To what extent are surnames words?

Comparing geographic patterns of surname and dialect variation in the Netherlands

Franz Manni,[1*] Wilbert Heeringa [2] and John Nerbonne [2]

(1) UMR 5145 CNRS, Musée de l'Homme MNHN – Paris, France.

(2) Alfa-Informatica, Faculty of Arts, University of Groningen, The Netherlands.

(*) Corresponding author:

Dr. Franz Manni

UMR 5145, Group of population genetics

National Museum of Natural History

MNHN - Musée de l'Homme

17, Place du Trocadéro

75016 Paris - France

Tel. 0033 1 44 05 72 84

Fax. 0033 1 44 05 72 41

Abstract

Since the early papers of Sokal (1988) and Cavalli-Sforza et al. (1989) there has been an increasing interest in depicting the history of human migrations by comparing genetic and linguistic differences that mirror different aspects of human history. Most of the literature concerns continental or macro-regional patterns of variation, while regional and micro-regional scales were investigated less successfully. In this paper we concentrate on the Netherlands, an area of only 40,000 km^2.

The focus of the paper is on the analysis of surnames, which have been proven to be reliable genetic markers since in patrilineal systems they are transmitted—virtually unchanged—along generations, similarly to a genetic locus on the Y-chromosome. We shall compare their distribution to that of dialect pronunciations, which are clearly culturally transmitted (children learn one of the linguistic varieties they are exposed to, normally that of their peers in the same area or that of their family). Since surnames, at the time of their introduction, were words subject to the same linguistic processes which otherwise result in dialect differences one might expect the distribution of surnames to be correlated with dialect pronunciation differences. But we shall argue that once the collinear effects of geography on both genetics and cultural transmission are taken into account, there is in fact *no* statistically significant association between the two. We show that surnames cannot be taken as a proxy for dialect variation, even though they can be safely used as a proxy to Y-chromosome genetic variation.

We work primarily with regression analyses, which show that both surname and dialect variation are strongly correlated with geographic distance. In view of this strong correlation, we focus on the residuals of the regression, which seeks to explain genetic and linguistic variation on the basis of geography (where geographic distance is the independent variable, and surname diversity or linguistic diversity is the dependent variable). We then seek a more detailed portrait of the geographic patterns of variation by identifying the “barriers” (namely the areas where the residuals are greatest) by applying the Monmonier algorithm.

We find the results historically and geographically insightful, hopefully leading to a deeper understanding of the role of the local migrations and cultural diffusion that are responsible for surname and dialect diversity.

1 Introduction

The aim of this study is to compare the geographic patterns of genetic variation with corresponding linguistic data in the Netherlands (Fig. 1). Family names can be regarded as genetic markers since they are transmitted along the male-line together with the Y-chromosome in patrilineal societies. Before becoming surnames with a strict rule of transmission, family names were also words and so they remain today (even if 'frozen' to meet the needs of administration), so we might expect them to pattern with other linguistic material, which is why our study also asks: *to what extent are surnames words?*

We investigate this dual nature of patronymic markers by comparing the geographic patterns of variability of 19,910 Dutch surnames accounting for 1,303,369 individuals with the linguistic differences of the Netherlands measured by Heeringa (2004). As we shall see, surnames are not distributed in the same way as dialect differences.

To assess how different surnames are in two locations we computed a specific pairwise surname distance ('Nei distance') between the 226 Dutch localities. Such measures were compared to Levenshtein distances that, analogously, assay linguistic diversity. We shall note that surname analyses have been implemented by excluding very common ('polyphyletic') surnames which otherwise lead to an underestimation of the actual levels of diversity.

1.1 Surnames

Male-transmitted family names simulate neutral alleles of a gene transmitted only through the Y chromosome (Yasuda and Morton 1967; Yasuda and Furusho 1971; Yasuda et al. 1974; Zei et al. 1984; King et al. 2005) and therefore satisfy the expectations of the neutral theory of molecular evolution (Cavalli-Sforza and Bodmer 1971; Crow 1980), which is entirely described by random genetic drift, mutation and migration (Kimura 1983). This property of surnames, together with their ready availability, has made them useful for the study of population structure since 1965, when Crow and Mange published the quantitative relation existing between isonymy[i] and inbreeding. Recently, the isonymy method was applied to a genealogical database (Gagnon and Toupance 2002) and consanguinity was estimated both from surnames and from true genealogies. Results indicate that random isonymy, estimated from family names, fits well with consanguinity estimates obtained from genealogical records.

Several papers have focused on surnames on account of their ready availability, from voters' lists or phone books. They are then useful in the investigation of genetic structures (meaningful differences in the geographic space) of populations. If the use of patronymic markers is easy and provides very large sample sizes, it also might suffers from limitations related to 1) non-paternity, 2) surname-change, 3) polyphyleticism and 4) limited temporal depth in generations. Non-paternity and surname-change are not at all major problems, infecting no more than 10% of the data, but polyphyleticism can decrease the reliability of surname studies.

By polyphyleticism we mean the circumstance that unrelated people may share the same surname. At the time of surname introduction the same surname (e.g. Woods, Grant, White, etc.) often came into use in different unrelated families, even those established in different geographic locations. In classical surname analyses, i.e. studies based on surname distances (Chen and Cavalli-Sforza 1983; Lasker et al. 1985), polyphyletic surnames decrease the value of pairwise distance measures between locations based on the numbers of families with the same surname. To avoid arbitrary exclusions of some family names, published studies were performed on the whole corpus of data by (unreliably) regarding polyphyletic surnames as monophyletic. We have recently shown (Manni et al. 2005) that it is possible to decrease this source of error via a neural network analysis (Kohonen 1995) of the geographic distribution of the surnames. In this way, the identification of some clearly polyphyletic surnames becomes possible, since they share the crucial properties *i)* the absence of a coherent geographic hearth of diffusion, *ii)* a high average number of people sharing the surname, and *iii)* a peculiar clustering in specific cells of the Kohonen map.

The second major constraint of surnames, as we mentioned, is related to their limited temporal depth. It is known that they provide no information for periods previous to the late Middle Ages (at best), when they first spread in most European countries. In the Netherlands surnames were not obligatory until the Napoleonic period. As a consequence, surname-inferred demographic phenomena–such as migrations, drift, and isolation–can be dated at best only within the last six centuries for most Europe, and only within the last two in the Dutch case. The distribution of family names deserves even more to be studied in comparison with linguistic variability since dialects evolve at rates detectable over similarly large time frames. This large-scale synchrony in the diffusion of surname and dialect variants justifies further the comparison that we are undertaking. Possible results might be: *i)* similar geographic patterns in surnames and dialects, thus suggesting that social and demographic processes were similar; or *ii)* genetic variability that differs from linguistic variability, which would show that the social contacts mirrored by dialects do not correspond to the demographic history of the populations speaking those dialects.

