linguistic innovations from a GIScience point of view. They took advantage of the fuzzy set theory (Zadeh, 1965) to propose models for 'indeterminate and crisp' linguistic boundaries, focusing on spatial change in dialect regions, with regards to topography. Their approach involved modelling dialectal diffusion based on fuzzy formal logic and concluded that the more isolating power a boundary has, the greater its impact on dialectal differences.

Multiple GIS applications have been developed in the past two decades for visualising dialect data and calculated dialectometric values, such as Goebl's (2004) *"Visual DialectoMetry"*, Peter Kleiweg's *"RUG/L04"* (http:// www.let.rug.nl/kleiweg/l04) and Nerbonne et al.'s (2011) "Gabmap" (http://www.gabmap.nl). The latter offers noisy clustering and multidimensional scaling beside association mapping (Snoek, 2014).

As a further overview of GIS methods, a joint publication with Philipp Stöckle (Stoeckle and Jeszenszky, 2017) demonstrates the potential power of GIS methods for linguistic geography and gives examples of the use of spatial statistical approaches, such as kriging interpolation and hot-spot analysis, as well as visualisation strategies.

2.6 Summary of the Research Gaps

Throughout the above review of related work, a number of research gaps potentially interesting for GIScience have been identified. Generally speaking, methods of GIScience and spatial analysis have not found a lot of attention in research using them on dialectal phenomena. As a further general observation, we note that although Swiss German dialects have been thoroughly researched, there are still significant research gaps that are not addressed yet, both at the level of individual variables and at the level of aggregate variation.

In the remainder of this chapter, the specific research gaps identified for this thesis are summarised. This will be done under the umbrella of the Research Objectives defined in Section 1.2.1, as these provide the motivation for the main research Chapters 4 to 6.

***Research Objective* 1. (RO1) Develop and experiment with methods to explore the relationship between geographic distances and dialectal variation, in particular Swiss German syntax.**

To address this objective requires determining what amount of variance in aggregate dialectal differences can be explained by geographic distances. As shown in the above review, this leads to the following research gaps concerning RO1:

- Little work has addressed the influence of geographic distance on syntax variation, and in particular not for Swiss German dialects.

- Dealing with multi-respondent, multivariate syntax data (such as SADS) requires developing a measure of aggregate linguistic distance between survey sites.
- Few studies have investigated the influence of barriers (e.g., topographic barriers, administrative borders) on language contact, and thus on the linguistic variation across space.
- Travel times, which implicitly incorporate topographic barriers, have rarely been explored in predominantly mountainous areas, in particular not in Switzerland.
- Previous studies have only considered correlation of linguistic and geographic distances at the level of entire data set, rather than regional subsets.

Research Objective **2. (RO2) Develop and evaluate methods to quantify the degrees of transition in interdialectal spatial variation.**

To address this objective, we validate prototypes of conceptual models defined in the dialectological literature to address the boundaries and transitions inherent to the variation patterns in individual variables. The above review leads to the following research gaps concerning RO2:

- Previous work only focused on the internal homogeneity of variant prevalence areas, devoting little explicit interest to the nature of dialectal boundaries and transitions at the level of individual variables.
- Although conceptual models have been proposed for the different kinds of interdialectal transitions (sharp to gradual), no mathematical prototype models were created based on which the fit of these conceptual models to survey data could be tested.
- No previous study has proposed a methodology to compare interdialectal boundaries with regard to their graduality.

Research Objective **3. (RO3) Automate the detection of interdialectal sharp boundaries and gradual transitions in a data-driven way.**

The approach for addressing this research objective lies in providing probabilistic boundaries by automating the traditional technique of drawing isoglosses and quantitatively adjusting it to the transitions found in the variation patterns of individual variables. This proposed approach is derived from identifying the following research gaps:

- The potential graduality of transitions is not concisely assessed by drawing sharp boundaries (i.e., isoglosses), but a data-driven emulation of the traditional isogloss drawing technique may overcome linguistic preconceptions.

- Previous quantitative work only focused on the areal classification or thresholding of dominant dialectal variants, without explicitly concentrating on the detection of boundaries and transitions.
- Studies proposing probabilistic areal classification exist, but the contribution of survey sites to positioning interdialectal boundaries has not been quantitatively investigated so far.
- Sharp and gradual boundary concepts are used in parallel in dialectology, prompting a need to look at these boundaries and transitions from a more holistic point of view, which would also allow comparison across linguistic variables.

Chapter 3

Data

The linguistic data used in this research stems from the Syntactic Atlas of the German-Speaking Switzerland (*'Syntaktischer Atlas der deutschen Schweiz'* – SADS[1]). This linguistic atlas and its database have been compiled by researchers at the German Department of the University of Zurich, under the leadership of Elvira Glaser and with contributions from Claudia Bucheli Berger, Guido Seiler, Matthias Friedli and Gabriela Bart, starting from 2000.

The atlas is viewed as a continuation of a preceding large language atlas project collected on Swiss German dialects, the Linguistic Atlas of Swiss German (*'Sprachatlas der deutschen Schweiz'* – SDS – Hotzenköcherle et al., 1962-1997). The SDS historically made a very important contribution to dialectology and made German-speaking Switzerland one of the best researched dialect areas in Central Europe (Scherrer and Stoeckle, 2016). Initiated in 1935, with data recorded between 1939 and 1958, the dialect maps were published between 1962 and 1997 in eight volumes. The vast volumes of the SDS cover the dialect phenomena of the German-speaking areas of Switzerland, containing 1548 maps in total (Scherrer, Leemann, et al., 2012). With its 625 survey locations and respondents, it is one of the most comprehensive dialect surveys on its scale. Although the authors initially had the intention of including syntactic phenomena, the SDS actually shows only half a dozen maps demonstrating geographical syntactic variation (Bucheli and Glaser, 2002).

3.1 The SADS Survey

3.1.1 Key characteristics

The SADS database (Bucheli and Glaser, 2002) has been compiled from surveys conducted on the linguistic level of syntax (Section 2.1), in order to fill this remaining gap in the dialectological research of Swiss German. Four surveys were conducted between 2000 and 2002 by questionnaires, aiming to capture the (morpho)syntactic diversity of Swiss German, focusing on phenomena with assumed notable geographic variation patterns. This latter criterion might admittedly cause a bias in the perception of spatial variation of dialects. Thus, it has to be noted that differences found

[1] http://www.dialektsyntax.uzh.ch

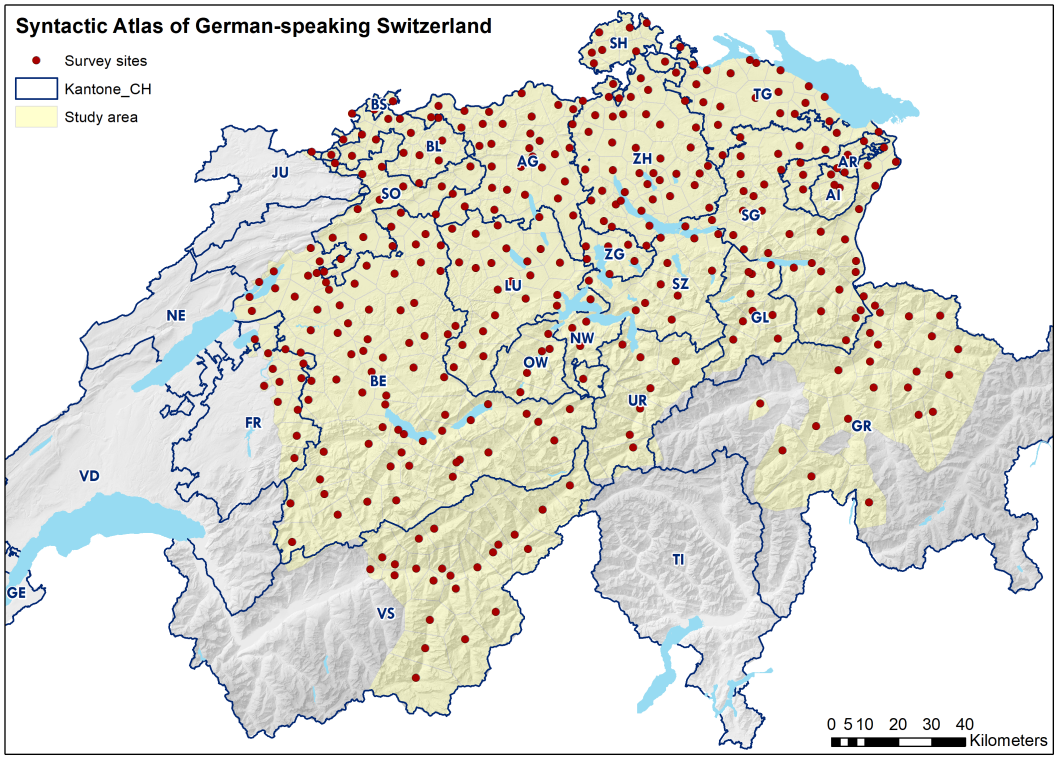

FIGURE 3.1: The survey sites of the Syntactic Atlas of German-speaking Switzerland (SADS), i.e., the study area of this thesis.

are not always representative of the whole dialectal landscape, as argued by authors pursuing dialectometry based on corpus data (e.g., Szmrecsanyi, 2012). The four questionnaires were sent to multiple respondents living in 383 Swiss municipalities (Figure 3.1). These sites are a subset of those used for the SDS (625 sites) half a century earlier. They correspond to approximately 25% of the German-speaking municipalities of Switzerland, resulting in a sample density that is remarkable among linguistic surveys conducted on such scale.

Throughout the duration of the survey, the four questionnaires were sent to 3187 respondents in 6 months intervals, and 2771 people returned all four questionnaires (Glaser and Bart, 2015). A key feature of the SADS is getting having multiple respondents at each survey site because it serves the goal of collecting more representative data, and aims to cover as much of the local dialectal variation spectrum as possible. To validate the data collected and to reduce the uncertainty inherent to the data, post-survey interviews were also undertaken at certain survey locations. The number of respondents varies by survey site and does not follow a spatial pattern. It ranges between 3 and 26, with a median value of 6 to 7 respondents per site, depending on the survey question, as not all respondents have given a valid answer to every question. Figure 3.2 shows the average number of respondents at each survey site.

The respondents have provided answers to 118 survey questions altogether. In

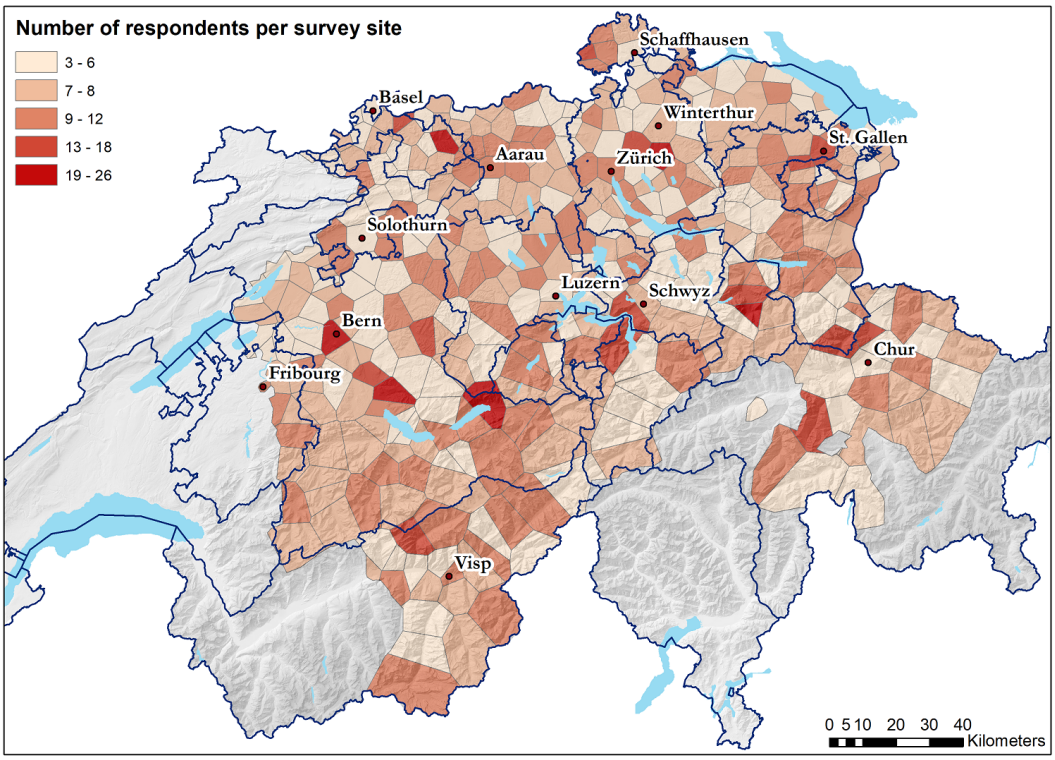

FIGURE 3.2: Mean number of respondents at each survey site of the SADS.

the database, these questions are linked to over 50 syntactic phenomena. Thus, in several cases multiple questions address the same syntactic phenomenon, aiming to increase explanatory power. Most questions in the survey were either translation questions, where the respondents had to translate a sentence from Standard German into their local dialect, or multiple choice (MC) questions, where respondents had to choose from a set of variants those that they considered *possible* to use in their local variety. For MC questions, respondents were also asked to specify a variant they *preferred* over the other variants they considered possible. In order to make the questionnaire more respondent-friendly and to consider the ambiguity within questions, short stories were often composed to frame the questions (Bucheli and Glaser, 2002).

Ensuring the endemic nature of the collected information was a key element of the sampling scheme used in the SADS project. Thus, it was required that the respondents had been raised and they had to live most of their lives in the surveyed village or town, along with at least one of their parents. In order to represent the demographics of the surveyed population, respondents of different ages and different professions were surveyed. This latter criterion contrasts the data collection conventions of traditional dialectology, where at each survey site usually only one respondent was taken into account, with a NORM (non-mobile, older, rural, male; cf. Chambers and Trudgill, 2004:29) profile. This type of speaker was often believed to be the best manifestation of authentic local linguistic information. According to

Bucheli and Glaser (2002), the goal in the SADS project originally was to include ten respondents at each survey site: men and women, different age groups and various professions. Apart from the general requirement of being a local, however, the respondents were eventually not recruited systematically, so their numbers as well as their socio-demographic backgrounds are not evenly distributed over the study area (Stoeckle, 2016b). Nevertheless, the SADS represents one of very few modern language atlases that capture some of the linguistic variation present *between* and *within speakers*.

As the population density in Switzerland is highly constrained by topography, the density of survey sites has been designed to be roughly uniform across the whole study area, in order to be representative of mountainous areas as well. The spatial distribution of the survey sites corresponds to neither the population density nor the settlement density as there are more survey sites in the densely populated Swiss Plateau (*'Schweizer Mittelland'*) than in the sparsely populated mountainous regions. The proportion of survey sites and settlements is higher in the Alps. This is in the interest of the survey, nevertheless, because the survey team aimed to capture the whole range of dialectal variation present in the investigation area. Due to the increased isolation in alpine areas, it was expected that a higher diversity in dialects is to be found within a relatively smaller population.

3.1.2 SADS example

Figure 3.3 shows an example map from the SADS atlas in preparation. Such *point symbol maps* make up the majority of the maps in the atlas. This map represents the variation of a comparative clause, where multiple answers are possible. The symbol sizes do not correspond to the actual proportion of variants occurring at the survey sites. Instead, they represent presence/absence of the variants on a short ordinal scale, thus emphasising the diversity of variants. The symbol sizes vary depending on whether a variant is preferred by a single or multiple respondents at the given survey site. It is visible that one variant (–, the *'als'* variant) is ubiquitously present. The circle (•, *'weder'*) variant covers a smaller area while the variants shown in red have only regional importance. As one of the aims of the atlas is to capture the full spectrum of variation in syntactic phenomena, it is important to emphasise minority occurrences of less frequently used variants.

3.2 Preprocessing the SADS Data

The original data of the SADS surveys were recorded in a *Filemaker database*, where the answers are stored for each respondent, along with an indicator of whether the answer is usable for the research aims. In order to process the data for this thesis, it had to be converted into a more commonly used format, and filtered by survey questions. For each question, *dbf* and *csv* files were produced, where responses are

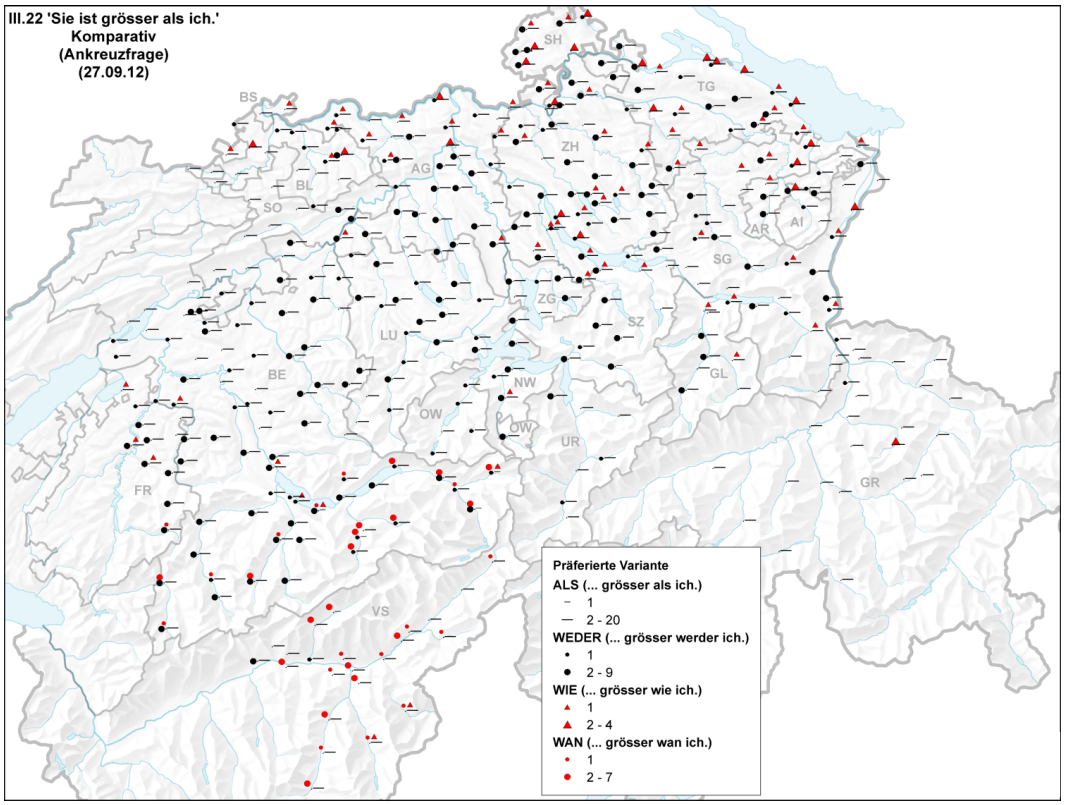

FIGURE 3.3: An example map from the actual SADS in preparation about a *'comparative clause'* (III.22) featuring several different variants.

sorted corresponding to survey sites. These (spreadsheet) files are used for visual and spatial analysis in *ArcGIS* and in *R*. For aggregate data, further files were created for different subsets of questions.

As mentioned above, the SADS database also stores for MC questions whether an answer variant is *accepted* by the respondent as usable in his or her local dialect and which one he or she *preferred*. For all the case studies in this thesis, the preferred variants were taken into account, as they better represent the authentic local language variation. Moreover, as most respondents accept several variants (Glaser and Bart, 2015), mapping them usually results in noisier patterns and in more mixture of variants with diffused areas of undetermined dominance. The patterns also differ between questions, due to differences in the number of potential answers, semantic categories and question types (Sibler, 2011). For every calculation, all the usable answers at each survey site have been used.

In the case studies presented in Chapters 4, 5 and 6, 60 questions (and subsets thereof) were used, those that had entirely and reliably been evaluated by the start of the research experiments. These survey questions are listed in Table 3.1. According to the SADS editors, they representatively cover the majority of the roughly 50 morphosyntactic phenomena investigated in the SADS. Each variable corresponds to one aspect of one survey question. Several survey questions address more than

one linguistic variable. Although a survey question can thus cover multiple variables (as also shown in Table 3.1), throughout the experiments reported in this dissertation, only one aspect per survey question is taken into account. This aspect is referred to as *variable*, in correspondence with the terminology introduced in Section 2.1.

For the three case studies presented in Chapters 4 to 6, respectively, different subsets of SADS questions were used. Details about these subsets are explained in the corresponding chapters.

3.3 Travel Time Data

As visible in Figure 1.1, Switzerland features a diverse surface topography that (despite the generally good Swiss transport infrastructure and recently built tunnels increasing accessibility) imposes constraints on transport routes and thus on possible contact paths between speakers. To account for the constraints of topography impacting the possibility of *dialect contact*, in Chapter 4 and 6, geographic distances were represented not only by Euclidean distance, but also by travel times. Travel times, as they capture an isotropic nature of geographic space, encompass the topographic and infrastructural features between the locations of interest. They translate to the effort needed for people located in two survey sites to meet, which is similar to the notion of *cost* in GIScience (e.g., Douglas, 1994; Rees, 2004; LaRue and Nielsen, 2008). However, as travel times cannot take into account the *interest* in *actually* meeting, using travel times as a proxy for dialect contact may not be completely accurate, but language contact is assumed to be influenced by social and cultural factors as well.

For our research, a database obtained from the Institute for Transport Planning and Systems at ETH Zurich (Fröhlich et al., 2004) has been used, consisting of travel times by individual and public transportation in a matrix format. The database contains comprehensive data for several points in time about travel times by public transport and by individual transport (for years 1850, 1888, 1910, 1930, 1950, 1960, 1970, 1980, 1990 and 2000).

Figure 3.4 shows the difference between Euclidean distance and travel times in 2000 (after normalisation). Red line colours indicate that travel times better express a greater 'isolation' between the locations than Euclidean distance. The green colours correspond to a negative effect, which means that these locations are 'brought closer together' when investigated by travel times. Many of these 'neighbours', connected by green lines, are found in mountainous areas and are linked by tunnels, however they lie in otherwise isolated valleys. It is visible that in most of the study area Euclidean distance 'underestimates' travel times (red colours). Note, however, that in the case of neighbours where the Euclidean distances are already considerably big, travel times might not mean a very big increase.

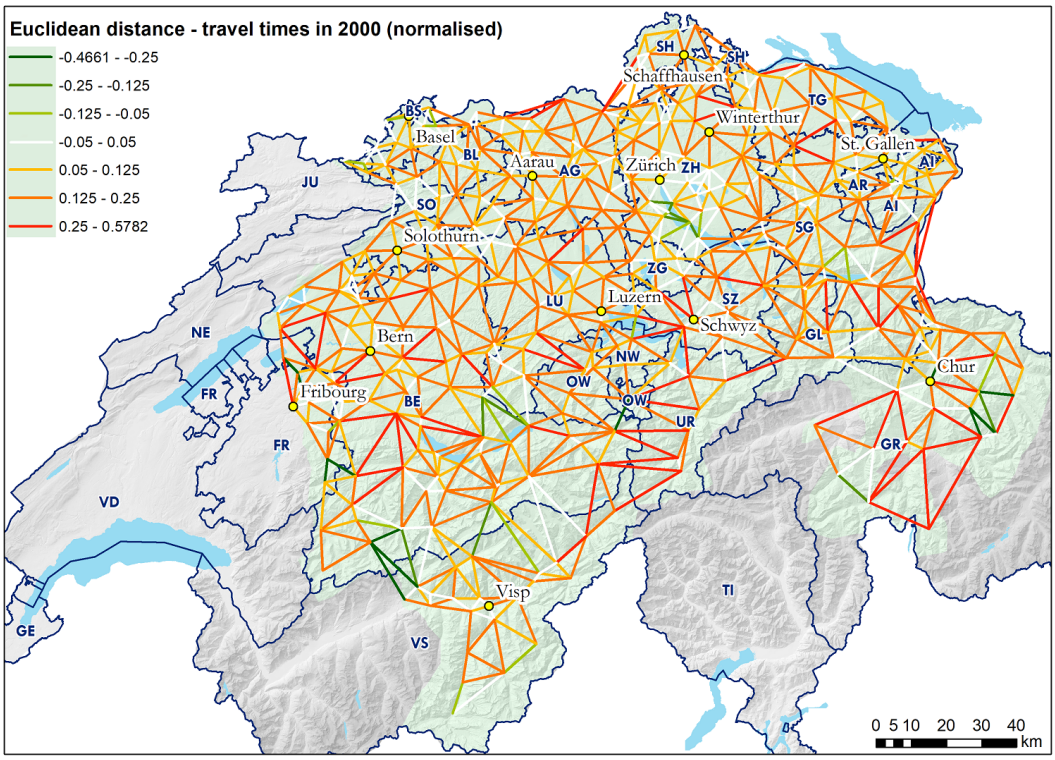

FIGURE 3.4: Natural neighbours among survey sites connected. The lines are coloured based on the normalised difference between the Euclidean distance and travel times by car in 2000.

For the case study presented in Chapter 4, travel times of individual transport for the years 1850, 1950 and 2000 were used. For 1950 and 2000, car travel time matrices covered all 383 survey sites of the SADS. Before 1950 – thus for the year 1850 – the data was available only for 120 locations, or roughly one third of the survey sites. However, as these places are equivalent to district (*'Bezirk'*) capitals, the spatial distribution generally follows that of the entire set of SADS points.

Chapter 6 uses travel times by car in 1950 as an input for the proposed clustering procedure included in its methodology. The clusters are thus formed based on the travel times between the points (survey sites) concerned.

TABLE 3.1: The linguistic variables used in the research experiments. In Chapter 4 all, and in Chapters 5 and 6 a subset of these variables were used, detailed in the respective chapters.

SADS ID	**Sentence (Standard German)**	**Sentence in English**	**Phenomenon in English (short)**
I.1	Entschuldigung, ich habe zu wenig Kleingeld, um ein Billett zu lösen.	Excuse me, I don't have enough change in order to buy a ticket.	infinitival purposive clause: linkage

I.2	Wem will er denn die schönen Blumen bringen?	To whom does he want to bring those beautiful flowers?	prepositional dative marking (PDM)
I.3	Oh, ich habe den Fritz kommen hören.	Oh, I heard Fritz coming.	perfect with 'hear': form and position of non-finite verb (IPP)
I.5	Der Korb ist umgekippt.	The basket is toppled over.	resultative: subject agreement
I.6	Wissen Sie, jetzt brauche ich sogar Tabletten zum einschlafen.	You know, now I even need pills in order to fall asleep.	infinitival purposive clause: linkage
I.7	Nein, das gehört meiner Schwester.	No, it belongs to my sister.	prepositional dative marking (PDM)
I.8	Aber ich habe im Fall schon gestern geholfen abzuwaschen.	But I already helped doing the dishes yesterday.	perfect with 'help': form and position of non-finite verb (IPP)
I.9	Also ich weiss auch nicht, ob er einmal heiraten will.	Well, I don't know if he ever wants to get married.	modal verb in subordinate clauses: position
I.11	Aber jetzt habe ich mich gerade hingesetzt, um ein Buch zu lesen.	But I just sat down in order to read a book.	infinitival purposive clause: linkage
I.12	Fischstäbchen muss man doch gefroren anbraten.	Actually, fish fingers should be fried while still frozen.	copredicative participle
I.13	Da wird gearbeitet.	People are working here.	expletive 'it' (impersonal passive)
I.18	Soll ich welche kaufen?	Should I buy some?	partitive object (pronoun)
I.19	Ich habe keine Ahnung, ob sie das Auto schon bezahlt hat.	I have no idea whether she has already paid for the car.	perfect auxiliary ('have') in subordinate clauses: position
I.20	Aber ich habe doch das Buch dir geschenkt.	But I gave the book as a present to you.	prepositional dative marking (PDM)
II.1	Hast du die Uhr flicken lassen?	Have you had the clock fixed?	infinitive particle (doubling/position) 'let'

II.2	Das ist doch die Frau, der ich schon lange das Buch bringen sollte.	This is the woman to whom I should have brought back the book long ago.	relative clause linkage: IO
II.3	Er lässt den Schreiner kommen.	He is going to call the carpenter.	infinitive particle (doubling/position) 'let'
II.4	Du hast sicher viel zu erzählen!	You must have a lot to tell.	non-finite form with 'have to' (gerund)
II.5	Ihr dürft alles liegen lassen.	You can leave everything.	infinitive particle (doubling/position) 'let'
II.7	Ich habe erst mit vierzig fahren gelernt.	I have only learnt to drive at forty.	perfect with 'learn': form and position of non-finite verb (IPP)
II.9	Nein, sie ist gerade verkauft worden.	No, it has just been sold.	passive auxiliary and agreement
II.11	Er hat die Hand immer noch eingebunden.	He has his arm still bandaged.	resultative: object agreement
II.13	Du musst die Milch aber heiss trinken!	But you have to drink the milk hot!	copredicative adjective
II.18	Das ist der Mann, dem ich gestern den Weg gezeigt habe.	That's the man to whom I gave directions yesterday.	relative clause linkage: IO
II.19	Und dann ist ein Fuchs geschlichen gekommen!	And then a fox came creeping around!	verbal construction 'come' + motion verb
II.20	Ich möchte aber ein Auto, das ich auch bezahlen kann!	But I want a car that I can actually pay for!	relative clause linkage: DO
II.22	Nein, das ist Peters [Dreirad].	No, that's Peter's. [tricycle]	predicative possessive
II.23	Nein, das ist Sandras [Dreirad].	No, that's Sandra's. [tricycle]	predicative possessive
II.28	Das ist der Mann, mit dem ich immer schwätze.	That's the man I always chat with.	relative clause linkage: PP
II.30	Der Hund des Lehrers	The teacher's dog	adnominal possessive
II.32	Ich habe Fritz gesehen	I have seen Fritz.	personal name: definite article and case inflection

III.1	Wenn es so warm bleibt, fängt das Eis an zu schmelzen!	If it stays this warm, the ice will start to melt.	infinitive particle (position/doubling) 'begin'
III.2	Wen suchst du?	Who are you looking for?	interrogative pronoun: case
III.3	Für wen sind denn die Blumen?	Who are the flowers for?	interrogative pronoun: case
III.4	Die sind nicht für dich!	They are not for you!	personal pronoun (2sg): PP
III.5	Ich habe schon angefangen zu kochen.	I have already started cooking.	infinitive particle (position/doubling) 'begin'
III.7	Sie hat es mir gestern erzählt.	She told me yesterday [about expecting a baby].	personal pronouns: position
III.8	Sie findet es nicht gut, dass ich angefangen habe zu rauchen.	She doesn't find it good that I started smoking.	infinitive particle (position/doubling) 'begin'
III.10	Wenn sie dich erwischen, bekommst du den Fahrausweis entzogen!	If they catch you, your driver's license will be taken away.	'get'-passive
III.11	Also mich erwischt keiner!	Well, no one will catch me!	personal pronoun (1sg): DO
III.12	Nimm die Suppe sofort weg, wenn sie zu kochen anfängt!	Take the soup off immediately, once it starts boiling.	infinitive particle (position/doubling) 'begin'
III.13	Er gibt sich einfach keine Mühe.	He just doesn't put any effort into it.	reflexive pronoun (3sgm)
III.16	Die Strasse ist schon seit einem Jahr aufgerissen.	The street has already been torn up for a year.	resultative: subject agreement
III.17	Wir müssen uns das überlegen.	We have to think about it.	reflexive pronoun (1pl)
III.20	Er schaut nur für sich selbst.	He only thinks about himself.	reflexive pronoun (PP)
III.22	Sie ist grösser als ich.	She is taller than me.	comparative clause linkage
III.23	Hinkend ist er gelaufen.	He went home limping.	converb
III.25	Sie gehen halt lieber schwimmen als laufen.	They would rather go for a swim than for a walk.	comparative clause linkage
III.28	Dann ist er ja älter, als ich gemeint habe.	So he is older than I expected.	comparative clause linkage

IV.3	Ich habe es ihm schon geschickt.	I have already sent it to him.	personal pronouns: position
IV.4	Wer ist das gewesen?	Who was it?	interrogative pronoun: case
IV.7	Jetzt kannst du anfangen.	Now you can start.	non-finite 'begin' with modal verb
IV.11	Doch, das ist im Fall er gewesen.	Yes, that must have been him!	personal pronoun (3sgm): subject
IV.14	Du musst das Licht anzünden, um zu lesen.	You have to turn the light on in order to read.	infinitival purposive clause: linkage
IV.17	Doch, das ist er sicher gewesen!	Yes, that was him for sure!	personal pronoun (3sgm): subject
IV.19	Ja, ich habe etwas ganz Schönes gekauft!	Yes, I have bought something really nice!	indefinite pronoun: position/doubling
IV.21	Ich habe nicht gewusst, dass er so spät fahren gelernt hat.	I didn't know that he learnt to drive only so late.	perfect with 'learn': form and position of non-finite verb (IPP)
IV.25	Das glaubst du ja selber nicht, dass sie so früh lesen gelernt hat.	No way she learnt to read so young!	perfect with 'learn': form and position of non-finite verb (IPP)
IV.28	Ich habe es Fritz gegeben.	I gave it to Fritz.	personal name: definite article and case inflection
IV.31	Das gefallen täte mir auch!	I would like it, too!	subjunctive auxiliary 'do' (position)

Chapter 4

Correlation of geographic distances and dialectal variation

4.1 Introduction

4.1.1 Motivation and hypotheses

The 'Fundamental Dialectological Postulate' (FDP), positing that 'geographically proximate varieties tend to be more similar than distant ones' is formulated by Nerbonne and Kleiweg (2007). It is very similar to the postulate known in geography, phrased by Tobler (1970:236) as: "Everything is related to everything else, but near things are more related than distant things", called the 'First Law of Geography'. Both postulates describe the effect that is called spatial autocorrelation in geography (Griffith, 1987). In dialectal variation, however, the strong presence of spatial autocorrelation is not guaranteed. For instance, Szmrecsanyi (2012) found only a very weak effect in his corpus-based study of English dialects, causing him to conclude that "geography is overrated."

This study explores the correlation between the variation in Swiss German syntax and geographic distances. However morphosyntactic dialect variation has recently witnessed an increase in interest after a long time of neglect, correlation analysis of syntax against geographic distances has so far only been carried out by Spruit (2006, 2008) and Szmrecsanyi (2012, 2014) on Dutch and English dialects, respectively. Owing to the peculiarities of the SADS data – most importantly the fact that it features contribution from multiple respondents in various numbers per survey site – we use a particular measure to represent linguistic distance, similar to Speelman, Grondelaers and Geeraerts (2003). We adopt an aggregate variation approach (e.g.,

This chapter is based on **Jeszenszky, Péter**, Philipp Stoeckle, Elvira Glaser, & Robert Weibel, (2017). Exploring global and local patterns in the correlation of geographic distances and morphosyntactic variation in Swiss German. *Journal of Linguistic Geography*, 5, 1–23. https://doi.org/10.1017/jlg.2017.5
Author's contributions: Conceived and designed the experiment: PJ RW. Performed the experiments: PJ. Analysed the data: PJ RW. Wrote the paper: PJ PS EG RW.

Nerbonne, 2009), aggregating 60 specific syntactic variables, as represented by 60 questions in the SADS (Table 3.1), to build a linguistic distance measure.

Switzerland features a diverse surface topography, which in mountainous areas places constraints on transportation and communication routes, and thus imposes barriers to potential language contact between speakers (Figure 1.1). We therefore use different geographic distance measures to operationalise the possibility of language contact and to calculate correlations with the linguistic distance. Besides Euclidean distance (distance 'as the crow flies'), travel times of different points in time (1850, 1950, 2000) are also included in the analysis. Finally, while previous studies by other authors were restricted to computing the correlation between linguistic variation and spatial distance at the global level only, we also study the correlation at local levels of smaller geographic areas, thus exploring local effects such as topographic barriers and, conversely, interconnectedness.

Besides physical barriers, other, (socio)cultural factors are considered to have influenced the evolution of Swiss German dialects (e.g., Hotzenköcherle, 1961), such as administrative subdivision and isolation in the times before the modern Swiss Confederation was formed, or religious borders (Roman catholic vs. protestant). Our interest is thus in finding out the degree to which geographic distance may explain linguistic variation, considering that a large part of the variation may be owing to the aforementioned factors. Using different points in time for the representation of travel times is motivated by the fact that convergence effects are noticeable in the evolution of Swiss German varieties (Christen, 1998). At the same time, syntax is assumed to differentiate at a slower rate over time than other linguistic levels (Longobardi and Guardiano, 2009). Thus, we expect the correlation to be best for the older dates of travel times.

Our work departs from the following hypotheses:

- **H1:** Geographic distance is responsible for, and thus explains, the majority of the variance ($R^2 > 0.5$) found in Swiss German syntax, as represented in the SADS data.
- **H2:** Among the geographic distance measures, travel time measures better reflect syntactic spatial variation than Euclidean distance.
- **H3:** Older travel times better represent syntactic spatial variation.

4.1.2 State of the art

With the generally relevant literature presented in Chapter 2 about the correlation of linguistic differences with geographic distances, in this section we mention only the immediately relevant studies that have impacted the study. Dialectometry has numerous studies where a certain linguistic distance is correlated with a value representing the effect of 'geography', usually Euclidean distance. The studies conducted

with syntactic data are also manifold. Spruit (2006) used an additive measure of differences to calculate a Hamming distance on the syntactic level, which he correlated with Euclidean distance. Spruit, Heeringa and Nerbonne (2009) investigated the influence of geography (expressed by Euclidean distance) for aggregate pronunciational, lexical and syntactic differences in Dutch. For the lexical and syntactic level, Goebl's (1982) Weighted Identity Value was used while for pronunciation, the Levenshtein distance.

Going beyond Euclidean distance, Gooskens (2004) was the first to compare Euclidean distances to modern and old travel times in order to establish which one correlates better with Levenshtein distances and perceptual distances calculated for Norwegian dialects. Euclidean distance and modern travel times produced the same correlation, while older travel times correlated better with both the Levenshtein and the perceptual distances. Haynie (2012) established the utility of cost distance modelling for studying historical language contact networks of Miwok languages in California, correlating it to a metric of recurrent sound correspondence. As mentioned above, the seeming universality of the correlation of increasing linguistic differences with increasing geographic distance is expressed in the Fundamental Dialectological Postulate (FDP; Nerbonne and Kleiweg, 2007). Several authors subsequently have tested this hypothesis, with different results. While Shackleton (2007) investigated the correlation of phonetic distance to Euclidean distance for the traditional English dialects, reporting an R^2 of up to 0.77 (i.e., explaining 77% of linguistic variance) for a regression model accounting for regional differences, Szmrecsanyi (2012) found much lower values, using corpus-based data on morphosyntax of English dialects. In Szmrecsanyi's results, Euclidean distance explains a mere 4% of morphosyntactic variance; travel times fare only slightly better at 8%; Trudgill's (1974) Linguistic Gravity explains 24%; and finally the maximum value of 32.5% is reached after clustering dialects into dialect groupings. Szmrecsanyi (2012:226) concludes that "It is fair to say that the FDP has failed this test."

From a geographical perspective all of the above studies suffer from the crucial drawback of restricting the analysis to the – geographically speaking – global level, computing correlations for entire study areas, rather than exploring linguistic variation in more detail at the local level. With the discovery of regional differences in correlation structures, one could deliver possible explanations of regionally different linguistic variation patterns. With additional local analyses, it may be possible to explain why high degrees of correlation had been reported in some studies, and low correlations in others.

4.2 Data

The SADS database, used in this study has been in detail described in Chapter 3. In this study, 60 variables in the SADS were used, with each variable corresponding to one aspect of one survey question (note that a survey question may cover more

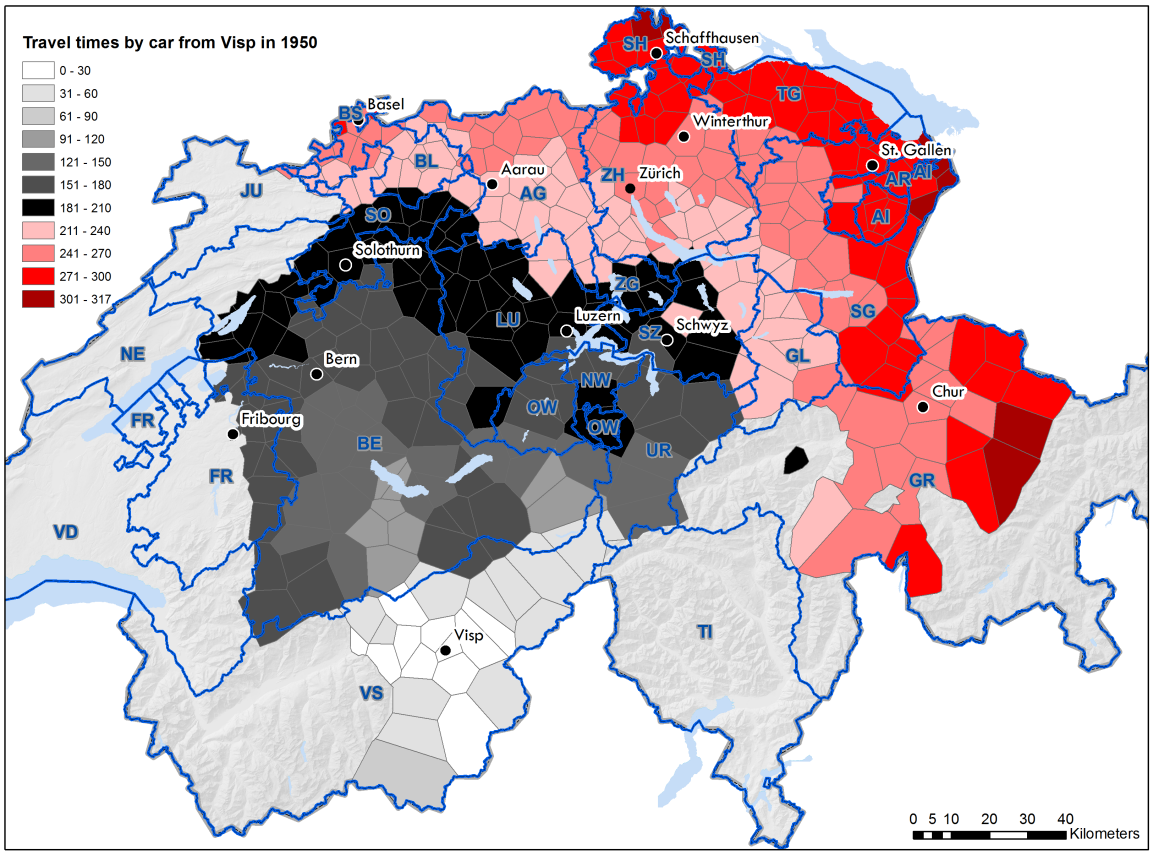

FIGURE 4.1: Travel times by car from Visp in 1950. Travel times that are longer than the ones that exist in 2000 are represented with a red colour scale, representing the development of 50 years in travel times.

than one variable). This subset, which was selected in collaboration with the SADS group, covers the majority of the 50 morphosyntactic phenomena investigated in the SADS. The variables used in the study are listed in Table 3.1.

