Structural analysis of dialect maps using methods from spatial statistics[*]

Jonas Rumpf, Simon Pickl, Stephan Elspaß, Werner König, Volker Schmidt

1 Introduction and Motivation

In 1898, Karl Haag (Haag 1898) introduced a new method of displaying dialect differences on a map. In the maps that he drew, the thickness of a line between two places indicates how different the dialects spoken on both sides of the line are, thus summing up the data of several dialect feature maps, each of which provides the realisations of one linguistic feature in space. This was the first step towards a quantitative dialectology, going along with other linguistic disciplines such as lexicography, phonetics, or historical linguistics (cf. Köhler et al. 2005, Best 2006). The dialectometrical approaches developed since then have put forth a wide range of methods, all of which are based on the measuring of dialect distance (the grade to which the dialects in two places differ), which is essential for all further dialectometrical investigation. Up to the 1970s, this was done exclusively by counting the dialectal differences between all neighbouring places, a method that Jean Séguy adapted from Haag for the *Atlas linquistique et ethnographique de la Gascogne* (Séguy 1965–1973), the first major dialectometrical project. In the 1970s, Hans Goebl was the first scholar to extend this method – the counting of differences – to all possible pairs of places, not only neighbours, and developed a broad spectrum of advanced methods of visual presentation and of analysis, such as different kinds of cluster analysis or coherence tests (see, for example, Goebl 1994, 2001, 2006, 2007). The subsequent dialectometrical approaches are based on Goebl's method (Kelle 1986, Hummel 1993, Schiltz 1996, and others). The next major step was made by John Nerbonne and Wilbert Heeringa, who introduced a new way to measure the dialectal distance into Goeblian dialectometry: they used an adapted version of the so-called Levenshtein-distance to measure phonetic distance and were thus the first to take into account gradual similarities between records, not only identity or non-identity, and they also greatly contributed new analysis methods such as multi-dimensional scaling or factor analysis (e.g. Nerbonne and Heeringa 1998, Heeringa 2004, Nerbonne 2006).

There are various other methods which are dialectometrical in the sense of 'quantitative

[*]The methods presented in this article are a first result of the DFG research project "New Dialectometry using Methods from Stochastic Image Analysis" ("Neue Dialektometrie mit Methoden der stochastischen Bildanalyse", http://www.uni-ulm.de/en/mawi/mawi-stochastik/forschung/projekte/sprachatlas.html) of the Chair of German Linguistics at the University of Augsburg and the Institute of Stochastics at Ulm University. The authors would like to thank Stefanie Eckel, Michael Böhm, and Matthias Kugler for their help in developing, implementing, and filling the database that was used for the maintenance of the data analysed in this study.

dialectology', e.g. techniques to measure the dialectality of a variety, i.e. its linguistic distance from a standard variety (a brief overview can be found in Schmitt 1992, 61–91).

The toolbox put together by dialectometricians over more than a century allows for substantiated and objective statements about the shapes and characteristics of dialect areas or the relations between linguistic landscapes. It provides a useful and reliable set of techniques that is now an important means of gaining information about dialects in space.

However, classical dialectometry is still restricted to the global question of the dialectal division of the areas under investigation. As all feature maps of the respective corpus are put together to contribute to the analysis of the spatial distributions of dialects, not of features, the structural characteristics of the single feature maps are not taken into account. When browsing through a dialect atlas, even a cursory glance at the different maps reveals that their distribution patterns are not just images of the partitioning obtained by global dialectometrical analysis, slightly distorted by stochastic noise, but that they show quite individual and highly variated patterns. The variants' spatial distributions can change so dramatically from map to map that this cannot be attributed to mere random fluctuation.

On some maps, the distributions of the variants form small, compact and hardly intermingled areas, on others, there are greater, overlapping areas, and there are maps on which we can make out nothing but utter chaos. Some geographically defined characteristics such as mountain ridges or rivers parallel borders between linguistic features, some geographical features seldom or never correlate with linguistic boundaries.[1] Several dialectologists have investigated and classified the highly diverse structures and patterns of such maps (e.g. Wenzel 1930, 107ff., Frings 1956, Bach 1969, 39–226, Hildebrandt 1983), but as yet, there is no quantitative approach that would allow for a systematic examination of the individual maps.

The structural characteristics of a linguistic feature map are a result of the way in which its geographical distribution developed. Several factors that could play a role in this development are imaginable: frequency, semantic field or geographical conditions are only a few. The way a map looks must in some way correlate or at least be somehow related to these variables. In a nutshell: there must be reasons why maps look so different. A quantitative analysis of a corpus of feature maps should help us answer the question of which variables play a role in the constitution of the maps, and how they affect the spread of variants.

[1]We refrain from using the term 'isogloss', as it is inaccurate for several reasons (cf. Händler and Wiegand 1982, Schneider 1988, 177ff.), and rather prefer the less problematic expression 'boundary'.

If we are to investigate what the mechanisms are that determine the geographical distributions of variants, we must first find a way to describe and quantify the structural characteristics of the maps. Concepts like “complexity” or “homogeneity” can be of help. As a first step, they must be defined and scaled. Next, a way has to be found in which to measure these characteristics, which has to be as objective as possible and in any event reproducible. If a large amount of maps has been analysed with regard to these characteristics, it is possible to try and correlate the values obtained with properties of the linguistic features, thus gaining insight into the mechanisms that are at work in language change in space.

2 Approach

The method presented in this article and its underlying theoretical approach are based on some general assumptions about the data in dialect atlases and the linguistic facts they represent. As the atlas data consist of the answers that selected persons have given at selected places to selected questions, it is clear that they reflect only a snippet of the reality. Therefore, the data obtained in the surveys do not faithfully reflect the "real" situation at the record locations. Even if we concede that the records are more than regular statistical samples, since an informant can speak for more persons than just for him- or herself,[2] we have to acknowledge the fact that the data are subject to a certain amount of random fluctuation. An approximation of the "real situation" can, however, be obtained by using statistical methods, which will turn out to be very useful for the calculations of the structural characteristics of the maps.

The estimation of the probability with which a certain answer should occur at a location, based on what the actual records reveal, helps us assign values to that location which assess the geographical distribution of variants in its environments. A look at the neighbourhood of each location gives us a clue to the validity of the record given there. If, for instance, a single outlier appears in an otherwise uniform area, it is very likely that, had another person at the same location been asked, he or she would have given the variant of the surrounding locations. At the same time, it is to be expected that, if enough persons had been asked in the surrounding area, a few of them would have come up with the variant of the outlier. This is a very simple example to illustrate our basic assumption: the more records of a certain variant appear in a location's environment, the more likely it is that this variant would have been given as an answer at the location itself if many people had been asked, even if the informant that actually was interviewed has given a different one. Abstractly speaking, the records of the neighbours of each record location are taken into account when assessing the location in question, their influence however declining with growing distance. Consequently, if sufficient differing variants around one location are given, they can "overpower" the actual record. On the other hand, this also weakens their "power" at their own locations, because the outlier has some minor effect on them, too. This expedient is legitimate, because the contact between speakers does not only occur within villages and towns, but also between them. The higher the geographical distance between two locations is, the lower is the degree of language contact to be expected.[3]

[2]Cf. "Der Informant als Experte" (SBS 1997–2005, vol. 1, 20f.)

[3]This is, of course, an over-simplification, because geographical proximity is not a direct indicator for language contact. It has indeed been one of the major goals of classical dialectology to find out what promotes language contact and what hinders it. As this question can not yet be regarded as resolved, at least not in quantitative terms, we have no reliable basis on which to estimate language contact. Therefore, we will preliminarily have to make do with geographical distance as a very rough, although not entirely unfit indicator for the expected degree of language contact between two locations.

The result of this method can be seen as an estimated percentage for each variant at each record location, specifying how likely it would be that a variant would be given as an answer at that location if all possible informants were asked, bearing in mind that the actual records are nothing but statistical samples. This estimation holds, of course, only for the part of the population that has been selected as eligible informants. The calculations for the estimation are based exclusively on the records given in the actual data (a reasonable amount of interpretation granted, as they have to be classified into variants) and the geographical coordinates of the locations. For the exact mathematical procedure, see Section 4.

How should the values obtained be interpreted? If the percentage for a certain variant at a location is fairly small, we can still assume that this variant is at least part of the passive vocabulary of a handful of possible informants. If no variant shows clear prevalence over the others, they are to be seen in a state of "competition", presumed that they are semantically equivalent.[4] Consequently, the percentages that are obtained for each location and each variant serve a dual purpose: firstly, they reflect better what the "real situation", i.e. the probability to get a certain answer at a location, can be presumed to be, and secondly, they allow us to assess the structural characteristics of a map locally. A location in which one variant dominates clearly over all the others is situated in a relatively homogeneous area, whereas a location in which no single variant shows a clear prevalence over the others lies in a rather intermingled, heterogeneous part of the map. These values can then, on the whole, give us an idea of the overall homogeneity of the map. The values also show us which variant is the prevalent one in a certain region, even if the region itself is scattered with three or four different variants, allowing us to divide the maps into prevalence areas – thus generating area-class maps automatically. They serve as a means for further analysis and provide – as a by-product – an objective way for the division of point-symbol maps into area-class maps.[5]

[4]Such a state of competition between two or more variants in a language community is usually seen as an indicator for ongoing language change: one variant is the older, the other the newer one. Eventually, one will prevail, either because the newer one supersedes the older one, or because the older one regains lost ground (cf. Haas 1978, 34–80, König 1982, 464). A state of stability is only given if one variant clearly dominates and has no serious competitors. A quantitative approach to language change as gradual replacement of older variants by newer ones is known as the so-called Piotrovskij-law (cf. Altmann 1983, Piotrowski et al. 1985, 81–100, Leopold 2005, Best 2006, 106–123).

[5]There is no question that the area-class maps generated with this method are simplified visualizations of geographical dialect data. Clearly, there are more elaborate and detailed ways of displaying linguistic data on a map. Our area-class maps serve first and foremost the assessment of the results of our method, and are not meant for a classical qualitative interpretation of a dialect map, which often deals with small scale variation. They are, however, reproducibly generated and provide a very accessible visualisation of medium scale variation in the SBS and atlases of similar size, illustrating the degree of variability in all parts of the map.

3 Data

All examples and results provided in this article have been gained from data from the *Sprachatlas von Bayerisch-Schwaben* (SBS 1997–2009), which was compiled under the direction of Werner König at the University of Augsburg in the years 1984–2005. It is the most comprehensive dialect atlas within the German-speaking area, comprising 14 volumes with altogether approx. 2,700 maps. The investigation area extends 90 km west to east and 150 km north to south, encompassing 272 record locations. Each item from the questionnaire was typically answered by one informant per location. Additional informants were only called in if the first informant proved insufficient. As this large amount of data is also available electronically, the SBS is a perfect candidate for the trial of new dialectometrical methods.

For each item of the questionnaire, the database of the SBS assigns to each locality one or more entries (depending on how many different answers have been given by the informants), consisting of a transcription of the actual record and a code for the symbol that has been used in the mapping process. This code is what we use primarily for the identification of variants, because it provides a pre-classification of the records following established linguistic principles, making it easy to decide whether two slightly differently pronounced forms belong to the same lexical item. It must be noted that the number of entries is not determined by the number of informants at a location, but by the number of different variants. The database does not yield any information about how many informants have contributed a certain variant, so that, for lack of other information, they have to be considered equally valid.

For the preliminary study presented in this article, we confine ourselves to lexical maps, as phonetic and morphological maps will pose additional problems that have yet to be dealt with, for instance, ordinal scale data. The corpus encompasses 823 maps, which is a major part of the word-geographical maps of the SBS.

4 Mathematical Methodology

4.1 Intensity Estimation and Area-Class Maps

The first practical step in our approach to the structural analysis of dialect maps is the automated generation of area-class maps on the basis of the raw data. Mostly, this raw data is charted in point-symbol maps, as is the case in the SBS. Since the data basically consist of specific symbols at specified points in the plane, it can be seen as a set of point patterns. Thus, it appears appropriate to employ methods from point process statistics. Such methods are well established and frequently used in many fields of science (see, for example, Baddeley et al. 2006, Diggle 2003, Illian et al. 2008, Stoyan and Stoyan 1994).