But, before addressing such comparisons, it is necessary to discuss an older criticism, related to the dual nature of patronymic markers. If surnames were words, they should mirror linguistic diversity (we note that it is often possible to guess someone's geographic origin by the sound and spelling of her/his surname). This motivates us to ask whether and to *what extent surnames are words*. If their variability really was related to their linguistic neighborhood, we would expect today to find patterns similar to those of dialects. This comparison is no longer only hypothetically interesting since dialectologists now process large amounts of data exactly, enabling the establishment of truly quantitative, statistically meaningful, correlations between linguistic and surname variability. This step is essential since the outcomes of genetic studies are frequency-vectors and distance matrices that deserve comparison with similarly exact information.

1.2 Dialects

In genetics the idea that genetic divergence increases with geographic distance is a well-accepted and established notion, and large-scale studies gave evidence of it (for an exhaustive introduction see Cavalli-Sforza et al. 1989). A similar (but in mechanical detail nonidentical) idea can be traced back to the 'wave theory' of Johannes Schmidt (1872) about Indo-European languages. From a distant perspective (following Isidore Dyen, personal communication) all languages chains of pairs of mutually intelligible speakers (or speech-types) where different varieties gradually shade one into another, where the extremes of the chains are the most different areas. The role played by geographic distance in the continuous increase of linguistic divergence is also the point of Chambers & Trudgill's "traveller's distance" (1998, p. 5). The idea is that a traveler going across a linguistic area will repeatedly encounter dialects whose features overlap to a large extent with those of the last dialect he heard and also the next one he will hear. He experiences in this way the *continuum* that is now frequently appealed to in dialectology: neighboring dialects are usually quite similar. A dialectometrical analysis of the traveler has been undertaken, on Dutch dialect data, by Heeringa and Nerbonne (2001), and the mathematical association between geographic and linguistic distance was so close that they summarized it in a mathematical regression between geographic and linguistic distance, an approach that was probably first applied to linguistics by Séguy (1971). Unlike authors who see the *continuum* just as an undulated landscape, Heeringa and Nerbonne have shown that the mean height of such ‘undulations’ is not constant through space, since pairwise comparisons of dialect variants lead to occasionally higher values as dialect borders are encountered.

If we were able to eliminate, from a dialectometric matrix of distances, the variance explained by geography we would be able to focus on the residual variance that probably is not related to contact between neighboring speakers. When interested in the historical evolution of dialect variation, large residual variance may signal a pattern of linguistic difference that is more ancient. We can also imagine that in ancient times, as a consequence of sparser population density, less contact between speakers and less reliable transportation, linguistic (dialect) differences were stronger than they are today.

In this paper, we compute a general regression model between Levenshtein dialect distances (see Heeringa 2004, pp. 121-144) and geographic distances between dialect locations, thus between pronunciation distances and the distances between pairs of dialect locations (in kilometers). We computed the Levenshtein distance between the sites in a pairwise fashion. Then expected distances are subtracted from the observed ones leading to the computation of residuals. Finally, we construct boundaries based on the residuals.

2 Methodology and data

2.1 Data

2.1.1 Surnames

From the original Dutch dataset (Manni 2001) of 51,578 surnames, corresponding to 2,294,154 individual telephone users (1997 data) in 226 locations we have eliminated very frequent surnames (those recorded in more than 100 locations) and very rare surnames (those recorded in fewer than 10 locations). The 226 locations are those as listed by Barrai et al. (2002), following Manni (2001). The exclusion of very frequent surnames relies on the confounding effect they have on analyses: polyphyleticism leads to inflated estimations of consanguinity (see introduction). Concerning the Netherlands, the demonstration that very frequent surnames are polyphyletic can be found in Manni et al. (2005). Rare surnames were excluded because their contribution to the overall picture is irrelevant (Manni, 2005; unpublished). For the 226 locations, the correlation (Mantel 1967; Manly 1997) between a surname distance matrix (whatever the distance) computed by retaining rare surnames and another obtained by eliminating them, approaches 1. This exclusion does not bias the dataset since removed surnames correspond to a similar fraction of individuals in each of the 226 samples.

From this new dataset, consisting of 19,910 surnames accounting for 1,303,369 individuals (8.1% of the entire Dutch population), we computed a matrix of Nei distances according to the formula:

$$\Sigma \mathrm{n}_{si}\mathrm{n}_{sj} \,/\, (\Sigma \mathrm{n}_{si}^2\ \Sigma \mathrm{n}_{sj}^2)^{\,½},$$

where n_{si} denotes the frequency of a given surname *s* in locations *i* while n_{sj} denotes the frequency of the same surname in location *j*. Note that the sums are done for all surnames. This is the accepted manner of calculating a measure of surname differentiation.

2.1.2 Dialects

Heeringa and Nerbonne (2001) analyzed Chamber's dialectal traveller (Chambers and Trudgill 1998) by sketching a line through the Dutch-Belgian area in which Dutch dialects are spoken. Naturally, this is a small sample of all the sites which one can compare in examining linguistic-geographic correlation, and Nerbonne et al. (1999) have computed regression models for all pairwise distances in the matrix of sampling sites, making the computation more stable than that of the monodimensional "traveller's distance" along a line.

In proceeding this way we are applying to linguistics the concept of "isolation by distance" that was first introduced in genetics by Malecot (1955) when he demonstrated that close populations are genetically more similar than distant ones. Interestingly, the similarity between genetic and linguistic data can be pushed further since, in both cases, the correlation with geographic distances is not linear and the same logarithmic transformation is applied to both datasets in order to obtain an improved linear model.

Levenshtein distances were computed over all pronunciations, using the same data as Heeringa (2004), although some technical constraints forced us to reduce the number of Dutch sites in the sample to 252. The Levenshtein algorithm calculates the least cost of operations needed to map one pronunciation string (phonetic transcription) into another (Nerbonne et al. 1999). The measurement is consistent for large samples of words (Cronbach's $\alpha > 0.96$ for 100-word samples from this set, see Heeringa, 2004, 170-177), and we used 125 words in the current study. The measurements have been validated with respect to scholarly tradition (Heeringa et al. 2002) and again with respect to lay dialect speakers' judgment of dialectal distance (Heeringa & Gooskens, 2004). The latter study showed that the measurement correlated highly with lay speakers' judgments (r=0.78). In addition, the technique has now been applied to Norwegian, American English, German, Sardinian, and

Bulgarian and Bantu languages of Gabon. Interestingly, the same Levenshtein algorithm has been applied extensively to measurement differences in long genetic strings (Sankoff & Kruskal, 1999).