In order to increase the comparability with other studies, we worked with the preferred variants in the case of MC questions.

The travel times used in the study are also discussed in detail in Section 3.3. As mentioned before, in this study we use travel times for years 1850, 1950 and 2000, given for individual transport.

Each of the three time points is representative of a particular point in the development of the Swiss transportation infrastructure. 1850 is representative of the transportation network before railroads were built in Switzerland (the only railroad line that opened earlier than 1850 measured a mere 20 km in 1847). 1950 is representative of a road and train network before motorways, new fast train connections and commuter train systems were built. Finally, in 2000 the network of national motorways was fairly complete, and 2000 also represents the year when the surveys for the SADS started. Over time, travel times were successively approximating the pattern of Euclidean distances, as some of the major topographic obstacles have been overcome. Figure 4.1 shows this effect. This map depicts the travel times by car in 1950,

centred on the alpine town of Visp. In 2000, the maximum travel time to the farthest place in the study area was 210 minutes. Any travel time that is higher than that is shown in red colours, representing how much longer travels took in 1950 compared to 2000. It should be noted that the majority of the population actually lives in the lowlands of the Swiss Plateau, which has a much denser transport network, with less impact of surface topography.

Swiss German syntax data is only available for the year 2000 (from the SADS), apart from some limited questions in former surveys (*Sprachatlas der deutschen Schweiz* (SDS) – Hotzenköcherle et al., 1962-1997), and thus absent for the other dates for which travel time data is available. For the purposes of this study, we assume that syntax changes at a slower rate over time than other linguistic levels (Longobardi and Guardiano, 2009), which means we expect historical language contact possibilities to be represented in today's syntactic landscape as well.

4.3 Methodology

As mentioned in the Introduction (Section 4.1), our methodology includes the development of a measure of linguistic distance (Section 4.3.1) suitable to deal with the SADS data, visualisation (Section 4.3.2) and the correlation of this measure (Section 4.3.3) with different geographic distances, including Euclidean distance as well as travel times. Furthermore, we conduct various further analyses (Section 4.3.4) to reveal local variation.

4.3.1 Calculating syntactic distance

Linguistic (dis)similarity in syntax data has often been computed using the Hamming distance (Spruit, 2006) or Goebl's (1982) Weighted Identity Value (Spruit, Heeringa, and Nerbonne, 2009). These measures define differences between two variants, assuming one variant per survey site. Since in the SADS multiple variants may be present at each survey site, we cannot assume the linguistic distance between two survey sites to be equal to the difference between the two variants.

In our study the occurrence of each variant at a given survey site is converted to ratios of the number of respondents using the particular variant divided by the total number of respondents answering at the given site. To calculate the syntactic differences for a given variable for a particular pair of survey sites, the proportions of answer variants at the two survey sites are subtracted from each other, resulting in a difference for the given survey site pair and given variant. The overall, aggregate syntactic distance is then calculated by adding up the differences for the variables of choice.

Figure 4.2 shows this procedure for two simplified survey questions (SADS Questions I.1, I.3) and two survey sites (Klosters, Flühli). For a single variable the maximum distance between two survey sites is 2, which may be reached if the two sites

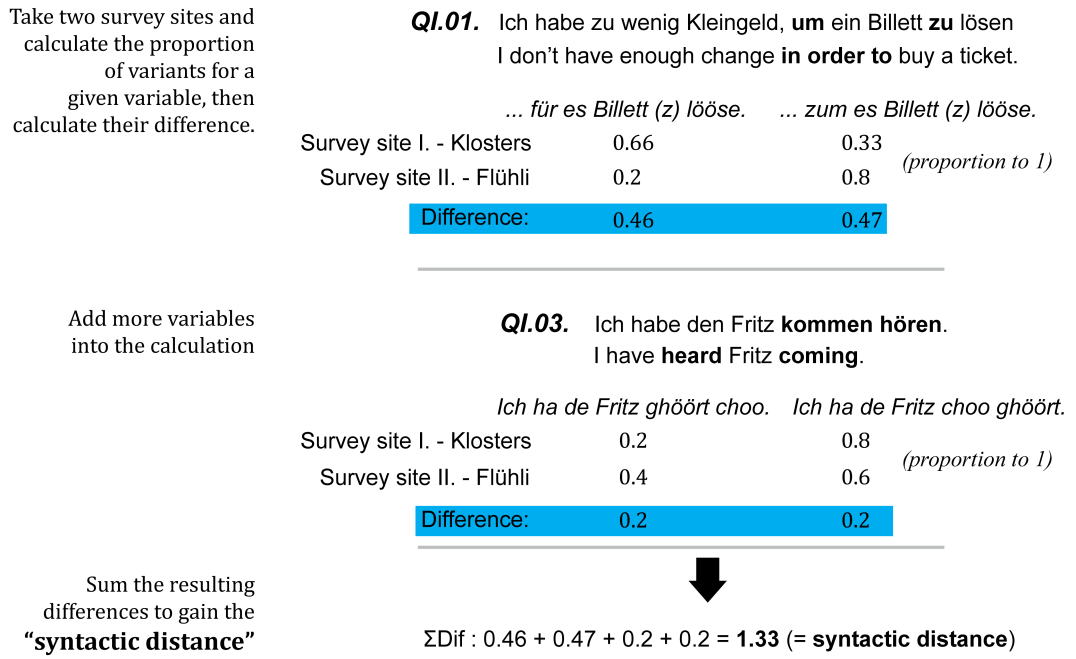

FIGURE 4.2: Flowchart of calculating the syntactic distance. Normally there are more than two answer variants for each question.

do not overlap in any variants at all (e.g., survey site A uses exclusively one variant while survey site B uses exclusively another). In our case, given 60 variables, syntactic distances for a survey site pair will range on a scale of 0 to 120. An equivalent method was first used by Speelman et al. (2003) for quantifying difference between language profiles and Pickl et al. (2014) to calculate linguistic distance between survey sites. In both works the resulting sum is divided by 2 to account for bidirectionality, but in our case this is not needed if only the linguistic distance is calculated.

4.3.2 Visualisation of syntactic distances

The results of this syntactic distance calculation can already reveal a lot about the relationship of geographic location and language variation simply by visualisation, comparable to maps of Goebl's identity values (e.g., Goebl, 2010). To this end, we create maps of syntactic distances centred on particular survey sites both for the entire study area as well as local subsets, and a global map of average syntactic distance per site.

4.3.3 Correlation of syntactic and geographic distances

Once the pairwise syntactic distances have been computed for all survey site pairs in the dataset, it is then possible to compute the correlation between the linguistic and the geographic distances over all survey site pairs. We use three correlation

measures: two measures of linear correlation – the Pearson product-moment correlation coefficient and the distance-oriented Mantel test (Mantel, 1967), similarly to Scherrer (2012), Haynie (2012) and Grieve (2014) – and logarithmic correlation. Furthermore, regression models of different types (linear, logarithmic) are fitted to the distributions of syntactic distance against geographic distance, separately for each type of geographic distance measure used. In most former dialectometric investigations not directed at syntax (e.g., Heeringa and Nerbonne, 2001; Nerbonne, 2009; Nerbonne, 2010b; Pickl, Spettl, et al., 2014) a logarithmic model better described the relationship between the linguistic and geographic distances, respectively. On the other hand, Spruit (2006), using syntactic data of Dutch dialects, found better agreement with linear correlation and Stanford (2012:273) found that the 'patterns of dialectometry' in general do not necessarily apply in smaller areas.

4.3.4 Local analyses

In order to further study the morphosyntactic variation at the local level, we carry out three types of local analyses. They enable the study of potential barrier and contact effects at the local level. Also, they allow the comparison of patterns of syntactic variation between different survey sites. The first two of these local analyses are identical to their global counterpart, but restricted to local subsets of the study area: First, we create maps of syntactic distances for local subsets, centred on a particular survey site and second, we carry out a correlation analysis for the subsets using the same linear and logarithmic methods as at the global level. In order to explore the effect of topography, we use spatial subsets in a mountainous area and in an area with gentle topography, respectively. The third analysis is local in the sense that it allows highlighting local deviations from the global regression model using geographic distance as a predictor of linguistic difference. To this end, we compute the residuals of syntactic distance and geographic distances (Euclidean, travel time) (Section 4.3.5), again centred on a particular survey site.

4.3.5 Residuals of syntactic and geographic distances

If the values of the syntactic distances as well as the geographic distances are both normalised to the interval [0..1], the differences (i.e., residuals) of the syntactic minus the geographic distance values can be calculated for each survey site, in relation to a reference site. These residuals can then be visualised either in scatterplots or in area-class maps for each particular reference site. The residuals are indicative of the degree of agreement between the syntactic and geographic distances.

4.3.6 Implementation

The statistics software R was used for the computation of syntactic distances, correlation, regression analysis and statistical testing, as well as diagrams. Packages *plyr*,

ggplot2 were used for preparing and plotting the data, while *ade4* was used for the Mantel test. The GIS software ArcGIS was used for producing the maps.

4.4 Results

The maps presenting the results of this study feature two areas. First, the whole investigation area of the SADS surveys (383 survey sites), which roughly covers the German-speaking area of Switzerland (referred to as the *global* area). Second, one of our specific *local* subsets, consisting of 46 survey sites in the Swiss cantons of Berne and Valais (German: *Wallis*), respectively. It is thus referred to as *BEOV* – short for 'Bernese Oberland and Valais'. This local region is dissected by a major ridge of the Swiss Alps, forming the border between the two cantons and an important topographic barrier. The Bernese part of this local subset represents the alpine part of the canton of Berne, also known as the *Bernese Oberland*. This area is characterised by a network of many valleys that are partially deep and secluded. The part in the Valais, located to the south, is dominated by a large valley (the upper part of the *Rhône Valley*), with some side valleys, most importantly the *Lötschental* valley.

The results for a third data set – a local subset of 46 survey sites located on the Swiss Plateau (*'Mittelland'*) between the cities of Aarau, Solothurn and Berne – are presented only for the local correlation analysis (Section 4.4.4), and not shown in maps. This data set, with the short name *ML46*, serves as an example with gentle and homogenous topography, which also allows for better transport connections, therefore possible direct contact throughout the area.

4.4.1 Maps of syntactic distance

Figures 4.3 and 4.4 show the survey sites of the entire study area as Voronoi polygons (similarly to e.g., Goebl, 1983, 2010) coloured according to their syntactic distance from a particular reference place, with the borders of the Swiss cantons overlaid. We chose two places in Switzerland to present different spatial patterns of syntactic distance variation over the topographically diverse investigation area. Schaffhausen, a city on the Swiss Plateau, close to the German border, serves as reference site in Figure 4.3, while the city of Freiburg (French: *Fribourg*), which is located at the far west end of the German-speaking area and actually has a French-speaking majority, is the anchor in Figure 4.4.

Figure 4.5 depicts for each survey site the average syntactic distance to all other survey sites, thus representing how different the given survey site is from all others, the darkest coloured polygons indicating the most dissimilar varieties.

4.4.2 Scatterplots and correlation analysis

Figure 4.6 depicts the syntactic distances plotted against the Euclidean distance for all the survey site pairs, with the linear and the logarithmic regression lines overlaid.

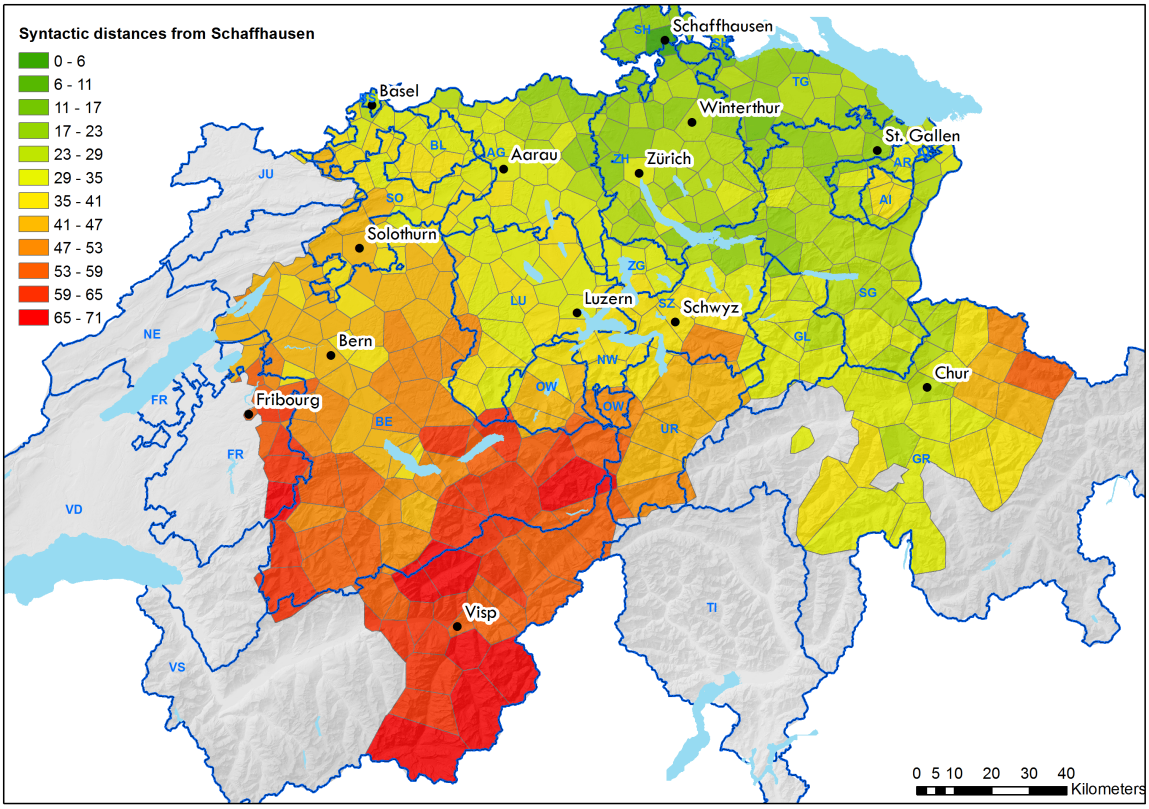

FIGURE 4.3: Syntactic distances of the survey sites compared to Schaffhausen.

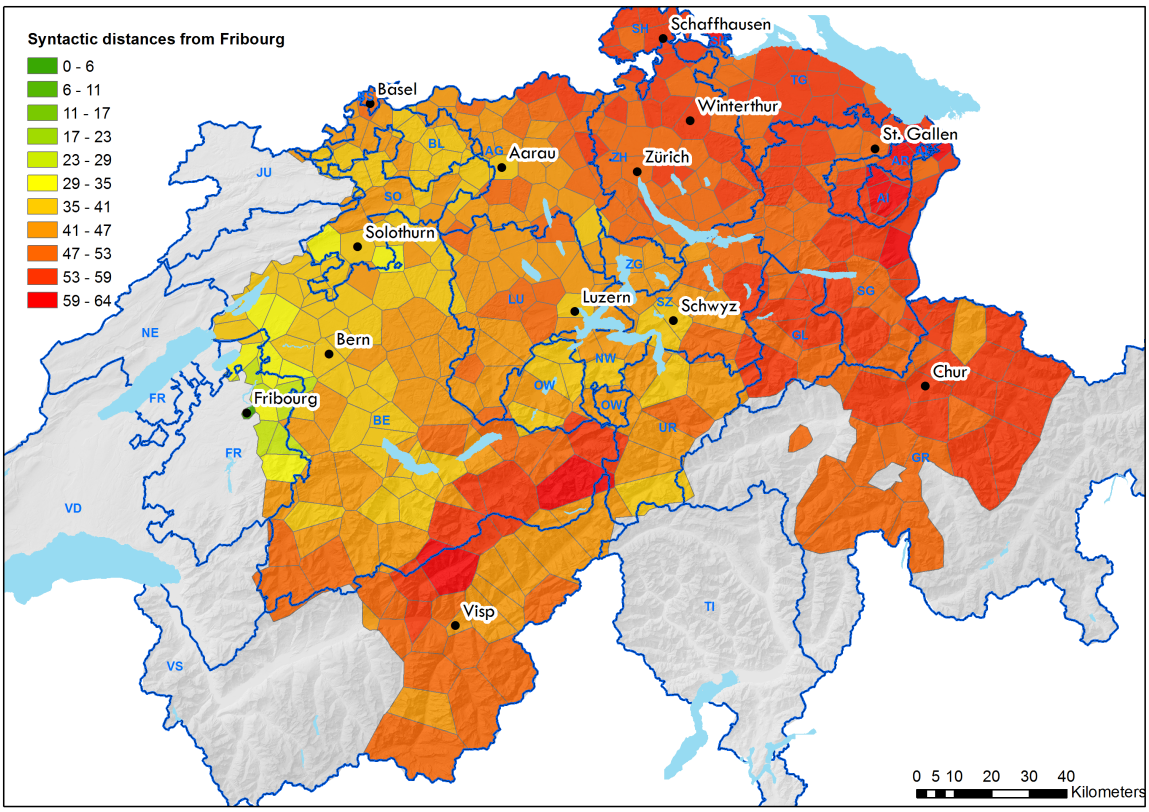

FIGURE 4.4: Syntactic distances of the survey sites compared to Freiburg (French: *Fribourg*).

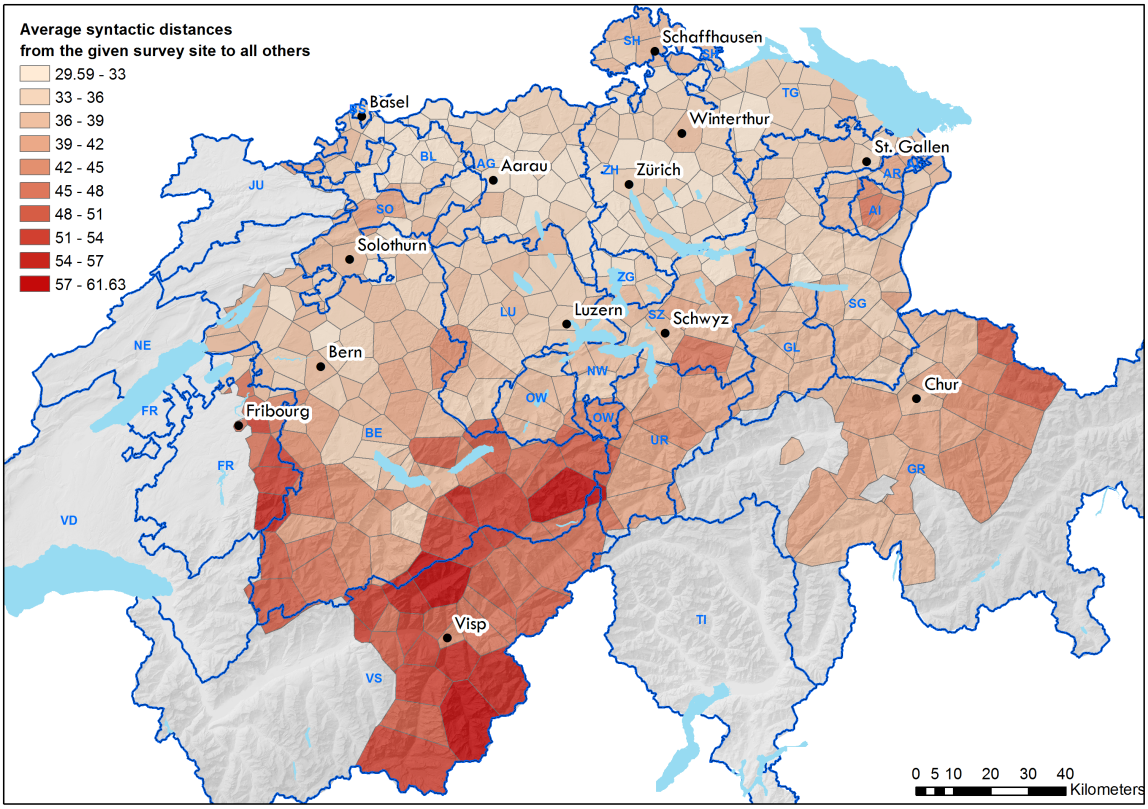

FIGURE 4.5: Average syntactic distances to all other survey sites.

The survey sites included in the BEOV subset are highlighted in green. Figure 4.7 shows three scatterplots, where the syntactic distance is plotted against the travel times for the three dates, 2000, 1950 and 1850, respectively, again with the linear and the logarithmic regression lines.

Table 4.1 presents the results of the correlation analysis of the syntactic distance with the geographic distances for the entire study area. The table shows the coefficients (r) obtained from the Pearson product-moment correlation and the Mantel-test methods, respectively, along with the coefficients from the logarithmic correlation. The resulting R^2 from the regression analyses is also shown, indicating the extent to which the geographic distances account for the variance of the syntactic distance.

4.4.3 Maps of syntactic distance for the BEOV subset

Figures 4.9 and 4.10 show the syntactic distances for the BEOV subset. For comparability the same colour scheme is used as for the corresponding maps of the whole study area (Figures 4.3 and 4.4). Figures 4.9 and 4.10 are centred on Blatten (Valais) and Grindelwald (Berne), respectively. Roads suitable for cars are also featured on these maps to give an impression of the main transport connections and mountain passes. In the past (i.e., 1850), these roads were used by horse carriages and stagecoaches, further mountain passes could have only been traversed by foot or mules.

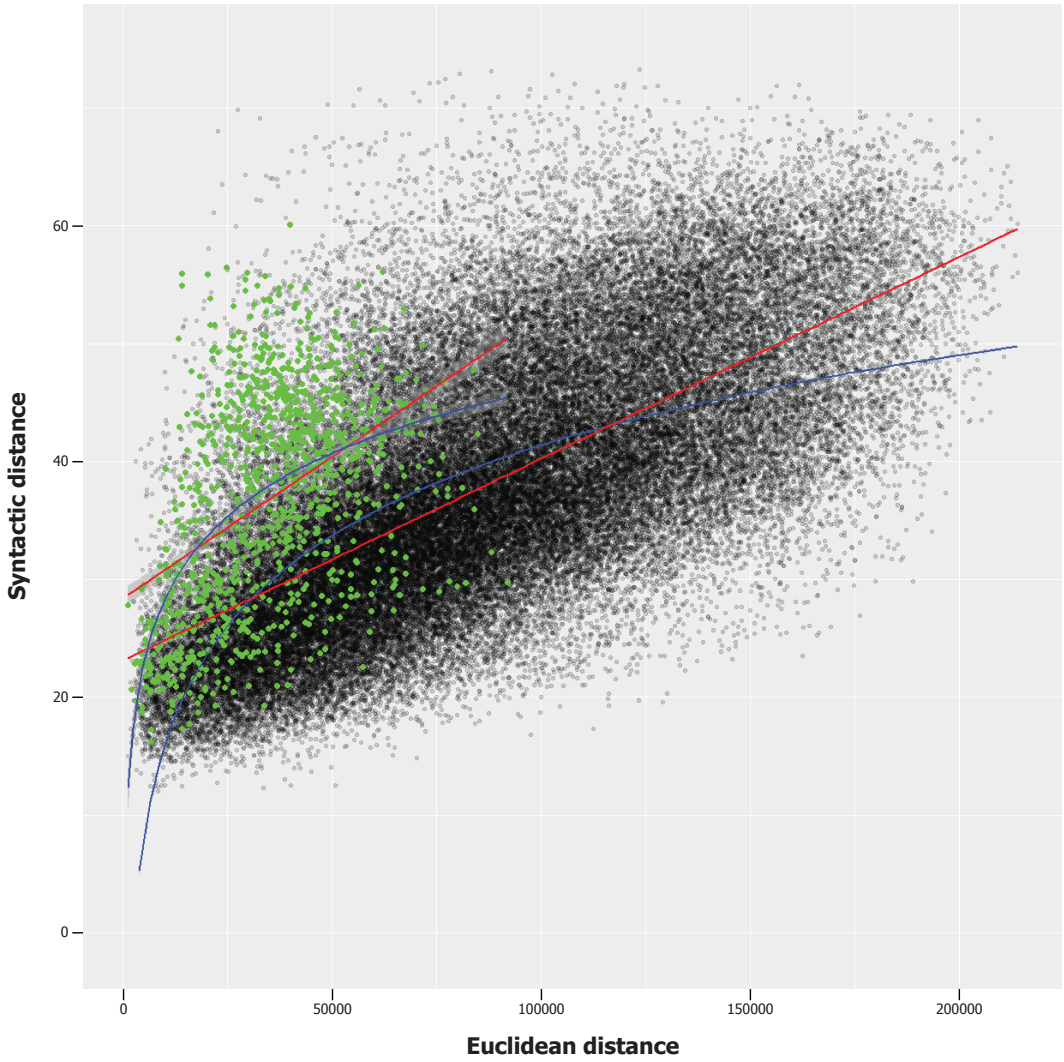

FIGURE 4.6: Syntactic distance plotted against the Euclidean distance [m]. The survey site pairs included in the BEOV subset are highlighted in green. Linear regression line is shown in red, logarithmic regression line in blue.

TABLE 4.1: Correlation coefficients of the syntactic distance with the different geographic distances, as well as the explained variance R^2, for both linear and logarithmic regression analyses. All 383 survey sites. *Travel time data for 1850 was available for only 120 survey sites in the investigation area.

Geographic distance	**Pearson's correlation**		**Mantel-test**		**Logarithmic correlation**	
	r	R^2	r	R^2	r	R^2
Euclidean distance	0.676	0.458	0.65	0.422	0.65	0.424
Travel times in 2000	0.775	0.599	0.76	0.577	0.744	0.553
Travel times in 1950	0.778	0.605	0.768	0.590	0.743	0.552
Travel times in 1850*	0.783	0.612	0.763	0.582	0.737	0.544

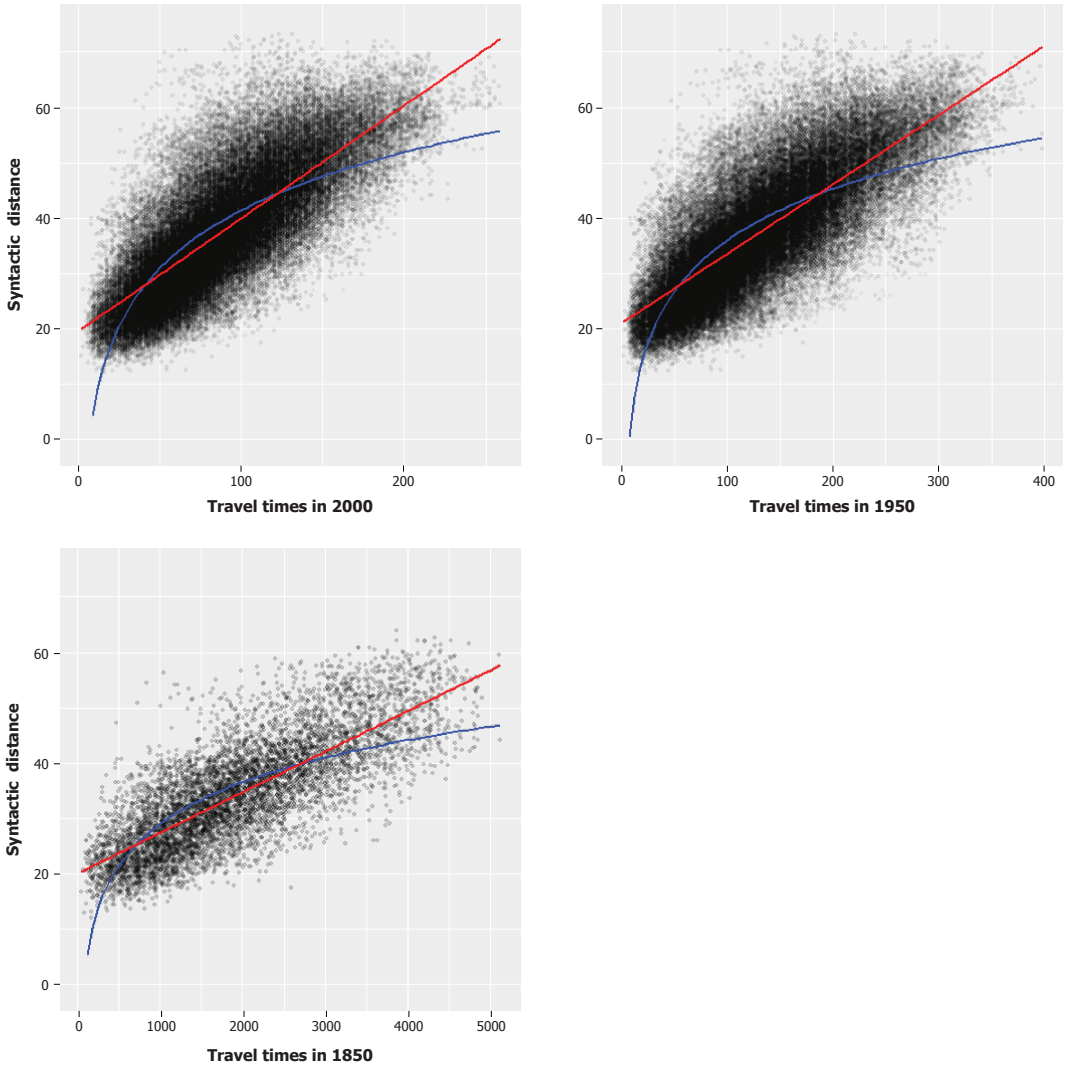

FIGURE 4.7: Syntactic distance plotted against the travel times in minutes. Top left: against travel times in 2000; top right: against travel times in 1950; lower left: against travel times in 1850. Linear regression line shown in red, logarithmic regression line in blue.

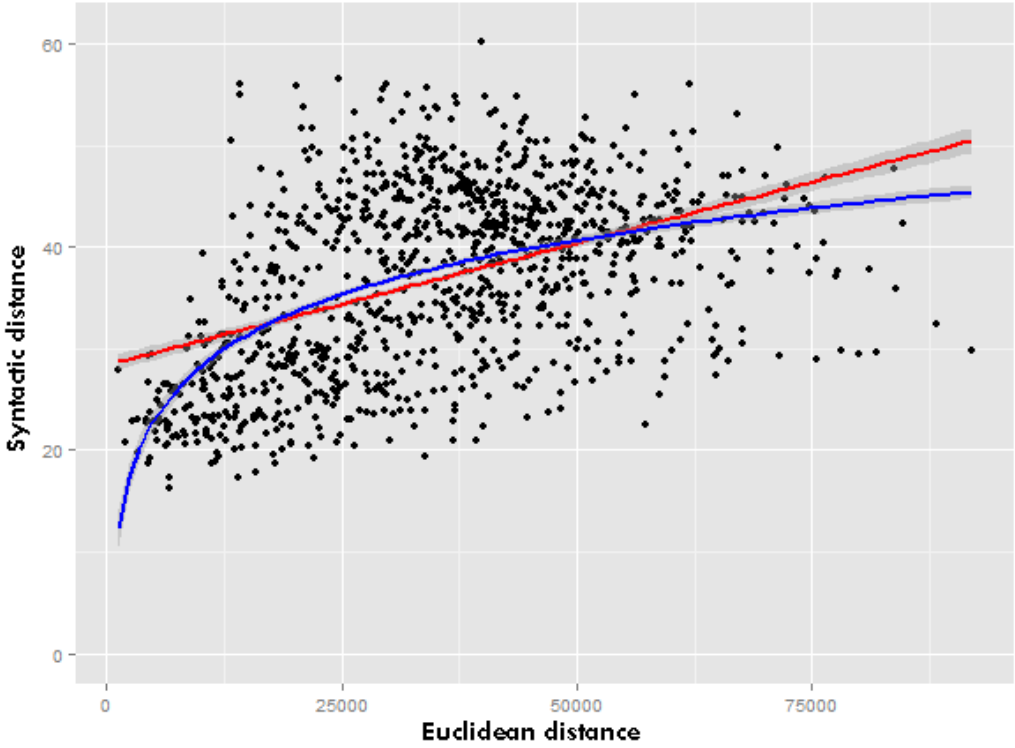

FIGURE 4.8: Syntactic distance plotted against the Euclidean distance [m] in the BEOV subset. Linear regression line shown in red, logarithmic regression line in blue.

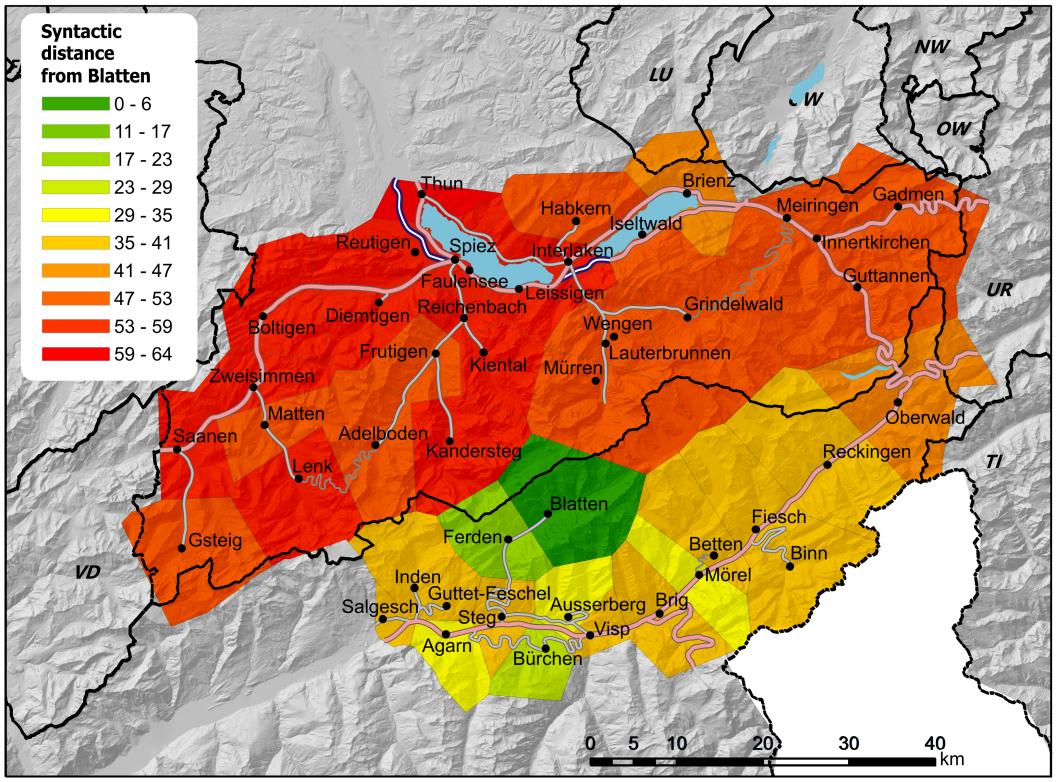

FIGURE 4.9: BEOV subset map centred on Blatten. Note that the Voronoi polygons used for area-class display do not respect the borders between cantons.

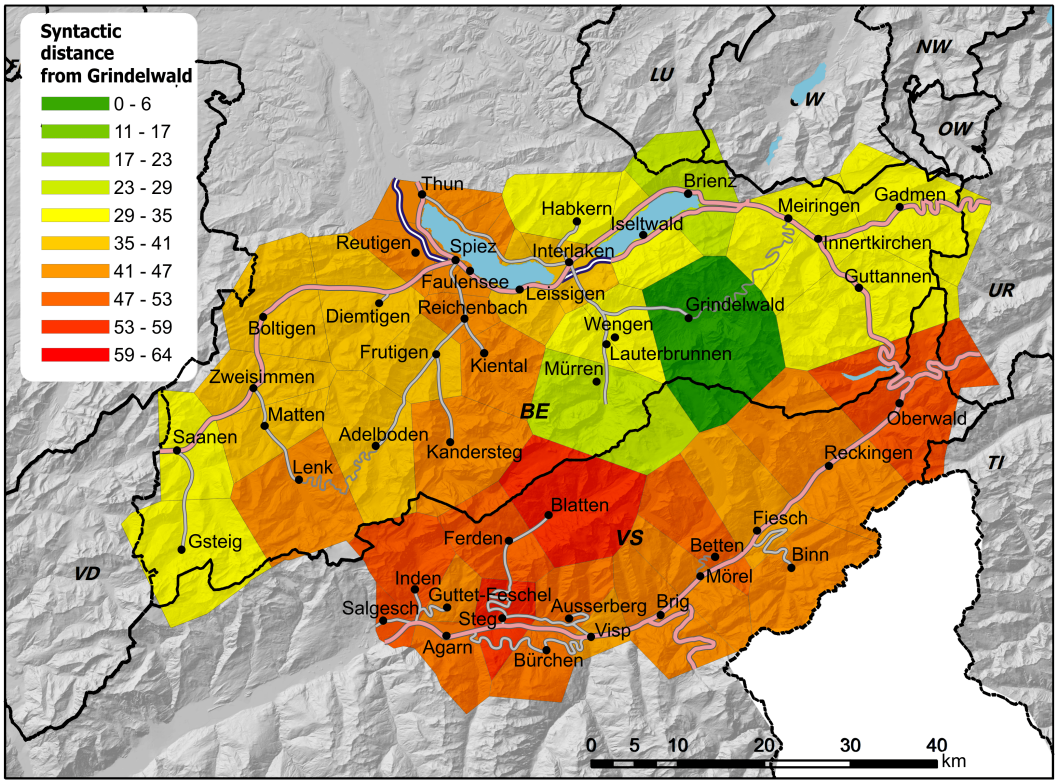

FIGURE 4.10: BEOV subset map centred on Grindelwald.

4.4.4 Scatterplots and correlation analysis of the local subsets

Figure 4.8 plots the syntactic distances between survey sites against the Euclidean distance while Table 4.2 presents for the BEOV subset the results of the Pearson correlation analysis of the syntactic distance with the geographic distances, and with their logarithms. In the same manner, Table 4.3 shows the results of the correlation analysis for the ML46 subset. In order to assess whether correlation coefficients differ significantly from each other, they were tested using Fisher's z-transformation (Lowry, 2000; Warner, 2013), with results shown in Table 4.4 for the global level and the BEOV subset.

4.4.5 Residuals of syntactic and geographic distances

Figure 4.11 plots the residuals of syntactic distance and Euclidean distance (shown on the y-axis) against the Euclidean distance of all survey sites relative to the alpine village of Obersaxen. Obersaxen was chosen as it is located in the periphery of the study area and scored a moderate average syntactic distance to all other survey sites (see Figure 4.5). Figures 4.12 and 4.13, respectively, then map the residuals to geographic space. Figure 4.12 does so for the residuals in Figure 4.11. Figure 4.13 shows the residuals of syntactic distance and travel times in 1950, relative to Freiburg. The patterns of residuals may differ considerably depending on the reference site and the type of geographic distance used; these maps are thus to be understood as examples.

TABLE 4.2: Correlation coefficients of the syntactic distance with the different geographic distances, as well as the explained variance R^2, for the linear and logarithmic regression analyses. BEOV regional subset. ** Travel time data for 1850 was available only for 11 survey sites in the BEOV subset.

Geographic distance	**Pearson's correlation**		**Logarithmic correlation**	
	r	R^2	r	R^2
Euclidean distance	0.445	0.198	0.519	0.27
Travel times in 2000	0.674	0.455	0.694	0.482
Travel times in 1950	0.727	0.53	0.749	0.562
Travel times in 1850**	0.815	0.665	0.811	0.657

TABLE 4.3: Correlation coefficients of the syntactic distance with the different geographic distances, as well as the explained variance R^2, for the linear and logarithmic regression analyses. ML46 regional subset. *** Travel time data for 1850 was available only for 19 survey sites in the ML46 subset.

Geographic distance	**Pearson's correlation**		**Logarithmic correlation**	
	r	R^2	r	R^2
Euclidean distance	0.543	0.295	0.535	0.287
Travel times in 2000	0.577	0.333	0.557	0.31
Travel times in 1950	0.558	0.312	0.547	0.3
Travel times in 1850***	0.607	0.369	0.554	0.307

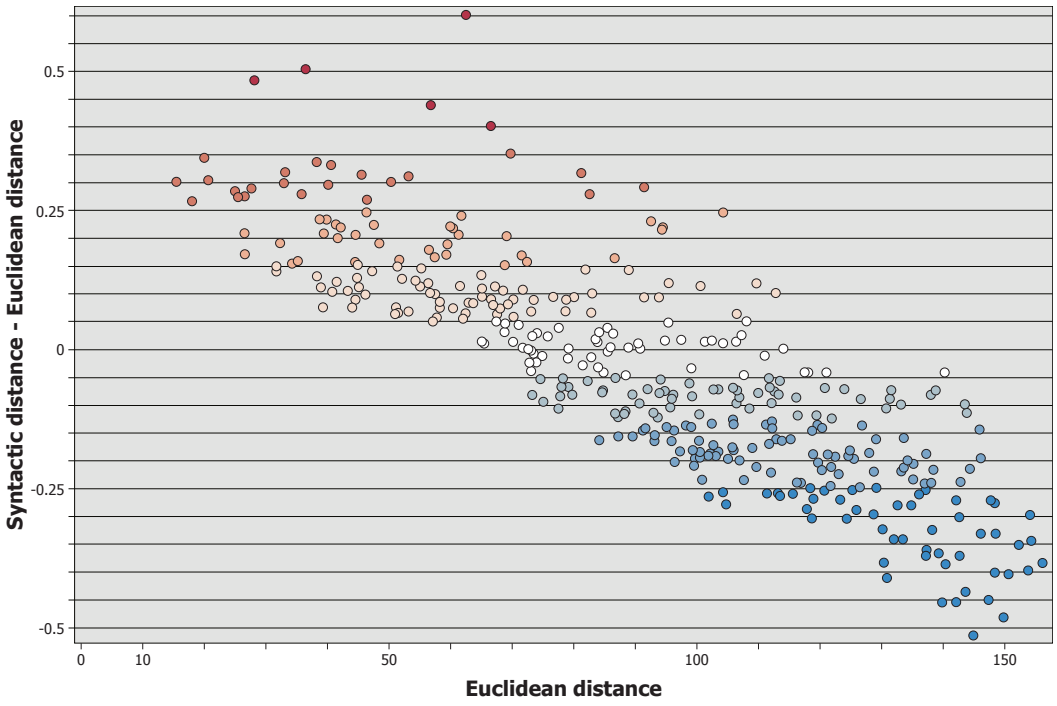

FIGURE 4.11: Scatterplot of the residuals of syntactic distance and Euclidean distance plotted against the Euclidean distance [km] from Obersaxen to all other survey sites.