The usual purpose of area-class maps, as explained above, is to provide the viewer with a division into areas that stand for the occurrence of certain variants, but here, they also help to prepare the data for further analysis. Therefore, as a first step, a point-symbol map is "separated" into as many so-called variant-occurrence maps as there are distinct symbols (representing distinct variants) on the point-symbol map. In each of the variant-occurrence maps, only one variant will occur, wherever it was originally recorded. At those locations where the respective variant does not occur in the point-symbol map, there will be no variant at all in the variant-occurrence map. In this way, a point pattern of occurrence points for this variant is obtained. Since it is possible that the point-symbol map features multiple symbols (i.e. variants) at a single location, each record location t is assigned an occurrence value $l_x(t)$ of the variant x: if x is the only variant occurring in the point-symbol map at location t, it is assigned the value $l_x(t) = 1$. Otherwise, it is assigned its relative frequency of occurrence, e.g. if x is one of three different variants occurring at location t, $l_x(t) = \frac{1}{3}$. Also, to simplify notation, all points of measurement t from the original data where x does not occur at all are assigned $l_x(t) = 0$ in the variant-occurrence map for x, which from equation (1) below can easily be seen to be equivalent to disregarding t completely when creating the variant-occurrence map.

Once a set of variant-occurrence maps is obtained from the raw data, a continuous intensity field is estimated for each variant-occurrence map. Informally speaking, at every location on the map, a variant's intensity field indicates the likelihood of that variant occurring at the respective location. This information is obtained by means of a two-dimensional so-called kernel estimator. For detailed discussions of all aspects of kernel estimation see, for example, Scott (1992) and Silverman (1986).

Here, we will only give an overview of the specific tools used for our investigations. The kernel estimate $u_x(t_i)$ of the intensity of a variant x at a location t_i is defined as

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

In this formula, $t_1, \ldots, t_n$ denote the n points of measurement, and $d(t_i,t_j)$ indicates the geographical distance between the locations t_i and t_j. The parameter $h > 0$ is called the bandwidth of the kernel estimator, while the monotonously decreasing function K is called the kernel. Basically, a kernel estimator works by assigning a certain "probability mass" to each location where the variant was recorded, and then spreading out this mass in a way that will give less mass to areas further away from that location. The shape of the kernel function determines the way in which this mass is spread out, while the bandwidth h can be interpreted as the scale. By adding up the mass created by all points of measurement at a certain location, the intensity estimate at that location is obtained. In our approach, the mass created at a location t of occurrence of variant x is additionally weighted with the variant occurrence value $l_x(t)$. Consequently, an occurrence of x in combination with one differing variant at one location will create only half the mass of a sole occurrence of only x.

Notice that the factor $\left(\sum_{j=1}^{n} K\left(\frac{d(t_i,t_j)}{h}\right)\right)^{-1}$ in equation (1) is a normalisation of the intensity estimate to avoid so-called edge effects. Without this factor, the value of $u_x(t_i)$ would depend on the position of t_i relative to all other locations: locations near the edge of the observation area would systematically be assigned less mass than those in the centre, since they are on average further away from the other locations. This factor is not dependent on the variant x, and only influenced by the location t_i. Thus, it does not change the relative proportions of the intensity estimates for different variants at a location, but it corrects only the proportions among the locations.

From equation (1), it is clear that there are two choices to be made when employing kernel density estimation: the choice of kernel and the choice of bandwidth. Preliminary investigations have shown that for our investigations, the standard normal kernel, i.e. the probability density function

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

of the (two-dimensional) standard normal distribution appears to be the most appropriate kernel.

This kernel is among the most frequently used kernels in practice; its most important characteristic is that of unbounded support, which means that the kernel function – and thus the mass created by any occurrence point – is strictly positive everywhere, although it will of course be close to zero at distances far away from an occurrence point.

There appears to be a consensus in the research literature that, when using a kernel estimator, the choice of bandwidth is much more important to the quality of the final estimate than the choice of kernel (see, for example, Diggle 2003, 118, Møller and Waagepetersen 2004, 37, and Stoyan and Stoyan 1994, 237ff.). In linguistic terms, the bandwidth indicates how strong the influence of a location is estimated for the assessment of its surroundings, i.e. how far a location with a certain variant can be expected to exude its influence into its environments. Tracing this back to the above mentioned language contact, the bandwidth indicates to what extent a certain linguistic feature is part of the communication between different locations. In the pertinent research literature, various techniques for the automated selection of an "optimal" bandwidth are discussed. For example, likelihood cross-validation (LCV) consists of determining the bandwidth that will result in that intensity estimate which best "predicts" the observed occurrences. Since this method turned out to be too sensitive to outliers (see also Silverman 1986, 88) for our purpose, we employed least-squares cross-validation (LSCV). With this procedure, a bandwidth h is found that minimises a certain score $M_x(h)$ for the variant x. This score is an approximation to the mean squared difference of the resulting intensity estimate from the "true" intensity. Note that in this way an individual bandwidth can be determined for each variant's intensity estimate. However, as the extent of mutual influence of locations cannot be expected to be dependent on the respective variant, we minimise the weighted sum

$$M(h) = \sum_x w_x M_x(h), \text{ where } w_x = \sum_{j=1}^{n} l_x(t_j) \tag{3}$$

to obtain a single bandwidth for all the variant-occurrence maps corresponding to one point-symbol map. The optimal h is found simply by evaluating $M(h)$ for a large number of values for h over a certain plausible range. For details on LSCV, see Silverman (1986, 87f).

Figure 1 shows an original point-symbol map from the SBS, in which variants for "Kartoffelkraut" ('potato haulm') are mapped. In Figure 2, the resulting intensity estimator for the variant *Kraut* is illustrated; in Figure 1, this variant is symbolised by a striped isosceles triangle. The record locations from the SBS are marked by a black dot in Figure 2. In this figure, darker shades of blue

in Figure 2 indicate areas with larger estimated intensities, paler shades indicate lower intensities, and areas that appear to be white have estimated intensities close to zero. Note that the estimated intensities are calculated for the $n = 272$ measurement locations of the SBS only, and not for the space between them. The surroundings of each measurement location, as defined by its cell of the corresponding Voronoi mosaic, are assigned the same estimated value. The Voronoi mosaic – also known as Thiessen polygons – assigns each part of the plane to that measurement point to which it is nearest. Since the data contain no information at all on the space between the measurement locations – there might be woods, fields or further villages – this is arguably the most natural way of partitioning the plane for the purpose of our investigation. For details on the definition and properties of Voronoi mosaics, see Okabe et al. (2000). Note that, for practical reasons, the outer boundaries of the investigation area have been chosen to be the edges of the convex hull of the points of measurement. This is a deviation from the original borders of the SBS, which could slightly distort the visual impression near the edges. These distortions are, however, arguably marginal. Also, and perhaps more importantly, since the SBS itself does not map all possible measurement locations in the observation area, neither choice of outer boundaries can be considered more correct than the other. The same argument can be made about the resulting lengths of edges between neighbouring points near the boundary, which will become relevant in the calculation of boundary lengths (cf. Section 4.2).

In comparing Figures 1 and 2, it can easily be seen that the highest estimated intensities for the variant *Kraut* occur in the north-eastern corner of the investigated area, where the corresponding symbol is the only symbol that occurs. In areas such as the middle or the north-west, where the variant *Kraut* is interspersed with other variants, the intensity is lower, and in the south, where this variant is not found, the estimated intensity is nearly zero. This result corresponds very well with the interpretation of the estimated variant occurrence intensity explained above.

From the set of estimated intensity fields, we then create a single area-class map in the following manner: for every point of measurement on the original map, we have one estimated intensity value for each of the variants occurring on the map. Thus, we simply assign each of the points to that variant which has the highest estimated intensity at that location. This is natural, since remembering the interpretation of an intensity field – this variant is the one predicted to be most likely used at that location. In this way, each location is assigned to exactly one variant and the whole observation area is partitioned into areas representing the different variants. An area can therefore be described as the area in which its assigned variant is the most likely to be used on every location in it, the boundaries marking the line where another variant becomes the most common. For any location *t*

on the map, the variant that t is assigned to will be denoted by $x(t)$. By $T(x)$, we will denote the set of locations assigned to the variant x, and by $|T(x)|$ the number of locations in $T(x)$. The boundaries between areas result automatically from this process as the edges between the Voronoi cells of points assigned to different variants. Note that an area standing for a particular variant that is created in this way does not necessarily have to be connected, and disconnected areas are not even uncommon in the SBS.

Figure 3 shows the area-class map corresponding to the point-symbol map depicted in Figure 1. Different variants are denoted with different colour hues; in this figure, the variant *Kraut*, whose estimated intensity map is shown in Figure 2, is marked in turquoise. The boundaries between the areas are marked in orange. Note that although there are 9 different variants mapped on the original point-symbol map, only 4 of them form areas in Figure 3. This is simply due to the fact that for the 5 other variants, there is no location on the map where the corresponding estimated intensity fields have the highest value of all estimated intensities. In other words, these variants are not prevalent enough – or arranged compactly enough – on the original map to form their own areas.

By simply assigning a location and its surroundings to a certain variant, all information on how high the estimated intensities for the other variants are at that location would be lost in the mapping. Therefore, the saturation and brightness of the colour denoting a certain variant are varied at all locations assigned to that variant: the higher the dominance of a variant at a location, the darker and more saturated the colour of this location. More precisely, saturation and brightness of the colour at location t are proportional to

$$b(t) = \frac{\max\limits_{x} u_x(t)}{\sum\limits_{x} u_x(t)}, \tag{4}$$

which can be explained as the fraction of the total estimated intensity at t that is due to $x(t)$. A juxtaposition of Figures 1 and 3 shows that the darkest and most saturated colour shades occur in regions where a certain variant occurs almost without any interference from other variants, such as the south-east and the north-east of the observation area. In areas where multiple variants intermingle, the colours are much lighter and paler. It is worth noticing that this is also the case in regions along the boundaries between areas. This is to be expected, of course, since the boundaries do "not mark a sharp switch from one word to the other, but the center of a transitional area where one comes to be somewhat favored over the other" (cf. FRANCIS 1983, 5).

All procedures described in this section have been implemented in the Java programming language, partially using methods available from the GeoStoch software library. This library contains classes and methods for the analysis and simulation of spatial data. For details, see MAYER ET AL. (2004) and GEOSTOCH (2009). This implementation allows for the automated generation of area-class maps for any number of point-symbol maps according to objectively identical standards, using a standard PC.

These maps provide a means for the quick comprehension of variability patterns on a map. In that they show generalised prevalence patterns of variants, they do not – and are not intended to – arrive at the richness of detail and faithfulness to the original data as point-symbol maps, which is due to their quantitative nature. Their primary purpose is the visualisation and assessment of the results gained with our method, which are the basis for further computational analysis, for example the measuring of similarities between maps.

4.2 Map Characteristics

The procedures described in Section 4.1 yield area-class maps, which are an established and useful means of linguistic investigation. Furthermore and perhaps more importantly, they can be seen as a preparation of the raw data for statistical investigation. By using some of the characteristics calculated in the process of creating the area-class maps – or numbers easily derived from those – one can describe the geometric features of the areas on a map. Obviously, the analysis of any such characteristic only makes sense when there is a plausible linguistic interpretation.