2.2 Visualization of Diversity

2.2.1 Multidimensional Space: Principal Component Analysis (PCA)

Principal component analysis (PCA) was applied to the data to graphically identify possible patterns of similarity between the 226 surnames samples and 252 dialect samples. The PCA method involves a mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called principal components. The first principal component accounts for as much of the variability (variance) in the data as possible, and each subsequent component accounts for as much of the remaining variability as possible.

Multidimensional relationships between items can be seen in a bivariate or trivariate plot (if two or three axes are plotted one against each other). The analysis was performed with the Excel applet GenAlEx, by R. Peakall and P.E. Smouse (2001), freely available at: **`http://www.anu.edu.au/BoZo/genAlEx`**

2.2.2 Geographic Analysis: the Monmonier Algorithm

When sampling locations are known, the association between genetic and geographic distances can be tested by regression methods. These tests give some clues about the shape of the genetic landscape. Nevertheless, regression analyses are by themselves unenlightening when attempting to identify where barriers may exist, namely *the areas where a given variable shows an abrupt rate of change*. To remedy this, we look to a computational geometry approach which uses computed distances (surname, linguistic), or the residuals of the regression procedure, to identify the locations of barriers, and which additionally can therefore also show where the geographic patterns of two or more variables are similar. Inspired by this idea we have implemented Monmonier's (1973) maximum difference algorithm anew (Manni 2004),[ii] in order to identify genetic and linguistic barriers, namely the areas where differences between pairs of populations are unexpectedly large (with respect to Nei and Levenshtein distance measures). To avoid ambiguity we stress that we use the term 'boundary' synonymously with 'barrier'.

To test the confidence with which we may view the barriers in genetic or linguistic landscape, a significance test was implemented in the software by means of bootstrap analysis. As a result, *i)* the noise associated in genetic or linguistic markers can be visualized on a geographic map and *ii)* the areas where barriers are more robust can be identified. Moreover, this multiple approach allows us to inspect the barriers in order to get an idea of *iii)* the patterns of variation associated with different markers in the same landscape. In this study bootstrap analysis is undertaken for surname data only. A manuscript focusing on a bootstrap approach toward Dutch dialects is in preparation.

2.2.2.1 The Triangulation

Delaunay triangulation (Brassel & Reif 1979) is the fastest triangulation method to connect a set of points (localities) on a plane (map) by a set of triangles. It is the most direct way to connect (triangulate) adjacent points on a map. It should be noted that Delaunay triangulation is the dual of Voronoi tiling (Voronoi 1908), which results in a set of polygons, each surrounding exactly one site and together covering the area studied. The Delaunay triangulation and Voronoi tiling may be derived from each other. Given a set of populations whose geographic locations are known, there is only one possible Delaunay triangulation. Once a network connecting all the localities is obtained, each edge of the tiling is associated with the distance between the sites in the tiles taken from a distance matrix. See Goebl (this volume) for a graphic sketch of these procedures.

2.2.2.2 The Algorithm

Monmonier's maximum-difference algorithm (1973) is used to identify boundaries. As we noted above, each edge in the Voronoi tiling is normally associated with the distance between the sites which the "tiles" surround. To apply Monmonier, we associate the edge both with the linguistic or genetic distance directly, or with the residual from the respective geographic regression. By repeated selecting edges associated with large residuals, we aim to identify coherent geographic boundaries. Boundaries are traced perpendicular to the edges of the network. Starting from the edge across which the genetic or linguistic distance value is maximum[iii] and proceeding across adjacent edges, the procedure is continued until the boundary under construction either has reached the limits of the triangulation (map) or has closed on itself by forming a loop around a population. In case of multiple barriers (constructed sequentially in an order in which the researcher has some choice), the construction stops at a previously computed boundary. In the unusual case where two edges have the same value, the one linking to a triangle with higher total values is included in the boundary.

2.2.2.3 Robustness of Barriers

The execution of Monmonier's algorithm recalls the splitting process seen in the construction of phylogenetic trees: once a barrier passes across the edges of a triangle it can be extended only across one of the two remaining edges, in what we will define a "right or left" decision. To assess the robustness of computed barriers, we have developed a test based on the analysis of resampled bootstrap matrices. We repeated the procedure of finding boundaries using matrices computed on datasets from which random elements had been removed while others, randomly selected as well, appear more than once. As with phylogenetic trees, a score is associated with all the different edges that constitute barriers, thus indicating how many times each one of them is included in one of the boundaries computed from the N matrices (typically $N \geq 100$). In other words, if we have 100 matrices and we want to compute the first barrier, 100 separate barriers will be obtained. Such 100 different barriers (different in the sense that they have been computed on matrices obtained from modified datasets) are displayed in a single picture by increasing the thickness of the edges of the Voronoï tiling in proportion to the number of times they belong to one of the 100 barriers. If a pattern exists, whatever the modification of the original dataset, barriers should repeatedly emerge in certain areas of the plot. If barriers emerge everywhere in the plot then the results may be not robust (in terms of geographic differentiation).

The bootstrap procedure is intended to test if a given 'signal' (let's say a North/South differentiation) is reliably perceptible in the original data or not. If a majority of the items (i.e. single surnames, single words or linguistic features) exhibit a geographic pattern (…North/South), then such pattern will repeatedly emerge even when some items of the original dataset are randomly deleted or over-represented. In contrast if only a minority of the items suggests a pattern, after a repeated random modification of the original dataset, only few barriers will display it. In the latter case the pattern is not robust.

This procedure recalls the use of bootstrap in phylogenetic trees (Felsenstein 1985) and similar advantages accrue to this way of computing barriers, notably the way in which the confidence of the postulation of the barrier is reflected in the visualization.

3 Results

3.1 Surnames

The PCA plot of surname distances (Fig. 2) distinguishes the geographic positions of the 226 sampled localities fairly well. It is possible to identify a well-defined Limburg cluster (see Fig. 1 for a geographic map of Dutch provinces) and a second cluster constituted by north Brabant samples. Remaining eastern and western samples are close to each other in a continuous swarm of points, while Zeeland samples are intermediate. A more detailed analysis of the topology of the plot reveals that, within the swarm, there is no overlap between northeastern and northwestern samples. It must further be noted that the topology of samples suggests more heterogeneity in the south of the Netherlands than in the north, where samples are plotted closer to each other. The two axes account for approximately 30% of the total variance, and further axes (even through the 10th) still account for significant fractions of the total variance. Even if the low fraction of variance explained by the first two axes—a frequent phenomenon when large numbers of samples are analyzed—means that Figure 2 is less than optimally representative and suggests a rather complex topology of samples in the multidimensional space, it still provides a reasonable first approximation of overall variability. Further axes point to the specificities of both Limburg and Zeeland and, more generally, to the differences existing between the northern and the southern part of the country.

