Modeling transitions between syntactic variants in the dialect continuum

Péter Jeszenszky
University of Zurich /
Department of Geography
Winterthurerstrasse 190
Zürich, Switzerland
peter.jeszenszky@geo.uzh.ch

Robert Weibel
University of Zurich /
Department of Geography
Winterthurerstrasse 190
Zürich, Switzerland
robert.weibel@geo.uzh.ch

Abstract

Although linguists have thoroughly studied the formation of language areas for given dialectal phenomena, little quantitative research has been conducted on how these areas relate to each other, and how the transition between these dominance areas of dialectal variants can be modelled. We propose gradient estimation methods used in GIScience to answer the key question to the analysis of such dialectal boundaries: to what extent we can find crisp boundaries in a dialectal landscape (termed 'isoglosses' in linguistics) or whether the transitions are rather gradual. Our methods are also aimed at the comparison of these boundaries. We apply trend surface analysis and regression analysis to Swiss German syntax data and test our methods on dialect phenomena with typical variation exhibiting the spatial competition of two variants. We conclude that subdividing the linguistic landscape of the given competing variants into three subregions (two dominance zones for each of the two variants and a transition zone between them) and calculating regression models in these subregions lets us quantitatively compare their relationships to each other and to other linguistic phenomena.

Keywords: dialectology; linguistic data; gradients; trend analysis; regression analysis

1 GIScience to Analyse Linguistic Data

Languages and dialects manifest in geographic space, and geographic factors are, among others, major explanatory variables in the formation of language areas [2]. However, from the perspective of GIScience, issues of linguistic research have not found a lot of attention so far, although the peculiarities inherent to linguistic data would provide many challenging problems. GIScience has a great number of methods that could make potentially valuable contributions to linguistic research.

Extending previous exploratory studies [5] we aim to contribute to modelling spatial variation in linguistic data using methods of GIScience. We will approach spatial variation in dialect phenomena by the notion of *gradients* to model transitions between dialect variants, using trend surface analysis and regression analysis in more dimensions of the data. We aim to account for the nature of boundaries often conceptualized by linguists in dialect landscapes and to quantify crisp and gradual spatial change present in linguistic phenomena. Although dialectologists have thoroughly studied the formation of dialect areas, quantitative modelling of the transitions between dominant dialect variants has not been undertaken before other than by qualitatively describing maps resulting from surveys.

The two key paradigms of dialectology to conceptualize dialect-internal boundaries are the *isogloss* and the *dialect continuum* [4], corresponding to the dichotomy of entities and fields, respectively, in GIScience.

An *isogloss* is a theoretical line drawn based on linguistic surveys, where the occurrence areas of the variants corresponding to a linguistic phenomenon are expected to be separated by a crisp boundary. In reality, however, single linguistic phenomena do rarely display this type of clear-cut regional pattern that are often claimed in traditional dialect classification studies.

On the other hand "modern dialectology recognizes that geographic distributions may involve continua" [11]. This implies that while dialect areas cannot be crisply delimited, also for single phenomena gradual transitions ought to be expected between areas of dominance of variants.

1.1 Data

Our database is the Syntactic Atlas of German-speaking Switzerland (SADS; [1]). Between 2000 and 2002 close to 3,200 respondents participated in a series of four surveys in 383 survey sites (i.e. one quarter of the German speaking Swiss municipalities), responding to questions about syntactic phenomena (survey sites visible in Figure 2). Among linguistic surveys, SADS is particular as it has multiple respondents per survey site. To capture the local linguistic diversity, at each survey site 3 to 26 respondents (median: 7) were involved in the survey. This wealth of data allows us to assess the usage variation of surveyed dialect variants.

1.2 Spatial variation characteristics

Patterns of spatial distribution in our data are very diverse, thus devising quantitative models for them is not easy. We aim to quantify the spatial change from the area of dominant usage of one variant – termed a *dominance zone* here – towards the dominant usage of another variant for the given phenomenon (usually corresponding to one survey question), that is, another dominance zone. These dominance zones are