The first characteristic we suggest to calculate for an area-class map is the total length of boundaries between the areas on the map. This can be easily interpreted as an indicator for the overall *complexity* of a map: the more boundaries there are on the map, the more frequently there is a change from the area of one variant to another, which makes a map more complex on a large scale, ignoring, however, the amount of smaller-scale "fluctuation" within the areas, which will be discussed below. Even maps that show the same number of areas can have vastly different amounts of boundaries: more irregularly shaped and disconnected areas will result in much longer boundaries than well-connected and smooth areas. The area-class map in Figure 3 shows a total length of boundaries of 240.8 km, all marked in orange. For comparison, the theoretical maximum length of all boundaries in the maps from the SBS is approximately 8,644 km. This, of course, is never attained in practice, since this would only be possible if not a single location on the map had a neighbour assigned to the same variant. The extensive results presented in Section 5 will help to understand these numbers better.

Secondly, we calculate the mean of all variant occurrence values l_x of x for each $T(x)$ on an area-class map:

$$\bar{l}_x = \frac{1}{|T(x)|} \sum_{t_j \in T(x)} l_x(t_j) \tag{5}$$

This value can be interpreted as the fraction of the total possible variant occurrence value within the area of variant x that actually has a record belonging to the variant x. The extreme $\bar{l}_x = 1$ applies only if at all locations assigned to $T(x)$, only one variant, variant x, occurs. In all other cases, some of the variant occurrence value at a location is not represented by the assignation to the area. For example, if a location t_i has variants x_1 and x_2 with $l_{x_1}(t_i) = 0.5 = l_{x_2}(t_i)$, and $x(t_i) = x_1$, then the occurrence value $l_{x_2}(t_i) = 0.5$ of x_2 at t_i is not represented by assigning t_i to $T(x_1)$. Roughly speaking, the fewer variants other than x occur at the locations belonging to $T(x)$, the higher the value of $\bar{l}_x$. Thus, $\bar{l}_x$ could be termed "area compactness of the area of variant x". We can also define a weighted mean of all $\bar{l}_x$:

$$\bar{L} = \sum_x \frac{|T(x)|}{n} \cdot \bar{l}_x = \frac{1}{n} \sum_x \sum_{t_j \in T(x)} l_x(t_j) \tag{6}$$

Here, the weights are given by the respective relative number of locations in the area of each variant. This is natural because this number multiplied by n is equal to the total possible occurrence value within this area, since the total occurrence value of all variants at a location is always 1. Extending the interpretation of $\bar{l}_x$ to a whole area-class map, $\bar{L}$ can be called the overall *area compactness* of the map, or, from a different point of view, the *fidelity* of the area-class map. In the map shown in Figure 3, the values of $\bar{l}_x$ range from 0.6 for the green area in the west of the observation window to 1.0 for the red area in the east. The overall area compactness, or fidelity, of the map is $\bar{L} = 0.72$, which means that 72 % of the records on the map are represented by the respective areas. Again, Section 5 will put these numbers into context.

Thirdly, we propose to calculate

$$\bar{b}_x = \frac{1}{|T(x)|} \sum_{t_j \in T(x)} b(t_j) \tag{7}$$

and the weighted mean

$$\bar{B} = \sum_x \frac{|T(x)|}{n} \cdot \bar{b}_x \tag{8}$$

as the indicators of the *homogeneity* of an area and the overall homogeneity of a map, respectively. These definitions are justified by the fact that large values of $b(t)$ indicate that the estimated intensity of variant $x(t)$ at t is much larger than that of other variants (see equation (4)). Thus, large values of $\bar{b}_x$ indicate that within the area of x, there is not much "interference" from other variants, which suggests calling that area homogeneous. In contrast to $\bar{l}_x$, $\bar{b}_x$ takes the estimated likelihood of the occurrence of the respective variants at each location into account, rather than the actual records. In Figure 3, the homogeneity of the green area is 0.44, whereas the purple area has a homogeneity of 0.75, and the overall homogeneity of this map is $\bar{B} = 0.70$. As before, these numbers will be given more meaning in Section 5.

Clearly, there are many further possibilities to characterise the area-class maps created with the method introduced in Section 4.1 geometrically and statistically, and we will mention some in Section 6. Not only for reasons of conciseness and clarity, however, we feel that the collection of characteristics presented is sufficient for our purposes. Also and more importantly, their simplicity and easy interpretation give them an advantage over other more complicated indicators.

5 Some Results

We have used the methods for the creation and characterisation of area-class maps proposed in Section 4 to analyse a large set of maps. This set contains 823 maps, a major part of all word geography maps from the SBS. Figure 4 shows the histograms of the characteristics "length of boundaries" (which we will denote by C), $\bar{L}$ and $\bar{B}$ of these maps resulting from our analysis. In Table 1, the values of these characteristics are listed for the sample maps that are depicted in Figures 3 and 5 through 8 and discussed in this section.

Figure 4(a) shows that while the average length of boundaries between areas on a map is roughly 389 km, there are quite a few maps with no more than 100 km of boundaries on them or none at all. This first column of the histogram of course also includes those maps where a single variant is so dominant that only a single area is formed, resulting in the lack of any boundaries between areas. The largest total length of boundaries on a map is less than 1,100 km, and values larger than 700 km are quite uncommon. Still, in this context, the value of $C = 240.8$ km in the map shown in Figure 3 is quite low. A different, extreme example is given in Figure 5, which maps the variants for "Rosenkranz" ('rosary'). Here, $C = 929.7$ km. The subsequent two examples (Figures 6 and 7) both have values of C that are above average but not extreme. Figure 8, showing the areas for "dürres Reisig" ('dry loppings'), has boundaries of 391.4 km in length, a value very close to the average. These numbers help us assess the complexity of a map. A comparison between Figure 3 and Figure 5 shows that there can be great differences between two maps, a fact that is not comprehended by classical dialectometry. An explanation of these differences is, however, still pending. A next step might be to group maps with similar values and determine what circumstances (possibly frequency or age of the linguistic item) comparable maps share that would account for their affinity.

The mean value of $\bar{L}$ over all maps is approx. 0.62, which means that on average, 62 % of the records in an area belong to the variant that the area is assigned to. The corresponding histogram is plotted in Figure 4(b). The example map given in Figure 3 has a value of $\bar{L} = 0.72$, indicating an overall area compactness that is above average. The map in Figure 6 has the very small value of 0.31 for $\bar{L}$, which means that the areas reflect only 31 % of the variants occurring at the locations they include. Again, the map in Figure 8 shows an average overall area compactness. These values give us a clue as to how apt a map is to be separated into areas, and thus they give us information about the structural characteristics of the maps. Maps whose areas are very compact allow for a high fidelity of the resulting area-class maps, which tells us that little abstraction from the raw data is required. Hence, this value primarily serves the assessment of the degree to which an area-class

map is suitable as a visual representation of the original point-symbol map.

A closely related, however more meaningful value is $\bar{B}$, the mean dominance of each location's decisive variant. It is an indicator for what we have dubbed the homogeneity of a map. More intermingling between variants on a small scale results in a smaller $\bar{B}$. If we recall the remarks in Section 2, we can also conclude that a small $\bar{B}$ means that the local competition between variants is rather high on the whole, which would hint at a less stable linguistic situation. For $\bar{B}$, an average value of 0.58 was obtained. The histogram of all values is given in Figure 4(c). From this figure, one can see that no homogeneity values below 0.2 are obtained. This was to be expected, of course: recalling equations (4), (7), and (8), it is clear that for $\bar{B}$ to tend to 0 on a map, the number of variants on that map would have to tend to infinity. Since obviously the number of variants is limited, so are the values of $\bar{B}$. The value of $\bar{B} = 0.71$ for the map in Figure 3 is significantly above average, which is reflected in the predominantly dark shades in the area-class map and only few brighter spots in the north-west and south-west and along the boundaries, telling us that the distribution of the variants is relatively stable. Figure 7 appears, on the whole, even darker than Figure 3, showing very few brighter spots within its areas, which corresponds to a higher $\bar{B}$. The fact that this is true despite much longer boundaries shows that their impact on $\bar{B}$ is not overwhelmingly great. Figure 6 on the other hand shows a rather brightly coloured map, which is reflected by a very low value of $\bar{B} = 0.24$. As before, the value of $\bar{B}$ in Figure 8 is close to the average.

When calculating different characteristics from the same underlying data, as it is done here, it is of interest to what extent these characteristics are related. A standard tool for answering that question for pairs of characteristics is the so-called Pearson product-moment correlation coefficient ρ, whose absolute values do not exceed 1. Large (absolute) values of ρ indicate that one of the characteristics is determined to a large degree by the other one, while values close to 0 indicate that such a relationship does not exist. For details on the correlation coefficient (see, for example, RODGERS AND NICEWANDER 1988). The values of ρ for the characteristics investigated here are given in Table 2. The rather large positive correlation between $\bar{L}$ and $\bar{B}$ means that maps with large values of $\bar{L}$ tend to exhibit also larger values of $\bar{B}$, and small values of $\bar{L}$ frequently appear together with small values of $\bar{B}$; i.e. homogeneity and compactness tend to have values of similar magnitudes. The values in Table 1 show this exemplarily. This, of course, is not surprising if one recalls the definitions of $\bar{L}$ and $\bar{B}$ (see Section 4.2). A map whose variants are distributed very homogeneously, which will be reflected by a large $\bar{B}$, can be transformed into an area-class map very easily, resulting in especially

accurate areas, which is expressed by a large $\bar{L}$. In other words: the fidelity of an area-class map depends on the homogeneity of the areas on it.

The total length of boundaries *C* on a map is negatively correlated with both $\bar{L}$ and $\bar{B}$. Although the absolute values of ρ are not quite as large as for the pairing of $\bar{L}$ and $\bar{B}$, this still means that maps with a higher complexity tend to have lower values of compactness and homogeneity, and vice versa. This effect is stronger with $\bar{B}$ than with $\bar{L}$: the fidelity of a map declines with increasing complexity, owing to the growing number of locations for which an assignation to one or the other area is equivocal, but the homogeneity is even more affected by *C*, as the locations on either side of every boundary influence each other. This can be explained by the however small amount of language contact that is to be expected for the linguistic item in question across the boundaries, even if the areas themselves are rather compact. Remembering the fact that locations *t* close to boundaries usually have lower values of *b*(*t*) (see Section 4.1), and that a higher complexity is indicated by more boundaries (and thus more locations close to boundaries), this is plausible.

6 Conclusions and Outlook

In this article, we have provided a methodology for the automatic assessment of the structural characteristics of dialect feature maps. The first substantial step of this assessment consists of the creation of area-class maps from raw dialect data by applying methods from spatial statistics. Subsequently, these area-class maps are evaluated by means of averaging certain values over a whole map. The three proposed characteristics, C, $\bar{L}$ and $\bar{B}$, allow for a quantitative description of a map, relating to the concepts 'complexity', 'area compactness', and 'homogeneity'. They render these concepts consistent quantities, facilitating further automatic investigation. We have used several examples to illustrate that the areas as well as the respective values of C, $\bar{L}$ and $\bar{B}$ correspond to visual impressions which a dialectologist would have had to phrase more or less intuitionally – and thus subjectively – before now. Hence, this new methodology is the first step towards a software system that should be applicable for all kinds of dialectographical data, and should give dialectologists the means to analyse dialect maps according to specific problems even without profound knowledge of its algorithms.

However, heading towards a quantitative, more objective assessment of linguistic feature maps, much remains yet to be done: maps should be classified automatically according to the values obtained, for which specific algorithms have to be developed. Additional characteristics, for example measures of variability such as the empirical standard deviations or coefficients of variation of $\bar{l}_x$ or $\bar{b}_x$ (cf. Section 4.2) on a map, could be of use in creating meaningful groupings within a given corpus of maps. These groupings can then help to determine what circumstances caused the areal distributions of the respective variants to develop similarly. Also, it is desirable to create methods to classify maps not only according to their overall characteristics, but also according to geographically defined, locally or regionally fixed patterns, for example circles of expansion around cities or graduated progression lines in the countryside. More advanced methods from spatial statistics might play an important role in obtaining this goal. This is also the case for the detection of shapes typical for certain types of spatial diffusion, e.g. funnels or wedges, whose position may vary. With all this achieved, it should be possible to make statements about the underlying linguistic variation and change that is reflected in the different distributional patterns.