To understand the geographic variability of surnames, given that general correlations are not informative about local variability, we analysed the surname distance matrix with Monmonier's algorithm (not shown). The barriers computed highlight some differentiation zones in the northeastern provinces and along the northern border of the Limburg and Dutch Brabant provinces. Moreover, the Zeeland province appears as very fragmented, suggesting that surnames are very heterogeneous in such area with important differences from one location to another. These conclusions are reinforced by the analysis of 100 bootstrap matrices computed by a resampling procedure of original surname data (thick black lines in Fig. 3). Bootstrap analysis leads to a clearer picture since some minor barriers in the northern part of the country and in Zeeland disappear. We also note the presence of a major barrier across the former IJsselmeer (the internal sea in the north of the Netherlands visible in Fig. 1).

To focus on the variance that is not explained by geographic distance, we computed a general regression between geographic and surname distances after a double logarithmic transformation log y = 0.155 log x + k, which is equivalent to log y = log $x^{0.155}$ + k, which in turn is equivalent to y = exp(log $x^{0.155}$ + k) = exp(log $x^{0.155}$) exp(k). If we represent exp(k) as c, then the relation is y = c$x^{0.155}$. Geneticists are thus accustomed to analyzing the relation between geography and genetic variation as a power law; in fact this is standard in the analysis of genetic data. It is interesting to note that Seguy (1973) analyzed the relation between linguistics and geography as ling = geo$^{0.5}$, and that Heeringa and Nerbonne (2001) as ling = log (geo). In fact it is difficult to distinguish the analyses based on the logarithms of geography from those postulating a power law with a fractional coefficient of this size, so that we also apply the double logarithm transformations to the linguistic data.

Using this model we computed the expected surname distance, according to geography, between two sampled localities. The residual distance between observed and expected values can be either positive or negative, reflecting the influence of phenomena other than geography (history, systematic errors in data recording, etc). In figure 3 (solid gray lines) we show the Monmonier analysis of the matrix of residuals. Besides some very local barriers (numbered as ‘2’; ‘5’; ‘6’; ‘8’; ‘9’; ‘14’; ‘16’), previously observed patterns are confirmed—with the exception of the IJsselmeer boundary, which disappears. Methodologically, it is interesting to note that this latter boundary was traced across some of the longest edges of the Monmonier triangulation. As a consequence, the IJsselmeer boundary mirrors a surname differentiation related to the longer geographic distances, if compared to the average length of the Delaunay triangulation edges, which naturally emerge when comparing the samples on the opposite sides of inland sea. Seen from this perspective, the IJsselmeer does not seem to have been a substantial geographic barrier to internal migrations.

3.2 Dialects

With an identical methodology we analyzed the dialect data of the Netherlands. The matrix of Levenshtein distances is visualized in the bidimensional PCA plot of Figure 4, which suggests very good structure in the dialect data. Low Saxon and Low Franconian dialects are grouped into separate clusters, while Frisian samples are represented by three different clusters that describe (rural) Frisian, archaic Frisian (Hindeloopen, Schiermonnikoog, Terschelling island), and Friso-Franconian varieties (Frisian cities, Midsland, Ameland island and Het Bildt). Intermediate between Friso-Franconian and Low-Saxon we find a small Friso-Saxon group (Westerkwartier and Stellingwerf). Gray dots represent varieties spoken in central Gelderland, while empty dots correspond to varieties of the Dutch province of Zeeland. The first and second axis account for 40,8 % and 36,7 % of total variance respectively. The second axis has been mirrored and the plot have been rotated to visually suggest the correlation between the topology of samples and their real geographic locations. We will not describe such classification in more detail since it has been already fully discussed (Heeringa 2004).

Not surprisingly, Monmonier boundaries (Fig. 5) confirm the PCA plot for the most part. We find a northwestern Frisian area (barriers ‘1’ and ‘2’), a small northeastern area (part of the province of Groningen surrounded by barrier ‘20’), a large northeastern area (Low Saxon), a large more or less southwestern area (Low Franconian), a small southwestern part (province of Zeeland, barrier ‘9’) and a small area in the southeast (part of the province of

Limburg encircled by boundaries ‘5’; ‘7’; ‘17’). The distinction between Low Saxon and Low Franconian is not unexpected. One of the best-known features which demonstrates this distinction is the pronunciation of the final [ən]-syllable. E.g. *lopen* ‘to walk’ is pronounced as [loːpm̩] in Low Saxon, and as [loːpə] in Low Franconian dialects. Fragmentation is found in Friesland due to the well-known cohesion among the urban, Friso-Franconian varieties.

As with surname data, we continued the analysis by computing, after a double logarithmic transformation, a linear regression model ($\log y = 0{,}287 \log x + \mathrm{k}$) between geographic and Levenshtein distances to obtain the matrix of residuals that is plotted in the PCA analysis of Fig. 6. This is a novel treatment of the linguistics data, which we discussed in 3.1 above. The residuals reflect variance that is unrelated to geographic distance in general, and in a way residuals correspond to the ideal case of linguistic differences that would obtain between locations that were equidistant. Therefore geographic-proximity or -distance virtually plays no role in residual distances. In this sense the proximity of samples in the PCA plot of figure 6 indicates that the same historical and social factors may be responsible for such similarity and vice versa. We find that the remaining structure in the multidimensional PCA plot, computed on residuals, is still striking and appears at some points to reflect geography after all, maybe suggesting that the influence of geography is not constant. See Goebl’s paper in this volume for a reflection on the variable effect of geography. Further research and an appropriate intellectual frame seem necessary to address such new issues, which might *a priori* be expected to shed light on the mechanisms of linguistic differentiation through space.