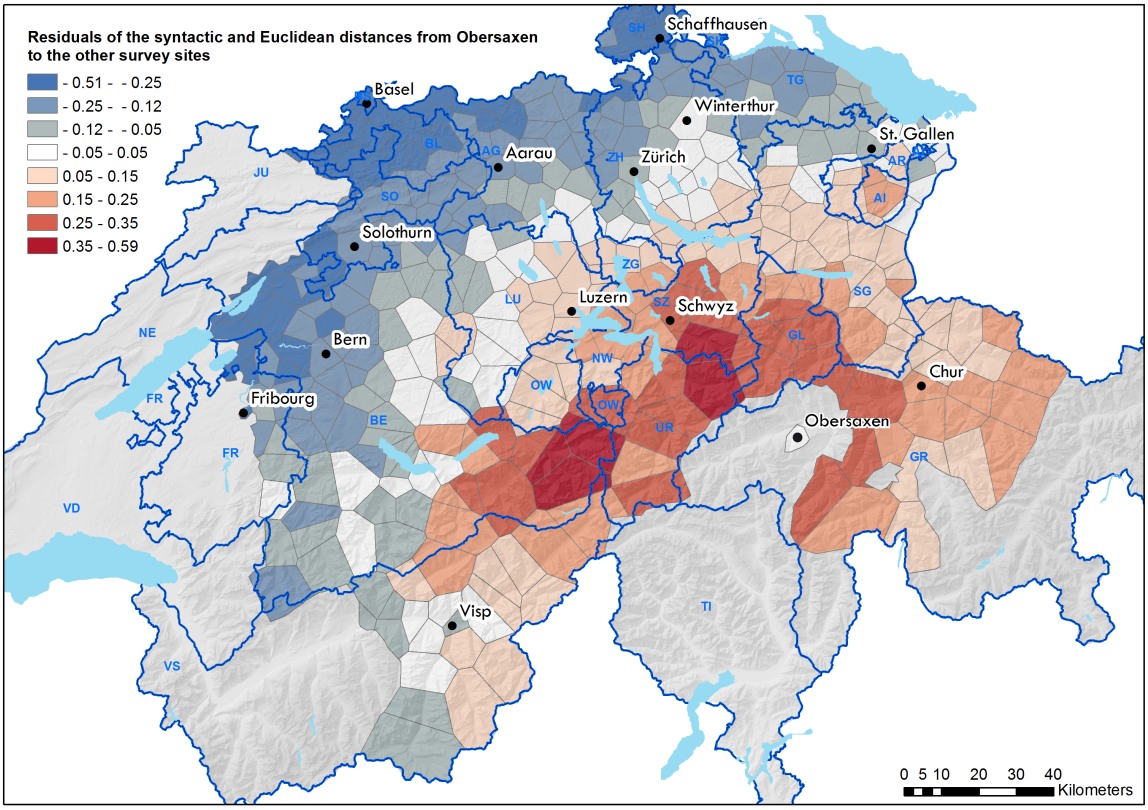

FIGURE 4.12: Residual map of Obersaxen showing the residual values between the normalised syntactic distance and the Euclidean distance. If the Euclidean distance is greater than the syntactic distance, it will yield a negative residual; conversely, if the Euclidean distance is lower than the syntactic distance, a positive residual is received.

TABLE 4.4: Results of the Fisher z-transformation (p-values) to test whether the difference between correlation coefficients is significant. Significant differences ($p < 0.05$) between the two values highlighted in green.

Similarity of the Pearson product-moment correlation of the given geographic distances to the syntactic distance; P-one-tailed values	Global Euclidean distance (0.676)	Global Travel times 2000 (0.775)	Global Travel times 1950 (0.778)	Global Travel times 1850 (0.783)
Global Euclidean distance (0.676)		0.002	0.0014	0.015
Global Travel times 2000 (0.775)			0.4562	0.4247
Global Travel times 1950 (0.778)				0.4522
Global Travel times 1850 (0.783)				
BEOV subset Similarity of the Pearson product-moment correlation of the given geographic distances to the syntactic distance; P-one-tailed values	BEOV Euclidean distance (0.4446)	BEOV Travel times 2000 (0.6738)	BEOV Travel times 1950 (0.7268)	
BEOV Euclidean distance (0.4446)		0.0571	0.0197	
BEOV Travel times 2000 (0.6738)			0.3121	
BEOV Travel times 1950 (0.7268)				

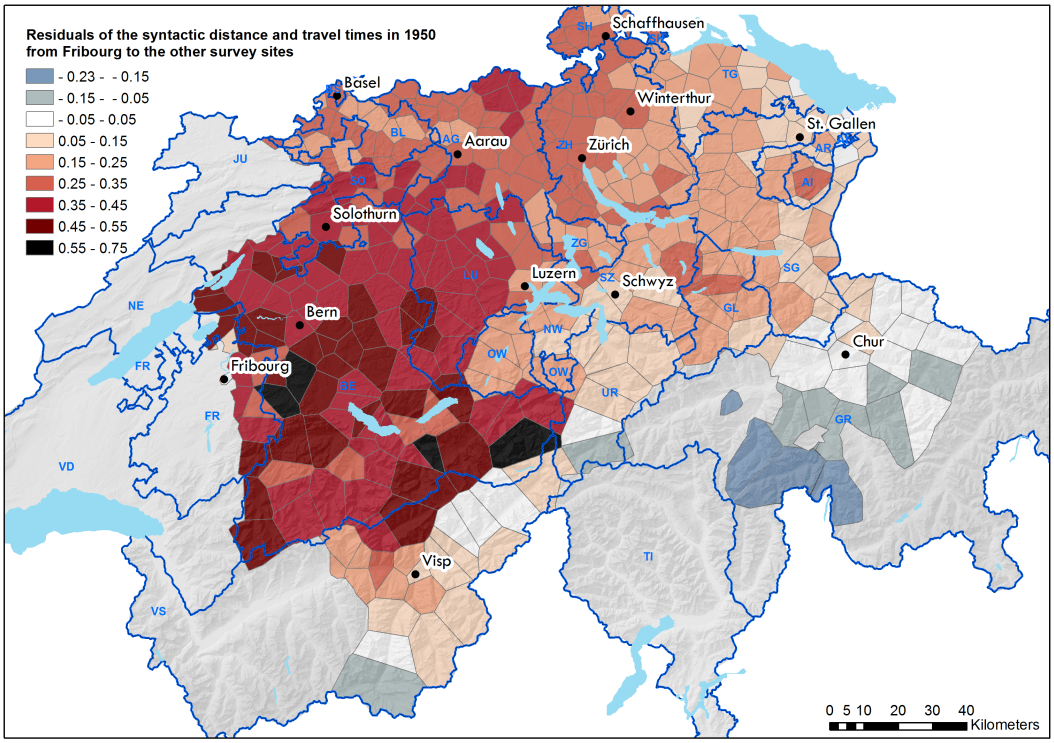

FIGURE 4.13: Residual map of Freiburg (French: *Fribourg*) shows the residual values between the normalised syntactic distance and the normalised travel times in 1950. For an explanation of the meaning of residuals, see Figure 4.12

4.5 Discussion

4.5.1 Syntactic distance measure

As explained in Section 4.3.1 we used an aggregative distance measure to express the linguistic (dis)similarity between survey sites, accommodating the fact that in the SADS survey multiple answers are provided per survey site. With this approach, our results are comparable to those by other authors who used similar measures in similar studies (Speelman, Grondelaers, and Geeraerts, 2003; Pickl, Spettl, et al., 2014). However, in our measure we did not assign weights to any of the phenomena, or to single answer variants for the survey questions, although it can be assumed that between certain types of answers, some of the differences are more pronounced and that some syntactic phenomena are more salient than others, at least from a perceptional point of view. We used a linear summation, not taking into consideration the potential mutual correlations between the answer matrices of the survey questions, essentially assuming independence between the variables. Establishing weights for each of the phenomena and each of their answer variants, however, would be a very tedious work and prone to subjective judgement. In practice, since we are forming our aggregate measure from a large number of variables, the resulting distance values are deemed to be realistic, as the weights and uncertainties of the various variables cancel out by aggregation, as posited in Nerbonne's (2009) work.

4.5.2 Global maps of syntactic distance

The maps of syntactic distance can be used to visualise patterns of variation of syntactic distance across the study area. Depending on the choice of the survey site on which the map is centred, the syntactic distances will show remarkably different patterns. Figure 4.3, which is centred on Schaffhausen, exhibits a largely concentrical progression of syntactic distance values with increasing geographic distance, suggesting support of the FDP. However, in that same map, we can also see breaks in this progression, particularly at canton borders, which often also form old historical and cultural (e.g., denominational) borders. The second effect that we may perceive is the influence of topography. For instance, in Figure 4.3 the progression of syntactic distance values is flatter in the lowlands of the Swiss Plateau (cf. Figure 1.1), and steeper towards the more mountainous areas, except for the upper Rhine valley, which provides good accessibility towards the city of Chur. This topographic effect, we would think, should also show in the correlation analysis, where we would expect higher correlation values for the travel time distances, as opposed to the Euclidean distance.

Figure 4.4, centred on Freiburg, shows a somewhat patchier pattern than the previous map. The two effects – canton borders and topographic effect – appear less pronounced, except for a relatively stark difference along the border between the cantons of Fribourg (German: *Freiburg*) and Berne, which also forms a strong

denominational border (Roman catholic vs. protestant). Clearly, the progression of syntactic distance with increasing geographic distance is steeper in Figure 4.4 than in Figure 4.3, suggesting that Freiburg is syntactically more different from other survey sites than is the case for Schaffhausen.

Figure 4.5 is depicting the average syntactic distances from each survey site to all other sites. Owing to the fact that average values are shown, we now see a different picture, but some of the effects visible in the preceding maps are still noticeable. The effect of topography becomes noticeable in two ways. First, differences in average syntactic distances on the Swiss Plateau are only rather subtle, hinting at better possibilities of communication. And second, the values are generally higher in the Alps, reaching peak values in the Bernese Oberland and the Valais, areas characterised by high mountains and topographic barriers, which foster isolation. The high values in the Canton of Fribourg further support the observation that Freiburg is syntactically distinct from the other Swiss dialects. The rather clear discontinuity between cantons of Fribourg and Berne might be less an effect of topography (which is little pronounced in this border area) than it might be caused by isolating denominational differences between a catholic canton (Fribourg) and a protestant (Berne) canton.

4.5.3 Global scatterplots and correlation analysis

Scatterplots of geographic distance against syntactic distance graphically show the strength and the direction of correlation between the two variables, and they also allow fitting regression lines to the data points. As mentioned in Section 4.3.3, we use linear and logarithmic regression. Figure 4.6 exhibits a positive correlation between syntactic distance and Euclidean distance for the entire SADS data set. As posited by the FDP, syntactic distance grows with increasing Euclidean distance. This is further underlined by the linear and the logarithmic lines of best fit, however it is hard to tell solely by looking at the figure whether the linear regression line or the logarithmic line fits better (which contradicts findings of earlier studies, e.g., Heeringa and Nerbonne, 2001; Nerbonne, 2009, 2010; Pickl, Spettl, et al., 2014, where the sublinear patterns were unequivocal). Although the correlation is obviously not perfect, as the data points deviate considerably from the regression lines, nevertheless a rather elevated strength of correlation is visible. The green points in Figure 4.6, highlighting the survey site pairs included in the BEOV subset, allow for making two qualitative observations. First, the syntactic distances in this subset tend to be higher than for other sites at equal Euclidean distance. And second, the syntactic distances appear to show a greater degree of variation, thus suggesting that the strength of correlation might be smaller in the BEOV subset than in the entire study area (cf. Section 4.5.5).

Figure 4.7 plots syntactic distance against the travel times in the years 2000, 1950 and 1850, respectively. Note that the length of the horizontal axis has been adjusted so that it fits the range of the travel times for the three reference years. Hence, each of the three subfigures of Figure 4.7 shows a different range of travel times in minutes, but on the same graphical length of the scatterplot, thus alleviating comparability.

Visually, one gets the impression that the strength of correlation and thus the fit to the regression line might be higher than in Figure 4.6. The direction of correlation remains the same.

The numerical correlation analysis for the entire SADS data set, summarised in Table 4.1, brings about further insights. The correlation coefficients of the syntactic distance with the geographic distances at the global level always show a strong positive relationship of at least 0.65 for both linear methods of correlation (i.e., Pearson product-moment correlation and the Mantel test). The R^2 are between a low of 42.19% of explained variance of syntactic distance for the Mantel test and the Euclidean distance, and a high of 61.24% for the Pearson correlation and the travel times of 1850. All linear correlation coefficients are significant ($p < 0.05$) or highly significant ($p < 0.01$), independently of the correlation measure used. The best correlation in the linear case is obtained for the travel times of 1850, with the other travel times showing similar values. However, there appears to be a marked drop in the correlation strength when the Euclidean distance is used. In Section 4.5.5, we will analyse this difference in more detail, but we already note that these results hint at the topographic effect, which translates directly into travel times.

There is always a positive correlation with the logarithms of the geographic distances too, and they are in fact higher than the patterns commonly reported in dialectometric studies on pronunciation and phonology (Nerbonne's comparative study (2010) accounts for correlation coefficients between 0.469 and 0.622 using Euclidean distance). In our case, a linear model describes the relation between linguistic and geographic distances slightly better than a logarithmic model (the maximum difference between R^2 values is 6.8%, for the 1850 travel times), but the difference is in no case statistically significant. This means that for our syntactic distance both the linear and the logarithmic model are equally good predictors. This finding aligns with the results of Spruit (2008), who found that the relationship between syntactic and geographic distances could be slightly more accurately described with a linear function than with a logarithmic transformation. The Mantel test results are also very similar to the Pearson correlation coefficients; their difference is also in no case statistically significant. For both linear and logarithmic cases, the Euclidean distance yields markedly lower correlation values than the travel time measures. On the other hand, the monotonic progression visible in the linear correlation measures for the travel times in 2000, 1950 and 1850, is no longer present for the logarithmic measures.

4.5.4 Maps of syntactic distance for the BEOV subset

To explore the correlations of the linguistic and geographic distances at the local level, the syntactic distances were also mapped for multiple subsets, with Figures 4.9 and 4.10 serving as examples on the BEOV subset. These maps also present the main roads, which provide the major routes of modern contact in this mountainous area. The major topographic feature in this area is the high mountain chain that forms

the border between the cantons of Berne and Valais. The Grimsel Pass between Guttannen and Oberwald, in use since the Middle Ages and open for cars, as well as the Lötschberg railway tunnel between Kandersteg and Ferden, opened in 1913 (with a new base tunnel in operation since 2007), provide connections across this topographic barrier.

Figures 4.9 and 4.10 use the same construction principle and colour scheme as the syntactic distance maps depicting the entire study area (Figures 4.3 and 4.4), and are centred on Blatten (Figure 4.9) and Grindelwald (Figure 4.10), respectively. Although the two maps show different patterns, we can clearly see the effect of the main mountain chain, which acts as a linguistic divider in both maps. The village of Blatten, the reference in Figure 4.9, is located at the back end of the Lötschental valley, a secluded side valley whose entrance from the main valley of the Valais is formed by a ravine with a vertical drop of some 700 metres, and which is separated from the canton of Berne by mountains exceeding 3000 and even 4000 metres. This particular location suggests how topography exerts an influence on linguistic differentiation. Ferden, which is shown to be syntactically close to Blatten, is located in the same valley, while there is already a marked difference visible to the other survey sites in the Valais, and a clear-cut difference to the sites across the mountain chain in the Canton of Berne. In general, the farther we follow the transportation routes, the higher the syntactic distances that we observe, in agreement with the FDP.

Figure 4.10 is centred on Grindelwald, a rather well accessible valley in the Bernese Oberland attracting many tourists. At first sight, the patterns we see are similar to those shown in Figure 4.9, only mirrored along the main mountain chain. However, at closer inspection, we see a different form of spatial variation, influenced by (so we believe) a different topography. Rather than a main valley with one major side valley, as in the case of the Valais, in the Bernese Oberland we find several separate valleys, as is clearly reflected in the road network. These valleys have quite different syntactic distances from Grindelwald, in some cases even reversing the colour scale with increasing geographic distance, and thus contradicting the FDP.

We also explored the potential 'bridging' effect of the Grimsel Pass and the Lötschberg tunnel. This effect would suggest that the syntactic difference between survey sites connected by a pass or tunnel should be smaller than between other sites. However, when we generated centred syntactic difference maps for all survey sites (the complete series not shown here), we found this effect to be rather weak and unsystematic.

Concluding the visual part of our analysis of the BEOV subset, we observe a mixed picture. On the one hand, topography appears to have a strong effect on linguistic differentiation, as shown very clearly through the main mountain chain, and also in the way survey sites of the same valley tend to be more similar than their neighbours in the next valley. This separating effect of topography was also observed in perceptual linguistics studies, where laypersons drawing dialect areas would intuitively link these to topographic features (Stoeckle, 2014:369, 519). On

the other hand, we also see deviations from this pattern, as seen on the example of Grindelwald in Figure 4.10, suggesting that other factors than geographic distance and topographic effects come into play.

4.5.5 Scatterplots and correlation analysis for the local subsets

Comparing Table 4.1, representing the results of the correlation analysis at the global level, and Table 4.2, representing the results for the local BEOV subset, we notice that the situation is quite different for the two data sets. With the exception of the travel times in 1850, correlation strength is higher for the global data set (Table 4.1) than for the local BEOV subset (Table 4.2). In Table 4.2 as well, syntactic distance correlates more strongly with travel times (which incorporate topography) than with Euclidean distance and that difference has become more marked, compared to the global data set (14.1% vs. 25.7% more variance explained by travel times). Lower correlation coefficients with Euclidean distances were indeed expected (as already suggested in the discussion of Figure 2), given the fact that throughout the BEOV region, topography crucially influences possibilities of contact. We further see that the more we move back in time, the correlation is monotonically increasing up to 0.815, suggesting that the separating effect of topographic barriers is decreasing as new, better transportation infrastructure is built. Note however, that travel times in 1850 were only available for 11 of the 46 survey sites in the BEOV subset.

In contrast to the global correlations represented in Table 4.1 (except for the case of 1850), in the BEOV subset the logarithmic correlation coefficients are higher than the linear Pearson correlation coefficients (Table 4.2). Differences between the respective correlation coefficients not being statistically significant however means that also at this local level a logarithmic model is equally good for describing the relation between syntactic and geographic distances as the linear model. This fact is also supported by Figure 4.8 where it is unclear whether the linear or the logarithmic regression line has a better fit. As Mantel-test results on the global level were not significantly different from the Pearson correlation coefficients, we did not include these results for the local subsets.

To further explore the impact of topography, the correlation analysis was also conducted for the ML46 subset, which features a very gentle topography, with results shown in Table 4.3. This subset aims to model the maximum direct language contact possible in a coherent spatial subset. The correlation coefficients obtained are very similar for the linear and logarithmic Pearson correlations and all geographic distance measures. In all but one instance, they are between 0.5 and 0.6, reaching a high of 0.607 for the Pearson correlation using the 1850 travel times. The values obtained for the Euclidean distance are the lowest for both correlation methods, however with very subtle differences to the travel time correlations. It is thus not surprising that none of the differences in correlations, when tested using Fisher's z transformation, came out as significant. We conclude that in this part of the study

TABLE 4.5: Correlation coefficients of the syntactic distance with the different geographic distances, as well as the explained variance R^2, for the linear and logarithmic regression analyses. Edge situation subset (Edge46). **** Travel time data for 1850 was available only for 10 survey sites in the edge situation subset.

Geographic distance	**Pearson's correlation**		**Logarithmic correlation**	
	r	R^2	r	R^2
Euclidean distance	0.692	0.488	0.675	0.456
Travel times in 2000	0.775	0.601	0.752	0.566
Travel times in 1950	0.775	0.600	0.750	0.563
Travel times in 1850**	0.734	0.539	0.687	0.472

area, characterised by a dense transportation network with no considerable topographic barriers, the effect of topography – represented by the difference between Euclidean distance and travel times – does not play out as much as on the global level, and not nearly as much as in the mountainous BEOV subset.

In turn we can also assume that the elevated level of contact possibilities also lends more opportunity for other, mainly socio-demographic variables to impact the linguistic differences, leading to lower correlations with geographic distances. When including all survey sites in the analysis, the majority of distances between survey site pairs are too big to have direct linguistic contact present. Thus, we posit that the effect of socio-demographic variables is suppressed by the effect of distance, resulting in higher correlations with geographic distances on the global level, than in local subsets. To test the assumption that geographic distance has a greater effect if no direct contact is possible, we analysed another subset of survey sites (termed *Edge46*, also *N*=46) where presumably little to no direct language contact is present. We systematically sampled the survey sites at the edge of the investigation area to simulate the distribution with the farthest distances possible. We assume that in this subset we can model the effect of geographic distances, undisturbed by direct contact (thus as clear of socio-demographic factors as possible). The resulting correlation coefficients (Table 4.5) are remarkably similar to those obtained at the global level (Table 4.1); logarithmic correlation coefficients are lower than the linear Pearson correlation coefficients. This suggests that geographic distances have similar effects at the global level as they have in the Edge46 subset, simulating elevated isolation. The results are in contrast with the findings in the ML46 subset, which features lower correlation coefficients in all cases, and where we assume more potential direct language contact.

4.5.6 Evaluating the hypotheses

In Section 4.1.1, we formulated three hypotheses for this study, H1 to H3. We will now discuss each of these hypotheses in turn.

H1 states that geographic distance explains the majority of the variance found in Swiss German syntax, as represented in the SADS data. However, $R^2 > 0.5$ really only holds for the travel times at the global level (Table 4.1), while in the case of Euclidean distance, the threshold of the coefficient of determination for both linear and logarithmic correlations is missed (R^2 = 45.78%, 42.18%, 42.35%). In the case of the more mountainous BEOV subset (Table 4.2) the Euclidean distances clearly did not reach the threshold (R^2 = 19.77% and 26.96%), while travel times in 2000 only slightly missed it for both the linear and the logarithmic case (R^2 = 45.46% and 48.2%). At the same time for the ML46 subset featuring gentle topography, geographic distances in no case explain the majority of variance (Table 4.3).

Despite the fact that some of the R^2 fell below the 50% threshold, the strength of correlation is considerable, with R^2 often reaching values greater than 60% (Tables 4.1 and 4.2). Particularly when comparing to the results of Szmrecsanyi's (2012) morphosyntactic study of English dialects, we obtained much higher coefficients of determination. This could be due to the different data source used: Szmrecsanyi used a frequency-based casual corpus dataset, while the SADS is a survey-based atlas aimed to discover the syntactic variation as deeply as possible, with a tendency of choosing phenomena that were assumed to show spatial variation patterns. It also might be due to the fact that Szmrecsanyi's data had much coarser spatial granularity (available for the former counties of Great Britain). The strength of correlation in our case is also higher than Spruit's (2006, 2008) findings for Dutch syntactic differences using the SAND Atlas data (Barbiers et al., 2005), where R^2 stays below 50%.

H2 posits that travel time measures better reflect syntactic spatial variation than Euclidean distance. Qualitatively, this hypothesis is very clearly supported by our results, as the correlation and determination values are always higher for the travel times than for the Euclidean distance. As shown in Table 4.4, the differences between Euclidean distance and travel times are also in all cases *statistically significant* for the global data set, and in one of two cases for the BEOV subset, with the other case just barely missing the 95% confidence threshold. The fact that the significance is lower for the BEOV subset than for the global dataset, despite the differences of correlation values being larger, can be explained by the much lower number of observations in the BEOV subset (N=46), compared to the entire study area (N=383). The effect of gentle topography inducing elevated transportation and communication possibilities in the ML46 subset is reflected in the overall lower correlation of geographic distances with the syntactic difference. Not surprisingly, travel times in this subset are not significantly better at explaining the syntactic difference, statistically speaking (thus values are not shown.)

The correlation analysis clearly supports H2 at the global level and in the more mountainous BEOV subset. These findings are in agreement with Szmrecsanyi's (2012) observation that geographic distance *per se* does not explain the vast majority of the syntactic variation. Geographic distance is only a proxy of potential language

contact or isolation, which presents itself nicely in the fact that travel times – which better reflect the actual effort that needs to be spent in order to establish contact – yield higher correlation values than Euclidean distance.

H3 states that older travel times better represent syntactic spatial variation. Based on Table 4.2, this hypothesis seems to be clearly supported; in Table 4.1 however, only the Pearson correlation values are systematically increasing from 2000 to 1950 to 1850. For Table 4.3, as mentioned above, there is no monotony to be observed. As seen in Table 4.4, differences between the correlations of different travel times with the syntactic distance are not significant. This means that, statistically speaking, for variation in Swiss German syntax, as represented in the SADS data, travel times are predictors of equal power, regardless of what year is taken.

4.5.7 Residuals of syntactic and geographic distances

We recall that in our case the residuals are not residuals of a regression analysis; they are instead obtained as the difference of the normalised syntactic distance minus the normalised geographic distance, centred on a particular reference site (Section 4.3.5). Positive residuals mean that the normalised geographic distance is smaller than the corresponding syntactic distance; negative residuals indicate the opposite relation. The relationship of syntactic and geographic distance can also be understood as a simple linear regression model with a single predictor variable, geographic distance. Positive residuals would then suggest that geographic distance underestimates syntactic difference between two survey sites, while negative residuals would suggest overestimation.

Figure 4.11 plots, for the reference site Obersaxen, the residuals of the normalised syntactic and Euclidean distance (y axis) against the Euclidean distance (x axis). If the syntactic distance from the reference site Obersaxen was in perfect linear agreement with the Euclidean distance, no residuals would show in this graph. The residuals, however, follow a decreasing (and linear) trend.

The residuals are positive at short ranges, meaning that the Euclidean distance underestimates short-range syntactic variation under the assumption that the syntactic distance will follow a growth linearly proportionate to the Euclidean distance from the reference site. Thus surrounding dialects are more different, than suggested by the Euclidean distance. The opposite is the case at long ranges, where Euclidean distance overestimates syntactic variation. This overestimation at long ranges is rational, as geographic distance increases continuously, whereas the syntactic distance may only increase to a certain level. If two dialects become too dissimilar, they will be considered two different languages, as mutual intelligibility is no longer maintained. Zero residuals, which would mean perfect correlation of the syntactic and the Euclidean distance, occur mostly in the range of 70 to 110 km.

The geographic patterns that the residuals exhibit become more apparent when maps are used to depict the residuals. Figure 4.12 presents the same residual values

for the reference site Obersaxen that Figure 4.11 showed in a scatterplot. We can again see the overall trend of underestimation at short ranges and overestimation at long ranges. This trend pattern is rather systematic, evolving in concentric circles outwards from the reference site. Furthermore, the numerical range of residuals is almost symmetrical (-0.51 to 0.59), with about the same number of negative and positive residuals (cf. Figure 4.11). As Figure 4.5 has shown, Obersaxen is fairly moderate regarding its average syntactic differences to all other dialects, which might explain this well-behaved pattern. Figure 4.13 is centred on the city of Freiburg (Fribourg), which in Figure 4.5 has shown to be among the survey sites with the highest average syntactic difference to all other sites. Furthermore, travel times of 1950 have been used to produce Figure 4.13 instead of Euclidean distances. The pattern visible in this map differs considerably from the map of Obersaxen. Positive residuals persist almost throughout the entire study area, meaning that syntactic variation is underestimated almost everywhere regardless of geographic distance. Only parts of the cantons of Valais and Grisons show negative residuals. As a consequence, the numerical range and distribution of the residuals is highly skewed towards the positive values. Also, the concentric progression seen in Figure 4.12 is almost not visible here, and high residual values are not restricted to short geographic distances. Indeed, the geographic range at which residuals decay to zero is much larger than in Figure 4.12. We attribute the patterns visible in Figure 4.13 to the special position of Freiburg in the Swiss dialect syntax landscape (Bucheli Berger, 2010; Scherrer and Stoeckle, 2016:109). Thus, we can claim that the analysis of residuals, both in scatterplots and even more so in maps, is an interesting tool to reveal such patterns and differences.

4.6 Conclusion

In this study, we compared different geographic distance measures (Euclidean distance and travel times for years 2000, 1950 and 1850) as an estimate of language contact possibility to a measure of linguistic difference between survey sites. To this end, we calculated syntactic distance based on survey data of the Syntactic Atlas of German-speaking Switzerland (SADS), involving multiple respondents per survey site, and computed different forms of correlation between syntactic distance and the different geographic distance measures, both for the entire SADS data set, as well as for local subsets. Furthermore, we generated different visualisations, again at the global and the local level.

The study set out from three hypotheses H1 to H3. Regarding H1, we showed that in most cases geographic distances explain the majority of variance inherent to the syntactic distance. Concerning H2, we have found that travel times are significantly better predictors for syntactic distance than Euclidean distance. Finally,

regarding H3, although older travel times seem to be better predictors for the syntactic distance, yielding higher correlation values, their superior performance did not prove statistically significant.

We further extended our analysis to the local level, enabling discovery of a more differentiated picture of the dialectal variation across space. At the local level, the effect of topographic barriers and the effect of potential direct (language) contact became more noticeable both in the visual representation of the maps of syntactic distance as well as in the correlation analysis. Building on our results, we can conclude that the aforementioned 'Fundamental Dialectological Postulate' (FDP) (Nerbonne and Kleiweg, 2007) seems to be true, especially on a global scale, that is, when little direct contact between speakers can be assumed. However, on a local scale linguistic distance (or similarity) depends much more on the particular characteristics of an area. This finding is in line with Stanford (2012:274) who states that "the issue of geographic size appears to be related to fundamental distance relationships in human interaction." On a local level, geographic distance may explain linguistic differences if, for example, topography is very pronounced and therefore actually poses a communicative boundary. If it is not, other factors (such as socio-demographic, cultural or attitudinal) may become more important.

At the level of the entire SADS data set, unlike in most other dialectometric research concerning other linguistic levels (where sublinear patterns were found unequivocal), a linear model described the correlations between geographic and linguistic distance better than a logarithmic model, although this difference is not statistically significant. Regarding the local subsets that we tested, logarithms of the geographic distances proved just as good a predictor for the syntactic distances, as the linear geographic distances.

Mapping the average syntactic distance to all other survey sites provided a way to find the local varieties that are most different from the others in the syntactic sense, comparable to maps of Goebl's identity values (Goebl, 2010). By computing residuals of normalised syntactic and geographic distances, we provided a way to show to what extent and by which pattern geographic distances predict the syntactic differences.

In order to account for linguistic variation more precisely, a number of other geographic and demographic factors should also be taken into account. Empirically, urban centres are important in the spread of linguistic innovations and might therefore cause dialects to converge to dialects spoken in economically, politically and culturally dominant places (Chambers and Trudgill, 2004:172). Finding measures for this gravity-like effect, possibly using Trudgill's linguistic gravity index (Szmrecsanyi, 2012) would be desirable along with testing measures known in GIScience, such as cost distances or terrain roughness.

As before the spread of individual transport public transport meant the only connection to other parts of the country for many people, using travel times of public rather than individual transport might lead to an improvement. Also, we plan to

investigate travel times from before 1850 for correlation with our syntactic distance measure.

Beyond the study of mere correlations at the aggregate level, ultimately it will be most interesting to compare spatial patterns of linguistic variation, such as local breaks in the dominance of dialectal variants, suggesting isoglosses, to geographic borders and extrageographical (political, cultural, historical; e.g., Hotzenköcherle, 1961) patterns.

Chapter 5

Sharp boundaries vs. gradual transitions: Quantitative models for transitions between areas of dialectal variants

5.1 Introduction

Some of the most important questions in dialectology address the areal distribution of dialectal variants and their relationship to so-called *dialect areas*. In order to understand the processes behind spatiotemporal patterns of dialect evolution and area formation, researchers traditionally investigated the distribution of individual linguistic phenomena in search of overlaps and correspondences between them. Based on point symbol maps derived from dialect surveys, isoglosses were drawn.

Due to the areal differences in their distribution, variants associated with a certain 'dialect' finally in most cases add up to a gradual transition, a *dialect continuum*, rather than sharp boundaries between 'dialect areas'. At the scale of individual linguistic variables, with a granularity of survey sites dense enough, it can be seen that individual variables usually do not show sharp boundaries between the corresponding variants either (cf. Heeringa and Nerbonne, 2001). Dialectal surveys relying on multiple informants per survey site such as the SyHD ('*Syntax of Hessian dialects*' – Fleischer, Kasper, and Lenz, 2012), the SBS ('*Sprachatlas von Bayerisch-Schwaben*', 1996-2009) and more recently, popular online and smartphone-based dialect surveys providing massive data with geolocations (e.g., Leemann, Kolly, Purves, et al., 2016), facilitate the discovery of local dialectal variation. Such data have the potential to

This chapter is based on: **Jeszenszky, Péter**, Philipp Stoeckle, Elvira Glaser & Robert Weibel, (in revision). Sharp boundaries vs. gradual transitions: Quantitative models for transitions between areas of dialectal variants. *Journal of Linguistic Geography*, 6 (2) special issue. **Author's contributions:** Conceived and designed the experiment: PJ RW. Performed the experiments: PJ. Analysed the data: PJ RW. Wrote the manuscript: PJ PS EG RW.

reveal different transition patterns between the usage areas of dialectal variants. As dialect areas are traditionally proposed based on the dispersions shared among different individual variables, spatially characterising and comparing the boundaries and transitions of these dispersions becomes essential. To our knowledge so far no research has been conducted on quantitatively representing such transitions in space between dialectal variants. This study, thus, focuses on introducing a methodology to quantitatively model the transitions between usage areas of dialectal variants.

The two main conceptualisations that address the spatial transitions in question, *isogloss* and *dialect continua* have been discussed in Chapter 2. In classical dialectology, maps of phenomena derived from survey data were used to find bundles of isoglosses in order to determine homogeneous areas, sometimes associated with dialect areas (Haag, 1898; Kurath, 1972; Maurer, 1942). When drawing the isoglosses, occurrences on the 'wrong' side of the boundary were noted as exceptions (such as in interpretation studies (e.g., Christen et al., 2010) of the point symbol maps in the Language Atlas of the German-speaking Switzerland (SDS – Hotzenköcherle et al., 1962-1998) and the maps based on the Wenker database (e.g., Wrede, Mitzka, and Martin, 1927)), and in several cases such outliers were smoothed away. However, individual linguistic variables do rarely display this type of clear-cut regional pattern. As Chambers and Trudgill (2004:104) point out: "Neighbours seldom differ absolutely in any respect."

Alternatives to clear-cut and categorical boundaries have been proposed already early on (e.g., Bloomfield, 1933), but to account for the spatial distribution of dialect areas in a *quantitative* manner was not initiated until Séguy (1971). Since then researchers strived to quantitatively account for the similarity of dialects through grouping locations along multiple dimensions (e.g., Goebl, 1982;Kessler, 1995; Heeringa, 2004). The concept of dialect continua implies that while dialect areas cannot be sharply delimited due to variables following different patterns of regional variation, gradual transitions ought to be expected between dominant usage areas of variants for individual variables as well (Heeringa and Nerbonne, 2001).

Quantitatively accounting for dialect areas. Numerous studies have been conducted with the aim of quantitatively accounting for areas of variants or, in an aggregate manner, for dialect areas. The studies mentioned in these regards in Chapter 2, however, feature sharp and fuzzy boundaries solely as implicit products; the focus was rather on the internal homogeneity of the spatial linguistic clusters found.

Research gap. To our knowledge, studies focusing explicitly on the quantification of boundaries and transitions between dialectal variants do not exist to date, although the studies by Sibler et al. (2012) and Scholz et al. (2016) offer first hints at methods that could be used to model boundary transitions. Today, however, modern dialect surveys exist with better spatial granularity and multiple respondents per survey site, offering a sound basis for the study of language transitions. Thus, this study proposes methods to quantitatively model the degree of graduality in the

transitions between dialectal variants, conceptualising the transitions as gradients. This allows to use the same approach to represent both boundary concepts prevailing in dialectology, the isogloss (denoted by an abrupt, near-vertical gradient) and the dialect continuum (which results in more gentle gradients) and it allows to test the validity of these concepts. The main motivation of this research is to support dialectology in the assessment of the nature of transitions, both in the variants of individual linguistic variables as well as comparatively across different variables, offering an alternative to the subjectivity of visual analyses. If it is possible to compare distribution and transition patterns across linguistic variables more objectively, it becomes easier to describe their (dis)similarity and formulate hypotheses regarding the reasons for differences in distributions and diffusions. Knowing the steepness of transition gradients may also contribute to formulating hypotheses regarding future diffusion scenarios of variants as well as to reconstructing historical linguistic changes.

The study is conducted at multiple spatial levels involving methods regularly used in spatial analysis: trend surface analysis and univariate regression analysis. First, the analysis extends over the whole study area (termed *global level*). Second, the analysis is conducted on spatial subsets dependent on the distribution characteristics of the linguistic variable in question. Third, transitions in *cross-sections* constructed in the direction of the main global change are studied to account for local patterns of transition. In this study a set of variables from the SADS are used (described in Section 5.3.2).

In the remainder of this study, Section 5.2 introduces the SADS data used in this study; Section 5.3 discusses conceptual models of transitions in syntactic variation; Section 5.4 proposes a set of methods to quantitatively model and describe transitions; Section 5.5 presents and discusses results and illustrates the application of the proposed methods on SADS data; and Section 5.6 concludes the study with a summary of insights gained and an outlook on potential future research.

5.2 Data

The SADS database, used in this study has been in detail described in Chapter 3.

Among linguistic surveys, SADS belongs to those that rely on multiple respondents at each survey site. To capture the local linguistic diversity, at each survey site 3 to 26 respondents (median = 7) were involved in the survey. Having multiple respondents per survey site allows to better capture the endemic linguistic diversity that might be present within a place owing to differences in gender, age and profession. Importantly, the availability of multiple responses per site enables accounting for the spatial distribution of variants in linguistic variables with a better attribute granularity and thus creates the foundations necessary to study dialectal transitions

in space. Besides, within-speaker variation is made possible through MC questions allowing each respondent to pick multiple answers. In case a respondent picked multiple answers, they were asked to specify what their most natural, "preferred" answer would be. In the case of MC questions, these preferred variants were used for this study.

5.3 Modelling Transition in Syntactic Variation

5.3.1 Spatial variation in syntactic variables

According to Glaser (2013:214), "[...] in syntax we often meet the situation that a particular feature characterises one area, but in a neighboring area we find variation between two features instead – and not another area defined by one typical variant. Isoglosses may thus mark off variation zones and not homogeneous areas." She also states that, in general, linguistic features are neither evenly nor randomly distributed in a speech community or within closely related neighbouring varieties; instead, they are clustered in certain regions and absent in others. (Glaser and Weibel, 2013).

Patterns of spatial distribution in the SADS survey database are very diverse. Glaser and Bart (2012, 2015) have described three principal types of spatial distribution of the syntactic questions surveyed in the SADS.

I. Two main variants with positive spatial autocorrelation are competing with each other, with a sharp or more gradual transition between their (dominant) areas of usage.

II. A more frequent, overarching variant occurs across most of the study area, with regional variants reaching dominance only in small areas. In these small areas most survey sites show at least two variants (one of them often being the overarching variant).

III. Highly variable spatial distribution, with no clearly distinguishable spatial patterns or only local patterns of dominant variants discernible.

The four maps in Figure 5.1 demonstrate different spatial distribution patterns in variables. In these maps, a Voronoi tessellation was used to spatially interpolate between the point locations of survey sites in order to obtain full spatial coverage, as often done in dialectometry (Goebl, 1982; J. Lee and Kretzschmar, 1993; Rumpf et al., 2009). The colours of the Voronoi polygons correspond to the locally dominant variant (except where no variant reaches dominance, in which case the colour is grey). The brightness of the colour corresponds to the proportion of the respondents using the locally dominant variant. This proportion is termed the intensity of the variant at the given survey site. Note that no smoothing was applied when assigning intensity values (in contrast to Rumpf et al., 2009). Three maps of Figure 5.1 represent examples of Type I variables (top left and right, bottom left): the two most frequent

variants form large and more or less homogenous areas, with gradual or sharp transitions in between, respectively. The map on the lower right, however, is an example of a Type II variable, with one overarching variant along with others having only local prevalence. No example is shown of a Type III variable, as the study of these variables (rarely occurring in the SADS) is less interesting from a spatial point of view.

The prevalence areas of variants are termed dominance zones here, while the mixing areas positioned between dominance zones are termed transition zones. Transition zones are often described and used in dialectology (Pickl, 2013; Scherrer, Leemann, et al., 2012; Scholz et al., 2016 and by the Meertens Institute project "Maps and Grammar" (2013-), e.g., Dros-Hendriks, 2018), but the concept lacks a clear definition. Besides, dominance zones have been quantitatively delineated in several studies, as mentioned in Section 2.3.2.

This study aims to quantitatively describe transitions between dominance zones of variants, using the concept of gradients. Hence, Type I variables of the SADS are used, as fitting gradient models is the most sensible if there are few dominance zones, separated by few transition zones. This type of spatial distribution accounts for about 40% of all questions represented in the SADS.

In order to model the transitions in three dimensional space, intensity landscapes are defined. As the third dimension, the intensity of each variant is assigned to every survey site, similarly to digital terrain models. These intensity values are visualised similarly to the 3-D models in Figure 5.2.

Figure 5.1 and Figure 5.2 allow making three observations. First, these illustrations support the existence of dominance zones and transition zones. For some variables the dominance zones have more clear-cut boundaries, while for some of them the transition is more gradual. Second, neither of these zones appear necessarily homogeneous. Third, the gradient of transitions may not be uniform; patterns of change may be different in every direction.