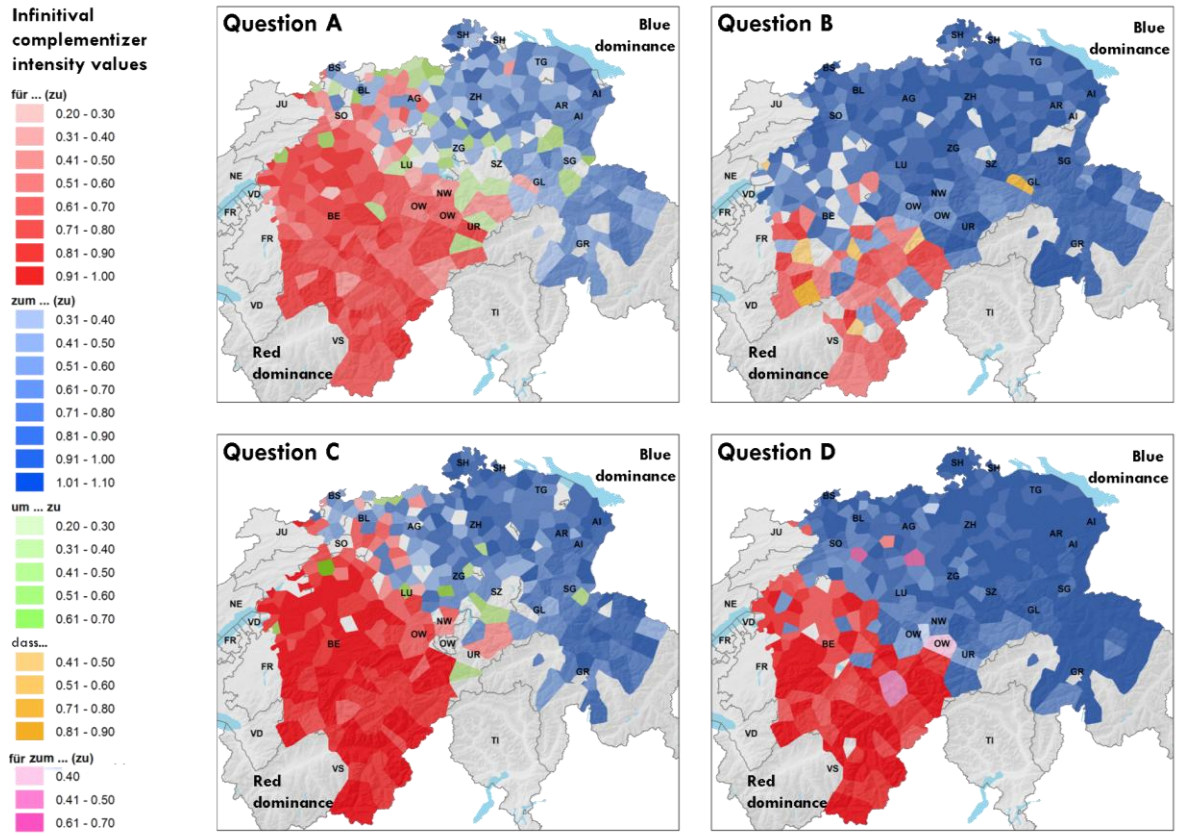
presumably separated by a zone where the usage of the two variants usage is mixed to some degree — the *transition zone*.

Linguists have described three prime types of spatial variation of the syntactic questions surveyed in the SADS [3].

1. Two spatially more autocorrelated main variants are competing with each other, their (dominant) usage zones crisply or more smoothly transitioning into each other in clearer patterns.
2. A more frequent variant occurs across most of the Swiss-German area, with regional variants in smaller areas. Most survey sites show at least two variants.
3. Highly variable pattern, with seemingly no spatial or only local patterns of dominant variants discernible.

Given these different types of variation it seems obvious that the relationship between the spatial distributions of given variants cannot be described in a satisfying way using one method for every phenomenon. Furthermore, transitions may not be uniform, and patterns of change may be different in every direction. The areal structures of dominant variants have been studied by e.g. Rumpf et al. [7] who developed three measures that characterise the spatial distribution of dominant variants focusing on homogeneity, border length and area complexity (of dominance zones).

Figure 1: Intensities of variants mapped. Colour hue of Voronoi-polygons indicates which variant received the majority among the respondents (if any), while lightness corresponds to the intensity, i.e. the proportion of respondents using the given variant.