The shape of Monmonier barriers, based on the matrix of residual Levenshtein distances (Fig. 7), confirms the barriers previously found in Zeeland (boundaries ‘11’; ‘15’; ‘19’ in Fig. 5) as well as the Saxon dialect area that is still surrounded by several barriers (‘1’; ‘3’; ‘10’; ‘14’). The surface of the northern part of the Saxon area seems less contoured when compared to Figure 5, since its northern part (corresponding to the province of Groningen) is now geographically continuous with Friesland, but is separated from the province of Drenthe by boundaries ‘1’ and ‘3’. As in the original matrix of Levenshtein distances (Fig. 5), Friesland is still fragmented (as shown by the shape of barriers ‘2’; ‘5’; ‘7’; ‘9’) because of the dialect islands of the urban Frisian mixed varieties (Friso-Franconian) in the Frisian dialect *continuum*. A completely new feature of Figure 7 is the boundary that begins on the left (west), follows the border between North and South Holland and then veers south to pass vertically through the provinces of Utrecht and North Brabant. Even if this border has not been discussed extensively in previous studies, so that we cannot easily compare alternative explanations about its meaning, the border is nonetheless interesting since it could be attributed either to heterogeneous transcriptions (Heeringa 2004) or to latent linguistic structures emphasized in some traditional maps (Lecoutere 1921).

4 Discussion

The major aim of the study was to evaluate to what extent the patterns of geographic variation of surnames overlap with those of linguistic diversity.

Family names are a specific part of language. Therefore, their interest as a proxy to Y-chromosome genetic diversity has sometimes been regarded as weak because they were also expected to be influenced by extra-chromosomal factors, i.e. the pressure of the linguistic environment. If this were true, such pressure would always be detectable—whatever the context. Following the geographical approach used here, and thus focusing on the barriers where geographic influence is insufficient as an explanation of genetic or linguistic difference, we note no striking correspondences between the two markers, e.g. in comparing the areas of differentiation in Figures 2 and 4; 3 and 7. We can then conclude that the pressure of the linguistic environment on surnames is absent or negligible and reasonably extend our claim to any future work addressing the comparison of surnames and linguistic markers. With respect to geographic distribution, *surnames are not words*.

To be sure distributions of linguistic and genetic variation correlate very significantly (r = 0.417***,[iv] where significance was calculated using Mantel test (1967) on the 74 sites common to the surname and dialect samples). But this just reflects the correlation existing between linguistic and geographic distances (N= 252; *r* = 0.546***) on the one hand, and between surname and geographic distances (N= 226; *r* = 0.507***) on the other. Because both pronunciation and surnames correlate strongly with geography, they seem to be correlated with each other (much as shoe size and reading ability correlate in children because both correlate strongly with age). But there is no correlation between matrices of residual Nei and Levenshtein distances, i.e. there is no correlation between surname and linguistic differences once their common dependence on geographic distance has been included in a statistical model.

If we describe the situation from the point of view of a multiple regression model in which we test geography and surname differences as independent predictors of linguistic distance, then the two predictors are collinear, leading a hasty analysis to attribute influence to both predictors, where a careful analysis in fact displays none. The correlation between linguistic and surname markers is entirely explained by their common collinearity with geography.

In fact we may strengthen our own conclusion that in the Netherlands there has been no demonstration of a relation between linguistics and surnames by noting the differences between the model used here and those used in our earlier dialectometric work. Nerbonne et al. (1999) calculated a correlation coefficient of (r=0.68**) using an overlapping 100-element set drawn from the same full data set (that includes the Flemish part of Belgium) from which the sample used in this paper was drawn, but they used a linear regression model rather than the power law (doubly logarithmic) model used here. The linear model clearly explains a great deal more linguistic variance than the power law model. Heeringa and Nerbonne (2002) use a logarithmic model, and although their data set yielded an unusually high correlation, we have found in general that logarithmic models function best. It appears that the optimal linguistic model takes a logarithmic form, in distinction to the power law relations favored by geneticists. This reinforces our main conclusion, viz. that the linguistic and genetic patterns of variation are different, even if they are both conditioned strongly by geography.

Our conclusions strikingly differ from those of a similar study comparing surnames and dialects in France by Scapoli et al. (2005). But we suspect that these authors failed to control their matrices of genetic and linguistic distances for common geographic conditioning, leading them to the incorrect conclusion that language similarity is an indicator of genetic kinship even at local levels. This may be occasionally true but needs to be systematically verified by analysing residuals.

Concerning the Netherlands, the only close match between the variation of surnames and dialects is found in the province of Zeeland, which is also geographically apart from surrounding areas (Figs. 3; 5; 7). This special status of Zeeland may be due to its geography, since it consisted until recently of several islands, which, starting in the XIV century (*Atlas van Nederland 1996*), increased in size and—thanks to land reclamation efforts—eventually turned into peninsulas at the beginning of 20th century. Relative social and geographic isolation, together with an economy based on fishing and trade, may have maintained and reinforced a closed social structure still visible in surname and dialect variability. A diversity that is also mirrored by the different agriculture practice between "insular" Zeeland and Zeeland Flanders (see Fig. 1). Finally, an additional and complementary explanation is represented by more intense contacts with the adjacent western Flemish area (Belgium).

The computation of a regression model leading to matrices of residuals is expected to better illuminate demography (surnames) and social patterns (dialects), both of which are related to history (in a broad sense) rather than to geography. As a consequence, we can interpret the surname barriers found along the northern borders of Zeeland, North Brabant and Limburg as the results of historical phenomena. The significance of such major separations is confirmed by bootstrap matrices visualized through the Monmonier algorithm and by the analyses of residual distances as analyzed with it (barriers '4';'18' in Fig. 3)—which brings up to new issues.

As we said the distribution of surnames only mirrors demographic phenomena,[v] without any influence from their linguistic environment. Therefore, when we seek explanations for such barriers, which we see that linguistic culture does not support, we must turn to other factors. In this case we are struck by the correspondence between the border induced by common surnames and the border of the Catholic area of the Netherlands (Fig. 8). The strength of the surname border suggests that the frequency of inter-marriages between Catholics and Protestants was very low. This religious distinction may have acted as a social boundary, thus increasing surname differences between populations on the border's sides. The fact that there is no linguistic evidence (Fig. 7) of such separation means that more casual social contacts and interchange were not diminished between Catholic and Protestant populations. Communication proceeded in spite of a profound social cleft.[vi]

Our article focused on patrilineal genetic differences (surnames) and their relation to culture and its transmission. Intuitively, the observed incoherence between patrilineal markers of genetic relatedness and linguistic space-distributions may be regarded as misleading, once culture (language) is assumed to be transmitted matrilinearly. This was a concern expressed by one of the scholars who reviewed our article. In other words, his question was *would our findings have been the same if surnames were maternally transmitted?* Some readers may remember a popular study pointing to the greater dispersion of females when compared to males (Seielstad et al. 1998). Such results, based on the comparison between specifically-paternal (Y-chromosome) and specifically-maternal (mitochondrial DNA) genetic markers, were explained in terms of patrilocality.[vii] Even if alternative explanations (Dupanloup et al. 2003) and different conclusions (Wilder et al. 2004) have been provided since, we note that this debate mainly concerns the deep time-frame of pre-historical times and not the time-frame of the surname data. Family names only portray the variability of populations as if "Adam(s)" and "Eve(s)" lived at the time of surname introduction (two centuries ago in the Netherlands). If surnames can be a proxy to genetics, they are effective only in the depiction of recent demographic phenomena.