5.3.2 Linguistic variables used in the study

In this study eight variables of the SADS database were used (which in this study correspond to eight survey questions), with the aim to develop a generalisable methodology. The variables were required to belong to Type I. Also, since the study aimed to capture the diversity of spatial variation patterns among these, the selected variables should represent different rates of change in transitions and different directions thereof, with regards to the diverse topography of Switzerland. The selection process was helped by exploratory visual analysis. The list of questions selected and their main answer variants are found in Table 5.1. Besides, the full sentences are presented in Standard German and English in Table 3.1.

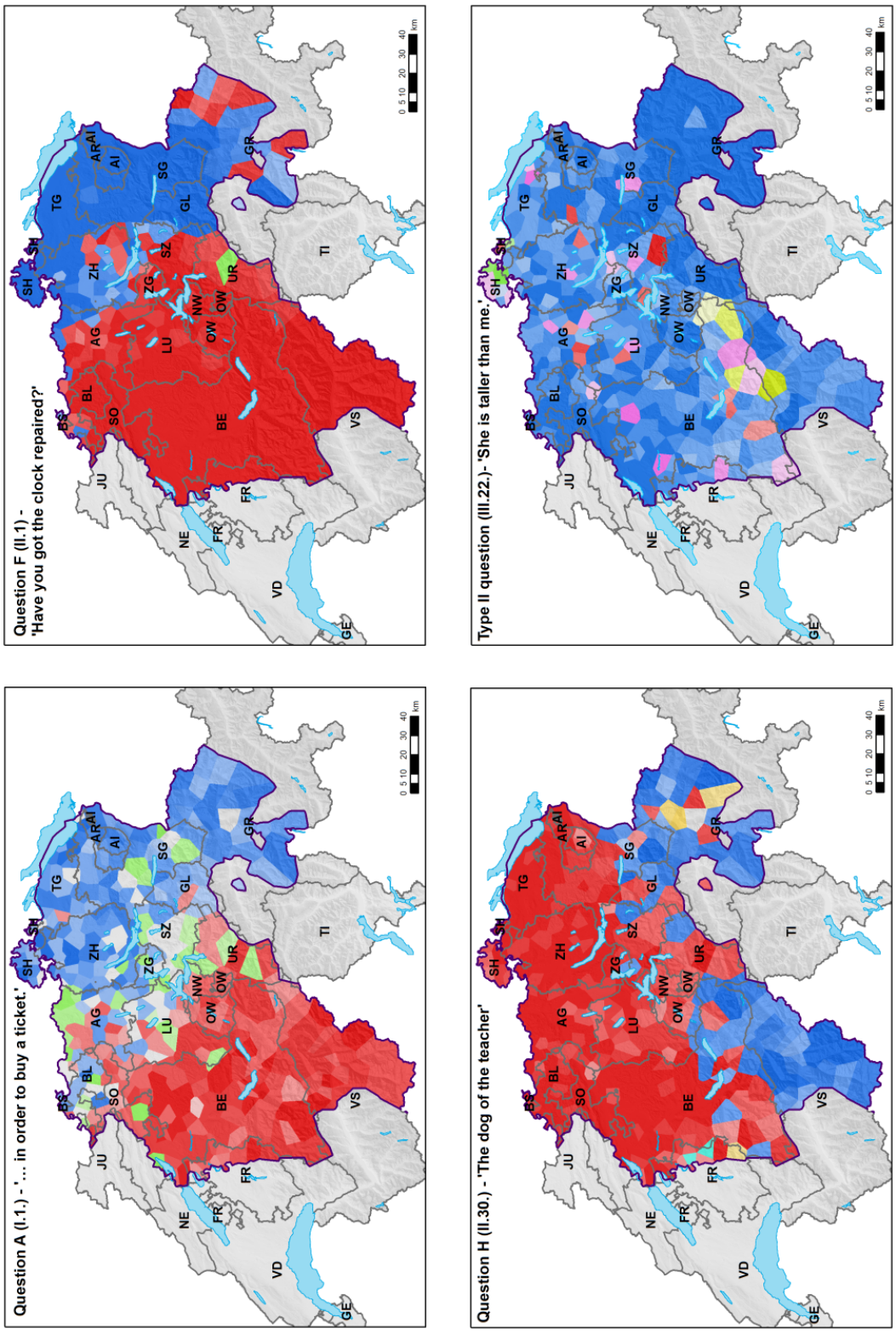

FIGURE 5.1: Intensity maps. Maps of distribution Type I: top left – Question A, top right – Question F, bottom left – Question H. Bottom right: a variable belonging to distribution Type II.

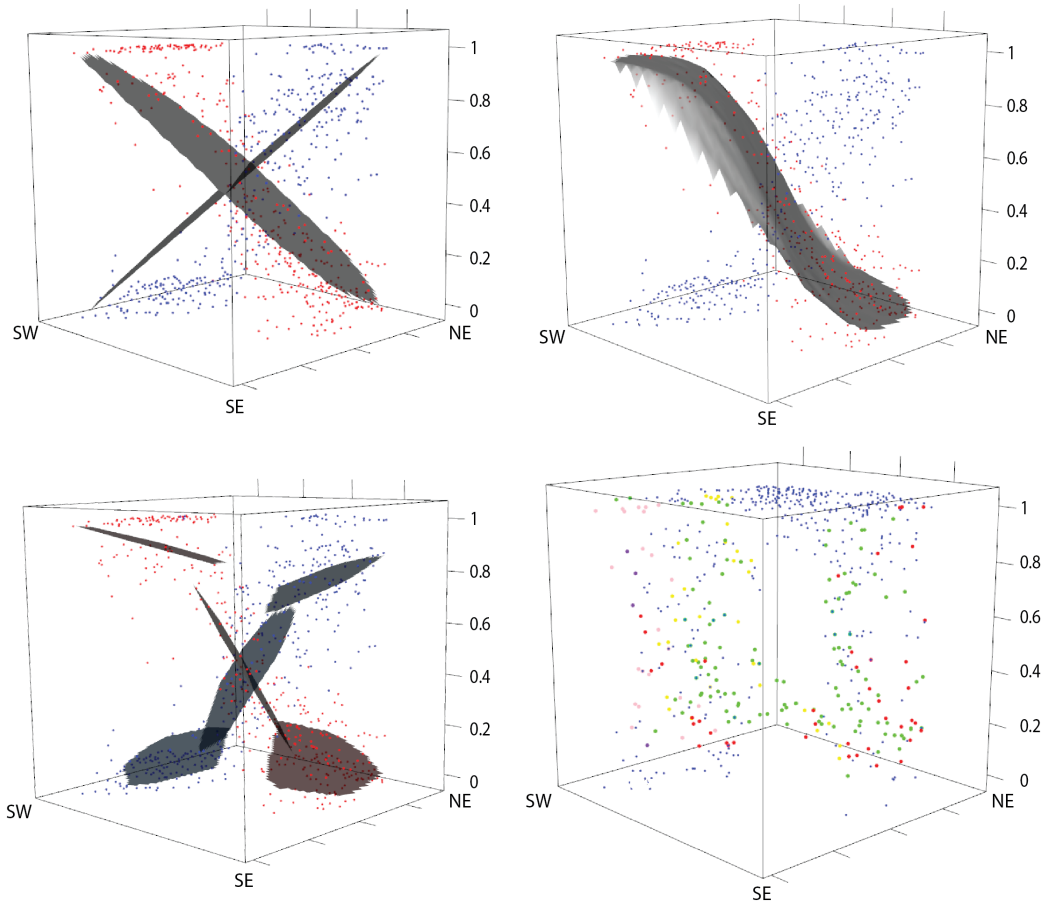

FIGURE 5.2: 3-D plots representing different regression strategies. Top left, top right and lower left plots show the intensity landscape of Question C. The benchmark surfaces for the inclined planes model, the logistic regression surface fitted on the global level, and planar trend surfaces fitted in the γ subdivision strategy are overlaid, respectively. For reference, in the bottom right plot the distribution of a Type II variable is shown, omitting zero values. The cardinal directions on the plots (SW, SE, NE) indicate the survey sites' geographic position.

Four out of the eight survey questions – addressing the phenomenon infinitival purposive clause, A, B, C, D – were already used in other studies involving the authors (Glaser and Bart, 2012, 2013; Sibler et al., 2012). The answers to the questions regarding the infinitival purposive clause reveal two main competing dominant dialectal variants *'für'* and *'zum'* (shown in red and blue in Figure 5.1 top left, respectively) and also some minor variants, including the standard German variant *'um...zu'* (shown in green). It is desirable for useful models to incorporate all data available, so the presence of further variants is allowed. All four infinitival purposive clause questions (with Question A-D shown in Figure 5.3) feature the dominance of the *'für'* variant (red) in the southwest, while the *'zum'* variant (blue) is dominant in the northeast of the study area. The standard German *'um...zu'* variant (green) and others seem to be spatially more randomly distributed. Transition occurs in different regions of the study area, with different rates of change observed.

The remaining four phenomena (Questions E to H in Table 5.1) relate to different syntactic phenomena. Question E and Question F, respectively, deal with the position of the non-finite verb in perfect tense and in the infinitive particle (Jeszenszky and Weibel, 2015; Scherrer, Leemann, et al., 2012). Question G addresses the occurrence of the infinitive particle (Stoeckle, 2018); finally, Question H relates to the adnominal possessive.

Based on exploratory analysis using visualisations similar to Figure 5.3, Figure 5.2 and Figure 5.6 the linguistic variables listed in Table 5.1 were classified according to the degree of graduality of their transition zones, using two types: 'gradual transition variables' and 'sharp transition variables'. The variables represented by Questions A, B, C and, to a lesser degree, D are seen as gradual transition variables, while Questions E and F are classified as typical representatives of sharp transition variables. Questions G and H are neither clear representatives of gradual nor of sharp transition variables. Moreover, Question H represents an interesting case, as the main trend of transition occurs in N-S direction.

TABLE 5.1: Variables of the SADS used in the study. MC = multiple choice question.

Linguistic phenomenon	**Standard German sentence**	**First main dialectal variant *(MV1)***	**Second main dialectal variant *(MV2)***	**English translation**	**Transition type**
A) Infinitival purposive clause	...*um* ein Billett *zu* lösen. SADS I.1	...z wenig Münz, *für* es *(z)* lööse	...z wenig Münz, *zum* es *(z)* lööse	... *in order to* buy a ticket.	gradual

B) Infinitival purposive clause (MC)	… *zum* einschlafen. SADS I.6	… *für* ii(z)schlaafe.	… *zum* ii(z)schlaafe.	… *in order to* fall asleep.	gradual
C) Infinitival purposive clause	… *um* ein Buch *zu* lesen. SADS I.11	… *für* es Buech *(z)* läse.	… *zum* es Buech *(z)* läse.	… *in order to* read a book.	gradual
D) Infinitival purposive clause (MC)	… s Licht anzünden *um zu* lesen. SADS IV.14	… s Liecht aazünde *für (z)* läse.	… s Liecht aazünde *zum (z)* läse.	… turn the light on *in order to* read.	gradual
E) Position of non-finite verb (perfect)	Ich habe den Fritz *kommen hören*. SADS I.3	Ich ha de Fritz *ghöört choo*.	Ich ha de Fritz *choo ghöört*.	I *heard* Fritz *coming*.	sharp
F) Infinitive particle	Hast du die Uhr *flicken lassen*? SADS II.1	Häsch du d Uhr *la flicke*?	Häsch du d Uhr *flicke (g)laa*?	Have you *got* the clock *fixed*?	sharp
G) Infinitive particle	Er *lässt* den Schreiner *kommen*. SADS II.3	Er *laat* de Schriiner *choo*.	Er *laat* de Schriiner *la choo*.	He calls the carpenter. (He *lets* the carpenter *come*.)	mixed
H) Adnominal possessive (MC)	*der* Hund *des* Lehrers SADS II.30	*em* Leerer *sin* Hund	*s* Leerers Hund	*The* dog *of the* teacher.	mixed

5.3.3 Modelling the conceptualisations of linguistic boundaries and transitions

As detailed above, in dialectology the concepts of the isogloss and the dialect continuum, respectively, address the nature of transitions between variants. In this study prototype models are proposed with a double aim. On the one hand to test the validity and presence of the two conceptualisations of linguistic boundaries and transitions on dense data. On the other hand to enable comparison of transitions across variables. Parker (2006) discussed the dynamic continuum of boundaries in space. The prototype models created can be considered as extremes along Parker's continuous spectrum of boundary qualities blending along the geographical landscape.

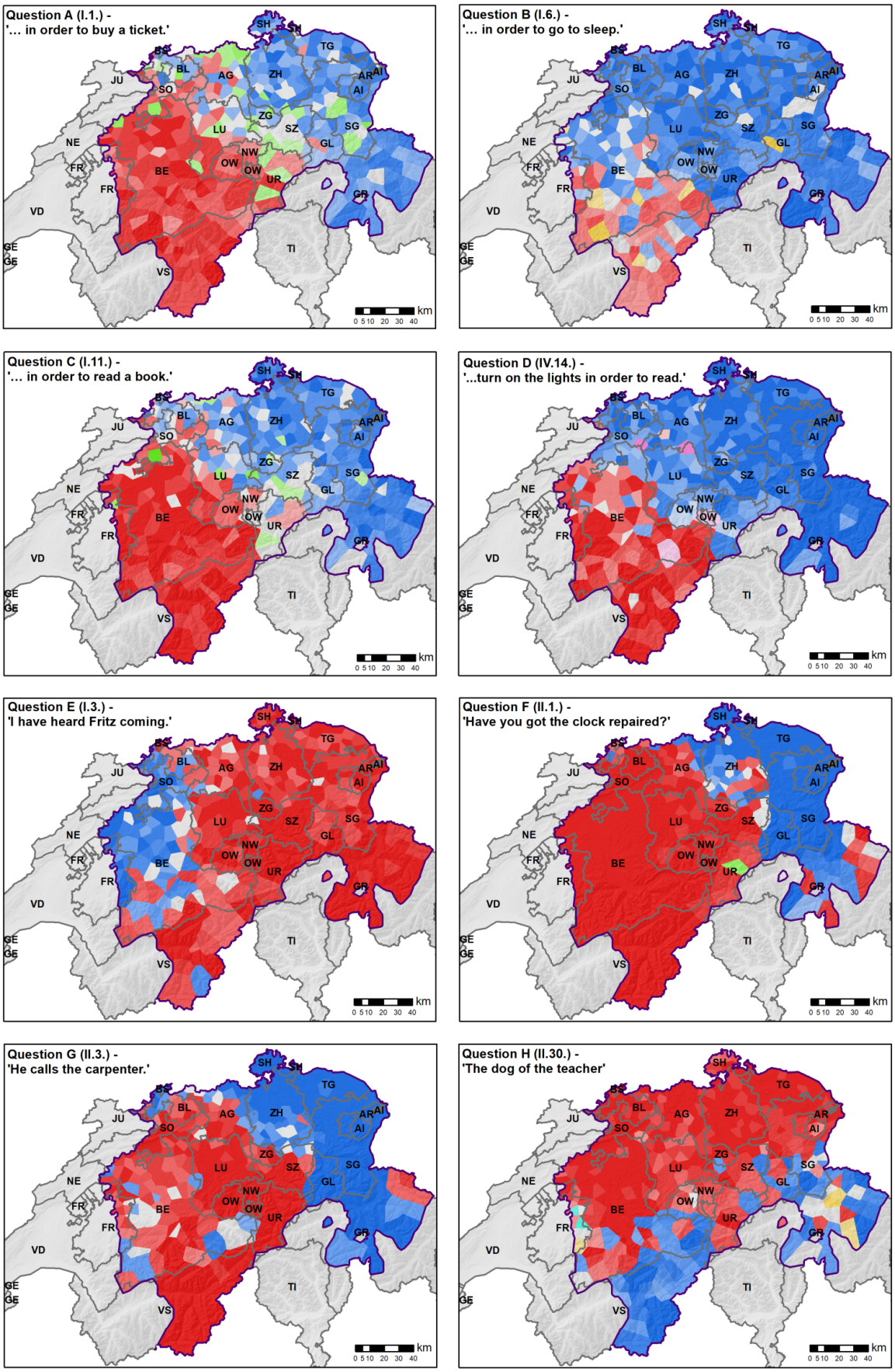

FIGURE 5.3: Intensity maps of the eight variables used.

As the vagueness of the linguistic conceptualisations impedes clear-cut mathematical definitions, the prototypical models are seen as quantitative representations of these qualitative conceptualisations. It has to be mentioned that due to their quantitative definition the prototype models do not entirely correspond to the original usage of the isogloss concept in classic dialectology, and Seiler's (2005) concept of 'inclined planes', which can be seen as a representative of the dialect continuum concept applied to individual linguistic variables. Seiler's concept is actually more complex than the mathematical specification given here, which abstracts his concept to a general model.

The two following prototype models are defined:

Isogloss model: Using the concept of isoglosses, in Type I variables a sharp boundary between the dominant usage areas of competing variants is expected, with vertical or near-vertical gradients perpendicular to the isogloss. To model this, at the global level (i.e., involving the entire study area) the mathematical model of a logistic surface is fitted to the intensity landscape, expecting a very abrupt drop between its maximum and minimum values 1 and 0, representing the exclusive usage and the lack of the variant in question, respectively.

Inclined planes model: In this work, the concept of the dialect continuum, which is usually applied at the level of aggregate linguistic variation, is applied to the level of individual variables. The inclined planes hypothesis suggested by Seiler (2005) is an example of the dialect continuum concept applied to individual variables. Discussing the spatial variation of the two main dialectal variants found in the answers to Questions A to C, Seiler posited that the transition between the usage areas of the two variants resembles two opposing inclined planes, each with constant declination from the maximum of the corresponding variant at one end of the study area to the minimum intensity values reached at the other end of the study area, similar to the two inclined planes shown in Figure 5.2, top left. For the syntactic variables with competing variants (Type I) forming the focus of this study, this could be best modelled by two planar surfaces (one for each variant) having a fixed maximum intensity of 1 at one end of the study area and fixed minimum intensity of 0 at the other end. This is the process used to produce the top left visualisation of Figure 5.2, to be introduced in Section 5.4.

As a result of preliminary visual analyses (cf. Section 5.3.2), investigation of transition patterns and intensity changes is proposed in *spatial subdivisions* as well. Based on the wave-theory and S-curve models of language spread in diachronic linguistics (Blythe and Croft, 2012; Yokoyama and Sanada, 2009) and the cascade model (Britain, 2010) it is assumed that in variables of Type I, the majority of the change in intensity happens in the transition zones, as opposed to dominance zones. This observation provides the motivation to also investigate the transitions zones and the dominance zones separately, defined in Section 5.4.3.

5.4 Methodology

The primary focus of this study is on proposing a methodology for the study of spatial transitions in linguistic variables, viewing transitions as gradients. Since few assumptions are made regarding the nature of the input data, the methodology has the potential of being generalisable to further linguistic or other spatial variables that have appropriate spatial granularity and for which intensities can be calculated. The methodology consists of several individual methods, which will be introduced one by one in the following subsections.

5.4.1 Exploratory visual analyses

In order to discover the prevailing patterns of transition present in a particular variable, determine appropriate prototype models, and evaluate the usefulness of the transition characteristics calculated, exploratory visual analyses of different kinds have been conducted using R and ArcGIS. Through the following steps each linguistic variable has been classified into the variation types defined in Section 5.3.1 and the prototype models used for representing the linguistic conceptualisations (Section 5.3.3) have been developed.

First, area-class intensity maps were created to visualise the spatial patterns of the main variants (Figure 5.1 and Figure 5.3). The construction of these area-class maps was introduced in Section 5.3.1. Due to the greater area available for visual perception, Voronoi polygons allow distinguishing more colours more accurately than point symbols do. Lighter colours indicate greater degree of mixture between variants.

Second, in order to reveal intensity values of non-dominant variants at each survey site, interactive 3-D plots were constructed (shown in a static version in Figure 5.2) using the *rgl* and *akima* packages in R. The fitted surfaces shown in Figure 5.2 are discussed in Section 5.4.2. Visualising the intensity values and surfaces in this fashion provides an intuitive approach to assessing the variation characteristics of the variants, in particular from the perspective of transition gradients, and allows developing hypotheses regarding the fit and adequacy of the prototype models proposed.

Third, the intensity landscapes were also investigated along cross-sections (similarly to Heeringa and Nerbonne, 2001; **Proll2015a**) in both the exploratory and the modelling phase, with regression curves of different types fitted to the cross-sections.

5.4.2 Surface and profile fitting

As the transition between two usage areas usually occurs at different rates of change (i.e., gradients) in different directions, describing it by a single number would not represent reality. Our aim is to fit models of transition that best characterise the

transition inherent to the data; models that are comparable across different variables and across different granularities. Usage of surface and curve fitting based on least-squares regression are proposed.

To represent the multi-layered and mixed nature of dialect areas, as well as the interaction of different, co-occurring variants in a particular spatial direction, Pickl (2013:70) and more specifically Pröll et al. (**Proll2015a**) used cross-sections, showing the overlap of 'dialect types' based on factor analysis results. A similar cross-section approach is used in this study at the level of individual variables.

However, first the dominant direction of the transition needs to be determined. For this purpose planar trend surfaces are fitted to the intensity landscapes of each variant, using ordinary least squares, and their aspect angles are calculated in the direction of the steepest gradient. Assuming that the aspect angles of two opposing variants are indeed pointing in the opposite direction, that is, their difference is roughly $180°$, the direction of the corresponding cross-section can be established as the average between the aspect angles of the main variants, representing the bisector of these aspect angles. The results for the dialectal variables used in this study are shown in Table 5.2, with a detailed discussion following in Section 5.5.1.

Having established the direction of the cross-sections for a particular linguistic variable, it is possible to proceed with the cross-section placement. Since the transition characteristics differ across space, it is advisable to construct several, parallel cross-sections to better cover the study area. In our study for each variable, four straight lines have been placed parallel to the bisector at four different quasi-random positions (Figure 5.4). Along these lines, cross-sections are constructed. As it becomes visible from Figure 5.3, the cross-sections of Questions C, D, and E have their main direction of change in the general direction Southwest – Northeast (SW – NE), which also marks the main direction of the densely populated Swiss Plateau, while for Question H the main direction of change is almost perpendicular to the direction of the other variables, i.e., South-southeast – North-northwest (SSE – NNW).

The final step of the spatial construction of cross-sections is shown in Figure 5.5. The cross-section line is first intersected with the Voronoi polygons. Then, the midpoint is computed for each cross-section-polygon intersection, and the intensity value of the corresponding survey site (represented by the associated Voronoi polygon) is assigned to the midpoint. In our study, each cross-section incorporates 16 to 35 survey sites into their respective linear subset (median = 26), with Question H having the shortest cross-sections and hence the least survey sites. Along the cross-sections two kinds of distances are registered with the midpoints. The topological distance is equal to the number of Voronoi polygons (or sites) that have been visited as a sequence along the given cross-section, disregarding their exact positions in geographic space. As an example, in Figure 5.5, the topological distance between the western-most site and the eastern-most site is 8, since 8 polygons had to be visited to get there. On the other hand, geographic distances are registered as 2-D Euclidean

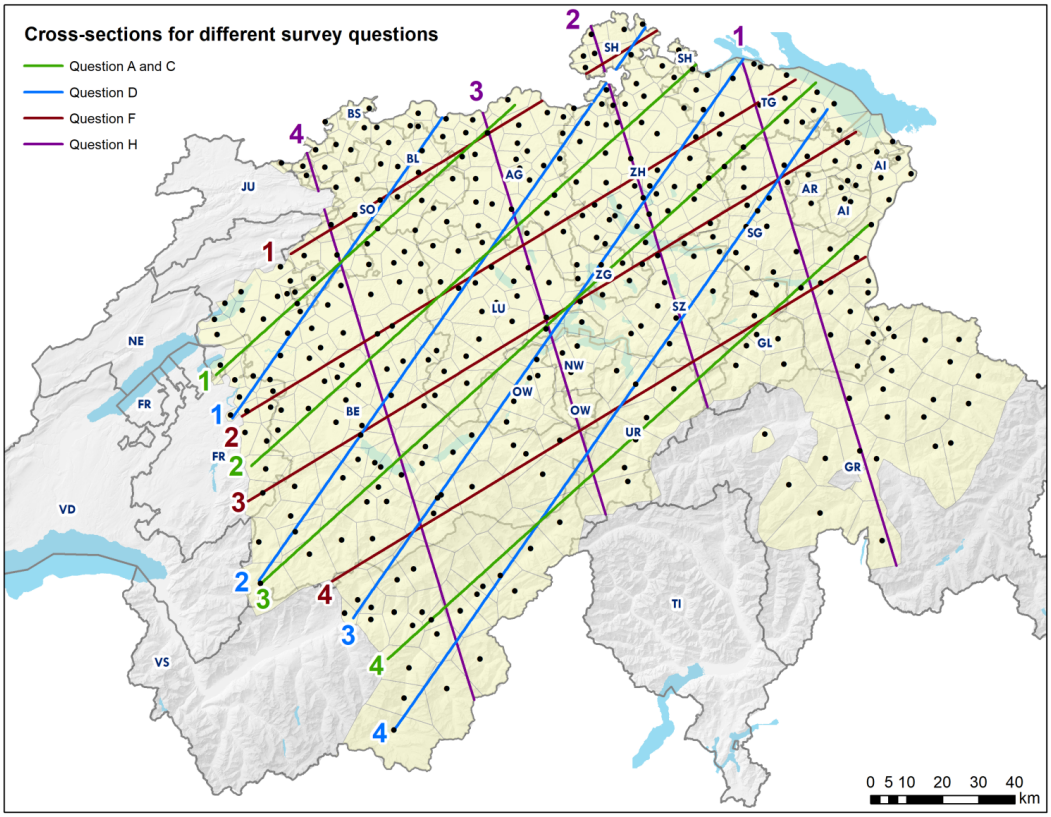

FIGURE 5.4: Cross-section locations overlaid on the SADS survey site map.

distance to the midpoints along the cross-section line (i.e., ignoring the third dimension). Examples of intensity profiles generated along three cross-sections are shown in Figure 5.6, with a detailed discussion following in Section 5.5.2.

5.4.3 Spatial subdivision strategies

Focusing on the overall transition patterns only, fitting one global transition (or trend) to the intensity values overlooks the regional variation and the vast differences transition patterns may exhibit. It is a general rule that globally fitted models smooth away details while locally fitted models give a more detailed picture, due to values being usually more similar in a local neighbourhood (Fotheringham, 1997:88), as also implied in the so-called 'First Law of Geography' (Tobler, 1970), describing spatial autocorrelation.

As proposed above, for Type I variables most of the gradient change is assumed to occur specifically in transition zones, as opposed to dominance zones. Taking this into account, three spatial subdivision strategies α, β and γ are proposed, aiming to test the validity of the prototype models with a better spatial granularity:

α A bipartite subdivision based on kernel density estimation (KDE) smoothing (Sibler et al., 2012), modelling the subdivision of the whole study area by the most plausible isogloss between the two main variants.

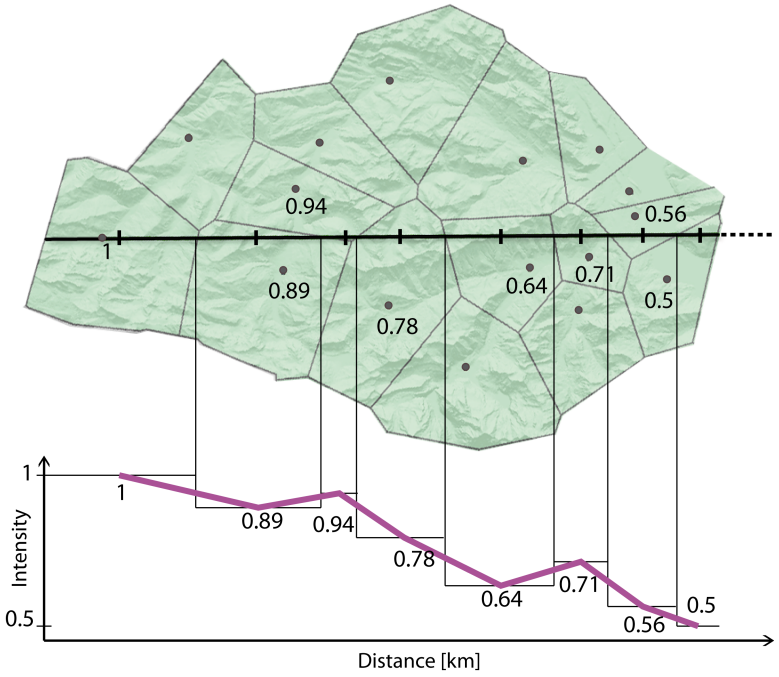

FIGURE 5.5: Constructing cross-sections in geographic space and registering the intensity values along the cross-section (2-D map view).

β A tripartite subdivision based on defining a transition zone where intensities of both main variants stay below a strict threshold of 62.5% (chosen based on visual analysis, corresponding to $^5/_8$ of the exclusive usage of a given variant); with the remainder forming the two dominance zones of the two main variants, respectively.

γ A tripartite subdivision where the transition zone is defined by more subjective cut-off values, placed at the (and including) survey site where the intensity of the given variant starts dropping markedly as we proceed along the direction of the main transition.

Using the α subdivision strategy, bipartite partitions are established on the level of the entire study area. These partitions are then passed to the level of cross-sections by coding the survey sites corresponding to the classification at the global level. As dialectology lacks a generally accepted definition of a transition zone, the tripartite subdivision strategies apply intuitive thresholding strategies based on their visual plausibility. While the β strategy uses a strict threshold, γ is more subjective, allowing to accommodate the local variation of the survey data and potential weaknesses of the cross-section positioning (cf. Section 5.5.2). The β and γ subdivision strategies are established based on the cross-sections. To construct the β spatial subdivisions the threshold points in the four respective cross-sections have been tied together in a subjective manner, allowing to take into account further local patterns of transition when constructing the spatial subdivisions. The subsets resulting from all three subdivision strategies are investigated in both the cross-sections and at the global level.

Regarding the validity of the prototype models and subdivision strategies, respectively, we expect the following:

- If the isogloss prototype model holds perfectly for a particular variable, then the α subdivision strategy should show an optimal fit. In that ideal case, the planar trend surfaces fitted to the intensity values in the two subsets are both expected to be level, with maximum and minimum intensity values on the two sides of the isogloss, respectively. In such a case using the tripartite β and γ strategies no transition zone could be found.
- If the inclined planes prototype model holds perfectly, then a global linear trend surface or trend line should achieve an optimal fit. Splitting the landscape into two or three partitions would not make a difference, as the monotonic and constant decline of the intensity landscape should seamlessly continue through the boundaries of any subdivisions.
- If the assumption holds that the majority of the change in intensity happens in the transition zones, then a tripartite spatial subdivision should provide the best fit. In this case planar trend surfaces are expected to be level nearing the maxima and minima in the dominance zones, respectively, and a steeper planar trend surface would fit well to the intensities in the transition zone.

5.4.4 Evaluating the validity of the prototype models

The validity of the prototype models may be evaluated using trend surface analysis, regression line analysis and analysis of residuals, as well associated measures. These measures can also be used to compare the characteristics of transitions across different variables.

Isogloss model

The validity of this scenario is assessed in the cross-sections and at the global level in trend surfaces by fitting a logistic function having one or two predictors, respectively: distance along the cross-section, and x- and y-coordinates, respectively. A logistic trend surface fitted to a main variant's intensity landscape is shown in the top right graph of Figure 5.2. In cross-sections the logistic regression curves (Figure 5.6, top row) are fitted independently of the global logistic trend surfaces. The steepness of the function (i.e., slope value at its inflection point) is used to measure the gradient of the transition in 3-D space and along the cross-section. The deviance (Hosmer, Taber, and Lemeshow, 1991) measure represents the fit in a logistic regression model analogously to sum of squares calculations in linear regression. Finding a steep and well-fitting logistic function is assumed to be a good indicator of a sharp transition, as indicated in Section 5.3.3.

As another measure, the correspondence to a binary spatial distribution is calculated. For both main variants the difference of intensity values from 100% in their

respective dominance zone and the difference from 0% in the dominance zone of the opposing variant are calculated, and sums of squared residuals are presented. Further, the width of the transition zones can be assumed to indicate the presence of plausible isoglosses. For cross-sections this width equals the number of survey sites contained in the transition zone (as detailed in Section 5.5.3). To compare transition patterns across different linguistic variables, slope values of planar trend surfaces fitted in transition zones are also calculated. In the graph to the lower left in Figure 5.2, planar trend surfaces have been fitted using the γ subdivision strategy.

Inclined planes model

The inclined planes prototype model is a plane surface reaching maximum intensity (100%) at the end of the study area where the given variant is dominant, and minimum intensity (0%) at the opposite end, fitting a plane between the two extreme intensity values, with the aspect angle of the corresponding variant. This surface is referred to as the *'benchmark'* of the variant concerned, and is similarly used for the evaluation of the global trend surfaces (e.g., Figure 5.2, top left) as well as cross-sections (Figure 5.6, top row). Its fit is assessed on the global level and using the corresponding α, β and γ spatial subdivisions. The goodness-of-fit is measured by the sum of squared residuals.

Furthermore, rates of change are calculated in the transition zones in the cross-sections, both for the β and γ subdivisions, in order to compare across different linguistic variables. Cumulative change in the intensity is calculated by summing up the differences in intensity along cross-sections, specifically to see the variation in intensity in the particular direction. Using residual maps (Figure 5.7) based on the differences between the observed intensity values and the fitted values of the benchmark surfaces and logistic regression surfaces, respectively, the spatial fit of the prototype models becomes visually comparable.

5.5 Results and Discussion

We will now present results of applying the methods described in the preceding section to the variables selected from the SADS database (Section 5.3.2, Table 5.1 and Figure 5.3). The various methods are part of the same methodology, which uses the concept of gradients to quantitatively model transitions in the spatial distribution of intensity values of dialectal variants. However, each method can also be seen as an individual tool, made for a particular purpose. Hence, results for each method are introduced, one by one, and a discussion of the corresponding results are provided. As the focus of this study is on methods development, the discussion will emphasise the evaluation of the performance of the proposed methods, rather than the linguistic interpretation of the results. Note also that for the sake of brevity, the presentation of results is restricted to a few selected examples.

5.5.1 Global patterns of transition

The strongest overall transition of a particular variable is expected in the direction of the average transition of opposing variants, which was modelled by the aspect angle of planar trend surfaces (Section 5.4.2). These aspect angles are presented in Table 5.2 for each variable's main variants (MV1, MV2), respectively. Most aspect values align with the SE – NW direction corresponding to the main transport direction of the densely populated *Swiss Plateau ('Mittelland')*. This is an artefact of the choice of variables, which nevertheless corresponds to a representative direction of transitions (Glaser and Bart, 2015; Hotzenköcherle, 1984) between variants. Another major transition direction in the SADS is close to N – S, represented in our set by Question H. Importantly, the aspect angles of the main variants deviate from the opposite direction (180°) usually less than 3°, meaning that the main variants' intensity surfaces are more or less exactly inclined against each other. This suggests that transitions in variables with two competing variants can indeed be modelled by planar surfaces, thus the inclined planes conceptualisation should be further investigated.

To demonstrate greater deviations from the opposition, the results for a less frequently used answer variant of Question A is also presented in Table 5.2, along with a variable that belongs to Type II (cf. Section 5.3.1), having one overarching and several regional variants. The regional occurrence of variants with no clear transition trends, renders planar surfaces poor models for such intensity surfaces. Not being able to model a transition between a regionally occurring and a majority variant with two planar surfaces points to the limitations of the methods proposed in this study, which assume linguistic variables with 'well-behaved' spatial variation of variants with few (typically two) dominance zones and few (typically one) transitions zones.

5.5.2 Cross-sections

Cross-sections capture the local patterns of transition in the direction of the average aspect angle for each variable, at different parallel positions as shown in Figure 5.4. Figure 5.6 depicts a typical gradual transition (Question C) in the left column, a sharp transition (Question F) in the right column and a somewhat undecided transition pattern (Question A) in the middle column. It is interesting to note that Question A in Table 5.1 was classified as a gradual transition variable, but the particular cross-section shown in Figure 5.6 indeed exhibits a rather noisy pattern. The plots present the intensities of all variants of the variable at each survey site along the cross-sections (using topological distance). In the top row, the benchmark of the inclined planes prototype model and the fitted logistic regression curve (to assess the validity of the isogloss prototype model) are depicted in black together with their residual sum of squares (RSS) and steepness. Intensity values tend to fluctuate along the cross-sections, forming local peaks and depressions, in most cases not matching the prototype models. In the middle and the bottom rows, respectively, subdivisions resulting from the α and the γ strategy are represented, with regression lines fitted in

TABLE 5.2: Aspect angles of planar trend surfaces fitted to the variable intensity landscapes used in this study (two main variants MV1 and MV2), as well as the difference of the two main variants from being exactly in opposite direction (i.e., difference from 180°). For reference, aspect angles of a minor variant in Question A and of a Type II variable are given.

Variable	**Aspect angle of the first main variant *(MV1)***	**Aspect angle of the second main variant *(MV2)***	**Difference** *(MV2 – 180°) – MV1*
A	47.75°	231.03°	3.28°
B	23.78°	203.172°	0.608°
C	46.401°	229.495°	3.094°
D	33.856°	216.48°	2.624°
E	99.154°	278.986°	0.168°
F	57.237°	234.269°	2.968°
G	49.2°	227.564°	1.636°
H	161.42°	338.302°	3.118°
A: minor variant (third most frequent) and MV2	Third most frequent variant (…um zu 19% of all): 202.25°	231.030°	28.78°
Type II variable:	86.108°	243.208°	22.9°

each subdivision, including 95% confidence intervals shown in grey. Subdivisions resulting from the α strategy result in two fairly homogeneous areas for Question F but are more inhomogeneous areas in the other cases. In the γ subdivision it is visible that by finding a transition zone the dominance zones can be kept more homogeneous and the undecided patterns are confined to a separate subdivision (the transition zone). Narrower transition zones imply that the transition is happening in a spatially more restricted area between the threshold values of intensity, suggesting a sharper transition.

As demonstrated by the examples depicted in Figure 5.6, the cross-section plots allow to visually assess in an intuitive manner the fit of benchmarks representing the two prototype models with the given linguistic variables. Adding to this the different subdivision strategies (none, α and γ) provide a further facility to visually explore the validity of the two prototype models. As was to be expected, Question C has the best fit with the linear benchmark (visible also in the RSS values), and hence appears to be a good (though not perfect) representative of the inclined planes model. This is further supported by the rather narrow confidence bands of the α subdivision, with homogenous gradients in both partitions, and the fact that the α subdivision is placed exactly in the middle of the cross-section. On the other hand, Question F shows the worst fit with the linear benchmark, but a visually good fit with the logistic regression lines, and both the α and the γ subdivision strategies show narrow confidence bands; this suggests a good correspondence to the isogloss

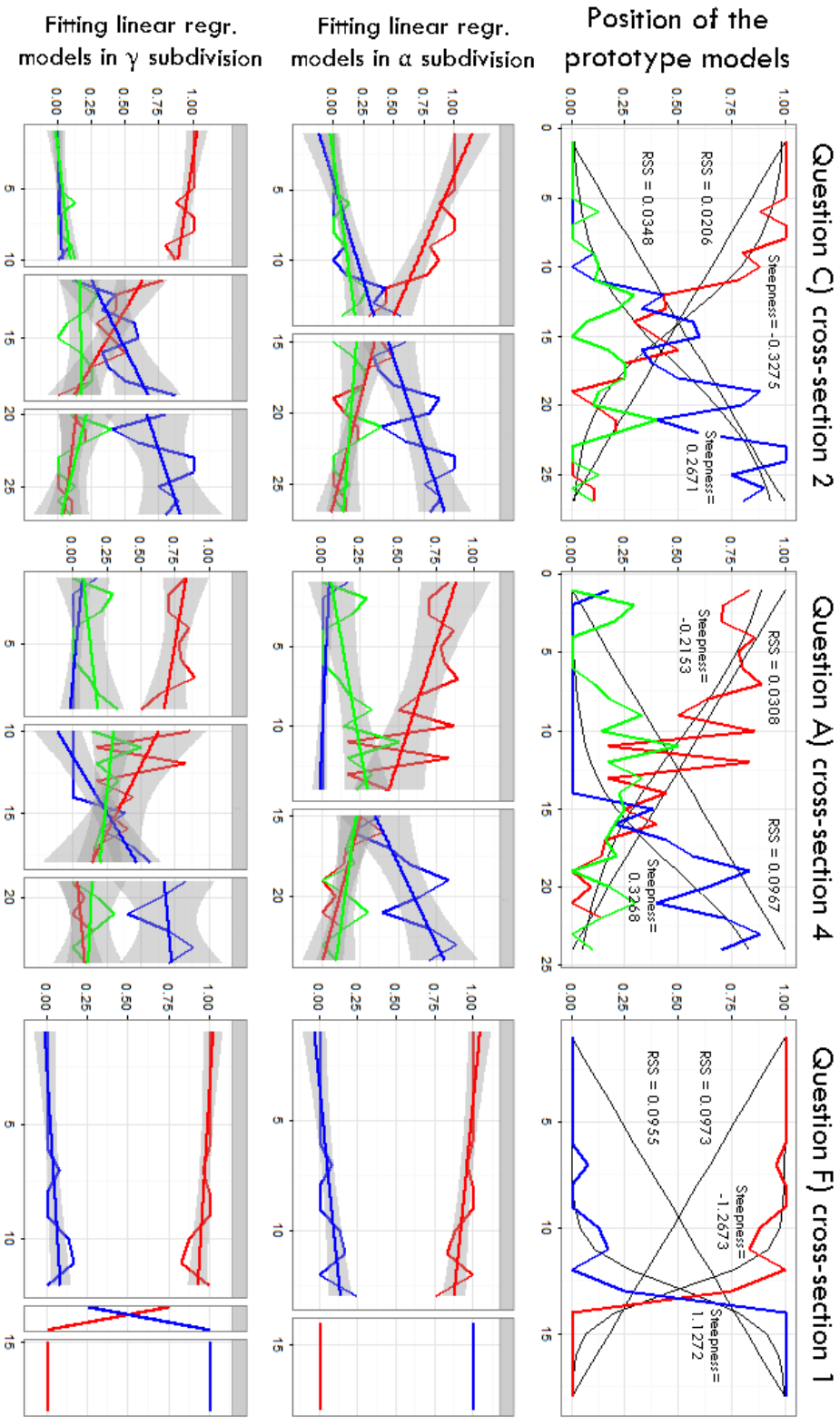

FIGURE 5.6: Typical cross-sections corresponding to the two main conceptualisations of boundaries and transitions used in linguistics: left column – inclined planes, right column – isogloss. The middle column shows a more varied cross-section. Red and blue lines, respectively, are used for the two main variants, while green colour is used for a third, minor variant (if it exists). The top row also depicts, in black, the benchmark line used to test for the inclined planes model and a logistic regression curve fitted to each main variant's intensities. The middle and the bottom row feature the spatial subdivision from the α and the γ strategy with linear trendlines fitted in each zone.

model. It has to be noted, though, that only the partitions on the left consist of several data points, while the other partitions each contain only two data points, making a perfect fit with a trend line trivial. Finally, the cross-section plots also allow to assess that Question A indeed takes the middle ground between a gradual and a sharp transition, though it appears to be leaning more towards the inclined planes model, as it shares more similarities with the patterns found for Question C.