The values obtained, namely $\bar{l}_x$ and $\bar{b}_x$, together with the division into areas, can also be the basis for research that goes beyond the assessment of particular feature maps. They can serve to verify certain hypotheses about linguistic borders and areas by testing whether given structures, such as rivers or landscapes, are paralleled by a statistically significant number of boundaries found in a

larger corpus of feature maps. Cumulative occurrence maps for linguistically defined groups of features, e.g. Romanisms or standardisms, can be generated easily, thus facilitating the search for horizontal or vertical spheres of influence.

References

Altmann, Gabriel (1983): Das Piotrovski-Gesetz und seine Verallgemeinerungen. In: Best and Kohlhase (1983), 59–102.

Baddeley, Adrian / Gregori, Pablo / Mateu, Jorge / Stoica, Radu / Stoyan, Dietrich (eds.) (2006): Case Studies in Spatial Point Process Modeling. New York: Springer (Lecture Notes in Statistics 185).

Bach, Adolf (1969): Deutsche Mundartforschung. Ihre Wege, Ergebnisse und Aufgaben. Heidelberg: Winter.

Besch, Werner / Knoop, Ulrich / Putschke, Wolfgang / Wiegand, Herbert Ernst (eds.) (1982–1983): Dialektologie. Ein Handbuch zur deutschen und allgemeinen Dialektforschung. Berlin/New York: de Gruyter (Handbücher zur Sprach- und Kommunikationswissenschaft. 1).

Best, Karl-Heinz / Kohlhase, Jörg (eds.) (1983): Exakte Sprachwandelforschung. Theoretische Beiträge, statistische Analysen und Arbeitsberichte. Göttingen: Edition Herodot.

Best, Karl-Heinz (2006): Quantitative Linguistik. Eine Annäherung. 3rd Edition. Göttingen: Peust & Gutschmidt.

Diggle, Peter J. (2003): Statistical Analysis of Spatial Point Patterns. 2nd Edition. London: Arnold.

Francis, W. Nelson (1983): Dialectology. An Introduction. London: Longman.

Frings, Theodor (1956): Sprache und Geschichte II. Halle (Saale): Niemeyer (Mitteldeutsche Studien. 17).

GeoStoch software library (2009): http://www.uni-ulm.de/mawi/mawi-stochastik/software/

Goebl, Hans (1994): Dialektometrie und Dialektgeographie. Ergebnisse und Desiderata. In: Mattheier, Klaus / Wiesinger, Peter (eds.): Dialektologie des Deutschen. Forschungsstand und Entwicklungstendenzen. Tübingen: Niemeyer, 171–191.

Goebl, Hans (2001): Arealtypologie und Dialektologie. In: Haspelmath, Martin / König, Ekkehard / Oesterreicher, Wulf / Raible, Wolfgang (eds.): Language Typology and Language Universals. An International Handbook. Berlin/New York: de Gruyter (Handbooks of Linguistics and Communication Science. 20), Vol. 2, 1471–1491.

Goebl, Hans (2006): Recent Advances in Salzburg Dialectometry. In: Literary & Linguistic Computing 21, 411–435.

Goebl, Hans (2007): Kurzvorstellung der Korrelativen Dialektometrie. In: Grzybek, Peter (ed.): Exact Methods in the Study of Language and Text. Dedicated to Gabriel Altmann on the Occasion of his 75th Birthday. Berlin: de Gruyter, 165–180.

Goossens, Jan (1969): Strukturelle Sprachgeographie. Eine Einführung in Methoden und Ergebnisse. Heidelberg: Winter.

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

Haas, Walter (1978): Sprachwandel und Sprachgeographie. Untersuchungen zur Struktur der Dialektverschiedenheit am Beispiele der schweizerdeutschen Vokalsysteme. Wiesbaden: Steiner (Zeitschrift für Dialektologie und Linguistik, Beihefte. 30).

Händler, Harald / Wiegand, Herbert Ernst (1982): Das Konzept der Isoglosse: methodische und terminologische Probleme. In: Besch et al. (1982–1983), Vol. 1, 501–527.

Heeringa, Wilbert (2004): Measuring dialect pronunciation differences using Levenshtein distance. Groningen.

Hildebrandt, Reiner (1983): Typologie der arealen lexikalischen Gliederung deutscher Dialekte aufgrund des Deutschen Wortatlasses. In: Besch et al. (1982–1983), Vol. 2, 1331–1367.

Hummel, Lutz (1993): Dialektometrische Analysen zum Kleinen Deutschen Sprachatlas (KDSA). Experimentelle Untersuchungen zu taxometrischen Ordnungsstrukturen als dialektaler Gliederung des deutschen Sprachraums. Tübingen: Niemeyer (Studien zum Kleinen Deutschen Sprachatlas. 4).

Illian, Janine / Penttinen, Antti / Stoyan, Helga / Stoyan, Dietrich (2008): Statistical Analysis and

Modelling of Spatial Point Patterns. Chichester: Wiley.

Kelle, Bernhard (1986): Die typologische Raumgliederung von Mundarten. Eine quantitative Analyse ausgewählter Daten des Südwestdeutschen Sprachatlasses. Marburg: Elwert (Studien zur Dialektologie in Südwestdeutschland. 2).

Köhler, Reinhard / Altmann, Gabriel / Piotrowski, Rajmund G. (2005): Quantitative Linguistics. An International Handbook. Berlin/New York: de Gruyter (Handbooks of Linguistics and Communication Science. 27).

König, Werner (1982): Probleme der Repräsentativität in der Dialektologie. In: Besch et al. (1982–1983), Vol. 1, 463–485.

Leopold, Edda (2005): Das Piotrowski-Gesetz. In: Köhler et al. (2005), 627–633.

Mayer, Johannes / Schmidt, Volker / Schweiggert, Franz (2004): A unified simulation framework for spatial stochastic models. In: Simulation Modelling Practice and Theory 12, 307–326.

Møller, Jesper / Waagepetersen, Rasmus Plenge (2004): Statistical Inference and Simulation for Spatial Point Processes. Boca Raton: Chapman & Hall / CRC.

Nerbonne, John / Heeringa, Wilbert (1998): Computationele vergelijking en classificatie van dialecten. In: Taal en Tongval, Tijdschrift voor Dialectologie 20, 164–193.

Nerbonne, John (2006): Identifying Linguistic Structure in Aggregate Comparison. In: Literary & Linguistic Computing 21, 463–475.

Okabe, Atsuyuki / Boots, Barry / Sugihara, Kokichi / Chiu, Sung Nok (2000): Spatial tessellations: concepts and applications of Voronoi diagrams. 2nd Edition. Chichester: Wiley.

Piotrowski, Rajmund G. / Bektaev, Kaldybay B. / Piotrowskaja, Anna A. (1985): Mathematische Linguistik. Bochum: Brockmeyer (Quantitative Linguistics. 27).

Rodgers, Joseph Lee / Nicewander, W. Alan (1988): Thirteen ways to look at the correlation coefficient. In: The American Statistician 42 (1), 59–66.

SBS: König, Werner (ed.) (1997–2009): Sprachatlas von Bayerisch-Schwaben. Heidelberg: Winter (Bayerischer Sprachatlas. Regionalteil 1). 14 volumes.

Schiltz, Guillaume (1996): Der dialektometrische Atlas von Südwest-Baden (DASB). Konzepte eines dialektometrischen Informationssystems. Marburg: Elwert (Studien zur Dialektologie in Südwestdeutschland. 5).

Schmitt, Ernst Herbert (1992): Interdialektale Verstehbarkeit. Eine Untersuchung im Rhein- und Moselfränkischen. Stuttgart: Steiner (Mainzer Studien zur Sprach- und Volksforschung. 18).

Schneider, Edgar W. (1988): Qualitative vs. Quantitative Methods of Area Delimitation in Dialectology: A Comparison Based on Lexical Data from Georgia and Alabama. In: Journal of English Linguistics 21 (2), 175–212.

Scott, David W. (1992): Multivariate Density Estimation: Theory, Practice, and Visualization. New York: Wiley.

Séguy, Jean (1965–1973): Atlas linguistique et ethnographique de la Gascogne. Paris: Centre National de la Recherche Scientifique.

Silverman, Bernard W. (1986): Density Estimation for Statistics and Data Analysis. New York: Chapman & Hall.

Stoyan, Dietrich / Stoyan, Helga (1994): Fractals, Random Shapes and Point Fields. Methods of Geometrical Statistics. Chichester: J. Wiley & Sons.

Wenzel, Walter (1930): Wortatlas des Kreises Wetzlar und der umliegenden Gebiete. Marburg: Elwert (Deutsche Dialektgeographie. 28).

Summary

In this article, we introduce a new methodology for the objective and automated assessment of dialect-feature maps. Unlike previous dialectometrical techniques, it is not aimed at the separation and analysis of dialects using large corpora of feature maps, but at the assessment of the structural characteristics of single feature maps, which are largely ignored by classical dialectometry. Thus, our approach is intended to provide a means of comparing linguistic feature maps rather than accumulating them. Using methods from spatial statistics, we estimate intensities, i.e. expected occurrence-frequencies, for all recorded variants of the respective feature in the whole observation area. By combining the obtained intensity fields of all variants, area-class maps are generated. The statistical analysis of certain characteristics of these area-class maps – such as the total length of boundaries between the areas of different variants – then yields objective information on the *homogeneity*, *complexity*, and *area compactness* of single feature maps. The methodology is exemplified by the analysis of several maps from the SBS and an interpretation of the results.

Zusammenfassung und Ausblick

In diesem Beitrag wird ein neues Verfahren zur objektiven und automatischen Analyse von Sprachkarten vorgestellt. Anders als bisherige dialektometrische Verfahren zielt es vorerst nicht auf die Einteilung und Analyse von Dialekten anhand von großen Korpora von Merkmalskarten ab, sondern auf die quantitative Beurteilung von strukturellen Eigenschaften einzelner Merkmalskarten, die von der klassischen Dialektometrie weitgehend ignoriert werden. So soll unsere Methode es ermöglichen, Merkmalskarten zu vergleichen und in ihnen bestimmte Strukturen zu finden, anstatt sie zu kumulieren. Dazu werden mittels räumlich-statistischer Methoden für alle Varianten des betreffenden Merkmals im gesamten Untersuchungsgebiet Intensitäten, d.h. die erwarteten Auftretenshäufigkeiten, geschätzt. Durch eine Kombination der erhaltenen Intensitätsfelder ergeben sich Flächenkarten, die als Grundlage für die weitere Analyse dienen. Dabei werden verschiedene Charakteristiken, wie z.B. die Gesamtlänge der Grenzen zwischen den Gebieten unterschiedlicher Varianten, herangezogen, um objektive Aussagen über die *Homogenität*, *Komplexität* und *Kompaktheit* von Gebieten bzw. Karten treffen zu können. Die Ergebnisse dieses Verfahrens werden exemplarisch anhand einiger Karten aus dem SBS vorgestellt und interpretiert.

Theoretische Grundlage für diese Vorgehensweise ist die Annahme, dass die Belege eines Sprachatlas Stichproben sind, die im Einzelfall nicht zwangsläufig zu hundert Prozent valide sind. So ist es vorstellbar, dass eine Gewährsperson an einem Ort eine andere Antwort gegeben hat, als es der Großteil der anderen möglichen Gewährspersonen getan hätte, hätte man die gesamte zu untersuchende Bevölkerungsschicht befragt. Die Validität der Belege kann aber eingeschätzt werden, wenn man die umliegenden Orte betrachtet: Haben alle dieselbe Antwort gegeben, so ist die Validität als hoch einzuschätzen; haben alle eine abweichende Antwort gegeben, so ist der Beleg in Frage zu stellen. Dabei ist die Bedeutung der anderen Orte für die Bewertung eines Beleges umso höher, je näher sie ihm sind, da geographische Nähe in Beziehung steht zu dem sprachlichen Kontakt, der zwischen ihnen stattfindet. (Die geographische Nähe ist natürlich nur einer von vielen Faktoren, doch der einzige, der objektiv und einfach zu quantifizieren ist. Der Einfluss von Verkehrswegen, Territorien, landschaftlichen Gegebenheiten u.ä. muss deshalb vorerst unberücksichtigt bleiben.) So kann ausgehend von den gegebenen Belegen – mit statistischen Methoden – für jeden Ort geschätzt werden, welche Variante die am wahrscheinlichsten zu erwartende ist, auch wenn der jeweilige Beleg tatsächlich ein anderer ist. So werden Flächenkarten generiert, die Ausreißer automatisch ausgleichen, die Dominanz der jeweils wahrscheinlichsten Varianten durch die Farbgebung darstellen und dadurch die tatsächliche linguistische Situation angemessener wiedergeben als die unmittelbar auf den Stichproben basierenden

Punktsymbolkarten.