1.3 Case study

The linguistic phenomenon presented in this article belongs to the first spatial distribution type. The so-called "*infinitival complementizer"* is covered by four survey questions in the SADS (*Questions A-D*). The answers to the four survey questions feature two main competing dominant dialectal variants '*für*' and '*zum*' (shown *red* and *blue* in Fig. 1, respectively) and also some minor variants, including the standard German variant '*um…zu*'. The *intensity* (the proportion of the respondents preferring the given variant) of these variants for each question regarding the phenomenon is shown in Figure 1. In general all four questions feature dominance of the red variant in the southwest, while the blue variant is dominant in the northeast, with the transition between these main variants occurring in different regions of the study area. The standard German variant (*green*) and others are spatially distributed more or less randomly.

In the following, we will use the terms 'red variant' and 'blue variant', to avoid confusion and as the focus is on methods for modelling dialectal variation rather than on linguistic interpretation.

In Figure 1 the colour hue of the polygons indicates that at the given survey site a variant reaches *relative majority* (i.e. is *dominant*), while the colour lightness shows the *intensity* of the variant. We see that both the red and the blue variants have an area where their hue is mostly dark and they are not very mixed with other variants. These vaguely defined areas denote the *dominance zones*, and the zones in between them, where more mixing is seen, are referred to as *transition zones*.

2 Patterns and Scales

As part of preliminary analysis of the data, spatial distributions of variants were mapped and consulting linguistic theories several types of relationships between the linguistic variants were conceived. We found crisp boundaries, wide gradual transitions and survey questions that share characteristics on different scales.

2.1 Global vs. local scale

As 'crispness' of a boundary is a matter of definition and highly dependent on scale, we have to consider different scales to investigate transitions between variants. At larger scales, it might often be appropriate to say that there is a crisp boundary between two dominance zones. According to former studies in Germany, for example, we might find that the *Appel/Apfel* variation (English: apple) indeed produces a sharp linear boundary, where deviating survey sites are present only up to about 30 km from the alleged isogloss [8]. On the *global* scale this could indeed count as a crisp boundary. Transferred to Switzerland, which is considerably smaller, a difference of 30 km would not count as a very sharp boundary on a global scale. (Throughout the paper the term *global scale* will refer to *all* survey sites concerned.)

The spatial distribution patterns mentioned before are recurring at all scales, featuring local maxima and sudden drops in between dominance peaks of variants, be it a small area or the whole area of investigation. As our general goal is to compare different linguistic phenomena based on characteristic transition patterns, we aim to find the appropriate scaling and fit models to best account for the transitions – i.e. gradients – between dominance zones of variants.

Comparing different strategies, we will argue that subdividing a variant's spatial intensity distribution into several spatial subsets and calculating regression models in these subsets is more meaningful for the quantification and comparison of competing variants than modelling the dialect landscape only on a global scale.

2.2 Linguistic theories and hypotheses

Different ideas concerning the nature of transitions between variants have been developed in linguistics. Their validity for different phenomena may be tested using quantitative models for transitions.

Based on the *isogloss* paradigm [2], a crisp boundary between the dominant usage areas of the competing variants would be expected. Splitting the landscape into two parts at this boundary, linear trend surfaces fitted to the respective subsets would be expected to be almost level, with maximum intensity values on one side, and minimum values on the other side.

The hypothesis of *inclined planes* that Seiler [9] has suggested (for the phenomenon used as an example in this paper) posits that the transition between the variants should be gradual and continuous. It assumes constant declination of one variant with the increase of the competing one. This could be best modelled by two planar, first-order trend surfaces (one for each variant) having maximum intensity in one end of the investigation area and reaching zero at the other end.

Contrasting these two theories with our concrete example's intensity data, we hypothesise that fitting a first-order trend surface to the global intensity values of a given variant would result in a steeper gradient and greater residuals than in a bipartite subdivision split at the assumed isogloss. As a best fit on the global scale we expect a third-order trend surface, whereas in linear cross-sections cut through the intensity surface, logistic regression models are assumed to fit best.

Based on preliminary analysis we expect that the transition is not continuous from one end of the investigation area towards the other, it rather occurs in a specific zone. However, if we subdivided the dialect landscape of the red and blue variant into two dominance zones and a transition zone between them, we expect linear regression models to fit quite well in the subsets, with a markedly steeper gradient in the transition zone. The gradient depends, however, a lot on how we define this transition zone, as it will affect its extent.