Some biases stem from the construction of the cross-sections. Topological distances between survey sites on a given cross-section disregard positions in space. Geographic distances between pivotal points of polygons (the midpoints of the transecting line of each polygon) on the other hand also do not correspond to the actual Euclidean distance between neighbouring survey sites. Also, the order in which survey sites are linked after one another along the cross-section (Figure 5.5) might not follow natural paths of contact between them, especially in mountainous areas. The more subjective γ subdivision strategy is aimed at resolving these biases, avoiding discrepancies stemming from incorporating outlying survey sites into cross-sections and into spatial subdivisions.

5.5.3 Evaluation measures

Several measures have been proposed above for the quantitative evaluation of the results of curve and surface fitting. We start with measures at the global level, followed by measures applied on subdivisions, including cross-sections.

Logistic regression surfaces. After fitting global logistic regression surfaces to the eight variables used in this study (e.g., Figure 5.2, top right), steepness values were calculated for each surface, along with deviance values to measure the fit of the surface (Table 5.3). For the validity of the isogloss model high positive steepness values are expected coupled with low deviance values. Deviance depends on the homogeneity of the variant's intensity values, thus on the distribution pattern, which varies on a larger spatial scale. Questions E to H have large deviance values, meaning poorer fit of the logistic regression surface. Regarding steepness values, save for very typical representatives of both concepts (Question A – 0.4339 and 0.4792 and Question F, 1.0531 and 1.1602 respectively) the values do not correspond to the baseline expectations (cf. Section 5.4.3). This can be due to varied patterns of distribution skewing the logistic surface fitted using least squares. Such (lack of) correspondence to the expectations on the global level indicates that a more local level analysis could prove more meaningful.

Linear benchmark surfaces. The validity of the inclined planes prototype model is tested at the global level by calculating the fit of the planar benchmark surfaces to the intensity landscape of both main variants. Table 5.4 contains the goodness-of-fit values of the benchmarks, given as residual sums of squares (RSS). In all cases low values mean a better fit of the benchmark, which hint at the validity of the inclined

TABLE 5.3: Steepness and deviance values of the global logistic regression surfaces.

Variable	Logistic regression steepness		Deviance	
	MV1	MV2	MV1	MV2
A	0.4339762	0.4792534	78.84715	80.45831
B	0.6250814	0.2684783	38.62499	85.98724
C	0.6330907	0.5122899	73.94815	85.38857
D	0.7324318	0.550227	67.49912	76.30624
E	0.4756827	0.4777913	107.7375	106.5889
F	1.053185	1.1602516	102.5939	102.2498
G	0.4211606	0.4155639	187.568	190.9918
H	0.5766196	0.5727836	108.4067	92.34029

planes model. High RSS values, thus poor fits, are expected for the sharp transition variables. Based on Table 5.4 the prototype model fairly well fits the benchmark for gradual transition variables (A – 0.0384 and 0.0831; B – 0.0444 for its more dispersed variant; C – 0.0405 and 0.0489; D – 0.0439 and 0.059). The worst fit corresponding to the baseline expectations is found for the sharp transition variable Question E – more than 0.123 for both main variants. It has to be noted though that the presence of more than two variants can hinder the intensity landscapes from reaching the benchmark, resulting in worse fits, as the example of the top-left and top-centre plots of Figure 5.6 shows. Furthermore, the fixed ends of the inclined planes benchmark assume the transition to be happening midway through the study area, which does not hold for Question B, for example.

Residual maps. To visually assess the fit of the prototype models at the global level, residual maps were generated, as the example of Figure 5.7 shows for the case of Question F, a sharp transition variable. On the left of Figure 5.7 the residuals of the MV1 intensity values and the inclined planes benchmark are shown, while on the right the residuals of the logistic regression surface for the same main variant in Question F are depicted. In the map to the right residuals are smaller and concentrated in the area where variants are mixed. Missing residuals indicate perfect fit of the model surface, expected from a sharp transition variable. In the map to the left the gradual change in the residuals suggests intensity values being constant, compared to the inclination of the planar benchmark surface. The spatial distribution of residuals thus suggests that Question F better fits the isogloss model (represented by the logistic surface) than the inclined planes model represented by the planar benchmark surface. Fits of prototype models alone, however, do not reveal much about their validity. Due to its two inflections in an S-shape, the logistic surface simply has more flexibility to fit to a spatial distribution than a planar surface. Hence, the results of the residual maps are not surprising. However, the spatial autocorrelation pattern of the residuals may serve as a good supplement for describing the

TABLE 5.4: Global goodness-of-fit of the inclined planes benchmark surfaces given by residual sum of squares.

Variable	Main variant	Residual sum of squares
A	MV1	0.0384
A	MV2	0.0831
B	MV1	0.1020
B	MV2	0.0444
C	MV1	0.0405
C	MV2	0.0489
D	MV1	0.059
D	MV2	0.0439
E	MV1	0.1237
E	MV2	0.1235
F	MV1	0.0941
F	MV2	0.1009
G	MV1	0.0822
G	MV2	0.0830
H	MV1	0.0667
H	MV2	0.0849

patterns of transition, making the residual map a powerful analytic tool, similarly to numerous implementations in geostatistics (Chorley and Haggett, 1965; Beale et al., 2010).

Fitting surfaces in subdivisions. For ideal sharp transition variables the exclusive presence of main variants in their dominance area, accompanied by their absence in the other variant's dominance area are supposed (i.e., a binary spatial distribution). Fits of planes in the spatial subdivisions are calculated in the dominance zones resulting from each subdivision strategy, using the residuals of intensity values to 100% in their respective dominance zones and the residuals to 0% in the dominance zone of the competing variant. Table 5.5 features these fits for Questions A, C and F similarly as Figure 5.6 does for the cross-section case. The goodness-of-fit is presented as sums of squared residuals divided by the number of survey sites in the subdivision, for comparison across subdivisions and variables. In all cases lower values mean a better fit. For an ideal sharp transition variable, values close to 0 should be achieved by both variants (MV1 and MV2).

For Question F, which was visually classified as sharp transition variable (Section 5.3.2), better (low) fit values of deviance are found, corresponding to the expectation. Gradual transition variables, such as Question C, and to a lesser degree Question A, usually achieve a better fit in the β and γ dominance zones than in the bipartite α dominance zones, supporting our claim through visual analysis that it is possible to delineate more homogeneous dominance zones and transition zones where the larger part of change in intensity could be captured. On the other hand the sharp transition variable Question F also shows (at least partially) good fits in β and

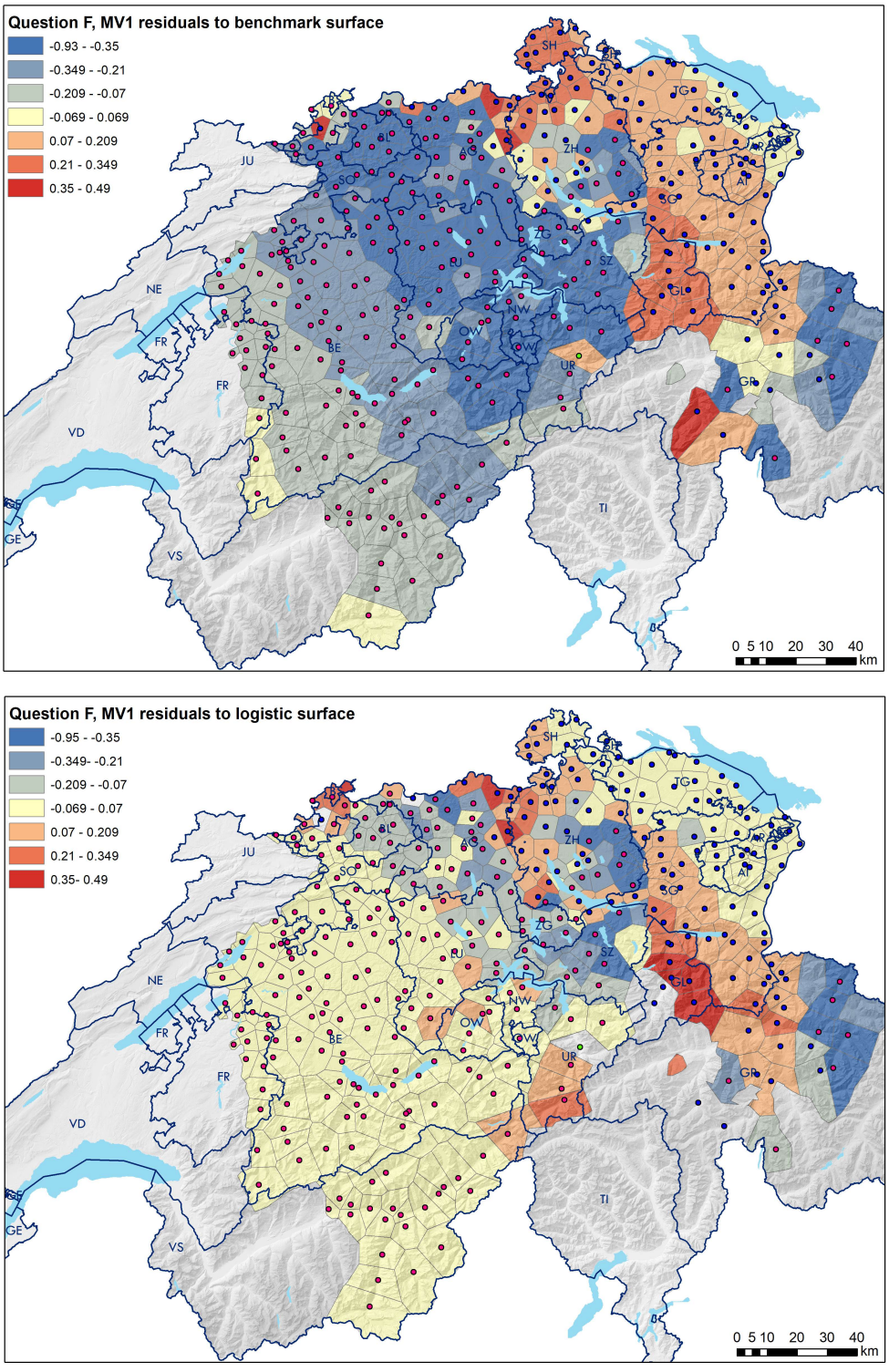

FIGURE 5.7: Maps representing the residuals of the MV1 intensities of Question F from the inclined planes benchmark surface (top), and from the fitted logistic regression surface (bottom).

TABLE 5.5: Correspondence to the binary spatial distribution. Fits of the planes of 100% and 0% in subdivisions, respectively, given by sums of squared residuals averaged by the number of survey sites in the dominance zone. Dom-1 = dominance zone of the first main variant, Dom-2 = dominance zone of the second main variant. α, β and γ denote the different spatial subdivision strategies.

Main variant	α **Dom-1**	α **Dom-2**	β **Dom-1**	β **Dom-2**	γ **Dom-1**	γ **Dom-2**
A) MV1	0.1854	0.0393	0.0765	0.0282	0.0614	0.0128
A) MV2	0.0285	0.2470	0.0028	0.2595	0.0011	0.1158
C) MV1	0.0866	0.0428	0.0401	0.0232	0.0234	0.0120
C) MV2	0.0485	0.1497	0.0305	0.1096	0.0332	0.0657
F) MV1	0.0483	0.0830	0.0170	0.1859	0.0167	0.1019
F) MV2	0.0484	0.0850	0.0150	0.2055	0.0143	0.1143

γ dominance zones, which does not completely meet our expectations, due to the presence of transition zones. However, as already noted above for the cross-sections (Figure 5.6) the tripartite subdivisions come very close to the bipartite subdivision for Question F, as the transition zone is so narrow.

In order to compare transitions across variables and to further assess the validity of the isogloss concept, slope values of planar trend surfaces were calculated in spatial transition zones of both tripartite subdivision strategies for main variants. In Table 5.6 these slope values are accompanied by the mean widths of the transition zones in their respective cross-sections. However, not in all cases were four transition zones found (corresponding to the four cross-sections per variable), as indicated by the numbers in parentheses. According to Glaser (2014), zones of mixture are assumed to be rather inhomogeneous, thus it is of interest to see whether trend surface analysis in transition zones is able to establish meaningful trends.

Transition zones of the β subdivision strategy contain the survey sites between the constant thresholds of 62.5%. Thus the distance between thresholds is assumed to account for the abruptness of the transition between the thresholds. The γ subdivision strategy sets more subjective thresholds, allowing to override small-scale 'noise' that stems from the construction of the cross-section lines. Thus, in γ subdivisions the width of the transition zones may not objectively correspond to the graduality of the transition.

Small average transition zone widths and lacking transition zones (i.e., not determinable for all four cross-sections) is supposed to indicate validity of the isogloss model, while wider transition zones are assumed to correspond to more gradual transitions. Typical gradual transition variables (Questions A, C) have widths above 50 km on average, while sharp transition variables (Questions E, F) stay typically below 30 km on average and have missing transition zones. Variables showing more irregular distribution (Questions G and H) also have rather narrow transition zones on average. The gradual looking but more undecided Question B and D have wider

TABLE 5.6: Slope values of the spatial transition zones accompanied by the mean value of the widths (geographic distance) of the transition zones in the four cross-sections, for the β and γ spatial subdivision strategies. β slopes range from green to red colours; γ slopes from blue to red. The number of cross-sections found using the subdivision strategy are given in parentheses (if lower than 4, which corresponds to the four cross-sections constructed for each variable).

	β slope	γ slope	Average width of the cross-section transition zones where they exist (km)	
			β	γ
A) MV1	0.3814	0.4842	60.1	72.9
A) MV2	0.2269	0.3939		
B) MV1	0.4216	0.2571	(2) 85.13	(3) 66.33
B) MV2	0.5018	0.3018		
C) MV1	0.5639	0.6004	50.02	66.08
C) MV2	0.4877	0.5491		
D) MV1	1.0572	0.8527	(3) 28.35	49.18
D) MV2	0.5335	0.7118		
E) MV1	0.0675	0.1280	22.21	26.52
E) MV2	0.0986	0.1300		
F) MV1	0.4325	0.8015	(2) 27.83	21.06
F) MV2	0.4289	0.9082		
G) MV1	0.2416	1.1421	(3) 6.1	24.2
G) MV2	0.2591	1.1928		
H) MV1	0.3280	0.3231	28.75	36.34
H) MV2	0.1441	0.1258		

transition zones, however some are missing. In general the transition zone widths have a fair correspondence to the expectations.

Slope values, however, in narrow transition zones are not as high as expected. For sharp transition variables (Questions E - H), the values are outright low in the β subdivision strategy, while for the γ strategy, both the highest and lowest values are found for them. A possible reason for this might be the subjectivity of the strategy: only lower values or values closer to each other are captured between the thresholds (γ strategy searches for the steepest breaks), resulting in less steep intensity surfaces.

In narrower transition zones steeper slopes are expected for the same thresholds. Pearson correlation of slope values with the average transition zone widths resulted in r = 0.2215 for the β subdivisions and r = -0.3164 for the γ subdivisions. These results only partially and only weakly confirm the expectation (negative correlation). Thus, slope values in the transition zones are not a function of the transition zone widths and consequently, transition zone widths are not a function of the threshold values. It has to be noted though that the width values are averages of four or less transition zones in cross-sections with considerable variation of the individual widths. In addition, the aspects of the planar regression surfaces in the transition zones are mostly highly different from their global level counterpart (not shown here, for reasons of brevity). The underlying reason might be that some of the variants (such as in the case of Question G) corresponding to more gradual transitions seem to form a subtype with regards to dispersion patterns, which was characterised by Glaser (2013), as mentioned in Section 5.3.1. This subtype is assumed to show a sharp transition between the transition zone and the dominance zone of the variant.

In conclusion the slope measure and relative slope differences in the transition zones are not powerful explanatory tools neither for comparison, nor for validating the prototype models, as these values crucially depend on the number and intensity values of survey sites captured. These are more diverse in the transition zones than in the dominance zones.

Cross-sections. Due to this diversity a more local analysis was conducted using the cross-sections and their subsets. Table 5.7 presents the steepness and deviance (measure of fit) values of logistic regression curves in the cross-sections of Questions A, C and F. Deviance is given as an absolute value for the variant and also as a proportional value, divided by the number of survey sites along the cross-section.

In general, the gradual transition variables Questions A and C have low steepness values coupled with relatively low deviance values. In Table 5.3, three rows are highlighted by black borders, as they correspond to the cross-sections shown in Figure 5.6. Question C Cross-section 2 (cross-sections are numbered from NW to SE on Figure 5.4) shows a steepness value higher than average, with low deviance values, even though it has been visually classified as a gradual transition variable. Question A Cross-section 4, representing a more undecided pattern in cross-sections, features one less steep and one steeper logistic curve, with good fits for both. Question

TABLE 5.7: Steepness and deviance (measure of fit) of logistic regression curves in the cross-sections (cross-sections are numbered as shown in Figure 5.4). The rows highlighted by black boxes correspond to cross-sections featured in Figure 5.6. Deviance is given as an absolute value along the cross-section for the variant and also as an average value, divided by the number of survey sites along the cross-section. The colours for deviance cells range between green and red, while colours for steepness cells range between blue and red.

Cross-section	Deviance MV1	Average deviance in MV1	Steepness MV1	Deviance MV2	Average deviance in MV2	Steepness MV2
A 1	5.4775	0.2608	-0.1218	3.2085	0.1528	0.1119
A 2	3.5974	0.1332	-0.2077	5.66	0.2096	0.2323
A 3	11.441	0.3365	-0.174	6.3168	0.1858	0.1806
A 4	3.3531	0.1397	-0.2153	3.6472	0.152	0.3268
C 1	5.6355	0.2684	-0.2294	2.7303	0.13	0.2854
C 2	2.7548	0.102	-0.3275	3.987	0.1477	0.2671
C 3	7.8044	0.2295	-0.2528	9.3509	0.275	0.1734
C 4	3.4635	0.1443	-0.3407	2.5897	0.1079	0.3402
F 1	2.0742	0.1152	-1.2673	2.4456	0.1359	1.1272
F 2	4.104	0.1368	-0.5202	3.6564	0.1219	0.547
F 3	6.0173	0.177	-0.4625	7.1346	0.2098	0.4134
F 4	4.1	0.1708	-0.8377	7.82E-09	3.26E-10	40.1053

F Cross-section 1 has the steepest logistic curves recorded, with a good fit, representing a typical sharp transition. For other sharp transition variables however, the picture is mixed, which might, again, point at a subtype in transition patterns.

The characteristics of the logistic regression curves do not correspond very well to the visual interpretation of Figure 5.3 (Section 5.3.2). Steepness values show a high variation and no clear patterns; due to the rather low number of survey sites contained in the cross-sections (17 to 36) the curve fitting and steepness estimation is susceptible to outliers. Furthermore, if the transition takes place closer to one side than to the centre, then the steepness values of the two curves are different even though the actual transition patterns in the cross-section seem visually symmetrical. In summary, solely based on the steepness of the logistic curves it is not possible to validate the isogloss model.

Rates of change. Having had little success using the steepness of logistic curves and slope values of linear regression, rates of change were calculated in the transition zones of the cross-sections, making use of actual intensity values rather than fitted models. Three kinds of measures are shown in Table 5.8 for the cross-sections and both main variants in Questions A, C and F (cf. Section 5.4.4): The absolute rate of change (the difference between the intensity values at the two end points) of the γ transition zone is given as per cent change in proportion per 10 km, for the sake of comparability across transition zones of different variables, which is usually on the scale of tens of kilometres. The cumulative change in intensity values along

the cross-section is also given as per cent change in proportion per 10 km for both tripartite subdivision strategies (β, γ). The width of each transition zone is given as geographic distance.

TABLE 5.8: Rates of change in the transition zones calculated for the β and γ subdivision strategies, represented by the cumulative change in intensity, and by the absolute rate of change, given in percents per every 10 km along the geographic distance. In addition the width (geographic distance) of each transition zone in km is given.

Cross-section and main variant	**β transition zone, cumulative change in intensity (%/10 km)**	**γ transition zone, cumulative change in intensity (%/10 km)**	**Absolute change in γ transition zone (%/10 km)**	**β transition zone width (km)**	**γ transition zone width (km)**
A 1 – MV1	55.44	37.61	10.03	66.37	39.88
A 1 – MV2	33.14	33.6	0		
A 2 – MV1	25.23	21.27	11.20	55.48	89.32
A 2 – MV2	33.34	29.11	11.20		
A 3 – MV1	24.45	35.49	10.82	62.58	92.43
A 3 – MV2	51.93	40.89	7.57		
A 4 – MV1	38.78	41.15	10.29	55.95	69.99
A 4 – MV2	14.30	13.29	8.143		
C 1 – MV1	37.94	59.96	12.05	33.21	66.37
C 1 – MV2	43.36	35.56	7.53		
C 2 – MV1	17.65	25.65	16.40	35.11	47.56
C 2 – MV2	23.06	31.75	16.19		
C 3 – MV1	24.11	24.64	11.62	80.87	86.02

C 3 MV2	–	52.05	56.72	11.62		
C 4 MV1	–	17.69	27.34	13.05	50.87	64.37
C 4 MV2	–	16.51	24.54	11.81		
F 1 MV1	–	0	115.63	115.63	0	6.48
F 1 MV2	–	0	115.63	115.63		
F 2 MV1	–	24.59	41.13	19.58	32.93	45.46
F 2 MV2	–	30.97	43.33	19.58		
F 3 MV1	–	47.96	50.11	1.86	22.72	26.94
F 3 MV2	–	61.61	56.42	1.48		
F 4 MV1	–	0	164.48	164.48	0	5.35
F 4 MV2	–	0	186.91	186.91		

Intuitively sharp transition variables are expected to have higher rate of change values for the absolute rate of change in γ, while gradual transition variables are expected to have lower values. The values obtained correspond to this expectation, with higher values (typically above 30) for the sharp transition variables (Question F in Table 5.8), while the gradual transition variables (Questions A, C) obtain lower values (typically below 20). Of the cross-sections depicted in Figure 5.6, Question C Cross-section 2 has low rates of change, as expected, while Question A Cross-section 4 has even lower absolute rates of change, showing the effect of subjectivity when choosing the γ threshold. Question F Cross-section 1 has a high rate of change, as expected, but there are only two sites contained in the transition zone, which generally is a possible source for skewed values. Nevertheless, few sites in the transition zone also means a narrow transition zone, which again supports the sharp transition model.

A possible artefact of the survey site distribution pattern is that a cross-section might be positioned in a direction of change that is atypical (e.g., for Question H at N – S direction; see also Maps 2 and 3), and may result in unexpected rate of change values too. Also, although the difference of the transition zones has been accounted

for by dividing the rates of change by the number of survey sites, a low number of survey sites in a transition zone skews the values more than a higher number of survey sites would. The cumulative change measure accounts for the fluctuations and spatial uncertainty in the spatial variation pattern of the given variant, but as several scenarios can result in the same cumulative rate of change value, it is less optimal for comparison.

In conclusion, the rate of change measures generally correspond to the expectations established in preliminary visual interpretation and thus the typical variables for both conceptualisations of dialectal transitions can be classified safely into either sharp or gradual transition variables.

Summary and recommendations. It is difficult to find an infallible combination of quantitative characteristics that allows to safely distinguish between sharp vs. gradual transition variables, and thus between cases where the isogloss concept holds, versus those where the dialect continuum concept applies. In particular, it proved rather hard to define a sharp transition, and thus suggest the presence of isoglosses, based on the models and methods used in this study. However, narrow or non-existent transition zones and high rates of change in transition zones are a definite prerequisite. Furthermore good fits of planar benchmarks in α subdivisions (as shown in Table 5.5) are good predictors as well.

It is more straightforward to establish the validity of the inclined planes conceptual model for a variable: a good fit of the benchmark in Table 5.4 is needed. Furthermore, a combination of wide transition zones, low absolute rate of change in the γ subdivision and relatively low rate of change in the transition zones suggests an inclined planes variable.

Based on the results of the study reported in this study, one may try to establish a ranking of the various evaluation measures proposed, from best to worst performing:

- Width of transition zones and missing transition zones (Table 5.6)
- Absolute rate of change in intensity in the γ subdivisions (Table 5.8)
- Cumulative rates of change in transition zones (Table 5.8)
- Goodness-of-fit to inclined planes benchmark for gradual transition variables (Table 5.4)
- Logistic curve steepness in cross-section – sensitive to outliers (Table 5.7)
- Steepness of global logistic surface – sensitive to the variation of the intensity values and the position of the transition relative to the middle (Table 5.3)
- Fits of the planar benchmarks – sensitive to the variation of the intensity values and the position of the transition (Table 5.5)

- Slope values of planar trend surfaces in transition zones – prone to errors due to high variability and small number of survey sites in the transition zones (Table 5.6)

5.6 Conclusion

In this study two quantitative prototype models were proposed for the two key conceptualisations of boundaries and transitions between dominance zones in dialectal variation: fixed inclined planes to model the *dialect continuum* concept, and logistic regression surfaces and lines as a model for the *isogloss* concept. A methodology was then proposed to quantitatively model such transitions as gradients in the spatial distribution of intensity values of dialectal variants. This methodology includes several methods based on curve and surface fitting using regression line analysis, trend surface analysis, and analysis of residuals, complemented by a variety of evaluation measures that are used to characterise the goodness-of-fit and gradient of the fitted lines and surfaces, as well as the existence and width of transition zones. The proposed methods were applied to data from the Syntactic Atlas of German-speaking Switzerland (SADS), testing the validity of the prototype models as well as the feasibility and performance of the proposed methods. The analyses have been conducted at different spatial levels: the whole study area (global level), in cross-sections cut into intensity landscapes of each variable, and in different spatial subdivisions at the global level and in the cross-sections alike.

We conclude that the intensity landscape of many dialectal variables is too noisy to be classified safely into either sharp or gradual transition variables. Nevertheless, the methods and evaluation measures proposed in this study allow to characterise different variables with regard to their prevailing transitions, and they allow comparing different variables using the computed characteristics. As it is clear that transitions in dialects cannot be compared using a single characteristic, a combination of different methods and measures has been proposed, similarly to the measures introduced by Rumpf et al. (2010) which they used for the characterisation of dominance zones. Furthermore, the spatial subdivision strategies proved useful in studying dominance and variation patterns in more detail. In particular, the tripartite subdivision strategies helped to demarcate transition zones between dominant usage areas of dialectal variants.

Several directions for future research may be noted. First, the proposed methods and measures could be applied to other dialect databases to further evaluate and possibly improve these tools. Second, to be more independent of subjective decisions, the detection of the boundaries of transition zones should be performed in a data-driven and automated manner. To improve the intuitive spatial subdivision strategy used in this study, the authors have developed a data-driven method to delineate transition zone boundaries (Jeszenszky, Derungs, et al., in prep. and Chapter 6). Such data-driven procedures might also be of interest for the analysis of other

kinds of intensity data with similar spatial distribution characteristics, in linguistics and beyond. And finally, it would also be interesting to compare the boundaries of dominance and transition zones to extralinguistic and geographical boundaries.

Chapter 6

Data-driven detection of transitions in dialectal variables

6.1 Introduction

Identifying transition areas and boundaries in interdialectal variation, as well as their characteristics, has always been an important research question in dialectology, linked to dialect area formation (Section 2.3.1; Chambers and Trudgill, 2004:104-105).

Dialectology usually conceptualises the spatial relationship between dialectal variants in two different ways: *dialect areas* and *dialect continua* (see Chapter 2 and Chapter 5). The concept of 'dialect areas' assumes relatively sharp boundaries between variants in individual variables or their aggregates, whereas 'dialect continua' are based on the assumption of boundaries being gradual transitions (Heeringa and Nerbonne, 2001). For individual variables, these boundaries are termed *'isoglosses'*. The premise, then, is that such boundaries in interdialectal variation are detectable in space (e.g., Bowern, 2013).

6.1.1 Isoglosses: The traditional classification method

Isoglosses (as introduced in Section 2.1) first appear as manual annotations on point symbol maps, traditionally found in dialect atlases, which are based on data from dialect surveys. In such cases, isoglosses were commonly used to demarcate boundaries where dialect change takes place. Isoglosses assume linguistic areas to be homogeneous and boundaries between them to be rather sharp and abrupt. Thus, they allow for global interpretation of the distribution patterns in a linguistic variable, but at the potential cost of losing local variation. Isoglosses are traditionally used in the delimitation of dialect regions. An early example is Haag's (1898) map (Figure 6.1), where isoglosses are drawn for individual dialectal variables with the aim of overlaying them and constructing what later was termed an isogloss bundle (Bloomfield, 1933).

Formal rules for drawing isoglosses are not discussed in detail in the literature. However, two competing requirements can be recognised (cf. Wrede, Mitzka, and Martin, 1927; Kurath, 1949; Chambers and Trudgill, 2004). First, isoglosses should separate the survey sites such that the areas of linguistic variants are homogeneous

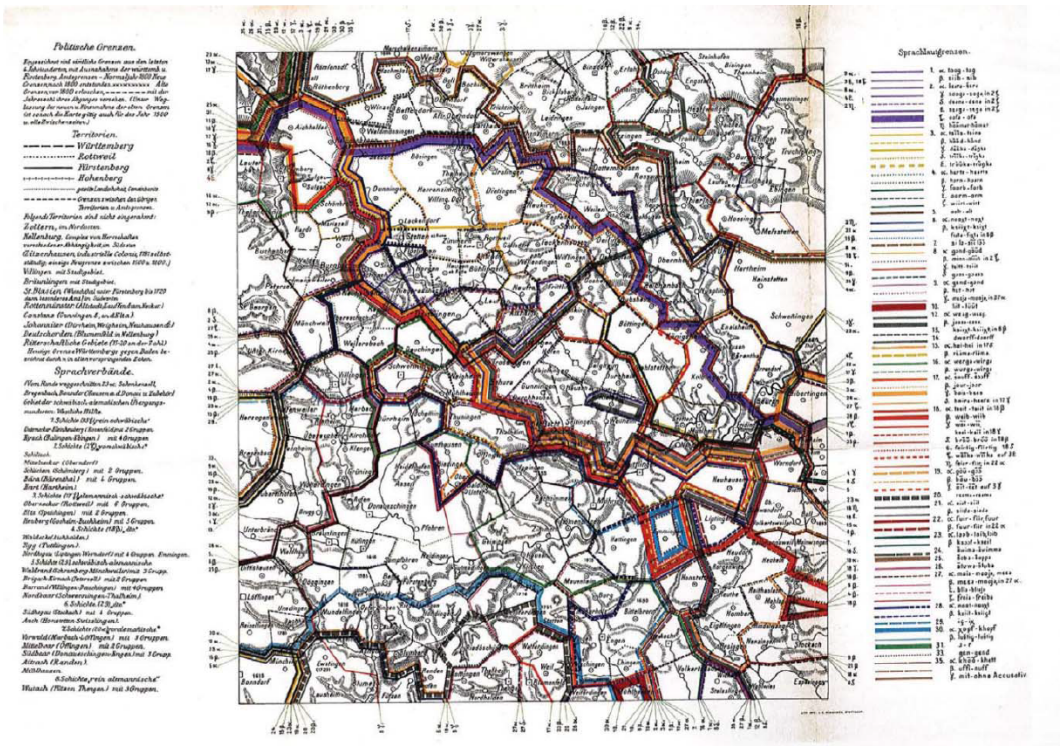

FIGURE 6.1: An early example map of isogloss bundles by Haag (1898).

on either side of the isogloss. Second, isoglosses should yield areas that simplify the often complex linguistic reality and disregard outliers to a certain degree. Balancing the trade-off between these requirements is a subjective decision of the analyst. Since isoglosses are by definition subjective and sharp they leave room for misinterpretations about the true nature of transitions in dialectal variables.

6.1.2 Continuous phenomena: Perception and reality

In dialect landscapes the transition between linguistic varieties is mostly gradual and not sharp (e.g., Heeringa and Nerbonne, 2001; Pickl and Rumpf, 2012). However, human perception about dialect landscapes is discrete. For example, people often categorise the Swiss German dialect continuum into entities such as 'Thurgau German' or 'Basel German'.

The mere fact that dialect maps often use isoglosses presumes that this discretisation is an intuitive human concept related to *classification*, exhaustively discussed in psychology (Rosch, 1973; E. E. Smith and Medin, 1981). Classification satisfies the inherent cognitive need for structuring continuous information into meaningful entities that, for instance, allow it to be used in communication (e.g., "Which dialect do you speak?").

Ultimately, recorded data and human perception may describe the same phenomenon in contradicting ways. Recorded data are capable of capturing the continuous nature of phenomena, while perception often leads to classifying the information for the sake of easier *interpretation*. The discrete perception of continua is also reflected in spatial cognition studies (Montello et al., 2003). For example, there is no exact definition to what a mountain is (B. Smith and Mark, 2003). The continuous nature of relief can be measured everywhere as elevation above sea level, but perception takes a more categorical approach. Our thinking about positioning ourselves in nature seems to be binary: *"I am on a mountain"* or *"I am not on a mountain"*; while categories such as *'50 % mountain'* or *'almost mountain'* do not exist (B. Smith and Mark, 2003). Spatial cognitive studies of language (e.g., Preston, 1989; Montgomery and Stoeckle, 2013) often use *'draw-a-map'* tasks to obtain cognitive information from laypeople about vague regions (such as a dialect area or 'downtown'; Montello et al., 2003). It is rare to find informants that draw zones of transition between the areas of perceived dialects.

Bounds (2015:36) summarises the relationship between isoglosses and the cognitive needs for discretisation: "The boundaries, or isoglosses, are generalisations created by our perceptions and facilitated by the methodologies used to study speech."

6.1.3 Quantitative characterisation of boundaries

There are only a few quantitative attempts to identify boundaries (i.e., *explicitly* boundaries) in dialect landscapes. Labov, Ash and Boberg (2006:42) proposed a method for the automatic delineation of isoglosses, however concerning only variables with two states (e.g., present/absent). Grieve et al. (2011) proposed quantitative methods analogous to finding isoglosses, isogloss bundles and dialect areas. Their focus, however, did not explicitly lie on the quantification of the boundaries, but on the delimitation of dialect areas. Lately Chagnaud et al.'s (Chagnaud et al., 2017) cartographic tool has implemented an automated isogloss drawing using smoothed Bézier curves. While their online tool *'ShinyDialect'* produces pleasing maps and includes the handling of outliers simulating interpolation in traditional dialectology, the tool only handles single answers per survey site.

Isoglosses are, however, not always adequate representations of linguistic boundaries. Labov, Ash and Boberg (2006) express the need for an automation that disregards the preconceptions about boundaries being sharp. Grieve et al. (2011:2) cast the above preconception into the quote "plotting an isogloss does not test if a regional pattern is present. A pattern is assumed to exist, and then an isogloss is plotted."

Several studies model the gradual transition of dialectal variants in space (see Section 2.3). Notably, the Augsburg/Ulm dialectology group has devised a 'dialectometric intensity estimation' technique based on kernel density estimation (KDE), which considers the local neighbourhood and thus smooths the intensity values and removes outliers, as a precursor to identifying dominance areas of variants in

individual variables (Rumpf et al., 2009) and to detecting and comparing spatial structures (Rumpf et al., 2010). Their method aims to overcome local uncertainties stemming from the scarcity of responses per survey site in their database (SBS). The method was further developed by Pickl et al. (2012; 2014) and Sibler et al. (2012). *Dialectometric intensity estimation* is probabilistic in the sense that it estimates the probability of a variant at each location. Thus, dialectal variation is initially modelled as a continuum with gradual transitions. However, eventually the boundaries between dominance zones are effectively made crisp: a boundary is the location where dominance changes, i.e., Variant A becomes more dominant than Variant B, and using a relative majority rule there is always only one dominant variant per survey site.

6.1.4 Research gap

Section 1.2.1 identified the following research objective for this chapter:

***Research Objective* 3. Automate the detection of interdialectal boundaries and transitions in a data-driven way.**

The literature shows that sharp isoglosses do not appropriately model dialect boundaries, as transition in dialect data is inherently gradual. Moreover, it was shown in Chapter 5 that there is a spectrum of different gradualities of transitions (Parker, 2006). Consequently, gradual models with predefined parameters might not capture the nature of a dialect boundary any better than sharp isoglosses. Survey data with appropriate levels of spatial and attribute granularity are capable of reflecting the true nature of dialect boundaries. Thus, the model to conceptualise boundaries[1] should not be defined *a priori*. This leads to the following research gap:

A method is missing that enables the identifcation of interdialectal transitions and boundaries that reflect the linguistic variation as conveyed in data. The methodology should neither assume sharp boundaries, nor superimpose a predefined gradual transition. The nature of the transition should thus be inferred from the data and should form the basis for subsequent boundary delineation.

In this chapter we propose such a methodology, which satisfies the following three requirements:

- **Infer the graduality of dialect boundaries from the data**

  The method should not superimpose sharp or gradual dialect boundary models, but infer the graduality from the data, along Parker's spectrum (2006). Boundaries should be inferred for each individual variable, thus allowing for a quantitative comparison of boundaries across variables.

- **Handle spatial bias**

[1] Note that 'boundaries' in this chapter (see also Section 2.1) and in the proposed delineation procedure does *not* represent a boundary *line*, that is, *not an isogloss*, but rather a boundary area. In that sense 'boundary' is synonymous with 'transition zone'.

The spatial distribution of survey sites is often irregular, which might lead to a bias when identifying dialect boundaries. The method proposed should account for the influence of this bias.

- **Emulate traditional isogloss delineation in an automated manner**

  The method should emulate the traditional technique of isogloss delineation in an automated manner. Although contested, the isogloss model offers a well-established method for delimitating boundaries for dialectal variables. Most importantly, isoglosses seem to correspond to human perception and interpretation of linguistic data. However, in contrast to traditional isoglosses, the proposed method should yield probabilistic results, reflecting the graduality of the boundaries and helping to overcome subjective decisions in spatial discretisation.

6.2 Data and Methodology

6.2.1 The SADS database

The *Syntactic Atlas of German-speaking Switzerland* (SADS) database, used in this study, has been described in detail in Chapter 3. The most important characteristic of this database is that multiple responses are present at each survey site for each variable. With regards to the general distribution patterns of variants, variables of the SADS have been classified into three *distribution types* by Glaser and Bart (2012, 2015), which are defined as Type I, II and III variables in Section 5.3.1. The method proposed in this chapter is optimally used with distribution Type I, where two main variants constitute large, relatively homogeneous dialect areas. These variants seem to compete with each other in space. The transition between their prevalence zones (see Section 2.1) shows differences in graduality ranging from sharp to completely gradual. Type I variables can be further subdivided into two *transition types*. Based on the apparent graduality of the transition between the usage areas of variants, *sharp transition variables* and *gradual transition variables* are identified (Table 5.1). It has to be noted, however, that both the distribution types and transition types are fuzzy categories.

6.2.2 Overview of the aims and the overall workflow

In this chapter a method for automatically identifying boundary regions between dialect variants is introduced. The spatial pattern of the identified boundary regions represents the degree of graduality and thus allows interpreting them as sharp boundaries (i.e., isoglosses) and more gradual transition zones.

- The intuition behind the manual delineation of isoglosses, as practiced in traditional dialectology, is formalized and implemented as a data-driven procedure.

- The approach is probabilistic in the sense that multiple parameter settings are randomly drawn from normal distributions and tested in independent model iterations. The results from all iterations are aggregated and thus allow the gaining of insight on the robustness of the resulting dialect boundaries.
- Dialectal variables are processed separately. However, aggregated dialect boundaries for multiple variables may in principle be computed by summing the results from individual features.

In the following, a detailed description of the methodological approach is given.

6.2.3 Workflow

The method for detecting boundary regions between individual dialect variants follows a four-step workflow.

1. Detection of *'potential boundary locations'* (bl_{pot})
2. Delineation of bl_{pot} into *'potential dialect boundaries'* (db_{pot})
3. Identification of *'baseline boundaries'* (bb)
4. Statistical comparison of db_{pot} and bb in order to detect *'dialect boundaries'* (db)

Step 1.

The detection of **potential boundary locations** (bl_{pot}) is a filtering step for survey sites with dominant variants, i.e., the variant used by the majority of respondents, gaining less than a certain threshold percentage of responses. Figure 6.2 shows a fictional example of a variable with two variants, represented using triangles and rectangles, respectively. Symbol size represents the *'intensity'* of the variant, i.e., the degree of dominance or the percentage of responses in favour for the respective variant, with small symbol size indicating that the respective variant is used by a small majority of people only.

Survey sites highlighted in blue represent bl_{pot}. Independently of the type of dominant variant, these locations have an intensity below a certain threshold value. In the proposed approach, threshold values are drawn at random from a normal distribution with $\mathcal{N}(0.7,0.2)$. Figure 6.3 shows bl_{pot} (orange polygons) generated with different threshold values for two variables in the SADS atlas. The two variables are expected to show gradual dialect boundaries and sharp boundaries, respectively (see Section 5.3.2). The bl_{pot} in Figure 6.3 are generated using the threshold values 0.5, 0.7 and 0.9, which are the mean $\pm$ one standard deviation of the normal distribution used in this study.