Grundlage für die Analyse ist eine Datenbank, die das Auftreten der verschiedenen Varianten an den Belegorten des Untersuchungsgebiets enthält. Sie wird zunächst nach den einzelnen Varianten separiert, so dass für jede Variante eine sog. Vorkommenskarte generiert werden kann, die nur verzeichnet, wo eine die Variante auftritt und wo nicht. Abhängig von der räumlichen Verteilung einer Variante wird nun für jeden Ort die Intensität der Verteilung geschätzt, die sich als Auftretenswahrscheinlichkeit der betreffenden Variante interpretieren lässt. Im nächsten Schritt werden die Intensitätsfelder der Varianten (siehe Abb. 2) zu einer Flächenkarte vereint (siehe z.B. Abb. 3), indem jeder Ort der Fläche zugerechnet wird, deren Variante die höchste Intensität an diesem Ort aufweist, d.h. deren Auftreten dort am ehesten zu erwarten ist.

Die so erzeugten Flächenkarten können auf verschiedene Weise hinsichtlich ihrer strukturellen Eigenschaften untersucht werden: Die Länge der Grenzlinien zwischen den Flächen kann als eine Maßzahl für die Komplexität der Karte betrachtet werden, die Intensitäten geben – gemittelt – Aufschluss über die kleinräumige Homogenität einer Karte. Weitere Maßzahlen sind vorstellbar, einige davon werden kurz vorgestellt. Insgesamt bietet das vorgestellte Verfahren die Möglichkeit, Karten von sprachlichen Merkmalen quantitativ und nach einheitlichen Maßstäben nach ihren Struktureigenschaften zu beurteilen, um so Aufschluss darüber zu gewinnen, nach welchen Gesetzmäßigkeiten sich bestimmte Merkmale räumlich entwickeln.

In erster Linie aber sollen die ermittelten Werte die Basis für weitere Forschung, die über die Analyse von Einzelkarten hinausgeht, sein. So können etwa bestimmte Hypothesen bzgl. Sprachgrenzen und -gebieten überprüft werden, indem getestet wird, ob vorgegebenen Strukturen wie Flüssen oder Territorialgrenzen eine statistisch signifikante Anzahl an Sprachgrenzen entspricht, die in einem größeren Korpus von Merkmalskarten bestimmt wurden. Des Weiteren können einfach kumulative Vorkommenskarten von linguistisch definierten Merkmalsgruppen wie Romanismen oder Standardismen erzeugt werden, um so auf einfache Weise horizontale oder vertikale Einflussbereiche festzustellen. Beim typologischen Vergleich der Karten wird es möglich herauszufinden, inwieweit vergleichbare sprachliche Erscheinungen auch vergleichbare Verbreitungsmuster zeigen; bzw. umgekehrt, was vergleichbare Verbreitungsmuster auch sprachlich gemeinsam haben.

Addresses of the Authors

Jonas Rumpf / Volker Schmidt
Institute of Stochastics
Ulm University
89069 Ulm
Germany

Simon Pickl / Stephan Elspaß / Werner König
Universität Augsburg
Lehrstuhl für Deutsche Sprachwissenschaft unter
besonderer Berücksichtigung des Neuhochdeutschen
Universitätsstraße 10
86159 Augsburg
Germany

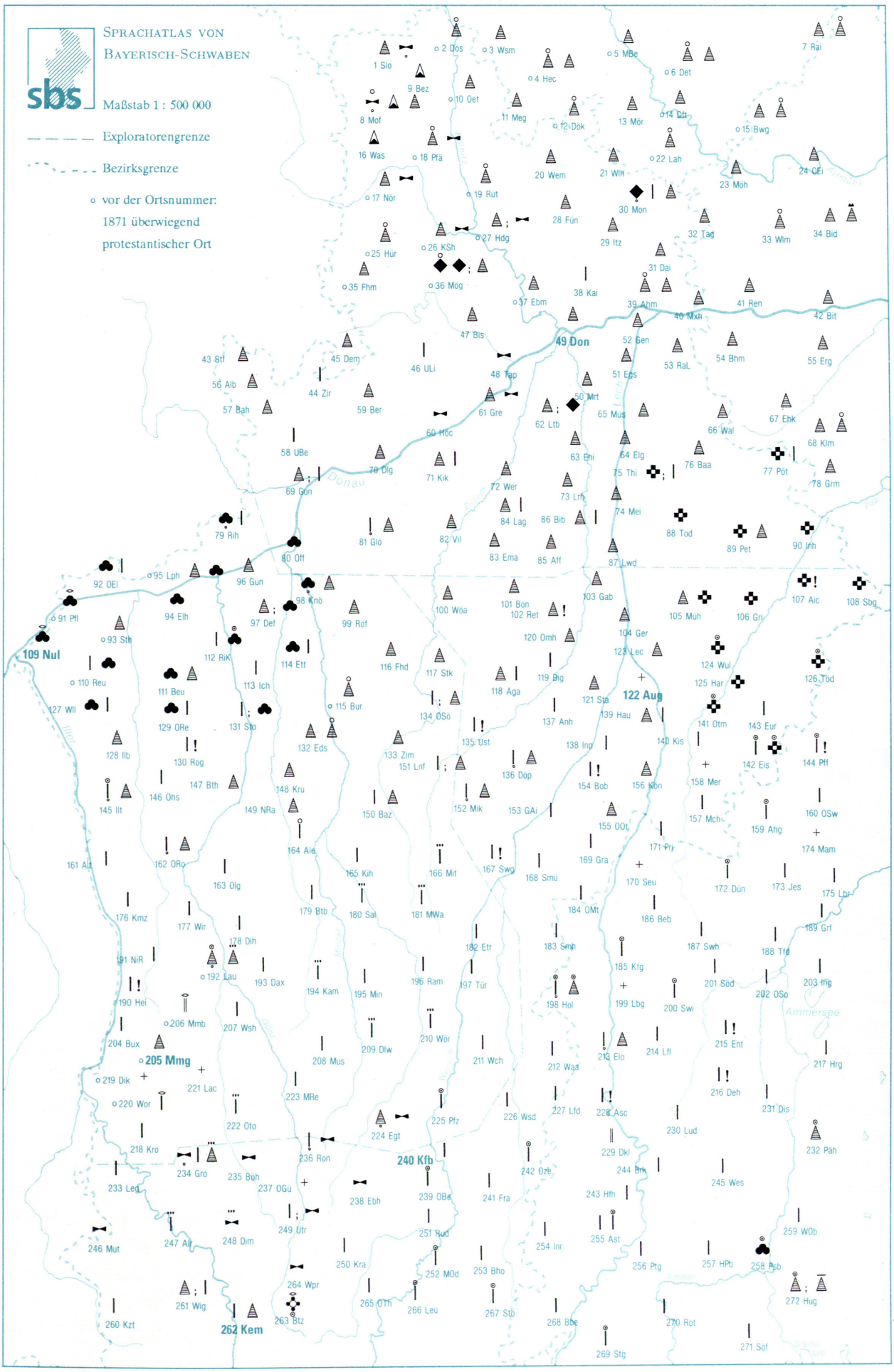


Figure 1a: Example: original point-symbol map 80 “Kartoffelkraut” from the SBS (1997–2005, vol. 8, 295)

Karte 80 (Frage 176.1)

Kartoffelkraut

Die Vorgabe im Fragebuch lautete: *Stauden der Kartoffel getrocknet?*
Suggerierungshilfen waren ›Pflichter‹ und »Kraut«.

Zum Material und zur Kartierung

- Da die GPs nur in seltenen Fällen in der Benennung nach dem Zustand (grün vs. getrocknet) unterschieden haben, werden hier alle Belege kartiert. Gegebenenfalls werden semantische Differenzierungen durch Strichpunkt zwischen den Symbolen angedeutet.
- Zu den Typen »Kraut« und »Staude« sind jeweils (kollektivierende) Ableitungen belegt, die mit abgewandelten Symbolen dargestellt werden.
- Numerusunterschiede sind nicht berücksichtigt.
- Bestimmungswörter sind durch Zusatzzeichen kartiert, außer »Kartoffel-«, das ignoriert wird.
- Der hier mit ›Pflichter‹ lemmatisierte Typ ist bei SCHMELLER (I, 785) für das Gebiet Ilm/Paar als Kollektivtyp *das Geflichter* zu *die Flichtern* 'Blätter von der weißen Rübe' vermerkt.

Zusatzmaterial

Der zweite kartierte Typ in 151 Lnf wird aus dem ZM bestätigt (*kχ̂artọvlgrêitr*).

Zeichenerklärung zu Karte 80

Simplizia oder Grundwörter in Komposita:

- »**Kraut**« n./»**Kräuter**« Pl. (z.B. *khrôut^h*, *ę̆bī̜əgrā̤da* Pl., *k^{ə}ɣ̈bškxrâod*)
- ›**Kräuterich**‹ n. (z.B. *grâ̜edri̜χ*)
- »**Stengel**« m. (z.B. *šdë̤ŋl*, *bō̤dəbī̜^{ə}radšteŋ̥l*)
- »**Staude**« f. (z.B. *štâ̜oda*, *ę̄ədë̤pfl̥štôuda* je Pl.)
- ›**G^{e}staude**‹/›**-g^{e}stäude**‹ n. (206 Mmb *gʁọmbī̜əʁəgštê̜id*, 229 Dkl *gštộuda*)
- »**Stock**« m. (z.B. *šdẹgχ̍^{h}* Pl.)
- ›**Pflichter**‹ n. (z.B. *ḅṿli̜χta*, 107 Aic *ḅṿli̜h^{x}ṭa d*)
- »**Rebe**« f. (z.B. *rę̄əbə*, *ę̄ədë̤pflrę̄əbə*, 98 Knö *rę̄əvə*)
- »**-laub**« n. (263 Btz *grụmbî̜ərəlâob*)

Zusatzzeichen über den Symbolen: Bestimmungswörter

»**Erdäpfel-**«	›**Bumser-**‹
»**Erdbirnen-**«	»**Bodenbirnen-**«
»**Grundbirnen-**«	»**Herbst-**«

Weitere Zeichen:

- ◦ unter dem Symbol: Beleg als „älter", „richtiger" u.ä. qualifiziert
- ⊙ unter dem Symbol: Beleg als E qualifiziert
- ; zwischen den Symbolen: semantische Differenzierung; vgl. Belegliste
- \+ nicht gefragt
- ! Hinweis auf Belegliste

Figure 1b: Legend of map 80 "Kartoffelkraut" from the SBS (1997–2005, vol. 8, 295)