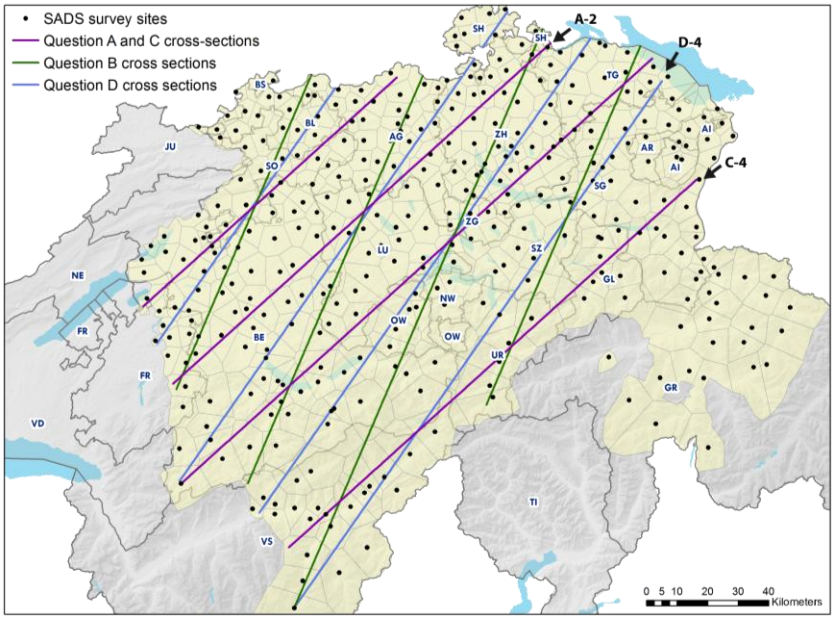
3 Methods

To assess whether the intensity values of the surveyed variants correspond to the theoretical models sketched above we used trend surface analysis (in the whole investigation area) and regression analysis in cross-sections cut through this landscape, respectively, both at a global scale (all data points) as well as in spatial subsets (i.e. subdivisions).

Trend surfaces — first-order (plane), third-order and logistic — were fitted by least-squares to the intensity values of the main variants (red and blue) at all survey sites for each of the four survey questions.

To expose the underlying variation of non-dominant variants that is concealed by the dominant variants shown on the intensity maps (Figure 1) and to account for the diverse patterns of transition at different places, we constructed *cross-sections* through the dialect landscapes, similarly to [6]. For the survey questions A-D, cross-sections were taken at four different positions to sample the entire study area. Each cross-section follows the direction of the bisector of the aspect angles belonging to the planar trend surfaces of the main variants. The cross-sections used for this study are shown in Figure 2. For calculations their lengths were normalized so that the steepness of their gradients would be comparable across variants and survey questions. Once constructed, the cross-sections served for both visual analysis and for regression analysis (linear, third-order, logistic).

Figure 2: Cross-sections taken through the variant landscapes of the different questions. Note that the colours do (deliberately) not correspond to the colours assigned to the linguistic variants in Figure 1.



Corresponding to the linguistic hypotheses outlined in Section 2.2 trend surfaces and cross-sections were further analysed at the *local* scale, using three *subdivision* strategies:

- *α*) a bipartite subdivision based on Sibler's optimized KDE smoothing [10] of the same phenomenon, similar to a tessellation by alleged isoglosses;
- *β*) a tripartite subdivision based on defining a transition zone where intensities of both main variants fall below 62,5%, the remainder forming two dominance zones;
- *γ*) a tripartite subdivision where the transition zone is defined between (and including) survey sites where the intensity of the given variant starts decreasing radically as we proceed along the line.

Lacking a clear definition in the linguistic literature, transition zones were assigned in the above two exploratory ways, based on prior visual analysis of the intensity maps and the cross-sections.

In the subdivisions only linear regression was conducted, since performing higher-order regression on such a small number of points (16-35 points) would cause overfitting.

For all trend surfaces *slope (gradient)*, *aspect* and R^2 associated with the goodness of fit were calculated. Similarly, for every linear regression model in cross-sections we calculated the *slope* of the function, a *p-value* associated with its significance, and R^2.

4 Results

Preliminary remarks. Figures 1, 2 and 3 are interrelated, with A-D referring to the investigated survey questions, numbers 1-4 to the cross-sections (the numbering starts from the NW) and red and blue denoting the colours used for the representation of the two main answer variants of the questions.

Trend surfaces. Table 1 presents the results of trend surface fitting. Planar trend surfaces, as expected, yield slope angles of the red and blue variants quite similar to each other, as at both ends of the study area intensity values are minimal and maximal, respectively. Also, the aspect angles of the red and blue variants are almost perfectly opposite to each other. In all cases, the third-order trend surfaces show a better fit than the planar trend surfaces.