Even considering that in Europe "matrimonial migrations" generally consist of only a few kilometers and that we are dealing with differences that can be traced back for eight generations only, it is likely that patrilocality plays a role in our dataset, meaning that females move more than males. A very recent paper (Gagnon et al. 2006, in press) based on the "core-fringe model" by Heyer (1993) suggests that sons inherit their propensity to migrate from their fathers, while such transmission is largely absent among women. The intergenerational dependency in the probability of migration implies that the pool of migrants is not a representative sample. The social explanation is that, once settled somewhere, the newcomers seldom become the owners of the land (or of other means of production) so their sons are more likely to move out. In this process their new Y-chromosome variants tend to disappear in the next generation, while daughters of immigrants can become part of the new community by marriage and, therefore, have higher chances to enrich the local pool of genetic diversity. If such migrational behavior partially counteracts the effects of patrilocality, females still migrate more than males. To answer the thorny question of our reviewer: if women transmitted Dutch surnames we would have computed pairwise surname distances smaller than patronymic ones. The picture would have been the same but with a lower level of detail (more migrations imply smaller local differences). Therefore there are no reasons to expect a higher correlation between surname and dialect variability if female lineages were taken into account. Moreover, concerning the role of the mother in language transmission, we also remind that most linguistic studies emphasize the importance of the peer group, outside the immediate family, in influencing adolescent patterns of speech, and the general suspicion is that these are normally then resistant to change in later life. This would be a valuable area for further research.

Besides the major research question of the article, we think that some methodological outcomes should be reviewed. First of all, the use of matrices of residual linguistic distances obtained after the computation of a regression between geographic and linguistic distances has been rewarding. This approach has enabled us to visualize computationally the geographic affinity of the province of Groningen to the Frisian speaking area (Fig. 7). This closer relation may mirror the early linguistic history of the Groningen area, where some Frisian varieties were last spoken in the early part of the 16th century (see Hoekstra, 2001, p. 139, Niebaum, 2001, p. 431). Besides some few contemporary phonetic features, there has been no linguistic evidence that a different language was once spoken in this area, thus underscoring the effectiveness of the methodological approach we undertook. But see Spruit (this volume) for an analysis of the syntactic variability in which the north of the Netherlands appears much less heterogeneous than it does in lexical and phonetic analyses.

We should also like to emphasize the value of Monmonier's algorithm for linguistic applications (see also Manni et al. 2004 for further discussion). The algorithm allows a geographical visualization of the variability in a distance matrix, showing where differentiation is located. Unless there is a perfect correlation between the variable under study and geographic distances (meaning that there are no major barriers), the Monmonier method adds geographical detail to the multidimensional analyses such as multidimensional Scaling or PCA, which are still the primary analytical tools for appraising linguistic variability. See also Goebl (this volume) for an examination of the variability of the influence of geography on dialect.

At first blush, barriers computed with the Monmonier's algorithm might remind linguists of bundles of isoglosses. While the Monmonier's approach may only be applied to dialectometrical data, since it requires numerical data, it is true that it mirrors the same goal of a synthetic representation of variability that isogloss bundles were likewise designed to operationalize.

Even though the methodologies for analyzing genetic and linguistic data are becoming very similar, at a conceptual level several differences still exist. The architecture of this paper reflects one of them: population geneticists are more interested in the differences between populations than in homogeneity or similarity. The main reason lies in the low differentiation of human populations on a global scale. Only 15% of the variance of the human genome is explained by differences between groups of populations, whereas individual differences explain 85% of the total variance (Barbujani et al. 1997). In other words, two individuals living in the same area are likely to be genetically more different than two individuals living in different continents.[viii] The above reasons explain why, at small geographic scales, the leading hypothesis of human population geneticists was to expect low or non-existent genetic differentiation. Since similarity (homogeneity) is expected, all the practical and conceptual work of the discipline has been focused on the detection of differences.

Linguists, on the other hand, have often focused on the geographic distribution of linguistic variety and its composition—regardless of its ultimate explanation. It is not unfair to say that the geographic conditioning of language variety has been studied as much for the light it sheds on the nature of linguistic structure as for the improved view of (social) history it enables. While linguistic studies, like genetic ones, are keen to catalogue the differences between language varieties that are really very similar, there has been no similar success in quantifying the degree to which language varieties (seen from the perspective of all existing varieties) might differ. Perhaps some further cross-fertilization from genetics into linguistics might be worthwhile.

In conclusion, the development of computational linguistics studies, as well as the application of spatial and statistical analyses enabled by this discipline, will tell us if dialect *continua* are a satisfactory view of linguistic variability or if more innovative interpretations of the geographic patterns of dialect variation are needed, especially when dealing with old or ancient linguistic patterns. We hope that future directions of investigation will be focused on interdisciplinary understanding, exhaustively discussed by Goebl (1996), of the interrelations existing between surnames and dialects. Since we are also investigating the varying degrees to which variation in different linguistic levels (pronunciation, vocabulary, and syntax) are geographically conditioned (Heeringa and Nerbonne, to appear), we shall keep in mind that vocabulary distributions may offer an interesting comparison to surnames.

Aknowledgements:

We are indebted to Bruno Toupance for providing help in data processing, to Pierre Darlu and to Bob Shackleton for commenting on the manuscript, to Isabelle Dupanloup for inspiring us and to Alain Gagnon for providing insightful references. We also thank George Welling for his critical remarks concerning the history of the Netherlands and Meindert Schroor for improving our geographic insight. A special thank to Evelyne Heyer who encouraged such study and provided the best working environment.

References

Atlas van Nederland (1986). Stichting Wetenschappelijke, ’s-Gravenhage.

Barbujani G., Magagni A., Minch E. and Cavalli-Sforza L.L. (1997) An apportionment of human DNA diversity. *Proceedings of the National Academy of Sciences of the United States of America*. 94:4516-9.

Barrai I., Rodriguez-Larralde A., Manni F. and Scapoli C. (2002) Isonymy and Isolation by Distance in the Netherlands. *Human Biololgy* 74:263-283.