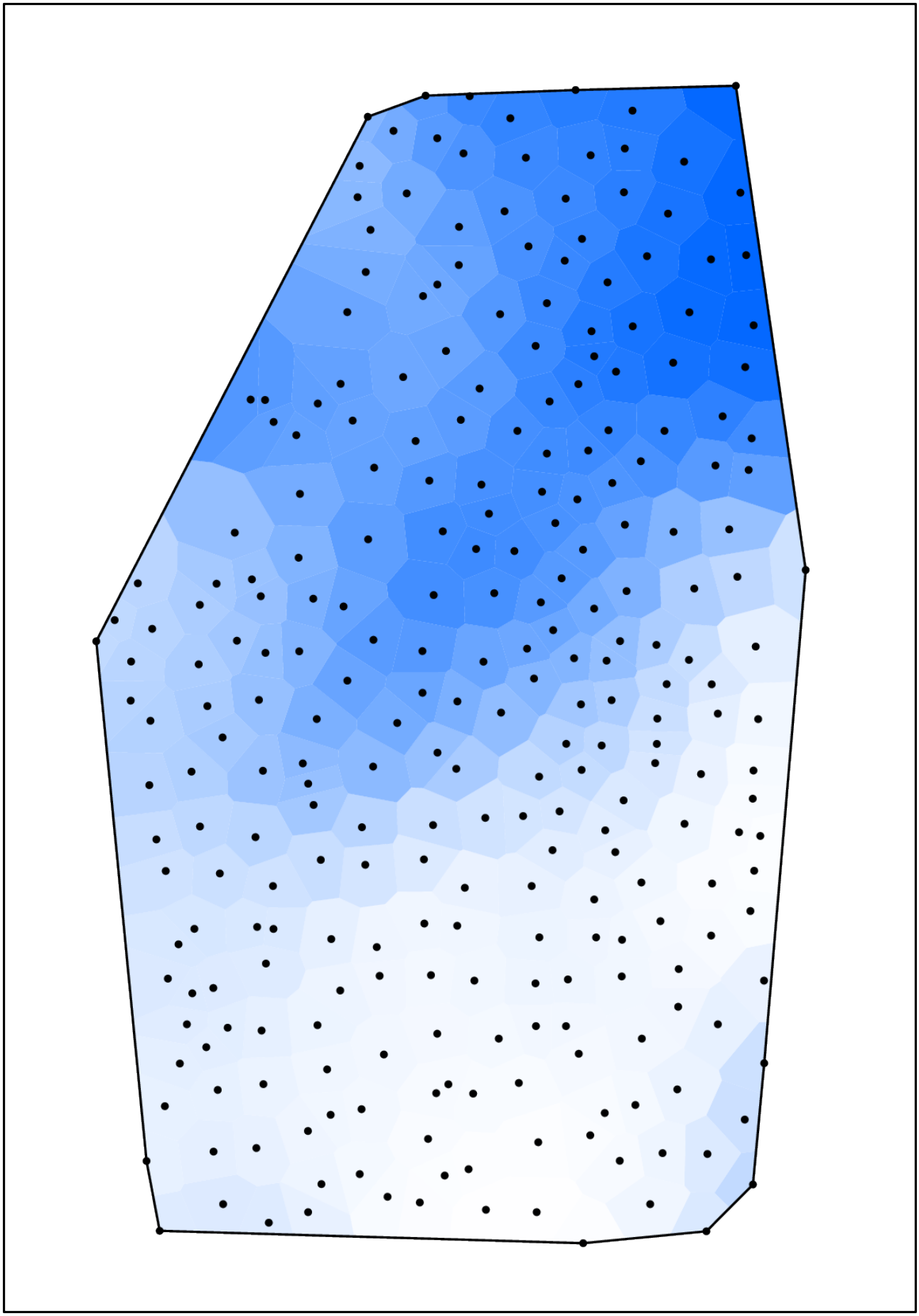


Figure 2: Example: estimated intensity field of variant Kraut in map 80 “Kartoffelkraut” (SBS 1997–2005, vol. 8, 294f.). This variant is represented by the triangular symbol in Figure 1. The corresponding area in Figure 3 is marked turquoise.

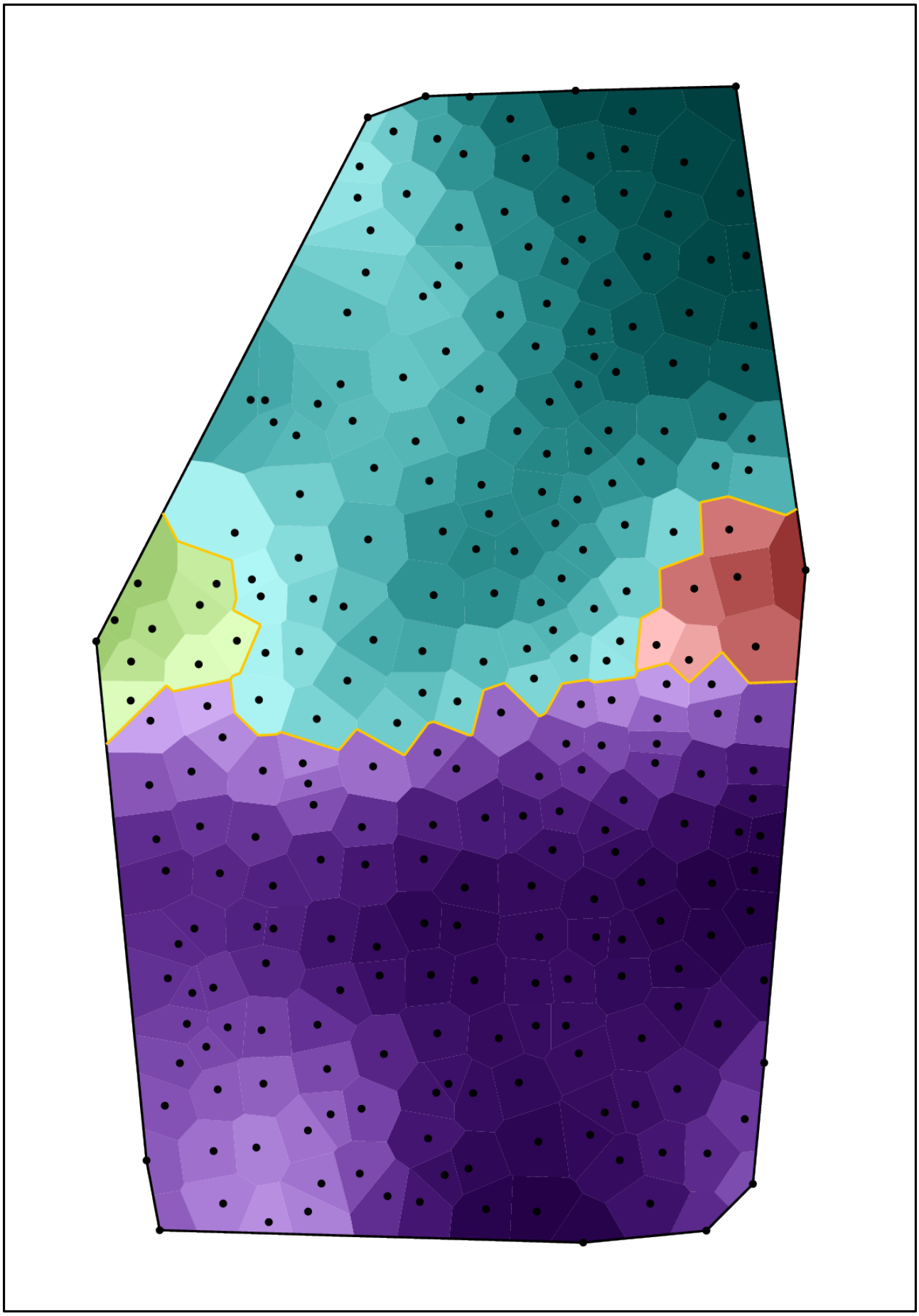
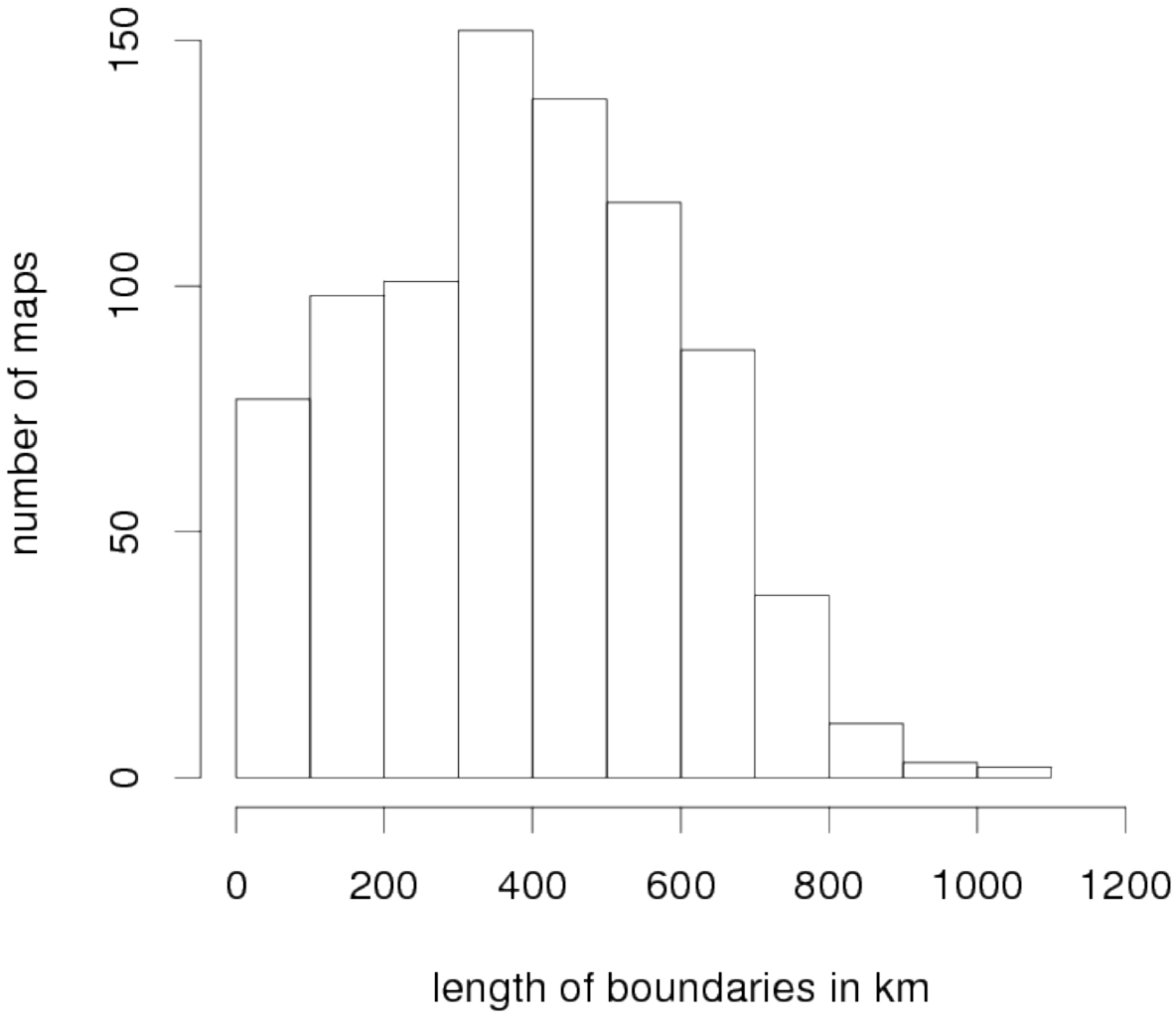
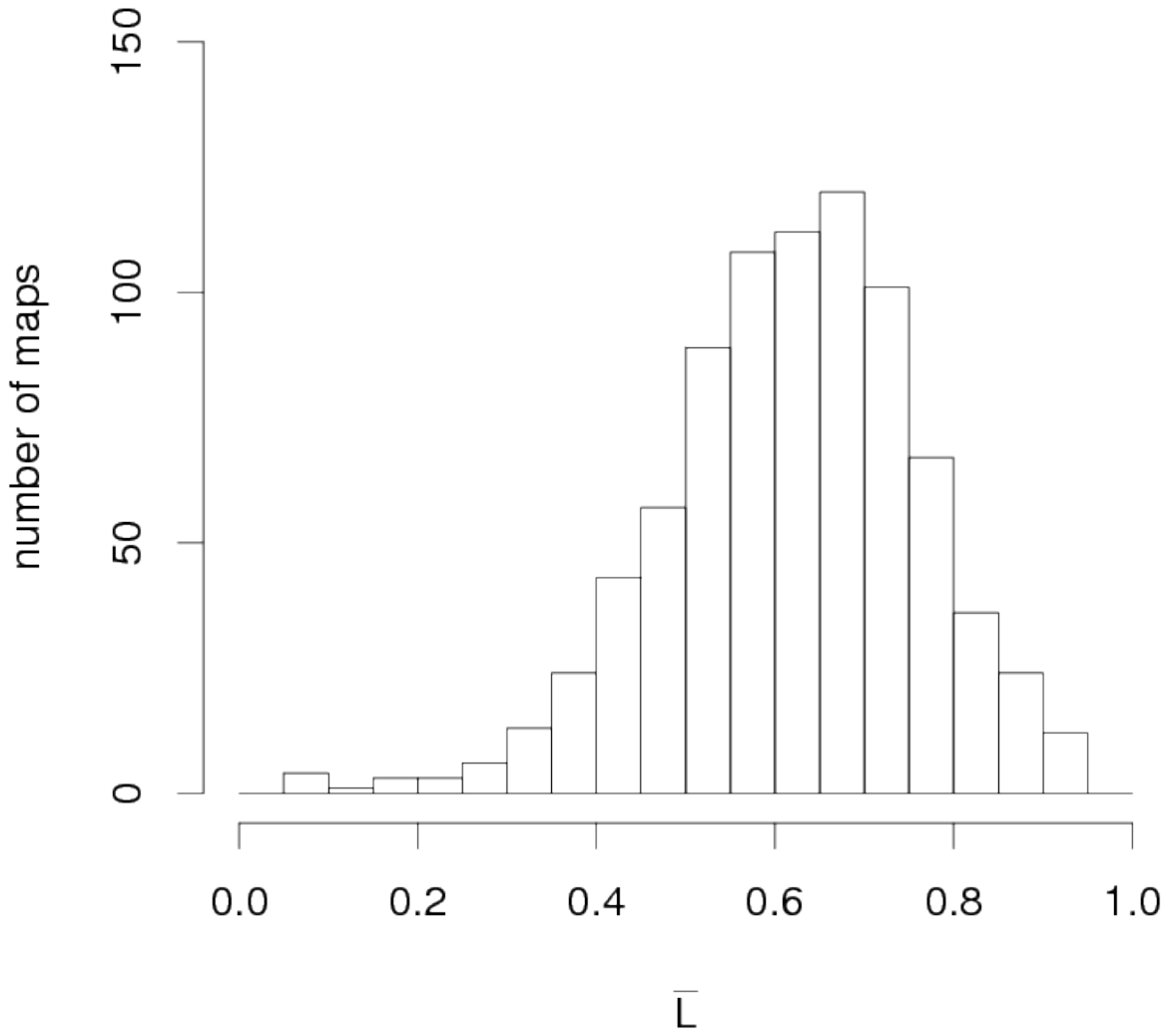