In the *α*-subdivision we see gradients flatter than on the global scale but not everywhere as flat as suggested by the isogloss theory (visible also in Figure 3). We discover more diversity in the slope angles in the transition zones of the *β*- and *γ*-subdivisions than among the trend surfaces fitted on the global scale. These slopes in transition zones are also steeper although not as symmetric concerning red and blue as expected. Steeper slopes mean a greater portion of gradual change is caught in the transition zone. Also, the differences between the slopes found for *β* and *γ* show how small differences in the definition of transition zones may influence our models.

Cross-sections. The results of the regression analysis in the cross-sections are presented in Table 2 and Figure 3. Three examples are used here as they were, by visual inspection, classified as a crisp transition (D-4), a typical gradual transition (A-2) and a more varied pattern (C-4).

On the global scale (i.e. taking all the points along the given cross-section) the best fitting model based on R^2 is either the third-order polynomial or the logistic model, depending on the intensity change along the line. From the slope values in Table 2 (and from Figure 3) it is visible that the isogloss model is appropriate for D-4 but the other two cross-sections show more gradual transitions as slopes in the *α*-subdivision are steeper. Their transition is comparable to one another as they both have steeper slopes in the transition zones of the *β*-

and γ-subdivisions, but their most significant regression lines are still found in different subdivisions.

Figure 3 shows the three cross-sections with different regression models. A 3rd order polynomial, a linear and a logistic model (the latter two not shown) were fitted on the global scale; linear models were used in the subdivisions.

Opposing the isogloss theory which presumes that spatial transition is abrupt and the inclined planes theory, by subdivision of the variant area generally we find a zone of varying size where most of the decrease in the intensity occurs. These transition zones are characterized by slopes of the regression lines steeper than in the respective dominance zones and linear regression slopes on the global scale (in Figure 3 only a portion of the investigated cross-sections is shown).

5 Conclusions

In this paper, we have used different forms of regression — trend surface analysis as well as linear regression along cross-sections — to model transitions between dialect variants that can be conceived as changes in *gradient*. Of the three types of spatial variation described in our dialect data by [3], we have focused on the one characterised by two dominant variants and an intermediate transition zone. For this type of variation, we have shown how regression analysis can be used to describe transitions between dialect variants quantitatively, and how this can be used to test linguistic theories that have explained patterns of dialect variation in qualitative and visual terms. We have furthermore shown that subdividing the study area into different zones — dominance and transition zones, in this case — and fitting regression models separately for these subsets, is preferable to a global approach. However, the global approach may still be warranted, for instance, to get an overview of variation patterns, among others, in the residuals of the regression surfaces or lines.

In future work, we seek to address the issue of transition zones in more depth. As their definition affects the sheer presence and size of transition zones, it is crucial to test multiple different definitions. We will also further explore the homogeneity and robustness of dominance zones [5]. Finally, while regression analysis worked well for the analysis of example dialect phenomena with gradual spatial variation, other methods of spatial analysis and statistics will have to be explored to deal with the other types of dialectal variation defined in [3].

References

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

[2] J. K. Chambers & P. Trudgill,. *Dialectology (2nd ed.)*. Cambridge: Cambridge University Press. 1998.

[3] E. Glaser & G. Bart. Dialektsyntax des Schweizerdeutschen. In R. Kehrein, A. Lameli, & S. Rabanus (Eds.), *Regionale Variation des Deutschen. Projekte und Perspektiven.* (pp. 79–105). Berlin: De Gruyter, 2015.

[4] W. Heeringa & J. Nerbonne. Dialect Areas and Dialect Continua. *Language Variation and Change*, 13(03), pages 375–400. 2001.

[5] P. Jeszenszky & R. 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 The 18th AGILE International Conference on Geographic Information Science Geographic Information Science as an Enabler of Smarter Cities and Communities (p. 5). Lisbon, 2015.

[6] S. M. Pröll, S. Pickl, A. Spettl, V. Schmidt, E. Spodarev, U. Stephan, W. König. Neue Dialektometrie mit Methoden der stochastischen Bildanalyse. In R. Kehrein, A. Lameli, & S. Rabanus (Eds.), *Regionale Variation des Deutschen. Projekte und Perspektiven* (pp. 169–190). Berlin, New York: Gruyter, 2015.