In the case of the gradual transition variable (left), multiple variants are used by respondents at most survey sites, which causes intensity values of locally dominant variants to be relatively low. The sharp transition variable, on the other hand (right),

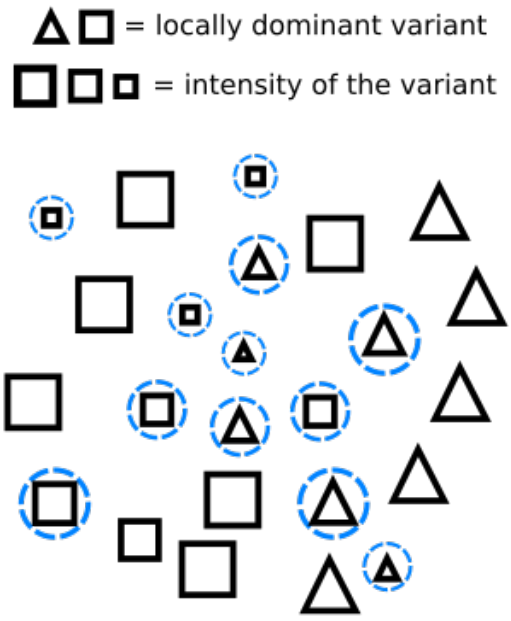

FIGURE 6.2: Determining *potential boundary locations* highlighted by blue circles in a fictional distribution of dialectal variants.

has most intensity values close to 1. As a consequence, more bl_{pot} are identified for the gradual variable, across all threshold values.

Step 2.

In this step, bl_{pot} are examined and, if the requirements are met, **delineated to form potential dialect boundaries** (db_{pot}). The logic behind this step is the following. For a survey site to be part of a dialect boundary, it is required to be a bl_{pot}. For a bl_{pot}, in turn, to be part of a dialect boundary, it is required to be in the proximity of other bl_{pot}. Spatial clustering is used to detect regions where multiple bl_{pot} in close proximity form db_{pot}.

The DBSCAN algorithm (Ester et al., 1996) is used for spatial clustering, with the locations in the of bl_{pot} set used as input. DBSCAN gains its popularity from its potential to find clusters of complex, non-concentric shape. This is an important prerequisite in this study, as isoglosses and transition zones are indeed expected to form complex spatial patterns. As input parameters, DBSCAN requires explicit information on the minimum number of points that constitute a cluster ($MinPts$) and the maximum allowed distance between points to be associated with the same cluster (ε). Guided by the results in Chapter 4 of this thesis, distance is measured as travel time in 1950. Travel time has been shown to outperform Euclidean distance in predicting syntactic distance and is thus expected to lead to better results in this study as well. This decision, however, is not further validated.

Figure 6.4 shows an example of one cluster, that is, a db_{pot} detected in the above fictional distribution of two variants. The bl_{pot} that are associated with one cluster are connected by red lines.

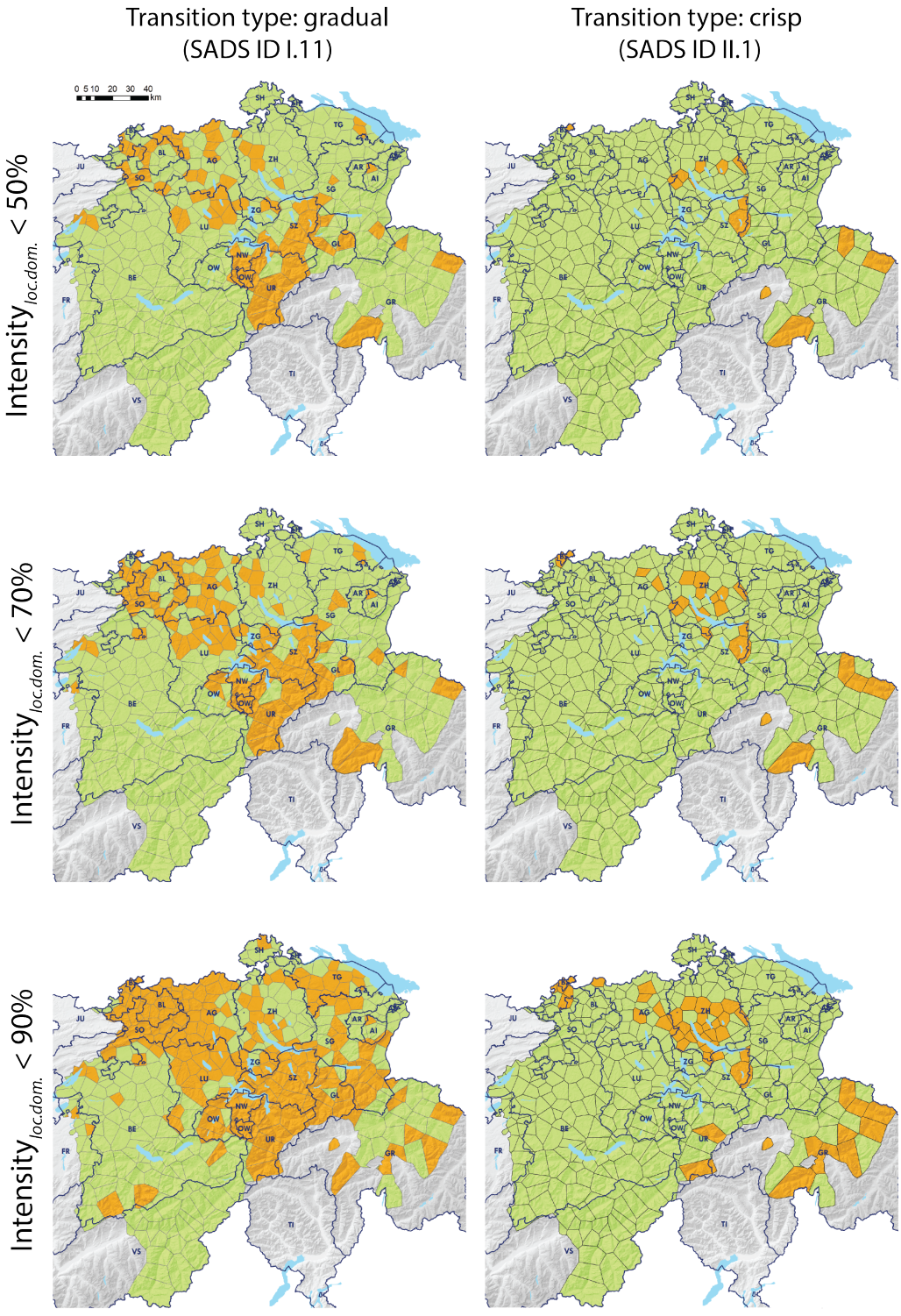

FIGURE 6.3: Potential boundary locations identified by using the intensity thresholds of 0.5, 0.7 and 0.9. On the left a gradual transition variable ('infinitival purposive clause' – I.11), while on the right a sharp transition variable ('position of the infinitive particle' – II.1) is shown. The resulting potential boundary nodes are displayed in orange.

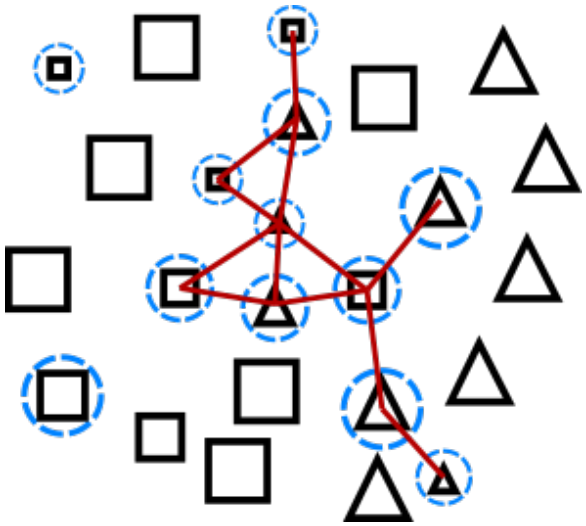

FIGURE 6.4: DBSCAN clustering applied to bl_{pot} (highlighted by blue circles). Survey sites (natural neighbours) that are clustered using one particular parameter setting are connected by red lines.

From this example it is clear that applying DBSCAN on bl_{pot} allows the distinguishing of those survey sites that constitute spatial patterns which resemble dialect boundaries from those bl_{pot} that should be considered as outliers. In DBSCAN outliers are defined as spatial isolates. Outliers detected are excluded from the follow-up analysis.

In theory, the 10 bl_{pot} in Figure 6.4 can be connected to 45 unique pairs. The SADS dataset consists of 383 survey sites. If each survey site occurs at least once in a cluster with each other survey site – which might not be the case –, approximately 70000 connections would need to be visualized in the final result. This large number clearly impedes interpretation. For this reason, only cluster associations between Voronoi neighbours are stored and visualised (as is the case in Figure 6.4). Survey sites that are Voronoi neighbours will be termed natural neighbours in the remainder of this chapter.

The two parameters $MinPts$ and ε have a crucial impact on the resulting db_{pot}. Therefore, the following series of parameter settings is incorporated in this study (Table 6.1). Results from all parameter combinations are subsequently aggregated and the number of times two bl_{pot} are grouped into one db_{pot} is counted.

ε (minutes)	10	17	25	180
$MinPts$	3	5	10	10

TABLE 6.1: The four pairs of ε and $MinPts$ values used in this study.

Large values for ε and $MinPts$ result in db_{pot} of large spatial extent, while small values lead to fine-grained db_{pot}. The combination of the above parameter pairs can therefore be considered a multi-scale approach (e.g., Fisher, Wood, and Cheng, 2004) for modelling dialect boundaries.

Step 3.

Clusters resulting from DBSCAN are now compared to **baseline boundaries** (*bb*). The likelihood of survey sites occurring in a cluster and thus constituting a db_{pot} is affected by the local dialectal intensity of variants, but also the geographic locations of bl_{pot}. The second effect is considered a bias and therefore should be accounted for in the final product of this analysis. For this reason, db_{pot} clusters are compared to *bb* clusters, computed exclusively from the geographic locations of survey sites. The computation of *bb* clusters is also carried out by DBSCAN, using the same combination of clustering parameters (Table 6.1) as for clustering db_{pot}. In contrast to db_{pot}, however, the computation of *bb* incorporates all survey sites, independent of the local intensity of variants. Figure 6.5 shows an example of *bb* computed for the above fictional example.

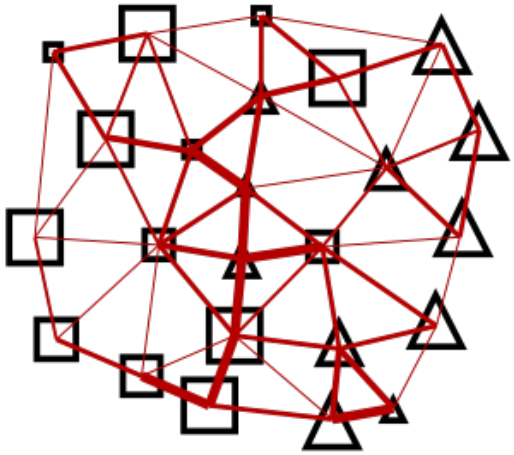

FIGURE 6.5: Visualisation of baseline boundaries (*bb*) computed from geographic locations of survey sites is visualized. Only cluster associations between natural neighbours are shown. Line width represents the number of times natural neighbours are associated in a *bb* across all DBSCAN parameter pairs (Table 6.1).

Line width represents the number of times two survey sites were associated in one *bb* across the iterations of DBSCAN with four pairs of parameters (Table 6.1).

Step 4.

Having computed db_{pot} and *bb*, it takes only a small step to generate the final product of this analysis, namely dialect boundaries (*db*). The number of times two survey sites (i and j) were associated in one db_{pot} (i.e., $\sum db_{pot(i,j)}$) is normalised with the number of times the two sites are associated in one *bb* (i.e., $\sum bb_{i,j}$), where $\sum bb_{i,j}$ is by definition greater or equal $\sum db_{pot(i,j)}$ (shown in Equation 6.1).

$$db_{ij} = \frac{\sum db_{pot(i,j)}}{\sum bb_{i,j}} \tag{6.1}$$

Figure 6.6 finally shows *db* values for the fictional example used throughout this methodology section. Line width now represents the likelihood of two survey sites occurring together in the dialect boundary for a fictional variable. As mentioned above, these likelihoods are generated by iterating over a series of threshold values and DBSCAN parameter settings incorporating a series of spatial scales.

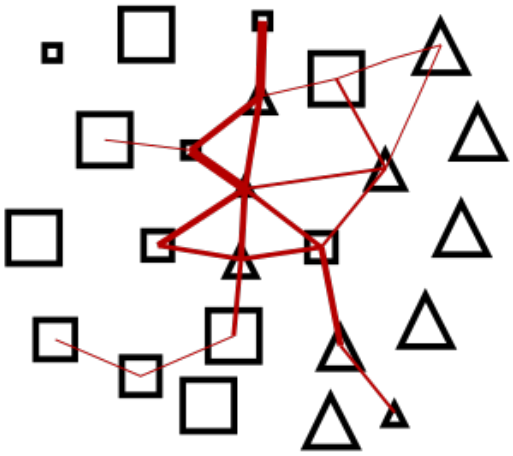

FIGURE 6.6: Representation of *db* in the fictional example. Line width represents the relative frequency of co-occurrence of survey sites in db_{pot}, as compared to *db*.

6.3 Results

6.3.1 Preliminaries

The proposed method was applied to 22 of the SADS variables[2], described in Table 3.1. The results shown in this chapter, however, are restricted to a small but representative subset of six variables, in order to demonstrate the principles and characteristics of the proposed method. Of the six variables, five are of Type I (as mentioned above, this is the preferred type for the proposed procedure), while one is of Type II.

For each variable, 20 runs were generated for the random threshold selection described in Step 1 of the proposed method, used to detect the potential boundary locations (bl_{pot}).

The results are presented using two types of maps. On the one hand, the *intensity maps* known from the previous chapters, conventionally used in other studies focusing on delineating dialect areas (Rumpf et al., 2009, 2010; Pickl, Spettl, et al., 2014; Sibler et al., 2012). On the other hand, a new type of map that we call *'transition maps'* is used, showing the links between natural neighbors, with the line width

[2]SADS Questions I.1, I.2, I.3, I.6, I.9, I.11, I.12, I.18, II.1, II.3, II.5, II.7, II.11, II.13, II.30, III.1, III.5, III.8, III.16, III.22, IV.7 and IV.14 were used in this study.

corresponding to their *db* value. The principle of this map type is shown in Figure 6.6.

In the following Section 6.3.2, we will first present results for individual variables exhibiting different variation patterns: a Type I variable with sharp transition, a Type I variable with gradual transition, and a Type II variable. Then, in Section 6.3.3, four Type I variables representing the same phenomenon, and thus expected to expose a similar variation pattern, will be compared to each other.

6.3.2 Individual variables

The presentation of results for individual variables serves to assess how the proposed method reacts to different types of dialectal spatial variation. We start with a typical Type I **sharp transition variable** (*'position of the infinitive particle'* – Question II.1), which is presented in Figure 6.7 in a transition map along with the corresponding intensity map. We make the following observations:

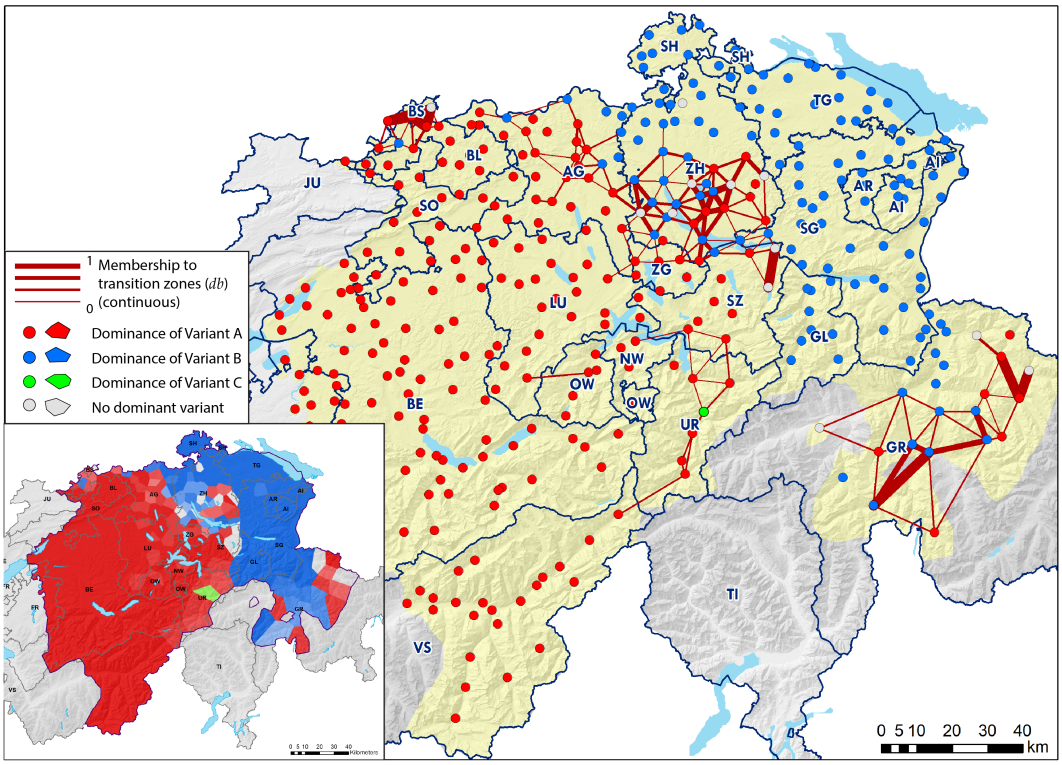

FIGURE 6.7: Transition and intensity maps of a sharp transition variable, II.1 – *'position of the infinitive particle'*.

- In the intensity map, two core areas belonging to the two main variants are visible with smaller areas of mixture. One mixture area is situated at the interface of the two variants, and a secondary in the southeast of the study area (Canton of Grisons; labelled 'GR'), where the locally dominant variants are more diverse.

- The same mixture areas are also present on the transition map. However, this time, the focus is on natural neighbour pairs that represent potential boundaries, highlighted by partially thick links connecting them.

- In the intensity map as well as in the coloured survey site points of the transition map, a survey site in the Canton of Basle-Country ('BL') and one in the Canton of Uri ('UR') appear to be outlier locations. The connection patterns, however, show a different picture around the two survey sites in the transition map. The node in BL is connected to its neighbours with links of different width, and there are thick links present in its vicinity as well. This suggests that the survey sites in the area were indeed often classified as parts of boundary clusters, while the node in UR shows only weaker connections to its neighbours, implying larger differences.

- Finally, in the transition map segments of what might be considered the main isogloss for the depicted variable are visible in the cantons of Aargau ('AG'), Zurich ('ZH') and Schwyz ('SZ'), respectively. However, we note that in areas where we might have expected an ideal configuration for detecting an isogloss, with a very abrupt change of intensity values (e.g., between Glarus 'GL' and UR, or GL and SZ), no links exist. Also, we note that in the mixture area in ZH, delineating a clear isogloss would seem difficult.

The above result is now contrasted with a Type I variable that represents a typical **gradual transition** (*'infinitival purposive clause'* – Question I.1), as shown in Figure 6.8. We observe:

- The number and distribution patterns of natural neighbour connections are very different from those of Figure 6.7, with connections present in a much larger region. The concentrated area where survey sites of different locally dominant variants are mixed, follows a northwest – southeast direction, as seen in both the intensity and the transition map. The presence of this large area of mixture indicates the gradual transition between the usage areas of the two most important variants.

- In the intensity map two core areas are visible with an area of mixed locally dominant variants in between, with a third variant reaching dominance in several local spots.

- This area of mixture of three variants also displays the most connections in the transition map. Additionally, some connections are present farther from the largest concentration of connections in the main transition zone, especially in the northeast. Areas around survey sites that, based on the locally dominant variant, appear to be outliers show different pictures in the transition map and in the intensity map, similarly to Figure 6.7. Good examples of such cases are

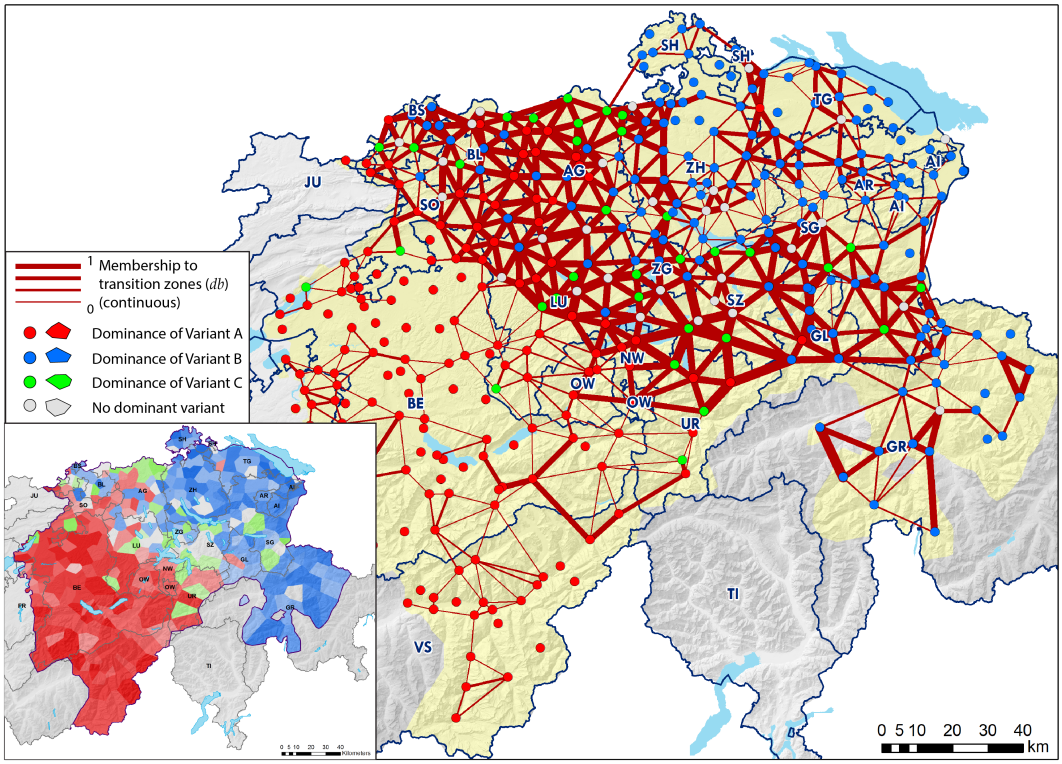

FIGURE 6.8: Transition map of a gradual transition variable, I.1 – *'infinitival purposive clause'*.

found in the south of the Canton of Lucerne ('LU') and in the west of Thurgau ('TG').

Figure 6.9 presents a **Type II variable**, (*'comparative clause linkage'* – Question III.22), the SADS map of which is shown in Figure 3.3. Such variables exhibit the dominance of one overarching variant along with other variants occurring and reaching dominance only in a certain local areas. Recall that in Chapter 5 the methodology for modelling transitions as gradients had been determined as unsuitable for such variables. Thus, the quantitative assessment of boundaries in Type II variables is revisited using the procedure proposed in this chapter.

- The intensity map shows the dominance of the blue variant almost everywhere, with only some survey sites indicating the dominance of another variant: clustered, such as the magenta variant, or scattered over a larger area, such as the green variant. The abundance of uncoloured and light blue polygons suggests, however, that generally a mixture of variants is present.
- In the transition map, clusters of connections always occur in the vicinity of survey sites where a minority variant is dominant. At the same time, dense connections are visible in larger areas of blue dominance, such as in Berne ('BE'), Aargau ('AG'), Zurich ('ZH') or Thurgau ('TG').

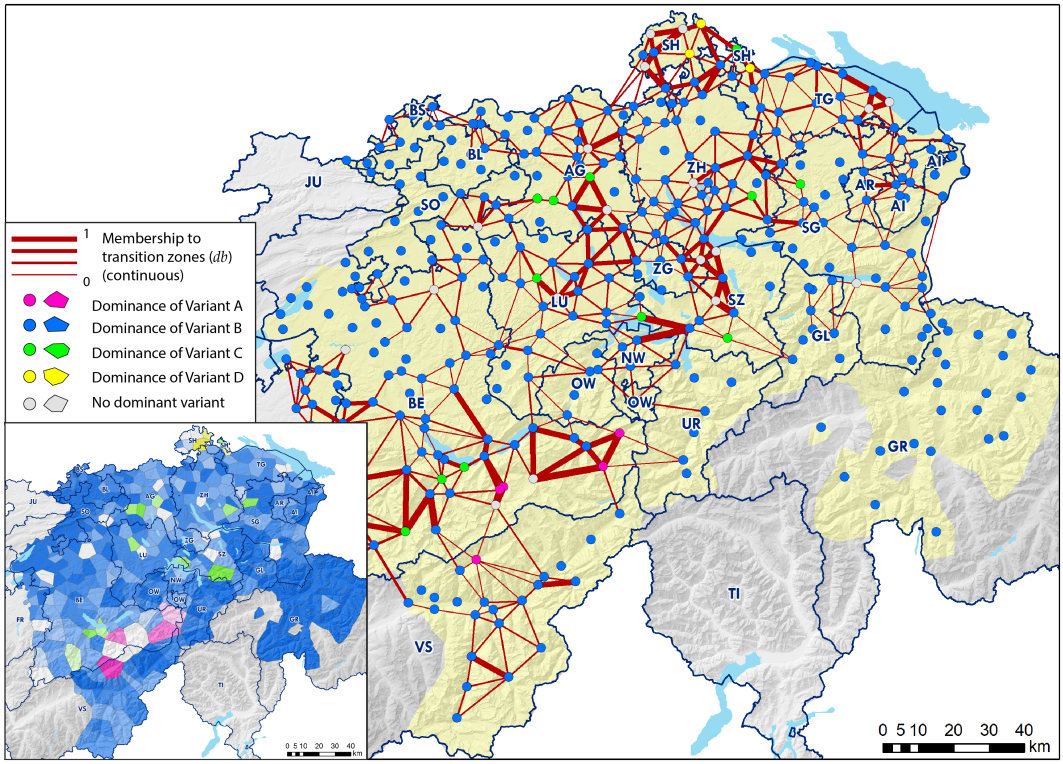

FIGURE 6.9: Comparing intensity map and transition map for III.22 – *'comparative clause linkage'* (a Type II variable).

6.3.3 Comparison across variables

Figures 6.10 and 6.11 are intended to display the characteristics of the proposed method for the comparison across variables. They present four maps of variables addressing the same linguistic phenomenon, the *'infinitival purposive clause'*. The four variables, corresponding to four survey questions of the SADS — Questions I.1, I.6, I.11, IV.14 (Table 3.1) – have been classified as variables of Type I exhibiting gradual transitions (Table 5.1), though with varying degrees of graduality. Note that Variable I.1 was already shown in Figure 6.8, above. The phenomenon mapped here has been involved in several linguistic studies (Seiler, 2005; Sibler, 2011; Bucheli Berger, Glaser, and Seiler, 2012; Glaser and Bart, 2012). Thus, the interpretation knowledge about its spatial variation is available and can inform the evaluation of the proposed method, although the above studies described boundaries and transitions only in a qualitative manner. Figure 6.10 shows the intensity maps for these variables. The colours used to depict a particular variant are the same in both Figures 6.10 and 6.11. Also, the display characteristics of intensity maps and transition maps, respectively, remain the same as above and will not be repeated. Regarding the comparison of the four maps in Figures 6.10 and 6.11, respectively, the following observations can be made:

- Both the intensity maps and the transition maps (through the point symbols at survey sites) show a core area for both of the main variants, one towards

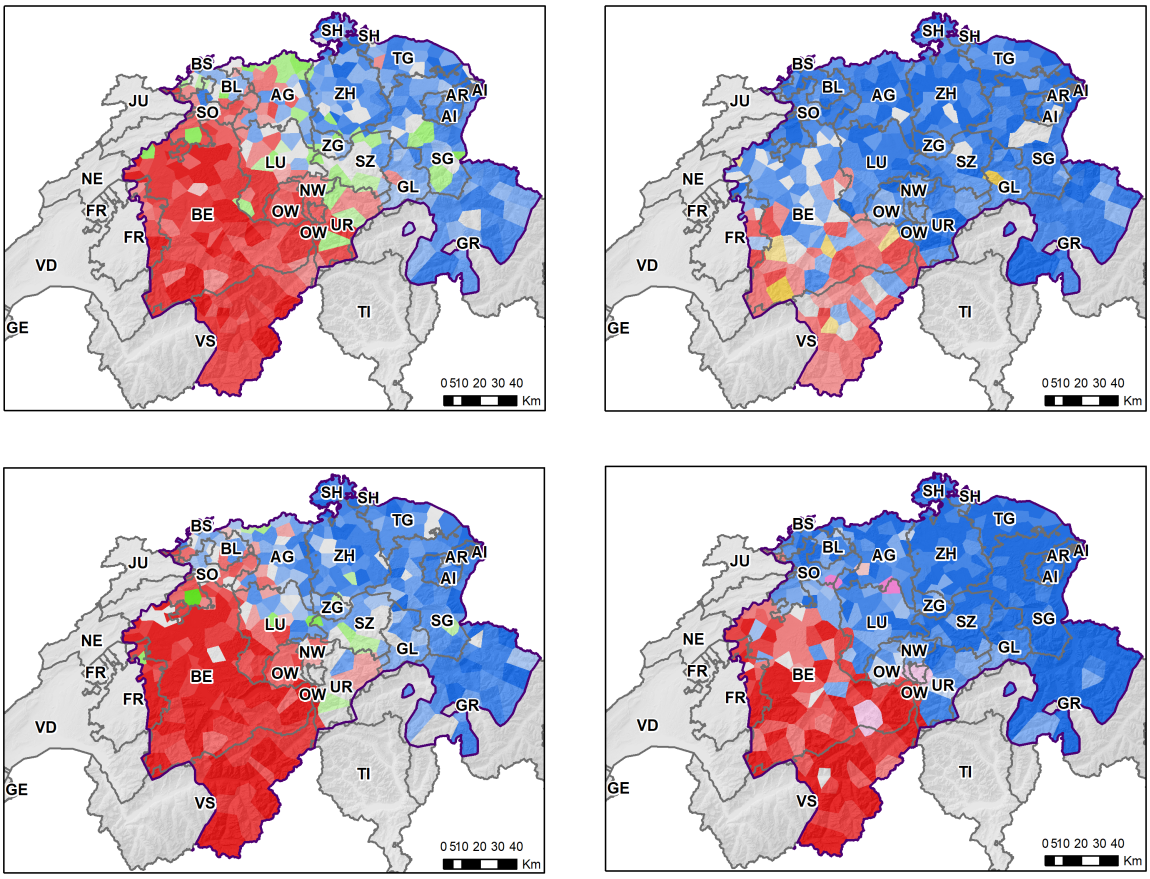

FIGURE 6.10: Intensity maps for the survey questions studying the *'infinitival purposive clause'* (top left, I.1; top right, I.6; bottom left, I.11; bottom right, IV.14).

the southwest for the red variant and one towards the northeast for the blue variant.

- Between these core areas transition zones can be seen, again visible in both map types. We can also see that the position of these transition zones differs between the variables mapped. The position is most to the east for variable I.1; followed by I.11, whose transition zone is shifted slightly west; followed by IV.14, which appears to be shifted even more westward; and finally I.6, whose transition zone has the most westerly position.

- In the intensity maps, it can be seen that there is considerable intensity variation not only in the transition zones, but also in the core areas. Again, the degree of variation differs between the four variables: I.1, I.11 and I.6 have a rather 'flat' distribution, with a considerable percentage of rather low intensity values, which might also be due to the presence of a third variant, besides the two main variants. Variable IV.14, then, seems to be much more clearly separated into two rather clear core areas, separated by a rather narrow transition zone.

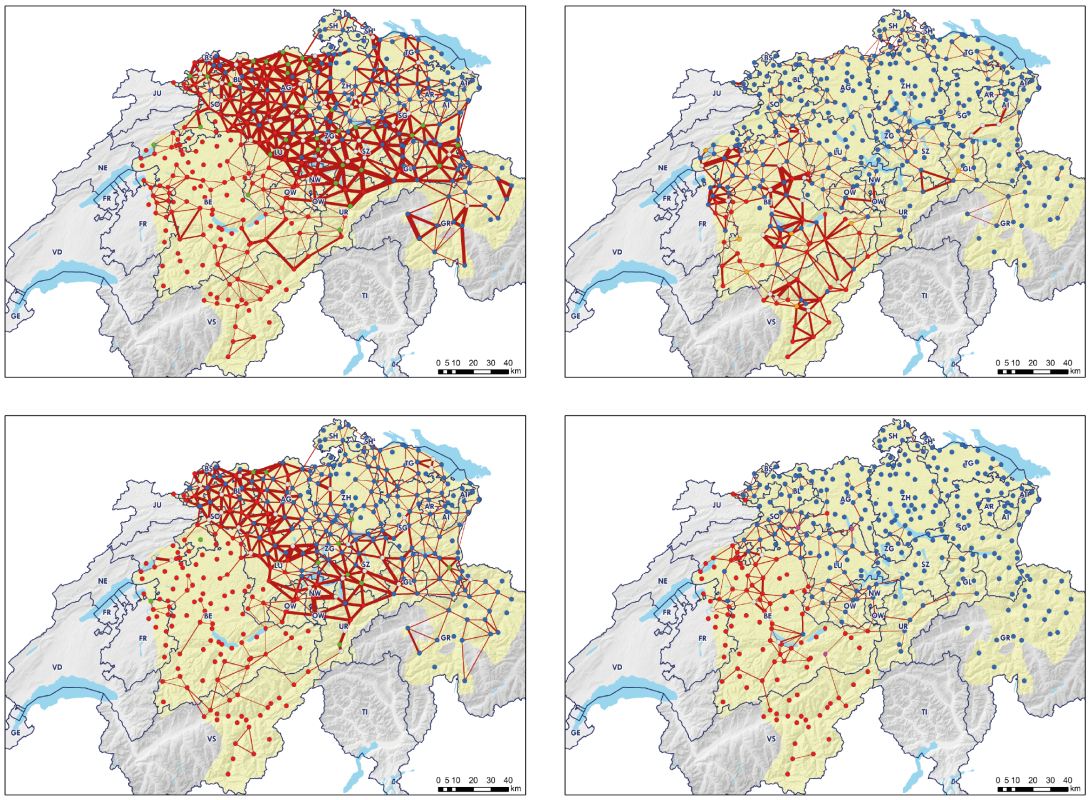

FIGURE 6.11: Comparison of transition maps for the survey questions studying the *'infinitival purposive clause'* (top left, I.1; top right, I.6; bottom left, I.11; bottom right, IV.14).

- This general picture is echoed in the transition maps. However, these maps show a more differentiated picture of the transitions present in the four variables. The transition areas already visible in the intensity maps are now more clearly highlighted for Variables I.1, I.11 and I.6, confirming the influence of the third variant on forming the transition zone. This is particularly true for I.1 and I.11, where the main transition zone shows clearly with strong connections, while secondary clusters of connections indicate increased local variation of the dominant variables. In the case of I.6, the transition pattern is more dispersed, as already suggested by the corresponding intensity map. For Variable IV.14, then, mostly thin links are found along the transition between the main variants. This is due to two main reasons: first, there is no third variant present, reducing the likelihood of differences of neighbouring sites; and second, the transition in this case is far less gradual than in the other three variables.

6.4 Discussion

The proposed method was designed to detect transitions in dialectal variation as potential locations of interdialectal boundaries. Thus, the aim of the method was to emulate the way a human interpreter would visually interpolate in a point symbol map or an intensity map to establish potential isoglosses.

In the above experiments, the results of the proposed method, presented in *transition maps*, have been compared to intensity maps displaying the same dialectal variables. This was done as intensity maps and point symbol maps are traditionally used as a basis for the delineation of isoglosses and transition zones, and more generally for the interpretation of spatial variation of dialectal variables. In comparison with intensity maps, several *strengths* can be noted:

- For each dialectal variable, the traditional isogloss drawing procedure is emulated multiple times (80 times in the case of the above experiments, i.e., 20 thresholds $*$ 4 parameter pairs in DBSCAN), using different combinations of parameters in the thresholding and the spatial clustering steps, respectively. By conducting several individual runs of the thresholding step, and by applying multiple scales of spatial proximity in the subsequent spatial clustering step, a probabilistic measure db is generated by summarising the resulting co-occurences in clusters. The db measure offers the potential for a differentiated interpretation, as demonstrated on Figures 6.7, 6.8, 6.9 and 6.11. Also, the use of multiple thresholds randomly drawn from a normal distribution renders the results more robust against spatial outliers contained in the intensity data.

- In contrast to existing procedures for dialect area classification, the proposed method takes space explicitly into account when determining if a potential boundary node is part of a transition. It handles relations among neighbouring survey sites encoded in the parameters of the spatial clustering. As the spatial distribution of survey sites in the SADS (and in most other linguistic databases) is not even, this potential spatial bias is accounted for by performing the spatial clustering with multiple parameter settings for $MinPts$. Note that this approach of dealing with varying density of survey sites is similar to the approach used in the OPTICS clustering algorithm (Ankerst et al., 1999), which automatically adapts the ε parameter to the density of the data points. However, our approach has the advantage that by using multiple runs the *proportion* of actually detected and potential cluster co-occurrences can be computed, rather than relying on absolute numbers. The resulting db can then be interpreted in a probabilistic sense as the degree of membership of a pair of natural neighbours to a transition zone.

- Thanks to its multi-scale approach the proposed method delivers a differentiated detection of outliers, in contrast to contemporary studies that map dialect

survey data with multiple responses per survey site (the dialectometric intensity estimation in Rumpf et al., 2009, 2010; Pickl, Spettl, et al., 2014 – Figure A.1 in Appendix A – and the diagram maps of SyHD[3] – Figure A.2 in Appendix A – Fleischer, Kasper, and Lenz, 2012). Sites that appear as outliers in intensity maps are further differentiated by the clustering step of the proposed procedure based on the intensities of the surrounding survey sites and their spatial distribution. Thus, the clustering procedure finds such sites as noise for some thresholds and ε, while for others they are classified as parts of db_{pot}. This process, engrained in the multi-scale approach, determines the nature of outliers.

- Another point that is novel from the methodological perspective is the use of travel times in the spatial clustering step. Using travel times to model the possibility of language contact seems more realistic than using Euclidean distances, and has also been confirmed by the study presented in Chapter 4 and Jeszenszky et al. (2017). Figure 6.8 shows a good example for the effect of using travel times. The thick connections present in ZH, TG and St. Gallen ('SG') in contrast with thinner connections visible in the Bernese Oberland (between BE and VS), despite their similar lightness in the intensity map. If clustering uses travel times, the distances between more isolated survey sites (such as those in the Bernese Oberland) will be bigger than perceived on the intensity maps, thus survey sites will have the potential to co-occur in less clusters. Nonetheless, the procedure can also be performed using other distance measures, such as Euclidean distance, cost distances or cultural distances, with the possibility of further incorporating other factors in the distance matrix, such as geographic boundaries or administrative borders (Sieber, 2017; Derungs et al., in review).

- Intensity maps display a differentiated picture of the spatial variation of the *intensities* of the dominant variants displayed. Similarly, the transition maps, through the point symbols representing the survey sites, are capable of displaying the dominant variant per survey site. On top of this, however, they also explicitly pinpoint the *transition zones* inherent to the variables that are mapped. This capacity is owing to the fact that the proposed method takes the local environment of each survey site into account, in contrast to intensity maps. It shows, for instance, nicely in Figure 6.11.

- Transition zones are also areas of *mixture* of co-occurring variants, characterised by a patchwork of different variants reaching dominance (as, for instance, visible in the main transition zone of Figure 6.8) or by areas of varying, weak dominance within the core area of a particular variant (e.g., in the area between ZH, TG and SG in Figure 6.8). In both cases, the intensity values will

[3] http://www.syhd.info/apps/maps/

be rather low and locally variable. Through the multiple thresholding operations of Step 1 of the proposed procedure, these areas can be detected and through the spatial clustering of Steps 2 and 3, they can be concatenated to form spatially coherent clusters.

As the result for Variable III.22 depicted in Figure 6.9 demonstrates, the proposed method is ill-suited to work with Type II variables. However, this result was to be expected, as the method has been designed primarily with variables in mind that have few variants with rather large areas of dominance and transitions, respectively. Typically, this would be the case in Type I variables. The dialectal variation inherent to Type II variables, such as III.22, can be far better analysed by Getis-Ord G^* statistics (Getis and Ord, 1992), which allow to highlight 'hot spots' of local relative dominance of minor variants. An example of this approach applied to SADS variables, including III.22, was shown in Sibler et al. (2012).

Apart from this 'restriction by design', the results of the above experiments have also manifested other shortcomings of the proposed method. In the following, these **limitations** are identified in the order of appearance, not suggesting a priority ordering; as an outlook on future work, possible **improvements** are proposed for each limitation:

- The *db* measure that is established to express the likelihood that a particular pair of natural neighbours belongs to a transition zone shares the strengths and weaknesses of all *aggregate* measures (with typical examples being the mean or standard deviation of an empirical distribution). It has as its strength the aggregation of a complex situation into a single, compact number – but also has the weakness of being hard to interpret, and possibly leading to misinterpretations. Just as different empirical distributions may yield the same mean, different spatial configurations in intensity values may lead to the same *db* outcome. A *possible solution* to this problem is to generate not one but several measures, just as an empirical distribution is usually characterised not only by its mean, but also by its standard deviation, skewness etc. Since the proposed method involves multiple runs with different parameter settings, it would be easy, for instance, to generate distribution measures for *db*. Also, further measures could be added that, for instance, focus more on the intensity differences between neighbouring survey sites. Naturally, while adding more measures would lead to more expressiveness and a potentially more differentiated picture of the dialect transitions mapped, the greater number of measures would also invariably lead to increased interpretation complexity. Eventually, solving this trade-off would only be possible by conducting user studies to establish whether more, and if so, which measures are needed.

- Further on the subject of interpretation, the transition maps depict links between natural neighbours (based on the so-called Delaunay triangulation).

With this representation, it is hard to visually infer potential locations of boundary lines or isoglosses. This process could be helped by using edges of Voronoi polygons, which happen to be the geometric dual of Delaunay edges, similarly to Goebl (2010:448) or Nerbonne (2010:13).