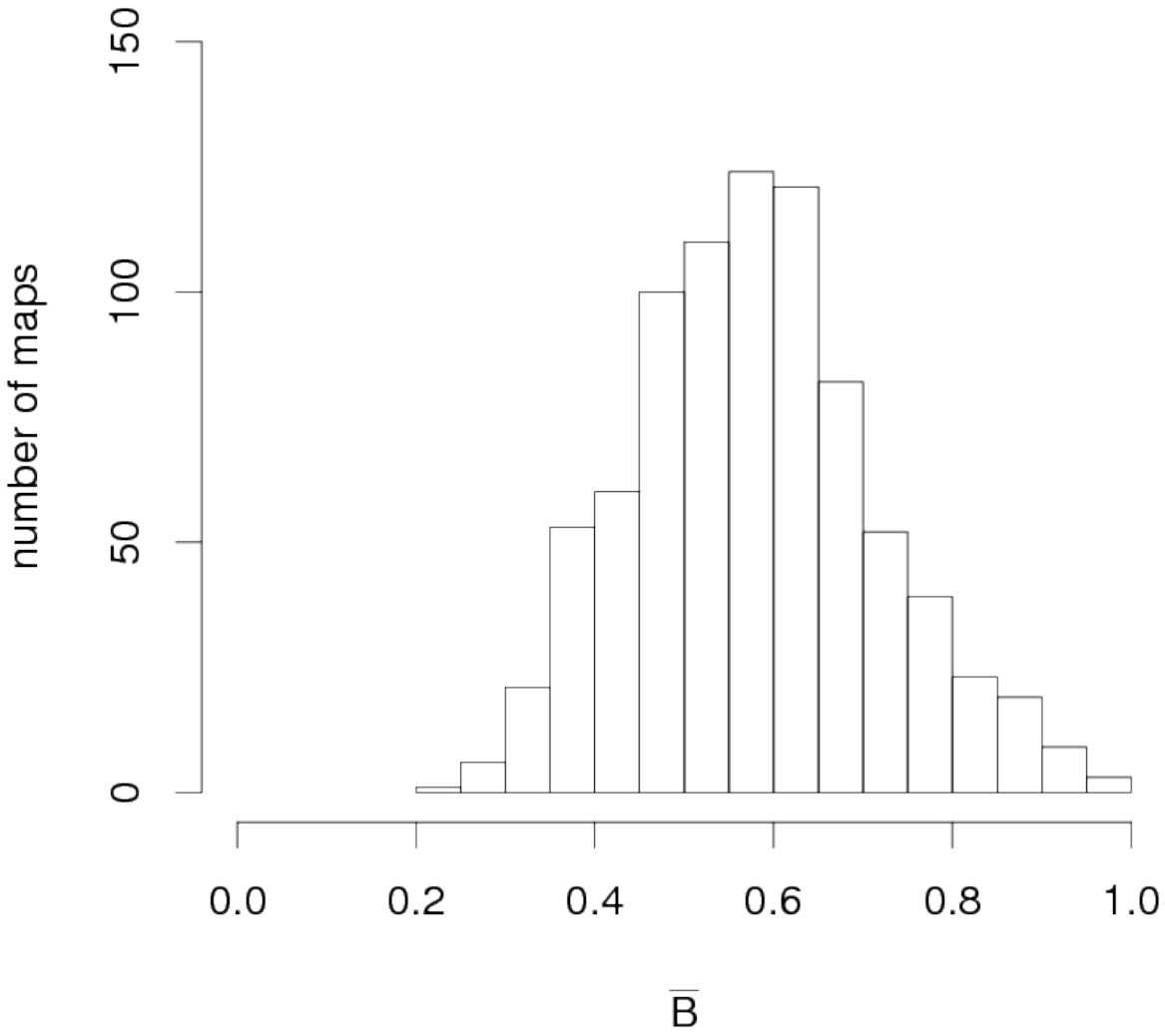
Figure 3: Example: area-class map of the data underlying map 80 “Kartoffelkraut” (SBS 1997–2005, vol. 8, p. 294f.)



(a) total length of boundaries *C* between areas on a map (*complexity*)



(b) overall *area compactness* $\bar{L}$ of maps



(c) overall homogeneity $\bar{B}$ of maps

Figure 4: Histograms of various map characteristics calculated from a set of 823 word geography maps from the SBS (1997–2005)

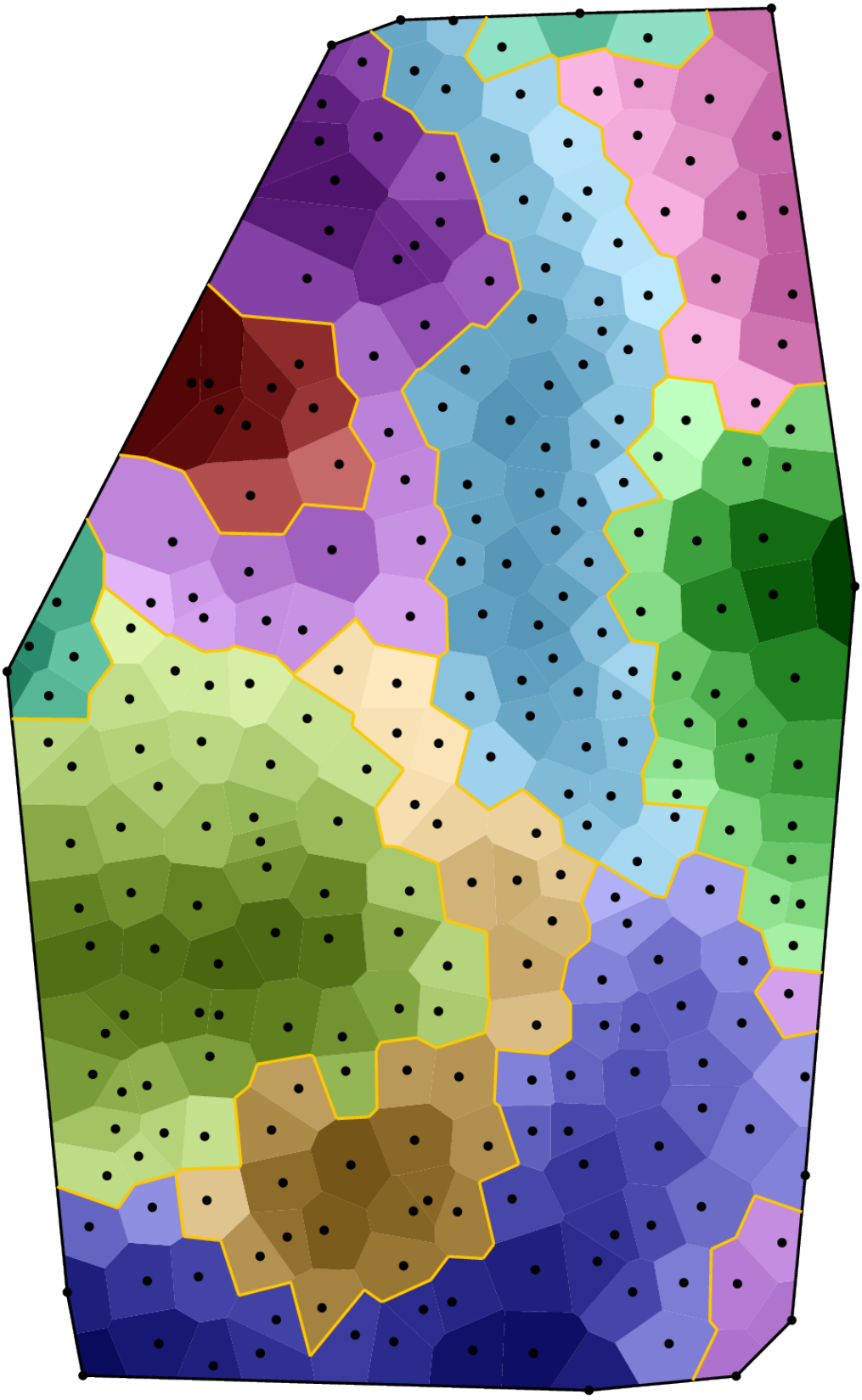


Figure 5: Example: area-class map of the data underlying map 126 “Rosenkranz” (SBS 1997–2005, vol. 2, 532f.)