[7] J. Rumpf, S. Pickl, S. Elspaß, W. König, & V. Schmidt. Structural analysis of dialect maps using methods from spatial statistics. *Zeitschrift für Dialektologie und Linguistik*, 76(3), 280–308, 2009.

[8] J. E. Schmidt. Formation of and change in regiolects and (regional) dialects in German. *Taal En Tongval*, *63*(1), 143–173, 2011.

[9] G. Seiler. Wie verlaufen syntaktische Isoglossen , und welche Konsequenzen sind daraus zu ziehen? In E. Eggers, J. E. Schmidt, & D. Stellmacher (Eds.), *Moderne Dialekte – Neue Dialektologie* (pp. 313–341). Stuttgart, 2005.

[10] P. Sibler, R. Weibel, E. Glaser, & G. Bart. Cartographic Visualization in Support of Dialectology. In *Proceedings - AutoCarto 2012* - Columbus, Ohio, USA, (p. 18), 2012.

[11] M. Wieling & J. Nerbonne. Advances in Dialectometry. *Annual Review of Linguistics*, *1*(1), 243 – 264., 2015.

Figure 3: Regression lines along selected cross-sections and within their subsets. 3rd order polynomial on the global scale and linear regression lines in subdivisions shown.

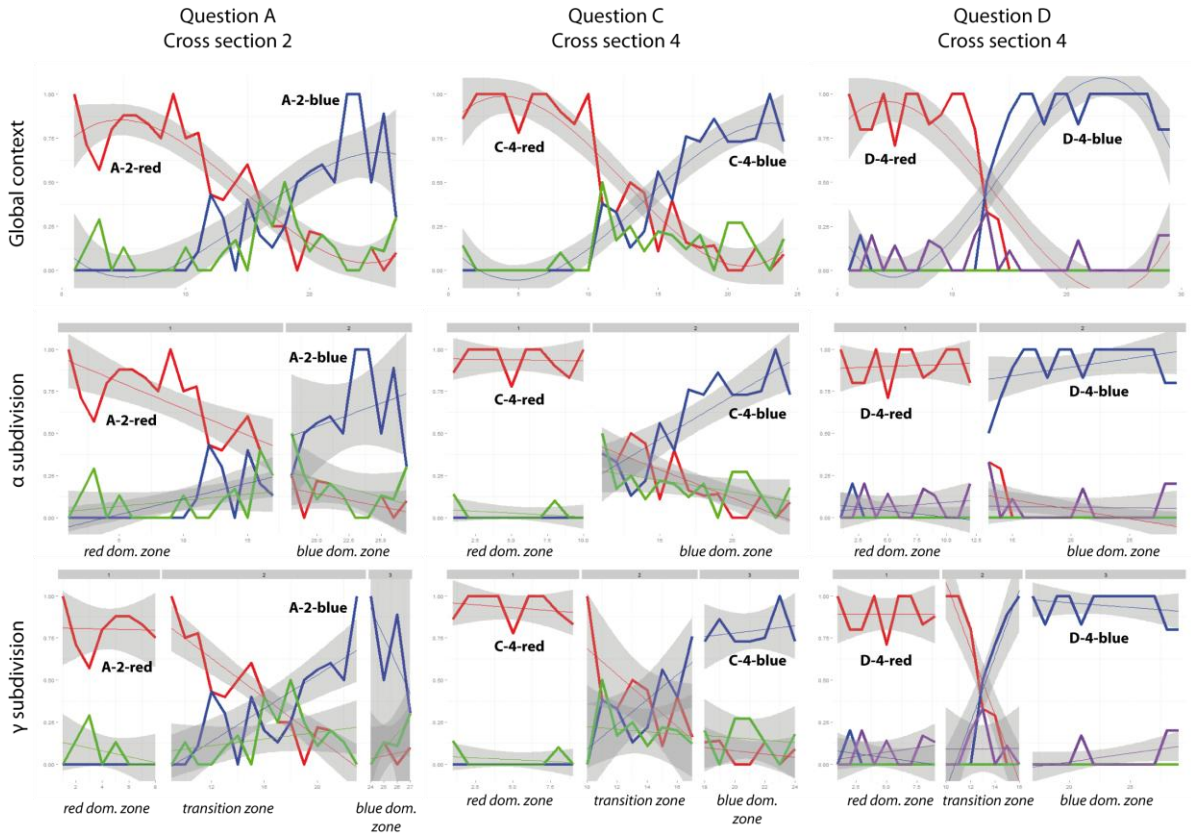


Table 1: Results of trend surface fitting

TREND SURFACE	**A-red**	**A-blue**	**B-red**	**B-blue**	**C-red**	**C-blue**	**D-red**	**D-blue**
Aspect angle	47.75°	231.03°	23.78°	203.17°	46.40°	229.49°	33.85°	216.48°
Divergence from opposition	3.28°		0.608°		3.09°		2.62°	
$\mathbf{R^2}$ - planar trend surface	0.7139	0.6829	0.5543	0.4609	0.7813	0.7562	0.7285	0.7400
$\mathbf{R^2}$ - 3rd order trend surface	0.7648	0.6897	0.6813	0.5599	0.8433	0.8075	0.8309	0.8077
Slope – global context	3.01E-4	2.66E-4	1.72E-4	1.68 E-4	3.61 E-4	3.31 E-4	3.48 E-4	3.35 E-4
Slope in α-subdivision								
red dominance zone	2.65 E-4	1.18 E-4	1.57 E-4	1.38 E-4	2.66 E-4	1.63 E-4	2.18 E-4	2.04 E-4
blue dominance zone	1.07 E-4	2.27 E-4	1.16 E-4	1.53 E-4	1.68 E-4	2.61 E-4	9.51 E-5	1.55 E-4
Slope in threefold subdivisions								
Slope in the transition zone of β-subdivision	3.60 E-4	2.50 E-4	2.10 E-4	2.42 E-4	5.28 E-4	4.73 E-4	9.53 E-4	5.50 E-4