Cavalli-Sforza L.L. and Bodmer W. (1971) Human population Genetics. San Francisco: Freeman.

Cavalli-Sforza L.L., Piazza A., Menozzi P. and Mountain J. (1989) Genetic and linguistic evolution. Science 244:1128-9.

Cavalli-Sforza L.L., Menozzi P. and Piazza A. (1994). *The history and geography of human genes*. Princeton (N.J.), Princeton University Press.

Chambers J.K. and Peter Trudgill (1998) Dialectology, 2nd edition, Cambridge University Press, Cambridge, UK.

Chen K.H., and Cavalli-Sforza L.L. (1983) Surnames in Taiwan: interpretations based on geography and history. *Human Biology* 55:367-374.

Crow J.F. and Mange, A.P. (1965) Measurements of inbreeding from the frequency of marriages between persons of the same surnames. *Eugenic Quarterly* 12:199-203.

Crow J.F. (1980) The estimation of inbreeding from isonymy. *Human Biololgy* 52:1-4.

Daan, J. and Blok D.P. (1969**).** *Van randstad tot landrand. Toelichting bij de kaart: dialecten en naamkunde*. Amsterdam, N.V. Noord-Hollandsche uitgeversmaatschappij.

Dupanloup I., Pereira L., Bertorelle G., Calafell F., Prata M.J., Amorim A. and Barbujani G. (2003) A recent shift from polygyny to monogamy in humans is suggested by the analysis of worldwide Y-chromosome diversity. *Molecular Ecololgy* 13:853-864.

Dyen I. (2004) Personal communication based on the discussion of the paper “Johannes Schmidt’s ‘Wave theory’ and the Homomeric method”, *Phylogenetic methods and the prehistory of languages* , workshop held at the McDonald institute for archaeological research, Cambridge (UK), 9-12 July. (All papers presented at the meeting will be published in the forthcoming book *Phylogenetic Methods and the Prehistory of Languages*, Peter Forster and Colin Renfrew eds., Oxbow books, 2006.)

Felsenstein J. (1985) Confidence limits on phylogenies. An approach using the bootstrap. *Evolution*, 39:783-91.

Gagnon A. and Toupance B. (2002) Testing isonymy with paternal and maternal lineages in the early Québec population: the impact of polyphyletism and demographic differentials. *American Journal of Physical Anthropology* 117:334-341.

Gagnon A., Toupance B., Tremblay M., Beise J. and Heyer E. (2006) Transmission of migration propensity increases genetic divergence between populations. *American Journal of Physical Anthropology* (in press, published online the 9 Dec 2005).

Goebl H. (1996). La convergence entre fragmentations géo-linguistique et géo-génétique de l’Italie du Nord, *Revue de Linguistique Romane* 60: 25-49.

Goebl, H. (this volume). The Salzburg Variety of Dialectometry.

Heek van, F. (1954) *Het geboorteniveau der Nederlandse Rooms-Katholieken.* Leiden.

Heeringa W.J. and Nerbonne J. (2001). Dialect areas and Dialect Continua. In: *Language Variation and Change*, Cambridge University Press, New York, volume 13, 2001, pp. 375-400.

Heeringa, W. and Nerbonne, J. (to appear) Taalvariatie in het Nederlandse dialectgebied: een analyse op basis van lexicon en uitspraak. in: *Nederlandse Taalkunde* 11(3), 2006.

Heeringa W., Nerbonne J., and Kleiweg P. (2002). Validating Dialect Comparison Methods. In: Wolgang Gaul and Gerd Ritter (eds.) *Classification, Automation, and New Media*. Proceedings of the 24th Annual Conference of the 'Gesellschaft für Klassifikation', University of Passau, Springer: Heidelberg, pp.445-452.

Heeringa W., and Gooskens C. (2004). Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data. In *Language Variation and Change* 16(3): 189-207.

Heeringa (2004). *Measuring dialect pronunciation differences*, Ph.D. Dissertation, University of Groningen, The Netherlands.

Heyer E. (1993). Population structure and immigration; a study of Valserine valley (French Jura) form the 17th century until present. *Annals of Human Biology* 20:565-573.

Hoekstra, E. (2001). Frisian Relics in the Dutch Dialects. In: H.H. Munske (ed.). *Handbuch des Friesischen/Handbook of Frisian Studies*. Niemeyer, Tübingen, 138-142.

Kimura M. (1983) *The Neutral Theory of Molecular Evolution*. Cambridge (UK): Cambridge University Press.

King T.E., Ballerau S.J., Schürer K.E. and Jobling M.A. (2005). Genetic signatures of coancestry within surnames. *Current biology* 16: 384-388.

Kohonen T. (1995) *Self-Organizing Maps*. Berlin: Springer.

Lasker G.W. (1985) *Surnames and genetic structure*. Cambridge University press, Cambridge (UK).

Lecoutere, C.P.F. (1921). *Inleiding tot de taalkunde en tot de geschiedenis van het Nederlandsch*. Brussel.

Malécot G. (1955b). The decrease of relationship with distance. *Cold Spring Harbor Symposia Quantitative Biology*, 20 , 52-53.

Manly, B.F.J. (1997) *Randomization, bootstrap and Monte Carlo methods in biology.* 2nd edition, Chapman and Hall.

Manni F. (2001) *Strutture genetiche e differenze linguistiche: Un approccio comparato a livello micro e macro regionale*. Doctoral thesis. Ferrara: University of Ferrara.

Manni F., Guérard E. and Heyer E. (2004) Geographic patterns of (genetic, morphologic, linguistic) variation: how barriers can be detected by using Monmonier's algorithm. *Annals of Human Biology* 76:173-90.

Manni F, Toupance B, Sabbagh A and Heyer E. (2005), New method for surname studies of ancient patrilineal population structures, and possible application to improvement of Y-chromosome sampling. *American Journal of Physical Anthropology* 126:214-28

Mantel, N.A. (1967) The detection of disease clustering and a generalized regression approach. *Cancer Research* 27:209-20.

Monmonier, M. (1973) Maximum-difference barriers: an alternative numerical regionalization method. *Geographical Analysis*, 3:245-61.

Nerbonne J, Heeringa W. and Kleiweg P. (1999). Edit Distance and Dialect Proximity. In: Sankoff D. and Kruskal J. (eds.). *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison*. Stanford: CSLI Press. pp.v-xv.

Niebaum, H. (2001). Der Niedergang des Friesischen zwischen Lauwers und Weser. In: H.H. Munske (ed.). *Handbuch des Friesischen/Handbook of Frisian Studies*. Niemeyer, Tübingen, 430-442.