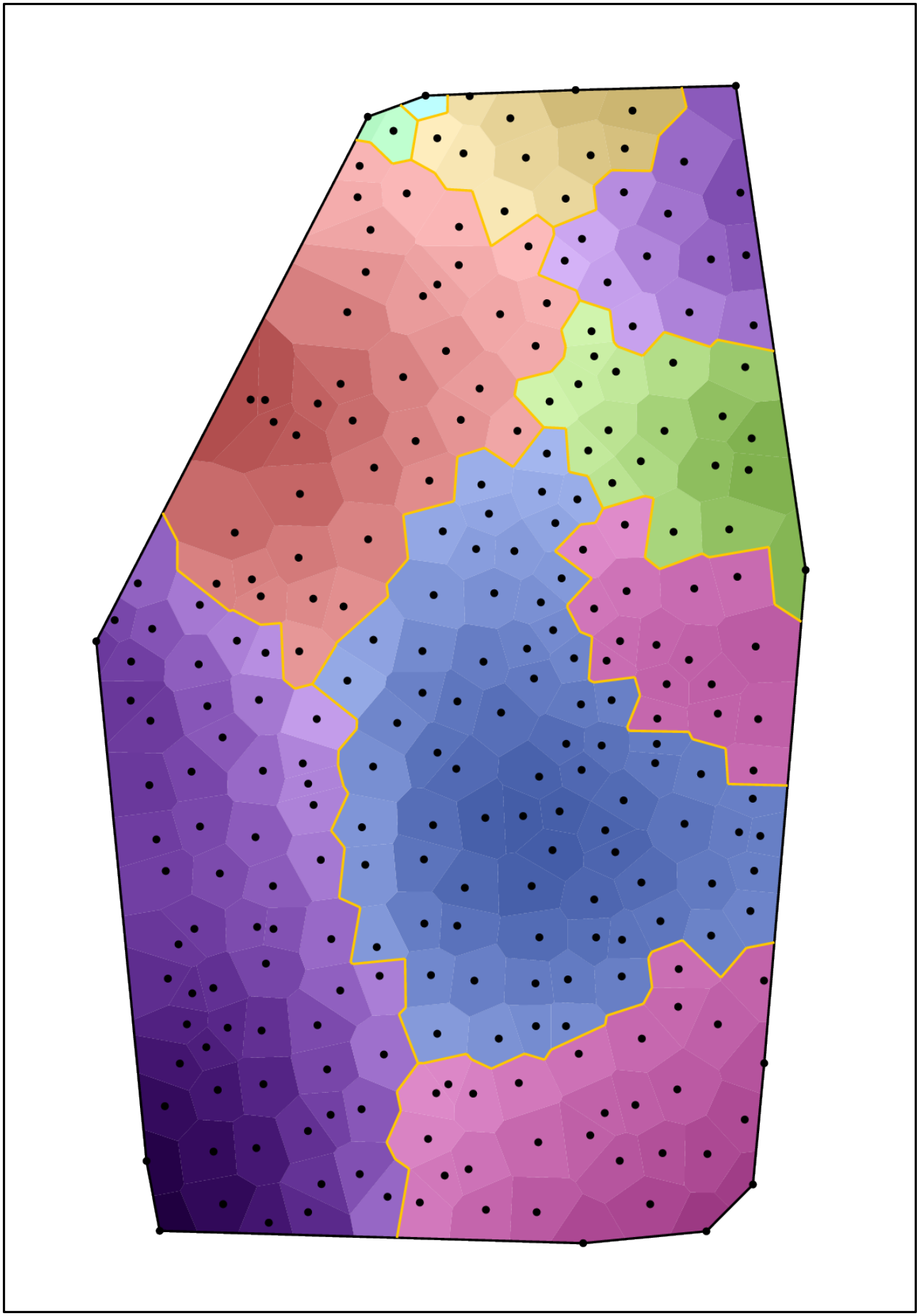


Figure 6: Example: area-class map of the data underlying map 15 “die kleinen Hinterklauen der Kuh” (SBS 1997–2005, vol. 11, 52f.)

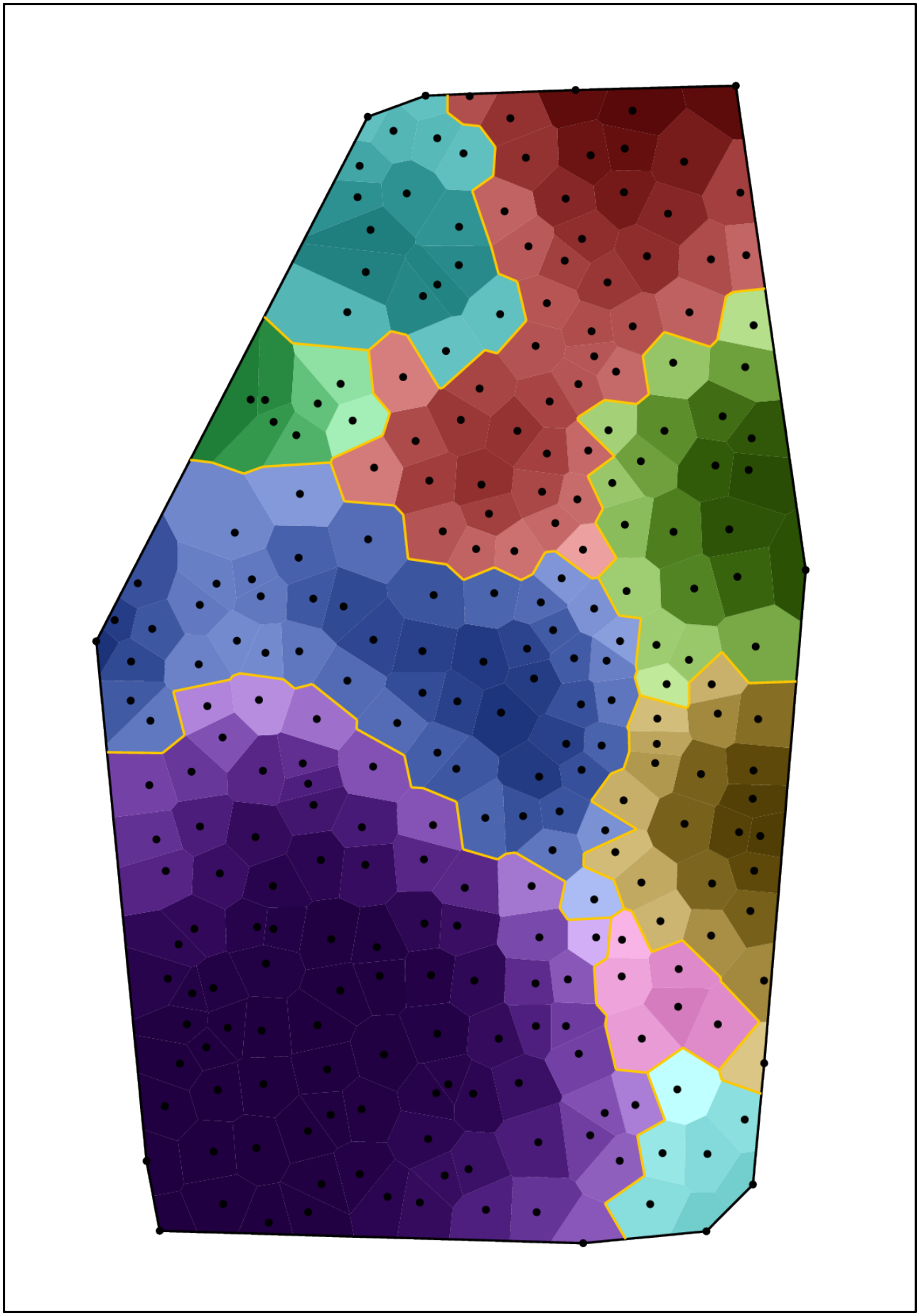


Figure 7: Example: area-class map of the data underlying map 71 “Heuhaufen bei drohendem Regen” (SBS 1997–2005, vol. 12, 220f.)

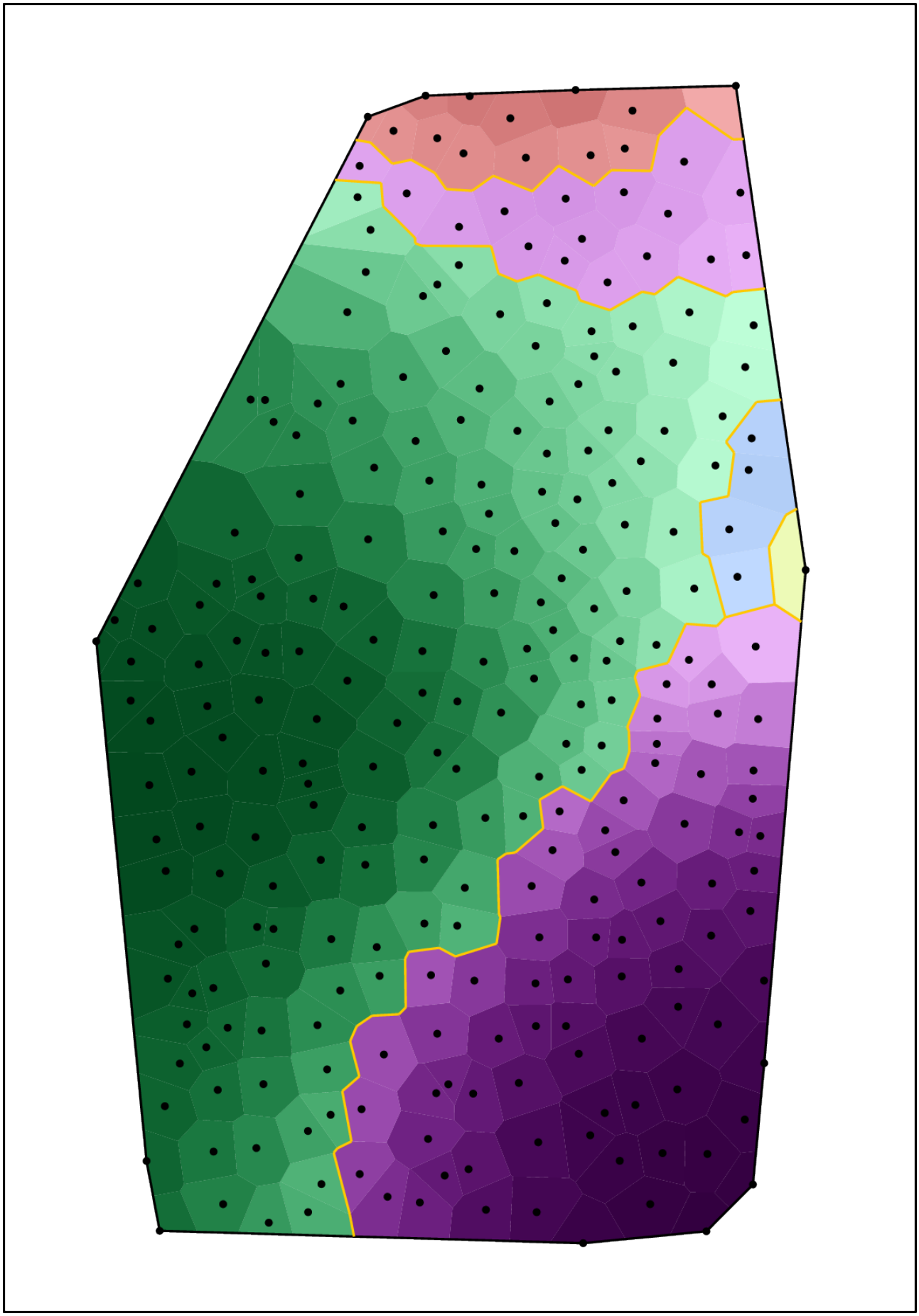


Figure 8: Example: area-class map of the data underlying map 13 “dürres Reisig” (SBS 1997–2005, vol. 13, 532f.)

Tables

figure	title	C	$\bar{L}$	$\bar{B}$
Figure 3	‘Kartoffelkraut’	240.8 km	0.72	0.71
Figure 5	‘Rosenkranz’	929.7 km	0.53	0.40
Figure 6	‘die kleinen Hinterklauen der Kuh’	637.2 km	0.31	0.24
Figure 7	‘Heuhaufen bei drohendem Regen’	653.3 km	0.80	0.67
Figure 8	‘dürres Reisig’	391.4 km	0.65	0.60
	average values	388.6 km	0.62	0.58

Table 1: Values of various characteristics for the example maps

	C	$\bar{L}$	$\bar{B}$
C	1	-0.261	-0.576
$\bar{L}$		1	0.745
$\bar{B}$			1

Table 2: Empirical correlation coefficients between various map characteristics calculated from a set of 823 word geography maps from the SBS (1997–2005)