- Continuing on the theme of boundary lines and isoglosses, the focus of the method on low intensity values to detect potential boundary nodes is, as discussed above, useful in highlighting transitions. However, this approach is less suited to detecting actual boundary lines as representations of the isogloss model. If the aim was isogloss detection, the current db measure would have to be modified to focus more on differences in dominant variants between natural neighbours, possibly coupled with a region growing step (Gonzalez and Woods, 2017) that attempts to expand the core areas of two variants towards the transition zone that separates them to establish a boundary line. Another alternative approach might mimic the way in which isolines are computed on elevation surfaces, by interpolating intensity surfaces and deriving isoglosses by interpolation.

- One particular limitation of the proposed method is that it will not, or only weakly, pick up sharp transitions. Imagine the 'ideal' situation of two natural neighbours p_1 and p_2, where p_1 has a normalised intensity of 1.0 for Variant A, and p_2 has an intensity of 1.0 for Variant B. Intuitively we would then expect that a very clear isogloss could be placed between these two sites. However, since both sites have maximum intensity, none of them will be classified as potential boundary location bl_{pot}, and hence no transition link will be drawn. This problem is noticeable in Figure 6.7, which depicts a sharp transition variable. A possible solution to this problem would add an 'auxiliary site' between the natural neighbours with low intensity in order to force the thresholding step to take effect.

6.5 Conclusion

In this chapter, a new method has been proposed to detect transitions in the intensity values of the variants of dialectal variables. The method is primarily designed for Type I variables. The development of this method was prompted by a multifaceted research gap. Existing techniques that allow the delineation of dialect areas – with the exception of Rumpf, Pickl and their colleagues' work (Rumpf et al., 2009, 2010; Pickl, Spettl, et al., 2014) – are essentially non-spatial in the classification used. Existing techniques also rarely focus on the transition areas proper but rather on the dominance areas, and they do not allow the connection of the detected transitions to boundaries.

The proposed method has shown several strengths. In particular, the resulting *transition maps*, in comparison with traditional point symbol and intensity maps, allow more effective highlighting of transition zones and also express the likelihood with which neighbouring survey sites participate in a transition. However, as the experiments reported in this chapter have shown, the method also exhibits some limitations. In its current version, the proposed method, and the resulting transition maps, are probably best used in combination with intensity maps. While intensity maps are capable of depicting the spatial variation in the intensities of the (dominant) variable per site, highlighting potential transitions, the transition maps can further characterise these transitions in space, formed by neighbouring sites, and how likely these sites are to participate in a transition. In cases of sharp transitions, where the proposed method exhibits a weakness, intensity maps may further assist the interpretation.

In the above discussion, possible improvements have been outlined for the limitations of the proposed method. Once these or similar improvements have been realised, several other points may be addressed in future work. First, the detected boundary and transitions links may also be tested against other boundary lines, such as political or cultural borders, as was recently done in a similar study by Sieber (2017) and later by Derungs et al. (in review). Second, and perhaps most importantly, the proposed method and its extensions should be evaluated in user studies involving experts from dialectology, and other categories of potential users, in order to establish the feasibility and usability of this new approach of quantitatively and visually interpreting dialect data. And finally, it would also be interesting to assess the transfer of this method to linguistic databases other than the SADS, and to linguistic levels other than syntax.

Chapter 7

Conclusion

The main research objective of the thesis was **developing spatial analysis methodologies to quantify the spatial variation in dialectal phenomena and account for underlying geographic effects.** This objective was developed from the collaborative research project between groups in dialectology and GIScience, respectively.

With its methodological focus and an emphasis on spatial linguistic processes, this thesis intended to involve GIScience and spatial information theory as a framework for approaching hypotheses in linguistics research, to explore the usability of methods related to GIScience and spatial analysis for the subfield of dialectology. The practical goal of the thesis was to quantitatively assess the validity of hypotheses and conceptual models developed in the linguistic literature. The development of methodologies aims to contribute to dialectology and, more generally, to linguistics, supporting the discovery and differentiated analysis of spatial variation present in dialect data and, ultimately, to better explain (geographic) processes governing the spatial variation present in dialects. There has been an advocacy for the involvement of spatial analysis in linguistics and specifically, in dialectology, but usually these methods have primarily been used as tools for visualising linguistic data, or applied for very specific purposes.

Conversely, for GIScience, the interest in working on problems of dialectology is motivated by the fact that linguistic data generally represent a source of (primarily) categorical, spatially referenced data, with particular characteristics and peculiarities unlike those of any other data source typically seen in GIS applications. Hence, in addition to the possibility of providing support for research in a neighbouring discipline, and that of getting increased publicity beyond the core application domain of GIScience, there is also a benefit to GIScience through a host of interesting and challenging methodological problems that help sharpen the methodological tools of the field.

GIScience has worked out the fundamentals for spatial analysis, which are now used in several branches of science. Its quantitative methodologies represent a crucial addition to the toolbox available in the study of the spatial distribution of linguistic, and more specifically, dialectal variables. The research in this thesis is meant to help quantify the fundamental understanding of space in dialectology by implementing the approach of GIScience.

Besides enabling the explicit representation of space GIScience enables repeatable and automated, and thus more efficient analyses that foster comparison and objectivity. At the same time, GIScience benefits from involving itself in dialect geography. A host of interesting and challenging methodological problems help sharpen the methodological tools of the field in the "the complex multi-attribute nature of dialect space" (Hoch and Hayes, 2010:29). In such an interdisciplinary setting GIScience gains a new area where its methods are put to the test, further developed and applied in an environment that is laden with a range of uncertainties where the vast variety in linguistic data also gives a lot of room for experiments. GIScience also gets inspiration from the shared interest in spatial variation and categorisation and the methods developed can be applied more generally, such as for other languages or in the broader research fields of humanities.

7.1 Contributions

7.1.1 Covariation of geographic and linguistic distances

The first steps in addressing the research gaps focusing on the exploration of the relationship between geographic influences and dialectal variation in Chapter 4 are ultimately applications and extensions of state of the art methods from dialectometry to a database of syntactic language atlas data (SADS).

A linguistic distance was expressed through a measure aggregated from 60 survey questions of the SADS. Going beyond most researchers' work, the possibility of language contact has been operationalised using more informed distance measures: travel times from different years (2000, 1950, 1850). This is a pioneering approach with regards to the study of Swiss German. With the diverse physical landscape of Switzerland having an impact on potential language contact, it has been found that travel times are significantly better predictors than Euclidean distance for the syntactic variation in Swiss German dialects. However, although older travel times yielded higher correlation values –in accordance with our hypothesis that they would be better predictors of syntactic variation – the difference to travel times of more recent years was not statistically significant. For the entire SADS data set, unlike in most other dialectometric research (on other linguistic levels), a linear model described the correlations between geographic and linguistic distance better than a logarithmic model. However, this difference is not statistically significant.

Mapping the average syntactic distance to all other survey sites provided a way of finding the dialects that are most different from the others in the syntactic sense. By computing residuals of normalised syntactic and geographic distances, it can be seen which geographic patterns predict the syntactic differences best, and to what extent.

For the first time, the correlations of syntactic and geographic distances have also been investigated at the local level, focusing on spatial subsets as well as centred on

individual survey sites. It has been found that at the local scale the difference between the explanation power of Euclidean distance and travel times is not always significant, depending on prevalent topography. These results have, thus, underlined even more strongly the interrelationship of topography and linguistic variation. A higher isolation effect found in mountainous areas results in the travel times being much better predictors, while better connected areas in the lowlands show no significant difference for the topography-related predictor variables.

The analysis in this study has tested the 'Fundamental Dialectological Postulate' (Nerbonne and Kleiweg, 2007) and Stanford's finding that "the issue of geographic size appears to be related to fundamental distance relationships in human interaction" (2012:274). It has been shown that linguistic distance (i.e., similarity) depends much more on the particular characteristics of an area on a local scale than at the global scale where the effects of local particularities tend to level out. With additional local analyses taking further different explanatory variables into account, however, it may be possible to explain why high degrees of correlation had been reported in some studies, and low degrees in others.

7.1.2 Hypothesis-driven modelling of interdialectal boundaries

Modelling interdialectal boundaries has implications for the subfields of dialect contact and language change. Having data of fine spatial granularity available, interdialectal boundaries may be viewed as gradients, aiming to quantitatively account for the transition patterns that are traditionally only implicitly inferred visually from maps. Chapter 5 models these boundaries in a hypothesis-driven way.

Prototype models have been proposed for the two key conceptualisations addressing transitions in dialectal variables: the isogloss and the dialect continuum concepts. Following this, regression models of different order and type have been introduced to describe ideal transition scenarios. The fit of these empirical regression models to the theoretical prototype models has been tested for several variables which had appropriate distribution properties. The model fitting has been conducted at the level of the whole investigation area and at local levels, that is, cross-sections and spatial subdivisions that represented 'transition zones' and 'dominance zones' for each variable.

It was found that the intensity landscapes of many dialectal variables are too noisy to be classified safely into either sharp or gradual transition variables. Nevertheless, a main contributions of this study are to allow on the one hand for characterising different variables with regards to the prevailing patterns in the transitions (i.e., fit with the isogloss vs. dialect continuum concept), and on the other hand for the comparison of different variables based on transition characteristics.

The properties that proved to be the most useful in characterising transitions between variants are the width and absence of transition zones, and the absolute rate of change in intensity in the subdivisions. Additionally, cumulative rates of

change in transition zones performed well and, for gradual transition variables, the goodness-of-fit value to inclined planes benchmarks as well.

It has been shown that neither end of a theoretical scale of boundary graduality (cf. Parker, 2006) is ideal to model all dialectal variables, due to the uncertainty inherent in linguistic data. This finding delivers strong evidence to dialectology of the gradual nature of dialectal boundaries at the level of individual variables. Furthermore, it shows that besides formulating hypotheses and proposing models on the spatial aspects of linguistic change, a holistic view of the variation present in the raw data is beneficial.

7.1.3 Data-driven detection of interdialectal boundaries

If it is possible to compare spatial variation and transition patterns in linguistic variables more objectively, it becomes easier to describe their (dis)similarity and formulate hypotheses regarding the reasons for differences in variation and diffusion. In order to overcome the overly constraining conceptualisations of linguistic boundaries as either isoglosses or dialect continua still predominant in dialectology today, in Chapter 6 a data-driven method has been proposed to delimit boundaries at the interface of variants' prevalence areas. This method aims to avoid subjective classification decisions that come at the price of generalisation and disregarding outliers ('smoothing'). Thus, instead of making binary decisions or determining sharp zones of dominance and transition, probabilistic boundary segments are proposed.

A procedure consisting of linguistic filtering and local spatial clustering is performed at different spatial scales using the original intensity values. Due to employing multiple thresholds (in the linguistic filtering step), the proposed method locates elements of sharp isoglosses in cases of abrupt changes in dialectal variation, whereas in cases of smoother variation the same algorithm detects more extended transition zones. The clustering procedure is performed using the travel time matrices, acting as a more realistic estimate for dialect contact possibilities, similarly to Chapter 4.

With the visualisations that can be derived from this method, a spatially more detailed comparison across variables is possible. Through the proposed procedure, the approach of the classical isogloss-drawing method has been automated and implemented, essentially emulating the working mechanism of human vision when performing a local categorisation and visual boundary interpolation in a point symbol map. Furthermore, the proposed approach accounts for variation underlying dominance areas and handles outliers appropriately.

7.1.4 Limitations

The main findings of Chapter 4 are outcomes only from an analysis of data from the SADS, thus focusing on Swiss German and syntax data. Applying the methodology to other languages with a different spatial extent, a different topography and a

different attitude towards dialects may show more (or less) diverse results. To establish the syntactic distance measure, a linear summation of 60 variables has been used, not taking into consideration the potential mutual correlations between the answer matrices of the survey questions, essentially assuming independence between the variables. A significant amount of variance in the syntactic distance still stays unexplained by geographic distances, suggesting that other effects are responsible for variation as well. This poses an interesting challenge, taking socio-demographic variables into account.

In Chapter 5, the data used shows stronger random patterns than that which would be expected under the idealised assumptions of the mathematical prototype models used. The presence of more than two variants can hinder the appropriate fitting of the mathematical models to the intensity surfaces. In terms of characterising transitions, the proposed set of evaluation measures is not yet capable of fully grasping the complexity of transitions as they exist in real data. Furthermore, spatial autocorrelation patterns of residuals may serve as a good supplement for describing the patterns of transition, yet this has not been implemented.

The methodology proposed in Chapter 6 for detecting transitions and boundaries has generated initial promising results, but also exhibited limitations. The discussion in Chapter 6 has highlighted several critical points and proposed extensions for future research. More generally, the proposed methodology is limited to certain configurations of linguistic variables, particularly of Type I (Section 5.3.1). In the case of Type II variables, for example, the proposed methodology basically detects an underlying mixture of variants, rather than finding actual boundaries, due to dominance areas being small. Multiple answers per survey site are crucial in order to build a linguistic landscape where intensity thresholding is possible. This requirement imposes further limits when trying to generalise the method to other kinds of data.

7.2 Outlook

Several natural extensions of this research are possible. With little effort, the methodologies could be altered to be used on different linguistic data (such as crowdsourced data, such as in Leemann et al. 2016), on different linguistic levels, and on other kinds of humanities-related data. For instance, the transition models and the boundary detection algorithm could be applied to analyse census data of different proportions as well as ecological data.

With regards to Chapters 4 and 6, using travel times of public transportation, historical trading routes (similarly to Lameli et al. 2015) and historical travel times preceding 1850 might lead to interesting insights. On the one hand, public transportation and the improvement in its travel times are indicators of the routes' importance

from the viewpoint of speakers' potential contact. On the other hand historical trading routes and travel times may better model the patterns of isolation and contact during the times of substantial dialectal differentiation in the late medieval period.

The quantitative analysis of sharp boundaries coinciding with fiat and bona fide geographic boundaries is appealed for by dialectology (Haag, 1898, (noted in Pickl, 2013); Trudgill, 1974; Glaser, 2013; Sieber, 2017). Input to such research can be gained using the boundary detection methods proposed in this thesis. This research direction could help determine the effects of different kinds of language-external boundaries on dialectal variation.

For quantitative comparison of variables at the global level, it would be possible to adopt measures similar to those used by Rumpf et al. (2009) and Labov, Ash and Boberg 2006.

For testing the usability of the boundary detection methodology in Chapter 6, the implementation of a user-based evaluation could prove useful. Thus, it could be assessed how well the resulting maps correspond to cognitive processes of trained linguists when compared with traditional point symbol maps.

Furthermore, exploiting the property of SADS of incorporating data originating from respondents of different age groups, it is possible to conduct apparent time analyses (Stoeckle, 2018) to predict future scenarios of evolution in different variables, taking into account the regional age and variation structure.

Finally, inspired by research from phylogenetics, and using road transportation network data, it may be possible to explore how inferences could be made on the spatial evolution of dialectal variables along potential communication routes, using route planning and route inference (Ranacher, Gijn, and Derungs, 2017).

References

Ankerst, Mihael, Markus M. Breunig, Hans-Peter Kriegel, and Jörg Sander (1999). "OPTICS: Ordering points to identify the clustering structure". In: *ACM Sigmod Record*, pp. 49–60. ISSN: 01635808. DOI: 10.1145/304182.304187.

Anselin, Luc, Anil K. Bera, Raymond Florax, and Mann J. Yoon (1996). "Simple diagnostic tests for spatial dependence". In: *Regional Science and Urban Economics* 26.1, pp. 77–104. ISSN: 01660462. DOI: 10.1016/0166-0462(95)02111-6.

Barbiers, Sjef (2013). "Microsyntactic variation". In: *The Cambridge Handbook of Generative Syntax*. Ed. by M. den Dikken. (Cambridge Handbooks in Language and Linguistics). Cambridge: Cambridge University Press, pp. 899–926.

Barbiers, Sjef, Hans J. Bennis, Gunther De Vogelaer, Magda Devos, and Margreet H. van der Ham (2005). *Syntactische Atlas van de Nederlandse Dialecten/Syntactic Atlas of the Dutch Dialects Volume I*. Amsterdam: Amsterdam University Press, pp. 1–80.

Bartholy, Heike (1992). *Sprache, kulturelle Identität und Unabhängigkeit, dargestellt am Beispiel Maltas*. Schuch.

Beale, Colin M., Jack J. Lennon, Jon M. Yearsley, Mark J. Brewer, and David A. Elston (2010). "Regression analysis of spatial data". In: *Ecology Letters* 13.2, pp. 246–264. ISSN: 1461023X. DOI: 10.1111/j.1461-0248.2009.01422.x.

Bickel, Balthasar (2007). "Typology in the 21st century: major developments". In: *Linguistic Typology* 11, pp. 239–251. ISSN: 1430-0532. DOI: 10.1515/LINGTY.2007.018.

Bielenstein, August Johann Gottfried (1892). *Die Grenzen des lettischen Volksstammes und der lettischen Sprache: In der Gegenwart und im 13. Jahrhundert : Ein Beitrag zur ethnologischen Geographie und Geschichte Russlands.* St. Petersburg: Eggers & Co.

Bloomfield, Leonard (1933). *Language*. New York: Holt, Rinehart & Winston.

Blythe, Richard A. and William A. Croft (2012). "S-curves and the mechanisms of propagation oin language change". In: *Language* 88.Number 2, pp. 269–304.

Börlin, Rolf (1987). *Die schweizerdeutsche Mundartforschung 1960-1982: bibliographisches Handbuch. Vol. 5.* Verlag Sauerländer.

Bouckaert, Remco, Philippe Lemey, Michael Dunn, Simon J. Greenhill, Alexander V. Alekseyenko, Alexei J. Drummond, Russell D. Gray, Marc A. Suchard, and Quentin D. Atkinson (2012). "Mapping the origins and expansion of the Indo-European language family." In: *Science (New York, N.Y.)* 337.6097, pp. 957–60. ISSN: 1095-9203. DOI: 10.1126/science.1219669.

Bounds, Paulina (2015). "Perceptual regions in Poland: An investigation of Poznań speech perceptions". In: *Journal of Linguistic Geography* 3.01, pp. 34–45. ISSN: 2049-7547. DOI: 10.1017/jlg.2015.1.

Bowern, Claire (2013). "Relatedness as a Factor in Language Contact". In: *Journal of Language Contact* 6.2, pp. 411–432. ISSN: 1877-4091. DOI: 10.1163/19552629-00602010.

Britain, David (2002). "Space and spatial diffusion". In: *Language and space: An international handbook of linguistic variation*. Ed. by Jack K. Chambers, Peter Trudgill, and Natalie Schilling-Estes. Oxford: Blackwell, pp. 603–637.

— (2010). "Conceptualisations of Geographic Space in Linguistics". In: *Language and space, an international handbook of linguistic variation, vol. 2: Language mapping*. Ed. by Alfred Lameli, Roland Kehrein, and Stefan Rabanus. Berlin: Mouton de Gruyter, pp. 69–97.

Bucheli Berger, Claudia (2010). "Dativ für Akkusativ im Senslerischen (Kanton Freiburg)". In: *Alemannische Dialektologie: Wege in die Zukunft. Beiträge zur 16. Arbeitstagung für alemannische Dialektologie in Freiburg/Fribourg vom 07.-10.09.2008*. Ed. by Helen Christen, Sibylle Germann, Walter Haas, Nadia Montefiori, and Hans Ruef. Stuttgart: Steiner, pp. 71–83.

Bucheli Berger, Claudia, Elvira Glaser, and Guido Seiler (2012). "Is a syntactic dialectology possible? Contributions from Swiss German". In: *Methods in Contemporary Linguistics*. Ed. by Andrea Ender, Adrian Leemann, and Bernhard Wälchli. De Gruyter Mouton, pp. 93–120. ISBN: 978-3-11-027568-1.

Bucheli Berger, Claudia and Christoph Landolt (2014). "Dialekt und Konfession in der Deutschschweiz". In: *Dialekt und Religion - Beiträge zum 5. dialektologischem Symposium im Bayerischen Wald, Walderbach, Juni 2012*. Ed. by Elisabeth Frieben, Ulrich Kanz, Barbara Neuber, and Ludwig Zehetner. Regensburg: edition vulpes, pp. 73–95.

Bucheli, Claudia and Elvira Glaser (2002). "The Syntactic Atlas of Swiss German dialects: Empirical and methodological problems". In: *Syntactic Microvariation*. Ed. by Sjef Barbiers, Leonie Cornips, and Susanne van der Kleij. Vol. 2. Amsterdam: Meertens Institute Electronic Publications in Linguistics, pp. 41–73.

Chagnaud, Clément, Philippe Garat, Paul-Annick Davoine, Elisabeth Carpitelli, and Axel Vincent (2017). "ShinyDialect : A cartographic tool for spatial interpolation of geolinguistic data". In: *Proceedings of GeoHumanities'17: 1st ACM SIGSPATIAL Workshop on Geospatial Humanities , Los Angeles Area, CA, USA, November 7–10, 2017 (GeoHumanities'17)*, pp. 23–30. ISBN: 9781450354967. DOI: 10.1145/3149858.3149864.

Chambers, Jack K. and Peter Trudgill (2004). *Dialectology*. 2nd. Cambridge: Cambridge University Press. ISBN: 0511034962.

Chorley, R. J. and P. Haggett (1965). "Trend-Surface Mapping in Geographical Research". In: *Transactions of the Institute of British Geographers* 37, pp. 47–67. DOI: 10.2307/621689.

Christen, Helen (1998). "Convergence and divergence in the Swiss German dialects". In: *Folia Linguistica* 32.1-2, pp. 53–68. ISSN: 0165-4004. DOI: 10.1515/flin.1998.32.1-2.53.

Christen, Helen, Elvira Glaser, Matthias Friedli, and Manfred Renn (2010). *Kleiner Sprachatlas der deutschen Schweiz*. Frauenfeld: Verlag Huber.

Cukor-Avila, Patricia and Guy Bailey (2013). "Real Time and Apparent Time". In: *The Handbook of Language Variation and Change: Second Edition*. Ed. by Jack K. Chambers and Natalie Schilling. Blackwell Publishing.

Daan, Johanna C. and Dirk Peter Blok (1969). *Van randstad tot landrand: toelichting bij de kaart; dialecten en naamkunde; met een kaart en een grammofoonplaatje.* Noord-Hollandsche uitg. maatschappij.

Derungs, Curdin, Christian D. Sieber, Robert Weibel, and Elvira Glaser (in review). "Borders in a dialect landscape – administration is more formative than economy or religion". In: *PLoS ONE*.

Douglas, David H. (1994). "Least-cost Path in GIS Using an Accumulated Cost Surface and Slopelines". In: *Cartographica: The International Journal for Geographic Information and Geovisualization* 31.3, pp. 37–51. ISSN: 0317-7173. DOI: 10.3138/D327-0323-2JUT-016M.

Dros-Hendriks, Lotte (2018). "Not another book on Verb Raising". PhD thesis. Utrecht. ISBN: 9789460932717.

Dürscheid, Christa (in press). "Internetkommunikation, Sprachwandel und DaF-Didaktik". In: *Sprachwandel - Perspektiven für den Unterricht Deutsch als Fremdsprache*. Ed. by Sandro M. Moraldo and Federica Missaglia. Heidelberg: Winter, pp. 139–157.

Egorova, Ekaterina (2018). "Investigating Space Conceptualizations in the Alpine Context". PhD thesis. Univeristy of Zurich.

Eisenstein, Jacob (2017). "Written dialect variation in online social media". In: *Handbook of dialectology*. Ed. by Charles Boberg, John Nerbonne, and Dominic Watt. Wiley.

Eisenstein, Jacob, Brendan O'Connor, Noah A. Smith, and Eric P. Xing (2014). "Diffusion of Lexical Change in Social Media". In: *PLoS one* 9.44, pp. 1–23. DOI: 10.1371/journal.pone.0113114.

Esser, Paul (1983). *Dialekt und Identität: Diglottale Sozialisation und Identitätsbildung*. Lang.

Ester, Martin, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu (1996). "A density-based algorithm for discovering clusters in large spatial databases with noise". In: *Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)*. Ed. by Evangelos Simoudis, Jiawei Han, and Usama M. Fayyad. Portland, pp. 226–231.

Fisher, Peter, Jo Wood, and Tao Cheng (2004). "Where Is Helvellyn? Fuzziness of Multi-Scale Landscape Morphometry". In: *Transactions of the Institute of British Geographers* 29.1, pp. 106–128.

Fleischer, Jürg, Simon Kasper, and Alexandra N. Lenz (2012). "Die Erhebung syntaktischer Phänomene durch die indirekte Methode: Ergebnisse und Erfahrungen aus dem Forschungsprojekt "Syntax Hessischer Dialekte" (SYHD)". In: *Zeitschrift für Dialektologie und Linguistik* 1.79, pp. 2–42.

Fotheringham, A. Stewart (1997). "Trends in quantitative methods II: stressing the computational". In: *Progress in Human Geography* 21.1, pp. 88–96. ISSN: 03091325. DOI: 10.1191/030913299667756016.

Fotheringham, A. Stewart, Chris Brunsdon, and Martin E. Charlton (2002). *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Wiley, p. 284. ISBN: 978-0-471-49616-8.

Francis, W. Nelson (1983). *Dialectology: an Introduction*. London: Addison-Wesley Longman. ISBN: 0-582-29117-8.

Freudenberg, Rudolf (1966). "Isoglosse: Prägung und Problematik eines sprachwissenschaftlichen Terminus". In: *Zeitschrift für Mundartforschung* 33.3/4, pp. 219–232.

Fröhlich, Philipp, Thomas Frey, Serge Reubi, and Hans Ulrich Schiedt (2004). "Entwicklung des Transitverkehrs-Systems und deren Auswirkung auf die Raumnutzung in der Schweiz (COST 340): Verkehrsnetz-Datenbank". Zürich.

Galton, Antony (2001). "A Formal Theory of Objects and Fields". In: *Spatial Information Theory: Foundations of Geographical Information Science*, ed. by Daniel R. Montello. Volume 2205 of Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, pp. 458–473. ISBN: 9783642231964. DOI: 10.1007/978-3-642-23196-4.

— (2003). "On the Ontological Status of Geographical Boundaries". In: *Foundations of Geographic Information Science*. Ed. by Matt Duckham. London: Taylor & Francis, pp. 151–171. ISBN: 0849382246. DOI: 10.1036/0071425799.

— (2004). "Fields and Objects in Space, Time, and Space-time". In: 4.1, pp. 39–68. DOI: 10.1207/s15427633scc0401{_}4.

Gemert, Ilse van (2002). *Het geografisch verklaren van dialectafstanden met een geografisch informatiesysteem (GIS): Master's thesis*.

Getis, Arthur and J. Keith Ord (1992). "The Analysis of Spatial Association". In: *Geographical Analysis* 24.3, pp. 189–207.

Giacomelli, Gabriela, Luciano Agostiniani, Patrizia Bellucci, Luciano Gianelli, Simonetta Montemagni, Annalisa Nesi, Matilde Paoli, Eugenio Picchi, and Teresa Poggi Salani, eds. (2000). *Atlante Lessicale Toscano*. Rome: Lexis Progetti Editoriali.

"ALF - Atlas linguistique de la France" (1902-1910). In: *Atlas linguistique de la France, 10 Volumes*. Ed. by Jules Gilliéron and Edmond Edmont. Paris: Champion.

Girard, Dennis and Donald Larmouth (1993). "Some applications of mathematical and statistical models in dialect geography". In: *American dialect research*. Ed. by Dennis R. Preston. Amsterdam, Philadelphia: John Benjamins, pp. 107–132.

Glaser, Elvira (2013). "Area formation in morphosyntax". In: *Space in language and linguistics: geographical, interactional and cognitive perspectives. (linguae & litterae 24*

). Ed. by Peter Auer, Martin Hilpert, Anja Stukenbrock, and Benedikt Szmrecsanyi. Berlin/Boston: Freiburg Institute of Advanced Studies (FRIAS), De Gruyter, pp. 93–119.

— (2014). "Wandel und Variation in der Morphosyntax der schweizerdeutschen Dialekte". In: *Taal en Tongval* 66.1, pp. 21–64.

Glaser, Elvira and Gabriela Bart (2012). "Discovering and mapping syntactic areas: old and new methods". In: *Proceedings of the International Symposium on Limits and Areas in Dialectology (LimiAr). Lisbon, 2011*. Ed. by Xosé Afonso Álvarez Pérez, Ernestina Carrilho, and Catarina Magro. Centro de Linguística da Universidade de Lisboa, pp. 345–363. ISBN: 9789729640490.

— (2015). "Dialektsyntax des Schweizerdeutschen". In: *Regionale Variation des Deutschen. Projekte und Perspektiven*. Ed. by Roland Kehrein, Alfred Lameli, and Stefan Rabanus. Berlin: De Gruyter. Chap. 4, pp. 79–105.

Glaser, Elvira, Philipp Stoeckle, and Sandro Bachmann (accepted). "Faktoren und Arten intrapersoneller Variation im Material des syntaktischen Atlas der deutschen Schweiz (SADS)". In: *Syntax aus Saarbrücker Sicht 3.: Beiträge der SaRDiS-Tagung zur Dialektsyntax 2018*. Ed. by Augustin Speyer and Philipp Rauth. Stuttgart: Steiner.

Glaser, Elvira and Robert Weibel (2013). *Modellierung morphosyntaktischer Raumbildung im Schweizerdeutschen (SynMod) - Swiss National Science Foundation project proposal*. Zürich.

Goebl, Hans (1982). *Dialektometrie: Prinzipien und Methoden des Einsatzes der Numerischen Taxonomie im Bereich der Dialektgeographie*. Wien.

— (1983). ""Stammbaum" und "Welle"". In: *Zeitschrift für Sprachwissenschaft* 2.1, pp. 3–44.

— (2004). "VDM—Visual DialectoMetry: Vorstellung eines dialektometrischen Software-Pakets auf CD-ROM (mit Beispielen zu ALF und Dees, 1980)". In: *Romanistik und neue Medien*. Ed. by W. Dahmen, G. Holtus, J. Kramer, M. Metzeltin, W. Schweickard, and O. Winkelmann. Tübingen: Narr, pp. 209–241.

— (2006). "Recent advances in Salzburg dialectometry". In: *Literary and Linguistic Computing* 21.4, pp. 411–427. ISSN: 02681145. DOI: 10.1093/llc/fql042.

— (2010). "Dialectometry and quantitative mapping". In: *Language and Space. Vol. 2: Language Mapping*. Ed. by Alfred Lameli, Roland Kehrein, and Stefan Rabanus. Vol. 2. Berlin: De Gruyter Mouton. Chap. xxii, pp. 433–457.

Goebl, Hans and Guillaume Schiltz (1997). "A dialectometrical compilation of CLAE 1 and CLAE 2: Isoglosses and dialect integration." In: *Computer developed linguistic atlas of England (CLAE)*. Ed. by Wolfgang Viereck and Heinrich Ramisch. Vol. 2. Tübingen: Max Niemeyer Verlag, pp. 13–21.

Goel, Rahul, Sandeep Soni, Naman Goyal, John Paparrizos, Hanna Wallach, Fernando Diaz, and Jacob Eisenstein (2016). "The Social Dynamics of Language Change in Online Networks". In: *The International Conference on Social Informatics (SocInfo)*.

Gonzalez, Rafael C. and Richard E. Woods (2017). *Digital Image Processing*. 4th. Pearson International. ISBN: 9781292223049.

Goodchild, Michael F. (2004). "The validity and usefulness of laws in geographic information science and geography". In: *Annals of the Association of American Geographers* 94.2, pp. 300–303. ISSN: 00045608. DOI: 10.1111/j.1467-8306.2004.09402008.x.

Gooskens, Charlotte (2004). "Norwegian dialect distances geographically explained". In: *Language Variation in Europe. Papers from the Second International Conference on Language Variation in Europe ICLAVE Vol. 2. 2004.* Ed. by Britt-Louise Gunnarson, Lena Bergström, Gerd Eklund, Staffan Fridella, Lise H. Hansen, Angela Karstadt, Bengt Nordberg, Eva Sundgren, and Mats Thelander. Uppsala, pp. 195–206.

Grieve, Jack (2013). "A statistical comparison of regional phonetic and lexical variation in American English". In: *Literary and Linguistic Computing* 28.1, pp. 82–107. ISSN: 0268-1145. DOI: 10.1093/llc/fqs051.

— (2014). "A comparison of statistical methods for the aggregation of regional linguistic variation". In: *Aggregating Dialectology, Typology, and Register Analysis: Linguistic Variation in Text and Speech*. Ed. by Benedikt Szmrecsanyi and Bernhard Wälchli. Berlin/ New York: Walter de Gruyter, pp. 1–34. DOI: 10.1515/9783110317558.53.

Grieve, Jack, Dirk Speelman, and Dirk Geeraerts (2011). "A statistical method for the identification and aggregation of regional linguistic variation". In: *Language Variation and Change* 23, pp. 1–29. ISSN: 0954-3945. DOI: 10.1017/S095439451100007X.

Griffith, Daniel A. (1987). *Spatial autocorrelation - A Primer*. Washington, DC: Association of American Geographers, pp. 317–331. ISBN: 0860942236. DOI: 10.1016/0166-0462(92)90032-V.

Guy, Gregory R. (1993). "The Quantitative Analysis of Linguistic Data". In: *American Dialect Research*. Ed. by Dennis R. Preston. Amsterdam: John Benjamins.

Haag, Karl (1898). *Die Mundarten des oberen Neckar- und Donaulandes (Schwäbisch-alemannisches Grenzgebiet: Baarmundarten)*. Reutlingen: Buchdruckerei Hutzler.

Haas, Walter (2010). "A study on areal diffusion". In: *Language and Space*. Ed. by Peter Auer and Jürgen Erich Schmidt. Berlin/ New York: Mouton de Gruyter, pp. 649–667.

Hägerstrand, Torsten (1952). "The propagation of innovation waves". In: *Lund studies in geography, Series B* Human Geography 4.

Händler, Harald and Herbert Ernst Wiegand (1982). "25 . Das Konzept der Isoglosse : methodische und terminologische Probleme". In: *Dialektologie. Ein Handbuch zur deutschen und allgemeinen Dialektforschung.* Ed. by Werner Besch, Ulrich Knoop, Wolfgang Putschke, and Herbert Ernst Wiegand. Berlin/New York: Walter de Gruyter, pp. 501–527.

Harmon, David and Jonathan Loh (2010). "The Index of Linguistic Diversity: A New Quantitative Measure of Trends in the Status of the World's Languages". In: *Language Documentation & Conservation* 4, pp. 97–151. ISSN: 1934-5275.

Haynie, Hannah Jane (2012). "Studies in the History and Geography of California Languages". PhD thesis. University of California, Berkeley, p. 284.

Heeringa, Wilbert (2004). "Measuring dialect pronunciation differences using Levenshtein distance". PhD thesis. University of Groningen, p. 315.

Heeringa, Wilbert and John Nerbonne (2001). "Dialect areas and dialect continua". In: *Language Variation and Change* 13.03, pp. 375–400.

Henzen, Walter (1927). *Die deutsche Freiburger Mundart im Sense- und südöstlichen Seebezirk*. Frauenfeld: Huber.

Hoch, Shawn and James Hayes (2010). "Geolinguistics: The Incorporation of Geographic Information Systems and Science". In: *The Geographical Bulletin* 51.1, pp. 23–36. ISSN: 0731-3292.

Hodler, Werner (1969). *Berndeutsche Syntax*. Bern: Francke.

Holman, Eric W., Christian Schulze, Dietrich Stauffer, and Søren Wichmann (2007). "On the relation between structural diversity and geographical distance among languages: Observations and computer simulations". In: *Linguistic Typology* 11.2, pp. 393–421. ISSN: 14300532. DOI: 10.1515/LINGTY.2007.027.

Horvath, Barbara and Ronald Horvath (2001). "A multilocality study of a sound change in progress: The case of /l/ vocalization in New Zealand and Australian English". In: *Language Variation and Change* 13.2001, pp. 37–57. ISSN: 09543945. DOI: 10.1017/S0954394501131029.

— (2002). "The geolinguistics of /l/vocalization in Australia and New Zealand". In: *Journal of Sociolinguistics* 6.3, pp. 319–346. ISSN: 1360-6441. DOI: 10.1111/1467-9481.00191.

Hosmer, David W., Scott Taber, and Stanley Lemeshow (1991). "The Importance of Assessing the Fit of Logistic Regression Models: A Case Study". In: *American Journal of Public Health* 81.12.

Hotzenköcherle, Rudolf (1961). "Zur Raumstruktur des Schweizerdeutschen". In: *Zeitschrift für Mundartforschung* 28.3, pp. 207–227.

— (1984). *Die Sprachlandschaften der deutschen Schweiz*. Aarau, Frankfurt a. M., Salzburg: Verlag Sauerländer.

Hotzenköcherle, Rudolf, Robert Schläpfer, Rudolf Trüb, and Peter Zinsli, eds. (1962). *Sprachatlas der deutschen Schweiz (SDS)*. Bern, Basel: Francke Verlag.

Jeszenszky, Péter, Curdin Derungs, Peter Ranacher, Philipp Stoeckle, Elvira Glaser, and Robert Weibel (in prep.). "Data-driven detection of transition zones in dialectal variables". In:

Jeszenszky, Péter, Philipp Stoeckle, Elvira Glaser, and Robert Weibel (2017). "Exploring global and local patterns in the correlation of geographic distances and morphosyntactic variation in Swiss German". In: *Journal of Linguistic Geography* 5.2, pp. 1–23. DOI: 10.1017/jlg.2017.5.

— (in revision). "Crisp boundaries vs. gradual transitions: Quantitative models for transitions between areas of dialectal variants". In: *Journal of Linguistic Geography* in revision.

Jeszenszky, Péter and Robert Weibel (2015). "Measuring boundaries in the dialect continuum". In: *AGILE Conference 2015, Lisbon 09.-12.06.2015 The 18th AGILE International Conference on Geographic Information Science Geographic Information Science as an Enabler of Smarter Cities and Communities*. Ed. by Fernando Bacao, Maribel Yasmina Santos, and Marco Painho. Lisbon.

Kelle, Bernhard (2001). "Zur Typologie der Dialekte in der deutschsprachigen Schweiz: Ein dialektometrischer Versuch". In: *Dialectologia et Geolinguistica* 2001.9, pp. 9–34. ISSN: 0942-4040. DOI: 10.1515/dig.2001.2001.9.9.

Kellerhals, Sandra (2014). "Dialektometrische Analyse und Visualisierung von schweizerdeutschen Dialekten auf verschiedenen linguistischen Ebenen". PhD thesis. Universität Zürich, p. 129.

Kellerhals, Sandra, Yves Scherrer, Elvira Glaser, and Robert Weibel (2014). "The distribution of aggregated syntactic construction types compared with other linguistic levels - A dialectometrical analysis of Swiss German dialects". In: *Methods in Dialectology XV*.

Kessler, Brett (1995). "Computational dialectology in Irish Gaelic". In: *Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics*. 1971. Dublin, pp. 60–66.

"SBS - Sprachatlas von Bayerisch-Schwaben" (1996-2009). In: *Sprachatlas von Bayerisch-Schwaben, 14 Volumes*. Ed. by Werner König (ed.) Heidelberg: Winter (Bayerischer Sprachatlas. Regionalteil 1).

Kortmann, Bernd (2002). "New Prospects for the Study of Dialect Syntax: Impetus from Syntactic Theory and Language Typology". In: *Syntactic Microvariation*. Ed. by Sjef Barbiers, Leonie Cornips, and Susanne van der Kleij. Amsterdam: Meertens Instituut, pp. 185–213.

Kretzschmar, William A., Virginia G. McDavid, Theodore K. Lerud, and Ellen Johnson, eds. (1993). *Handbook of the linguistic atlas of the middle and south Atlantic states*. Chicago: The University of Chicago Press, p. 454.

Kretzschmar, William A. and Edgar W. Schneider (1996). *Introduction to Quantitative Analysis of Linguistic Survey Data: An atlas by the numbers*. Thousand Oaks: SAGE.

Kronenfeld, Barry Joel (2007). "Triangulation of Gradient Polygons: A Spatial Data Model for Categorical Fields". In: *Spatial Information Theory*. Ed. by Stephan Winter, Matt Duckham, L. Kulik, and B. Kuipers. New York: Springer, pp. 421–437.

Kurath, Hans (1949). *A Word Geography of the Eastern United States*. Ann Arbor, MI: University of Michigan Press.

— (1972). *Studies in Area Linguistics*. Bloomington/London: Indiana University Press.

Kurath, Hans, Miles L. Hanley, Bernard Bloch, Guy S. Lowman Jr., and Marcus L. Hansen, eds. (1939). *Linguistic Atlas of New England*. Providence: Brown University (sponsored by the American Council of Learned Societies [ACLS], assisted by universities, and colleges in New England).

Kürschner, Sebastian and Charlotte Gooskens (2011). "Verstehen nah verwandter Varietäten über Staatsgrenzen hinweg." In: *Dynamik des Dialekts - Wandel und*

Variation. Akten des 3. Kongresses der Internationalen Gesellschaft für Dialektologie des Deutschen (IGDD). Ed. by Elvira Glaser, Jürgen Erich Schmidt, and Natascha Frey. Stuttgart: Steiner.

Labov, William (2001). *Principles of Linguistic Change, Volume 2: Social Factors*. Wiley-Blackwell, p. 592.

Labov, William, Sharon Ash, and Charles Boberg (2006). *Atlas of North American English: Phonetics, phonology, and sound change*. New York: Mouton de Gruyter.

Lakoff, George (1987). *Women, fire, and dangerous things: What Categories Reveal about Thought*. Chicago and London: University of Chicago Press. ISBN: 0-226-46804-6.

Lameli, Alfred (2013). *Strukturen im Sprachraum: Analysen zur arealtypologischen Komplexität der Dialekte in Deutschland*. Vol. 54. Walter de Gruyter.

Lameli, Alfred, Roland Kehrein, and Stefan Rabanus (2010). *Language and Space: An International Handbook of Linguistic Variation. Volume 2*. Berlin: De Gruyter Mouton.

Lameli, Alfred, Volker Nitsch, Jens Südekum, and Nikolaus Wolf (2015). "Same same but different: Dialects and trade". In: *German Economic Review* 16.3, pp. 290–306. ISSN: 14656485. DOI: 10.1111/geer.12047.

LaRue, Michelle A. and Clayton K. Nielsen (2008). "Modelling potential dispersal corridors for cougars in midwestern North America using least-cost path methods". In: *Ecological Modelling* 212.3-4, pp. 372–381. ISSN: 03043800.