Slope in the transition zone of γ-subdivision	4.73 E-4	3.95 E-4	2.10 E-4	2.43 E-4	5.79 E-4	5.31 E-4	7.40 E-4	6.99 E-4

Slope values *per se* are not meaningful but they can be compared to one another. However, they are not comparable to slope values in Table 2.

Table 2: Results of different types of regression analysis in selected cross-sections

VARIABLE IN THE GIVEN CROSS-SECTION	**A-2-red**	**A-2-blue**	**C-4-red**	**C-4-blue**	**D-4-red**	**D-4-blue**
3rd order regression line – global context	***	***	***	***	***	***
$\mathbf{R^2}$ (variance in intensity explained by the geogr. distance (%))	0.8777	0.7204	0.9068	0.9049	0.8947	0.9046
Logistic regression	***	***	***	***	***	***
$\mathbf{R^2}$	0.9759	0.8690	0.8927	0.9273	0.8633	0.8141
Linear regression						
Slope – global context	-0.4824***	0.4142***	-0.5417***	0.4813***	-0.5935***	0.5949***
Slope in α-subdivision – red dominance zone	-0.4085**	0.2422**	-0.0175 (NS)	0.0000	0.0299 (NS)	-0.0939 (NS)
Slope in α-subdivision – blue dominance zone	-0.1788 (NS)	0.3428 (NS)	-0.3830***	0.5892***	-0.1890**	0.1858 .
Slope in the transition zone – β-subdivision	-0.7707**	0.3741 (NS)	-0.1723 (NS)	0.2305 (NS)	no such zone	no such zone
Slope in the transition zone – γ-subdivision	-0.7148***	0.6058***	-0.6715 .	0.6275*	-2.1897***	2.1700**
$\mathbf{R^2}$ global context	0.8173	0.6644	0.8711	0.8511	0.8032	0.7795
$\mathbf{R^2}$ in α-subdivision – red dominance zone	0.5064	0.3956	0.0029	NA	0.0063	0.1626
$\mathbf{R^2}$ in α-subdivision – blue dominance zone	0.1903	0.0957	0.6514	0.6943	0.4067	0.2055
$\mathbf{R^2}$ in the transition zone of β-subdivision	0.6942	0.2340	0.0029	NA	no such zone	no such zone
$\mathbf{R^2}$ in the transition zone of γ-subdivision	0.8157	0.6404	0.4863	0.5434	0.9113	0.8864

Significance p-value: 0 < *** < 0.001 < ** < 0.01 < * < 0.05 < . < 0.1, NS – not significant. The slope values *per se* are not meaningful, because the intensity is basically a ratio value between 0 and 1, while the predictor variable is in km. Nevertheless it is meaningful to proportionally compare the different variants and phenomena to each other.