Lee, Jay and William A. Kretzschmar (1993). "Spatial analysis of linguistic data with GIS functions". In: *International Journal of Geographical Information Systems* 7.6, pp. 541–560. ISSN: 0269-3798. DOI: 10.1080/02693799308901981.

Lee, Sean and Toshikazu Hasegawa (2014). "Oceanic barriers promote language diversification in the Japanese Islands". In: *Journal of Evolutionary Biology* 27.9, pp. 1905–1912. ISSN: 14209101. DOI: 10.1111/jeb.12442.

Leemann, Adrian, Marie-José Kolly, and David Britain (2018). "The English Dialects App: The creation of a crowdsourced dialect corpus". In: *Ampersand* 5.August 2017, pp. 1–17. ISSN: 22150390. DOI: 10.1016/j.amper.2017.11.001.

Leemann, Adrian, Marie-José Kolly, S. Grimm, S. Robert, Stephan Elspaß, R. Möller, Jürg Fleischer, and Roland Kehrein (2015). *Grüezi, Moin, Servus*.

Leemann, Adrian, Marie-José Kolly, Ross Purves, David Britain, and Elvira Glaser (2016). "Crowdsourcing language change with smartphone applications". In: *PLoS ONE* 11.1, pp. 1–25. ISSN: 19326203. DOI: 10.1371/journal.pone.0143060.

Leung, Yee (1987). "On the Imprecision of Boundaries". In: *Geographical Analysis* 19.2, pp. 125–151. ISSN: 1538-4632. DOI: 10.1111/j.1538-4632.1987.tb00120.x.

Löffler, Heinrich (2003). *Dialektologie - Eine Einführung*. Tübingen: Gunter Narr, p. 160. ISBN: 3-8233-4998-8.

Longobardi, Giuseppe and Cristina Guardiano (2009). "Evidence for syntax as a signal of historical relatedness". In: *Lingua* 119.11, pp. 1679–1706. ISSN: 00243841. DOI: 10.1016/j.lingua.2008.09.012.

Lötscher, Andreas (1983). *Schweizerdeutsch*. Huber.

Lowry, Richard (2000). *VassarStats: website for statistical computation*.

Lucas, Christopher (2015). "Contact-induced language change". In: *The Routledge Handbook of Historical Linguistics.* London: Routledge, pp. 519–536.

Manni, Franz, Etienne Guérard, and Evelyne Heyer (2004). "Geographic Patterns of (Genetic, Morphologic, Linguistic) Variation: How Barriers Can Be Detected by Using Monmonier's Algorithm". In: *Human Biology* 76.2, pp. 173–190.

Mantel, Nathan (1967). "The Detection of Disease Clustering and a Generalized Regression Approach". In: *Cancer Research* 27.2, pp. 209–220. ISSN: 00280836 (ISSN). DOI: 10.1038/212665a0.

Mark, David M. and Ferenc Csillag (1989). "The Nature Of Boundaries On 'Area-Class' Maps". In: *Cartographica: The International Journal for Geographic Information and Geovisualization* 26.1, pp. 65–78. ISSN: 0317-7173. DOI: 10.3138/D235-3262-062X-4472.

Maurer, Friedrich (1942). *Oberrheiner, Schwaben, Südalemannen*. Strassburg: Hünenburg.

Montello, Daniel R., Michael F. Goodchild, Jonathon Gottsegen, and Peter Fohl (2003). "Where's downtown?: Behavioral methods for determining referents of vague spatial queries". In: *Spatial Cognition and Computation* 3.2-3, pp. 185–204. ISSN: 13875868. DOI: 10.1080/13875868.2003.9683761.

Montgomery, Chris and Philipp Stoeckle (2013). "Geographic information systems and perceptual dialectology: a method for processing draw-a-map data". In: *Journal of Linguistic Geography* 1.01, pp. 52–85. ISSN: 2049-7547. DOI: 10.1017/jlg.2013.4.

Nerbonne, John (2006). "Identifying linguistic structure in aggregate comparison". In: *Literary and Linguistic Computing* 21.4, pp. 463–475. ISSN: 02681145. DOI: 10.1093/llc/fql041.

— (2009). "Data-Driven Dialectology". In: *Language and Linguistics Compass* 3.1, pp. 175–198. ISSN: 1749818X. DOI: 10.1111/j.1749-818X.2008.00114.x.

— (2010a). "Mapping aggregate variation". In: *Language and Space. An international Handbook of Linguistic Variation. Vol 1. Theories and Methods*. Berlin/ New York: Mouton de Gruyter, pp. 476–495.

— (2010b). "Measuring the diffusion of linguistic change". In: *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 365.1559, pp. 3821–8. ISSN: 1471-2970. DOI: 10.1098/rstb.2010.0048.

Nerbonne, John, Rinke Colen, Charlotte Gooskens, and Peter Kleiweg (2011). "Gabmap - a Web Application for Dialectology". In: *Dialectologia* II.special issue, pp. 65–89.

Nerbonne, John and Wilbert Heeringa (2007). "Geographic distributions of linguistic variation reflect dynamics of differentiation". In: *Roots: Linguistics in Search of its Evidential Base*. Ed. by Sam Featherston and Wolfgang Sternefeld. New York: Mouton de Gruyter, pp. 267–297.

Nerbonne, John, Wilbert Heeringa, and Peter Kleiweg (1999). "Edit Distance and Dialect Proximity". In: *Time Warps, String Edits and Macromolecules: The Theory*

and Practice of Sequence Comparison. Ed. by D. Sankoff and J. Kruskal. Stanford, CA: Cent. Study Lang. Inf., pp. v–xv.

Nerbonne, John and Peter Kleiweg (2007). "Toward a dialectological yardstick". In: *Journal of Quantitative Linguistics* 14.2, pp. 148–167.

Nguyen, Dong and Jacob Eisenstein (2017). "A Kernel Independence Test for Geographical Language Variation". In: *Computational Linguistics* 43.3, pp. 567–592.

Nichols, Johanna (1992). *Linguistic diversity in space and time*. Chicago: University of Chicago Press. ISBN: 0-226-58057-1.

— (2013). "The vertical archipelago: Adding the third dimension to linguistic geography". In: *Space in Language and Linguistics*. Ed. by Peter Auer, Martin Hilpert, Anja Stukenbrock, and Benedikt Szmrecsanyi. Berlin, Boston: Walter de Gruyter, pp. 38–60. ISBN: 978-3-11-031196-9.

Niebaum, Hermann and Jürgen Macha (2006). *Einführung in die Dialektologie des Deutschen. 2.* Tübingen: Aufl. Niemeyer (Germanistische Arbeitshefte 37).

O'Grady, William Delaney, John Archibald, Mark Aronoff, and Janie Rees-Miller, eds. (2016). *Contemporary Linguistics. An Introduction*. Eighth edi. London and New York: St. Martin's Press.

Parker, Bradley J. (2006). "Toward an Understanding of Borderland Processes". In: *American Antiquity* 71.1, pp. 77–100.

Pickl, Simon (2013). "Probabilistische Geolinguistik". PhD thesis. University of Salzburg, p. 265.

Pickl, Simon and Jonas Rumpf (2012). "Dialectometric concepts of space: Towards a variant-based dialectometry". In: *Dialectological and Folk Dialectological Concepts of Space - Current Methods and Perspectives in Sociolinguistic Research on Dialect Change*. Ed. by Sandra Hansen, Christian Schwarz, Philipp Stoeckle, and Tobias Streck. Linguae &. Berlin/ New York: Walter de Gruyter, pp. 199–214.

Pickl, Simon, Aaron Spettl, Simon Magnus Pröll, Stephan Elspaß, Werner König, and Volker Schmidt (2014). "Linguistic distances in dialectometric intensity estimation". In: *Journal of Linguistic Geography* 2.01, pp. 25–40. ISSN: 2049-7547. DOI: 10.1017/jlg.2014.3.

Preston, Dennis R. (1989). *Perceptual dialectology: Nonlinguists' views of areal linguistics.* Providence: Fortis Publications.

Preston, Dennis R. and Gregory C. Robinson (2005). "Dialect Perception and Attitudes to Variation". In: *Language in Society* 36.133.

Pröll, Simon Magnus (2013a). "Detecting structures in linguistic maps–Fuzzy clustering for pattern recognition in geostatistical dialectometry". In: *Literary and Linguistic Computing* 28.1, pp. 108–118. ISSN: 0268-1145. DOI: 10.1093/llc/fqs059.

— (2013b). "Raumvariation zwischen Muster und Zufall". PhD thesis. Universität Augsburg, p. 216.

Pröll, Simon Magnus, Simon Pickl, and Aaron Spettl (2014). "Latente Strukturen in geolinguistischen Korpora". In: *Deutsche Dialekte. Konzepte, Probleme, Handlungsfelder. Akten des 4. Kongresses der Internationalen Gesellschaft für Dialektologie des*

Deutschen (IGDD) in Kiel. (Zeitschrift für Dialektologie und Linguistik, Beihefte, 158.) Ed. by Michael Elmentaler, Markus Hundt, and Jürgen Erich Schmidt. Stuttgart: Steiner, pp. 247–258.

Ranacher, Peter, Rik van Gijn, and Curdin Derungs (2017). "Identifying probable pathways of language diffusion in South America". In: *Societal Geo-innovation. Selected Papers of the 20th AGILE Conference on Geographic Information Science, Wageningen, The Netherlands*. Ed. by Arnold Bregt, Tapani Sarjakoski, Ron van Lammeren, and Frans Rip. Cham, Switzerland: Springer.

Rees, W. G. (2004). "Least-cost paths in mountainous terrain". In: *Computers and Geosciences* 30.3, pp. 203–209. ISSN: 00983004. DOI: 10.1016/j.cageo.2003.11.001.

Reid, Scott A. and Sik Hung Ng (1999). "Language, Power, and Intergroup Relations". In: *Journal of Social Issues* 55.1, pp. 119–139. ISSN: 0022-4537. DOI: 10.1111/0022-4537.00108.

Rogers, Everett M. (1995). *Diffusion of innovations*. 4th editio. New York: Free Press.

Rosch, Eleanor H. (1973). "Natural categories". In: *Cognitive Psychology* 4.3, pp. 328–350. ISSN: 00100285. DOI: 10.1016/0010-0285(73)90017-0.

Rumpf, Jonas, Simon Pickl, Stephan Elspaß, Werner König, and Volker Schmidt (2009). "Structural analysis of dialect maps using methods from spatial statistics". In: *Zeitschrift für Dialektologie und Linguistik* 76.3, pp. 280–308.

— (2010). "Quantification and statistical analysis of structural similarities in dialectological area-class maps". In: *Dialectologia et Geolinguistica* 18.1, pp. 73–100.

Scherrer, Yves (2012). "Generating Swiss German sentences from Standard German: a multi-dialectal approach". PhD thesis. Université de Genève, p. 446. DOI: 10.13097/archive-ouverte/unige:26361.

Scherrer, Yves, Adrian Leemann, Marie-José Kolly, and Iwar Werlen (2012). "Dialäkt Äpp - A smartphone application for Swiss German dialects with great scientific potential". In: *SIDG, Wien*. July 2012. Wien: SIDG 2012, Wien, p. 29.

Scherrer, Yves and Philipp Stoeckle (2016). "A quantitative approach to Swiss German – Dialectometric analyses and comparisons of linguistic levels". In: *Dialectologia et Geolinguistica* 24, pp. 92–125. DOI: 10.1515/dialect-2016-0006.

Scholz, Johannes, Thomas J. Lampoltshammer, Norbert Bartelme, and Eveline Wandl-Vogt (2016). "Spatial-temporal Modeling of Linguistic Regions and Processes with Combined Intermediate and Crisp Boundaries". In: *Progress in Cartography: EuroCarto 2015*. Ed. by Georg Gartner, M. Jobst, and Haosheng Huang. Springer International Publishing, pp. 133–151. ISBN: 978-3-319-19601-5. DOI: 10.1007/978-3-319-19602-2{_}9.

Schrambke, Renate (2010). "Language and space : Traditional dialect geography". In: *Language and Space. An international Handbook of Linguistic Variation. Vol 1. Theories and Methods*. Ed. by Peter Auer and Jürgen Erich Schmidt. De Gruyter Mouton, pp. 87–107. ISBN: 9783110220278. DOI: 10.1515/9783110220278.87.

Schreier, Daniel (2002). "Past be in Tristan da Cunha: The Rise and Fall of Categoricality in Language Change". In: *American Speech* 77.1, pp. 70–99. ISSN: 0003-1283, 0003-1283.

— (2009). "Language in isolation, and its implications for variation and change". In: *Linguistics and Language Compass* 3.2, pp. 682–699. ISSN: 1749818X. DOI: 10.1111/j.1749-818X.2009.00130.x.

Séguy, Jean (1971). "La relation entre la distance spatiale et la distance lexicale". In: *Revue de Linguistique Romane* 35.138, pp. 335–357.

Seiler, Guido (2005). "Wie verlaufen syntaktische Isoglossen, und welche Konsequenzen sind daraus zu ziehen?" In: *Moderne Dialekte – Neue Dialektologie*. Ed. by Eckhard Eggers, Jürgen Erich Schmidt, and Dieter Stellmacher. Stuttgart: Franz Steiner Verlag, pp. 313–341. ISBN: 3515087621.

Shackleton, Robert G. Jr. (2005). "English-American Speech Relationships: A Quantitative Approach". In: *Journal of English Linguistics* 33.2, pp. 99–160. ISSN: 0075-4242. DOI: 10.1177/0075424205279017.

— (2007). "Phonetic Variation in the Traditional English Dialects: A Computational Analysis". In: *Journal of English Linguistics* 35.1, pp. 30–102. ISSN: 0075-4242. DOI: 10.1177/0075424206297857.

Sibler, Pius (2011). *Visualisierung und Geostatistische Analyse mit Daten des Syntaktischen Atlas der Deutschen Schweiz (SADS): Master's thesis*.

Sibler, Pius, Robert Weibel, Elvira Glaser, and Gabriela Bart (2012). "Cartographic Visualization in Support of Dialectology". In: *Proceedings - AutoCarto 2012 - Columbus, Ohio, USA - September 16-18, 2012*. 2000, p. 18.

Sieber, Christian D. (2017). *Einfluss von scharfen und unscharfen Grenzen auf syntaktische Dialektunterschiede in der deutschen Schweiz: Master's thesis*.

Smith, Barry and David M. Mark (2003). "Do mountains exist? Towards an ontology of landforms". In: *Environment and Planning B: Planning and Design* 30.3, pp. 411–427. ISSN: 0265-8135. DOI: 10.1068/b12821.

Smith, Barry and Achille C. Varzi (1997). "Fiat and Bona Fide Boundaries: Towards an Ontology of Spatially Extended Objects". In: *Spatial Information Theory: A Theoretical Basis for GIS*. Berlin, Heidelberg: Springer, pp. 103–119.

— (2000). "Fiat and Bona Fide Boundaries". In: *Philosophy and Phenomenological Research* 60.2, pp. 401–420.

Smith, Edward E. and Douglas L. Medin (1981). *Categories and Concepts*. Cambridge, MA / London, UK: Harvard University Press, p. 208.

Snoek, Conor (2014). "Review of Gabmap: Doing Dialect Analysis on the Web". In: *Language Documentation and Conversation* 8.June, pp. 192–208.

Sonderegger, Stefan (1962). *Die schweizerdeutsche Mundartforschung*. Frauenfeld: Verlag Huber.

— (2013). *Appenzeller Namenbuch. Vol. 2: Die Orts- und Flurnamen des Landes Appenzell: Herkunft und Bedeutung der Orts- und Flurnamen des Landes Appenzell. Teilbd. 1: Einführung und historisches Namenlexikon A - G*. Frauenfeld.

Speelman, Dirk, Stefan Grondelaers, and Dirk Geeraerts (2003). "Profile-Based Linguistic Uniformity as a Generic Method for Comparing Language Varieties". In: *Computers and the Humanities* 37, pp. 317–337. ISSN: 00104817. DOI: 10.1023/A:1025019216574.

Spruit, Marco René (2006). "Measuring Syntactic Variation in Dutch Dialects". In: *Literary and Linguistic Computing* 21.4 - Progress in Dialectometry: Toward Explanation, pp. 493–506.

— (2008). "Quantitative perspectives on syntactic variation". PhD thesis. Utrecht: Universiteit van Amsterdam, p. 157. ISBN: 9789078328483.

Spruit, Marco René, Wilbert Heeringa, and John Nerbonne (2009). "Associations among linguistic levels". In: *Lingua* 119.11 The forests behind the trees, pp. 1624–1642. ISSN: 00243841. DOI: 10.1016/j.lingua.2009.02.001.

Stanford, James N. (2012). "One size fits all? Dialectometry in a small clan-based indigenous society". In: *Language Variation and Change* 24.02, pp. 247–278. ISSN: 0954-3945. DOI: 10.1017/S0954394512000087.

Stark, Elisabeth, Simone Ueberwasser, and Anne Göhring (2014). *Corpus "What's up, Switzerland?". University of Zurich.*

Stoeckle, Philipp (2014). "Subjektive Dialekträume im alemannischen Dreiländereck". PhD thesis. Albert-Ludwigs-Universität, Freiburg im Breisgau, p. 632.

— (2016a). "Horizontal and vertical variation in Swiss German morphosyntax". In: *The future of dialects: Selected papers from Methods in Dialectology XV (Language Variation 1).* Pp. 195–214. DOI: 10.17169/langsci.b81.150.

— (2016b). "Horizontal and vertical variation in Swiss German morphosyntax". In: *The future of dialects: Selected papers from Methods in Dialectology XV (Language Variation 1).* Ed. by Marie-Hélène Côté, Remco Knooihuizen, and John Nerbonne. Berlin: Language Science Press, pp. 195–215.

— (2018). "Zur Syntax von afa (‚anfangen') im Schweizerdeutschen – Kookkurrenzen, Variation und Wandel". In: *Syntax aus Saarbrücker Sicht 2. Beiträge der SaRDiS-Tagung zur Dialektsyntax*. Ed. by Augustin Speyer and Philipp Rauth. Stuttgart: Steiner, pp. 173–203.

Stoeckle, Philipp and Péter Jeszenszky (2017). "Sprachgeographie und Geographische Informationssysteme (GIS)". In: *Deutsche Dialekte in Europa. Perspektiven auf Variation, Wandel und Übergänge.* Ed. by Timo Ahlers, Susanne Oberholzer, Michael Riccabona, and Philipp Stoeckle. Hildesheim: Olms (Kleine und regionale Sprachen, Bd. 3)., pp. 261–287. ISBN: 978-3-487-15419-0.

Stucki, Karl (1917). *Die Mundart von Jaun im Kanton Freiburg: Lautlehre und Flexion*. Frauenfeld: Huber.

Sui, Daniel Z. (2004). "Tobler's first law of geography: A big idea for a small world?" In: *Annals of the Association of American Geographers* 94.2, pp. 269–277. ISSN: 00045608. DOI: 10.1111/j.1467-8306.2004.09402003.x.

Swadesh, Morris (1955). "Towards Greater Accuracy in Lexicostatistic Dating". In: *International Journal of American Linguistics* 21.2, pp. 121–137.

Szmrecsanyi, Benedikt (2012). "Geography is overrated". In: *Dialectological and Folk Dialectological Concepts of Space - Current Methods and Perspectives in Sociolinguistic Research on Dialect Change*. Ed. by Sandra Hansen, Christian Schwarz, Philipp Stoeckle, and Tobias Streck. Berlin, Boston: De Gruyter, pp. 215–231. ISBN: 978-3-11-022912-7.

— (2014). "Methods and objectives in contemporary dialectology". In: *Contemporary approaches to dialectology: The area of North, Northwest Russian and Belarusian vernaculars*. Ed. by Ilja A. Seržant and Björn Wiemer. 1. Bergen: Department of Foreign Languages, University of Bergen, pp. 81–92.

Thomas, Alan Richard (1980). *Areal Analysis of Dialect Data by Computer: A Welsh Example*. Cardiff: University of Wales Press.

Tobler, Waldo R. (1970). "A computer movie simulating urban growth in the Detroit region". In: *Economic Geography* 46.2, pp. 234–240.

Trudgill, Peter (1974). "Linguistic change and diffusion : Description and explanation in sociolinguistic dialect geography". In: *Language in Society* 2, pp. 215–246.

Uiboaed, Kristel, Cornelius Hasselblatt, Liina Lindström, Kadri Muischnek, and John Nerbonne (2013). "Variation of verbal constructions in Estonian dialects". In: *Literary and Linguistic Computing* 28.1, pp. 42–62. ISSN: 02681145. DOI: 10.1093/llc/fqs053.

Valls, Esteve, Martijn Wieling, and John Nerbonne (2013). "Linguistic advergence and divergence in north-western Catalan: A dialectometric investigation of dialect leveling and border effects". In: *Literary and Linguistic Computing* 28.1, pp. 119–146. ISSN: 02681145. DOI: 10.1093/llc/fqs052.

Vogt, Lars, Peter Grobe, Björn Quast, and Thomas Bartolomaeus (2012). "Fiat or Bona Fide Boundary - A Matter of Granular Perspective". In: *PLoS ONE* 7.12. ISSN: 19326203. DOI: 10.1371/journal.pone.0048603.

Voronoi, Georges (1908). "Nouvelles applications des parametres continus a la theorie des formes quadratiques." In: *Journal für die reine und angewandte Mathematik* 133, pp. 97–178. ISSN: 0075-4102. DOI: 10.1515/crll.1908.134.198.

Wang, William S.-Y. and Luca L. Cavalli-Sforza (1986). "Spatial distance and lexical replacement". In: *Language* 62, pp. 38–55.

Warner, Rebecca M. (2013). *Applied Statistics - From Bivariate Through Multivariate Techniques*. 2nd. Los Angeles / London / New Delhi / Singapore / Washington D.C.: SAGE. ISBN: 978-1-4129-9134-6.

Wattel, Evert and Pieter van Reenen (2010). "Probabilistic maps". In: *Language and Space An International Handbook of Linguistic Variation Volume 2: Language Mapping*. Ed. by Alfred Lameli, Roland Kehrein, and Stefan Rabanus. De Gruyter Mouton, pp. 496–505. ISBN: 9783110180022. DOI: 10.1515/9783110220278.

Weinreich, Uriel, William Labov, and Marvin I. Herzog (1968). "Empirical foundations for a theory of language change". In: *Directions for historical linguistics*. Ed. by Winfred P. Lehmann and Yakov Malkiel. Austin: University of Texas Press, pp. 95–195.

Wenker, Georg (2013). *Schriften zum Sprachatlas des Deutschen Reichs. Band 1: Handschriften: Allgemeine Texte, Kartenkommentare 1889–1897*. Ed. by Alfred Lameli, Johanna Heil, and Constanze Wellendorf. Hildesheim, Zürich, New York: Olms. (Deutsche Dialektgeographie 111.1.)

Wieling, Martijn (2012). "A Quantitative Approach to Social and Geographical Dialect Variation". PhD thesis. Groningen, p. 178. ISBN: 9789036755214.

Wieling, Martijn, Simonetta Montemagni, John Nerbonne, and R. Harald Baayen (2014). "Lexical differences between Tuscan dialects and standard Italian: Accounting for geographic and sociodemographic variation using generalized additive mixed modeling". In: *Language* 90.3, pp. 669–692. ISSN: 0097-8507. DOI: 10.1353/lan.2014.0064.

Wieling, Martijn and John Nerbonne (2015). "Advances in Dialectometry". In: *Annual Review of Linguistics*, pp. 243–264. ISSN: 2333-9683. DOI: 10.1146/annurev-linguist-030514-124930.

Willis, David (2017). "Investigating geospatial models of the diffusion of morphosyntactic innovations: The Welsh strong second-person singular pronoun chdi". In: *Journal of Linguistic Geography* 5, pp. 41–66. ISSN: 2049-7547. DOI: 10.1017/jlg.2017.1.

Wolk, Christoph (2014). "Integrating aggregational and probabilistic approaches to dialectology and language variation". PhD thesis.

Wrede, Ferdinand, Walther Mitzka, and Bernhardt Martin (1927). *Deutscher Sprachatlas auf Grund des Sprachatlas des Deutschen Reiches von Georg Wenker.* Marburg (Lahn): N.G. Elwert'sche Verlagsbuchhandlung.

Yokoyama, Shoichi and Haruko Sanada (2009). "Logistic regression model for predicting language change". In: *Studies in Quantitative Linguistics 5, Issues in Quantitative Linguistics*. Ed. by R. Köhler. Lüdenscheid (D): RAM-Verlag, pp. 176–192. ISBN: 978-3-9802659-9-7.

Zadeh, L A (1965). "Fuzzy Sets *". In: *Information and Control* 8, pp. 338–353.

Appendix A

Appendix A

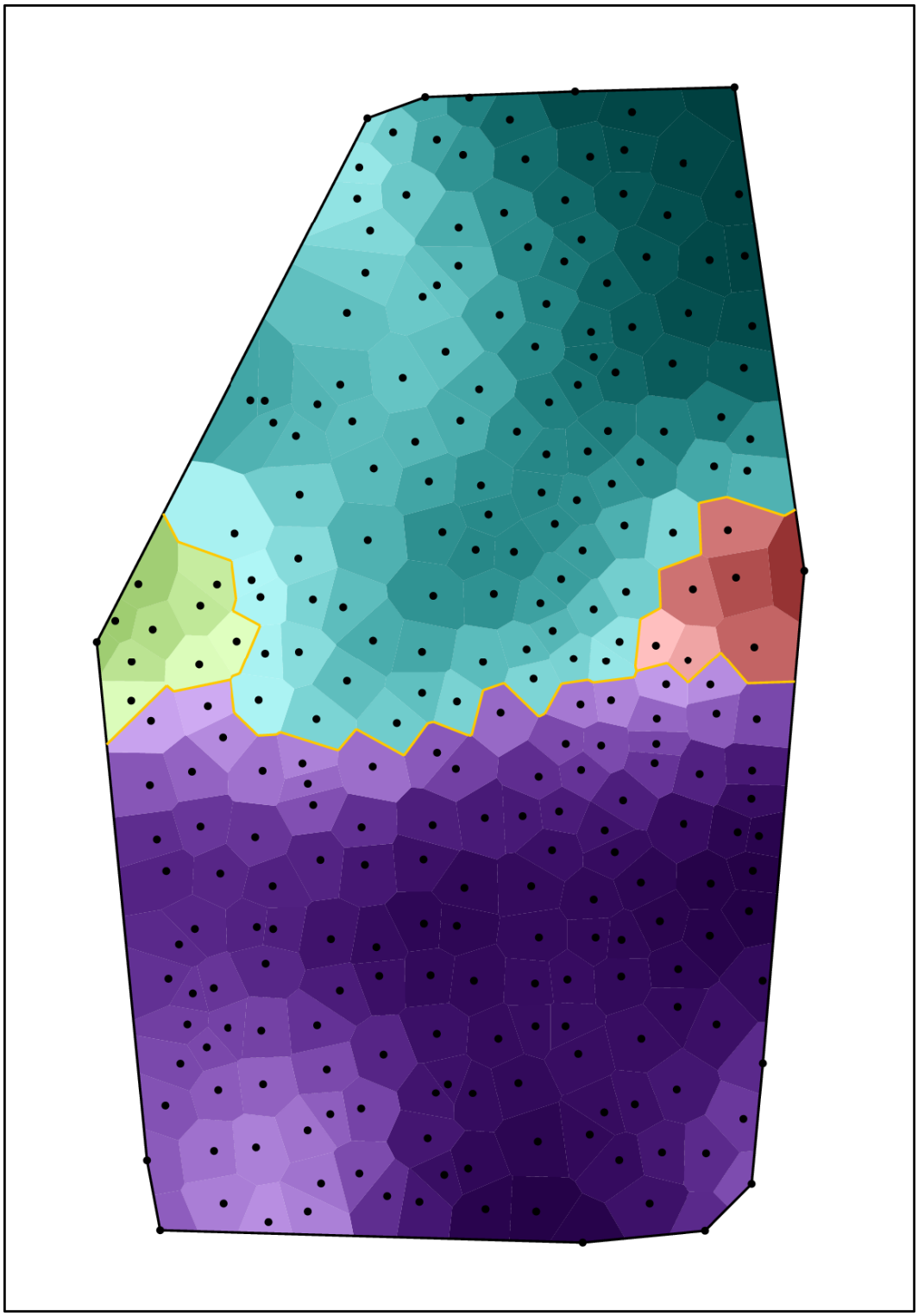

FIGURE A.1: Example coropleth map based on SBS map 80 *'Kartoffelkraut'* (SBS 1997–2005, vol. 8, p. 294f.), as published in Rumpf et al., 2009.

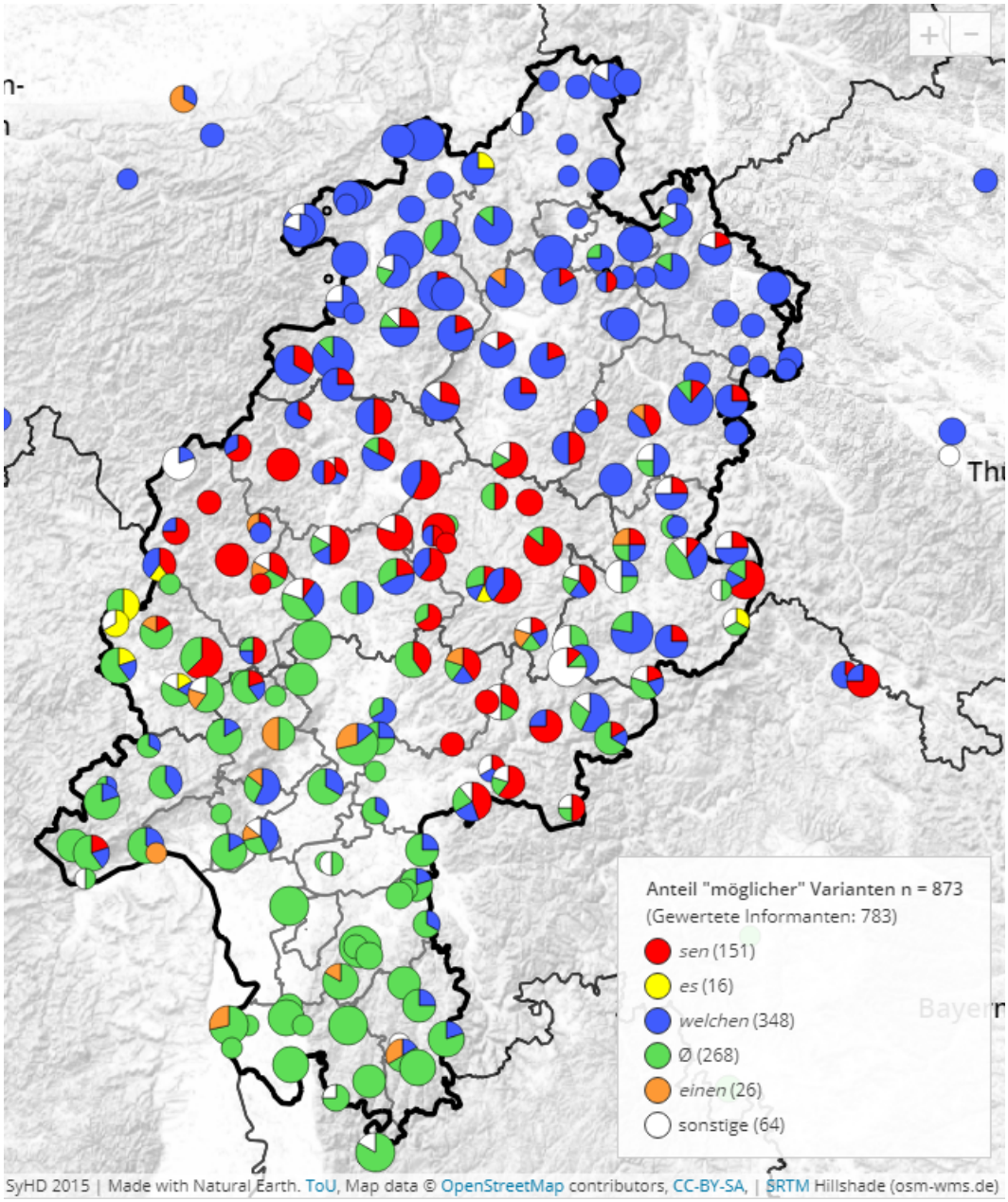

FIGURE A.2: Example diagram map in the SyHD - *'Indefinite partitive pronoun'* - Translation task: *"Ich habe keinen Zucker mehr. Hast du X noch?"* http://www.syhd.info.

Appendix B

Curriculum Vitae

Curriculum Vitae

PÉTER JESZENSZKY

Geographic Information Systems Unit

Department of Geography

University of Zurich

Education

- **2013-2018**
  Ph.D., Geographic Information Systems Unit, Department of Geography, University of Zurich, Zurich, Switzerland
  Thesis: "Quantitative Modelling of Geographic Influences on Spatial Variation in Swiss German Dialects", supervisor: Prof. Dr. Robert Weibel
- **2010-2012**
  College Certificate, International freighting and logistics,
  Budapest Business School (University of Applied Sciences),
  Faculty of International Management and Business,
  Thesis: "The packaging system of a car parts factory – Analysed through the introduction of a new product"
- **2005-2011**
  University Diploma (equivalent to Masters' Degree), Cartography,
  Eötvös Loránd University, Budapest, Hungary
  Faculty of Informatics
  Thesis: "Creating an interactive 3D tourist guide using Javascript and Google Earth". Supervisor: Mátyás Gede
- **1999-2005**
  Diploma, Teleki Blanka Secondary Grammar School (Teleki Blanka Gimnázium), Székesfehérvár, Hungary

Professional experience

- **PhD Student / Research Assistant**
  Diploma, Geographic Information Systems Unit,
  Department of Geography, University of Zurich, Zurich, Switzerland
  02.2013 – 04/2017
- **Packaging Manager**
  Logistic Department, Valeo Auto-Electric Hungary, Veszprém, Hungary
  02.2012 – 02/2013
- **Freelance Cartographer**
  10.2009 – 02.2013

Teaching and work experience during the doctoral studies

- Lecture on making effective posters, for the class 'Advanced Applied GIS' (GEO 372), Fall Semester 2014
- Teaching assistant for 'Advanced Applied GIS' (GEO 372), Fall Semester 2014, Fall Semester 2015, Fall Semester 2016
- Teaching assistant for 'Getting started with R for spatial analysis' (GEO 812) Support at practicals; September 2016
- Project assistant for the project 'LiMiTS', Department of Comparative Linguistics, University of Zurich, July – August 2017
- Project assistant for the project 'MOASIS', University Research Priority Program (URPP) Dynamics of Healthy Aging, Department of Geography and Department of Psychology, University of Zurich, January – February 2018
- Project assistant for the project 'evertools', Chair of Cognitive Science, ETH Zurich, March – April 2018

Publications during the doctoral studies

Jeszenszky, Péter & Robert Weibel. (2014). Correlating morphosyntactic dialect variation with geographic distance : Local beats global. In K. Stewart, E. Pebesma, G. Navratil, P. Fogliaroni, & M. Duckham (Eds.), *Extended Abstract Proceedings of the GIScience Conference 2014, Vienna* (pp. 186–191). Vienna: GeoInfo Series Vienna 2014.

Jeszenszky, Péter & Robert Weibel. (2015). Measuring boundaries in the dialect continuum. In F. Bacao, M. Y. Santos, & M. Painho (Eds.), *AGILE Conference 2015, Lisbon 09.-12.06.2015 Proceedings of The 18th AGILE International Conference on Geographic Information Science Geographic Information Science as an Enabler of Smarter Cities and Communities*. Lisbon.

Jeszenszky, Péter & Robert Weibel. (2016). Modeling transitions between syntactic variants in the dialect continuum. In T. Sarjakoski, M. Y. Santos, & T. L. Sarjakoski (Eds.), *Proceedings of the AGILE Conference 2016, Helsinki 14.-17.06.2016, The 19th AGILE International Conference on Geographic Information Science*. Helsinki.

Stoeckle, Philipp & **Péter Jeszenszky**. (2017). Sprachgeographie und Geographische Informationssysteme (GIS). In Timo Ahlers, Susanne Oberholzer, Michael Riccabona & Philipp Stoeckle (eds.), *Deutsche Dialekte in Europa. Perspektiven auf Variation, Wandel und Übergänge.* Hildesheim: Olms (Kleine und regionale Sprachen, Bd. 3).

Jeszenszky, Péter, Philipp Stoeckle, Elvira Glaser & Robert Weibel. (2017). Exploring global and local patterns in the correlation of geographic distances and morphosyntactic variation in Swiss German. *Journal of Linguistic Geography 5*(2), 1-23. https://doi.org/10.1017/jlg.2017.5

Jeszenszky, Péter, Philipp Stoeckle, Elvira Glaser & Robert Weibel. (under revision). Crisp breaks vs. continuous transitions : finding quantitative models for transitions between syntactic variants. *Journal of Linguistic Geography.* 6(2) special edition

Jeszenszky, Péter, Curdin Derungs, Philipp Stoeckle, Elvira Glaser & Robert Weibel. Data-driven detection of transition zones in dialectal variables. *Transactions in GIS* (in preparation).

Jeszenszky, Péter, Philipp Stoeckle, Anja Hasse. Dialect Areas. In: R. van Gijn & H. Ruch (Eds.), *Language Contact: Bridging the gap between individual interactionsand areal patterns* (in preparation)

Presentations during the doctoral studies (reverse order)

Jeszenszky, Péter, Curdin Derungs, Robert Weibel. "Automatically Detecting Boundaries and Transition Zones in Swiss German Morpho-syntactic Variation". *Sixteenth International Conference on Methods in Dialectology*, Tachikawa, Japan, 2017.08.07-11.

Jeszenszky, Péter, Philipp Stoeckle, Elvira Glaser, Robert Weibel. "Visualising and Analysing the Impact of Geographic Factors on Linguistic Variation in Dialects", *28th International Cartographic Conference*, Washington DC, USA, 2017.07.02-07.

Jeszenszky, Péter, Philipp Stoeckle, Elvira Glaser, Robert Weibel. "Analysing the Effects of Geographic Factors on Syntax Variation in Individual and Aggregate Phenomena of Swiss German". *Spatial Boundaries and Transitions in Language and Interaction: Perspectives from Linguistics and Geography Conference*, Monte Verità, Ascona, Switzerland,
2017.04.23-28.

Jeszenszky, Péter, Robert Weibel. "Comparing Models of Transitions Between Syntactic Variants in the Dialect Continuum". *Borderland Linguistics Conference*, Bristol, UK,
2016.06.27-28

Jeszenszky, Péter, Robert Weibel. "Modeling Transitions between Syntactic Variants in the Dialect Continuum". *19th AGILE (Association of Geographic Information Laboratories for Europe) International Conference on Geographic Information Science*, Helsinki, Finland,
2016.06.14-17.

Jeszenszky, Péter. "Are There Really Isoglosses? – Possible Methods to Quantify Boundaries of Variants". *ZüKL (Zürcher Kompetenzzentrum Linguistik) Linguistischer Nachmittag – Afternoon of Linguistics* at the University of Zurich, Zurich, Switzerland. Short popular presentation.
2015.11.06.

Stoeckle, Philipp, Péter Jeszenszky, Elvira Glaser, Robert Weibel. "SynMod: Modelling Morphosyntactic Area Formation in Swiss German". *URPP Language and Space site visit,* University of Zurich, Zurich, Switzerland. Poster presentation.
2015.10.27.

Jeszenszky, Péter. „Dialektunterschiede und räumliche Distanz". *Podcast for ZüKL: „angesprochen – der Linguistik-Podcast"*, recorded by Robert Schikowski and Juliane Schröter.
2015.07.25

Jeszenszky, Péter, Robert Weibel. "Measuring boundaries in the dialect continuum – Modeling variation and area formation in linguistic data". *18th AGILE (Association of Geographic Information Laboratories for Europe) International Conference on Geographic Information Science*, Lisbon, Portugal,
2015.06.09-12.

Jeszenszky, Péter. „Wirkungen der geographischen Faktoren auf die schweizerdeutsche syntaktische Variation – Zwei Fallstudien: Reisezeiten für die Vorhersage linguistische Distanzen und Quantifizierung der intralinguistische Grenzen". *3. gemeinsames linguistisches Kolloquium der Universität Kioto und der Gakushuin Universität (3rd Joint Linguistic Colloquium of the Kyoto University and the Gakushuin University)*, Kyoto, Japan,
2015.04.17

Jeszenszky, Péter, Robert Weibel. "Correlating morphosyntactic dialect variation with geographic distance: Local beats global". *GIScience 2014: Eighth International Conference on Geographic Information Science*, Vienna, Austria,
2014.09.25

Jeszenszky, Péter, Philipp Stoeckle, Robert Weibel. "Exploring global and local patterns of the correlation of geographic distances with morpho-syntactic variation in Swiss German dialects". *Methods in Dialectology XV Conference*, Groningen, The Netherlands,
2014.08.12

Jeszenszky, Péter. "Correlating geographic distances with morphosyntactic variation in Swiss German dialects". *8th Days of Swiss Linguistics*, University of Zurich, Zurich, Switzerland,
2014.06.19

Jeszenszky, Péter. "Quantifying effects of geographic factors on morpho-syntactic variation in Swiss German dialects". *Maps and Grammar Workshop*, University of Zurich, Zurich, Switzerland,
2014.03.21

Jeszenszky, Péter. „SADS – Syntaktischer Atlas der Deutschen Schweiz / SynMod – Modellierung morpho-syntaktischer Raumbildung im Schweizerdeutschen". *ZüKL Linguistischer Nachmittag – Afternoon of Linguistics* at the University of Zurich, Zurich, Switzerland. Poster presentation.
2014.02.21

Jeszenszky, Péter, Philipp Stoeckle, Robert Weibel. "Correlating geographic distances with morpho-syntactic variation in Swiss German dialects". *Forum for Germanic Language Studies 11*, Newnham College, Cambridge, UK,
2014.01.09

Jeszenszky, Péter. “Correlation of Geography and Syntactic Distances in Swiss German Dialects”, *Workshop Morphologie und Syntax deutscher Dialekte (Workshop on the Morphology and Syntax of German Dialects)*, Albert-Ludwigs-Universität Freiburg, Freiburg, Germany,
2013.11.08-9.