Peakall R. and Smouse P.E., (2001) *GenAlEx vs. 5: Genetic analysis in Excel. Population genetic software for teaching and research.* Australian National University, Canberra, Australia.

Sankoff D. and Kruskal J. (eds.) (1999). *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison*. Stanford: CSLI Press.

Scapoli C, Goebl H, Sobota S., Mamolini E., Rodriguez-Larralde A. and Barrai I. (2005) Surnames and dialects in France: Population structure and cultural evolution. *Journal of Theoretical Biology* 237:75-86.

Schmidt J. (1872) *Die Verwandtschaftsverhältnisse der indogermanischen Sprachen.* H. Böhlau, Weimar, Germany.

Séguy J. (1971) La relation entre la distance spatiale et la distance lexicale. *Revue de Linguistique Romane*, 35:335-357.

Seielstad M.T., Minch E. and Cavalli-Sforza L.L. (1998) Genetic evidence for a higher female migration rate in humans. *Nature Genetics* 20:278-280.

Sokal, R. (1988) Genetic, geographic and linguistic distances in Europe. *Proceedings of the National Academy of Sciences of the United States of America.* 85 (March), 1722-1726.

Spruit, M. (this volume) Measuring Syntactic Variation in Dutch Dialects.

Voronoi M.G. (1908) Nouvelles application des paramètres continus à la théorie des formes quadratiques, deuxième mémoire, recherche sur le paralléloedres primitifs. *Journal für die reine und angewandte Mathematik,* 134**:** 198-207.

Wilder J.A. Kingan S.B., Mobasher Z., Pilkington M.M. and Hammer M.F. (2004). Global patterns of human mitochondrial DNA and Y-chromosome structure are not influenced by higher migration rates of females versus males. *Nature Genetics* 36: 1122-1125.

Yasuda N, and Morton N.E. (1967) Studies on human population structure. In: Crow JF, Neel JV, editors. *Third International Congress of Human Genetics*. Baltimore: Johns Hopkins Press. p. 249-265.

Yasuda N., and Furusho T. (1971) Random and non random inbreeding revealed from isonymy study. I. Small cities in Japan. *The American Journal of Human Genetics* 23:303-316.

Yasuda N., Cavalli-Sforza L.L., Skolnick M. and Moroni A. (1974) The evolution of surnames: an analysis of their distribution and extinction. *Theoretical Population Biology* 5:123-142.

Zei G., Matessi R.G., Siri E., Moroni A. and Cavalli-Sforza L.L. (1983) Surnames in Sardinia. I. Fit of frequency distributions for neutral alleles and genetic population structure. *Annals of Human Genetics* 47:329-352.

Captions to figures:

Figure 1

Map of the Netherlands showing the location of the Dutch provinces.

Figure 2

Principal component analysis (PCA) of the surname differences in the Netherlands (Nei's distances). Two almost distinct clusters corresponding to North Brabant and Limburg samples can be identified. Remaining samples, belonging to the other provinces, cluster in a same swarm of points. Further details can be found in the text. 1st axe explains 17,6 % of the total variance, while the 2nd axe accounts for 11,7 % of it. The third and fourth axes (not shown) explain 5.8% and 4.3% of the total variance and highlight the diversity of Limburg and Zeeland provinces respectively. Further axes (5th, 6th, 7th) point, in different ways, to the differences existing between the north and the south of the Netherlands.

Figure 3

1) Black thick lines correspond to barriers obtained with the Monmonier algorithm on 100 matrices of surname distances computed according to the Nei's method. The different matrices were computed by a bootstrap resampling of original surnames. Only the first 20 barriers are shown (2000 barriers in total). The thickness of barriers is proportional to their bootstrap score and barriers whose score is lower than 50% are not shown (see scale).

2) Gray lines, correspond to barriers obtained from a matrix of residual surname distances. After a linear regression between the logarithms of geographic and Nei's distances, we computed the expected surname distance according to the regression. Such values were subtracted from observed ones, thus leading to the residuals. The first 20 barriers are shown (numbered at both extremes from '1' to '20').

The Delaunay triangulation is visualized by a thin gray network.

Figure 4

Principal components plot on the basis of 252 Dutch dialects. Low Saxon and Low Franconian dialects are grouped into separate clusters, while Frisian samples are represented by three different clusters that describe (rural) Frisian, archaic Frisian (Hindeloopen, Schiermonnikoog and Terschelling island), and Friso-Franconian varieties (Frisian cities, Midsland, Ameland island and Het Bildt). Intermediate in between of Friso-Franconian and Low-Saxon we find a small Friso-Saxon group (Westerkwartier and Stellingwerf). Gray dots represent varieties spoken in central Gelderland, while empty dots correspond to varieties of the Dutch province of Zeeland. The first and second axis account for 40,8 % and 36,7 % of total variance respectively. The second axis has been mirrored and the plot have been rotated to visually suggest the correlation between the topology of samples and their real geographic locations.

Figure 5

Barriers (solid black lines) obtained with the Monmonier algorithm on a matrix of dialect (Levenshtein) distances between 252 localities. The first 20 barriers are shown (numbered from ‘1’ to ‘20’). A thin gray network visualizes the Delaunay triangulation.

Boundaries identify areas corresponding to Friesland (local barriers corresponding to different Frisian varieties are displayed in gray to provide a clearer representation) and to parts of Zeeland and Limburg. On a wider scale, it appears that some major barriers well depict the geographic locations where Low Franconian and Low Saxon varieties are spoken (see labels). Further details can be found in section 3.2.

Figure 6

(A) Principal Components plot of the variability of residual dialect distances after a linear regression between the logarithms of Levenshtein distances and their corresponding geographic distances. Both axes have been mirrored for a better display.

(B) Map of the Netherlands showing the correspondence between the multidimensional position of samples (A) and their real geographic location. Different symbols do not necessarily correspond to clusters; they are just intended to help the comparison between the topology of the PCA plot and the geographic map.

Figure 7

Barriers (solid black lines) obtained with the Monmonier algorithm on a matrix of residual dialect distances (to be compared with the identical analysis on the original matrix in figure 5). The provinces of Friesland and Groningen appear as linguistically continuous but see the text for further details. The first 20 barriers are shown (numbered from '1' to '20'). As in Fig. 5 barriers corresponding to different Frisian varieties are displayed in gray. The Delaunay triangulation is visualized as a gray network.

Figure 8

Map showing the frequency of Roman Catholics in the Netherlands in 1954. Redrawn from van Heek (1954).

Fig. 1

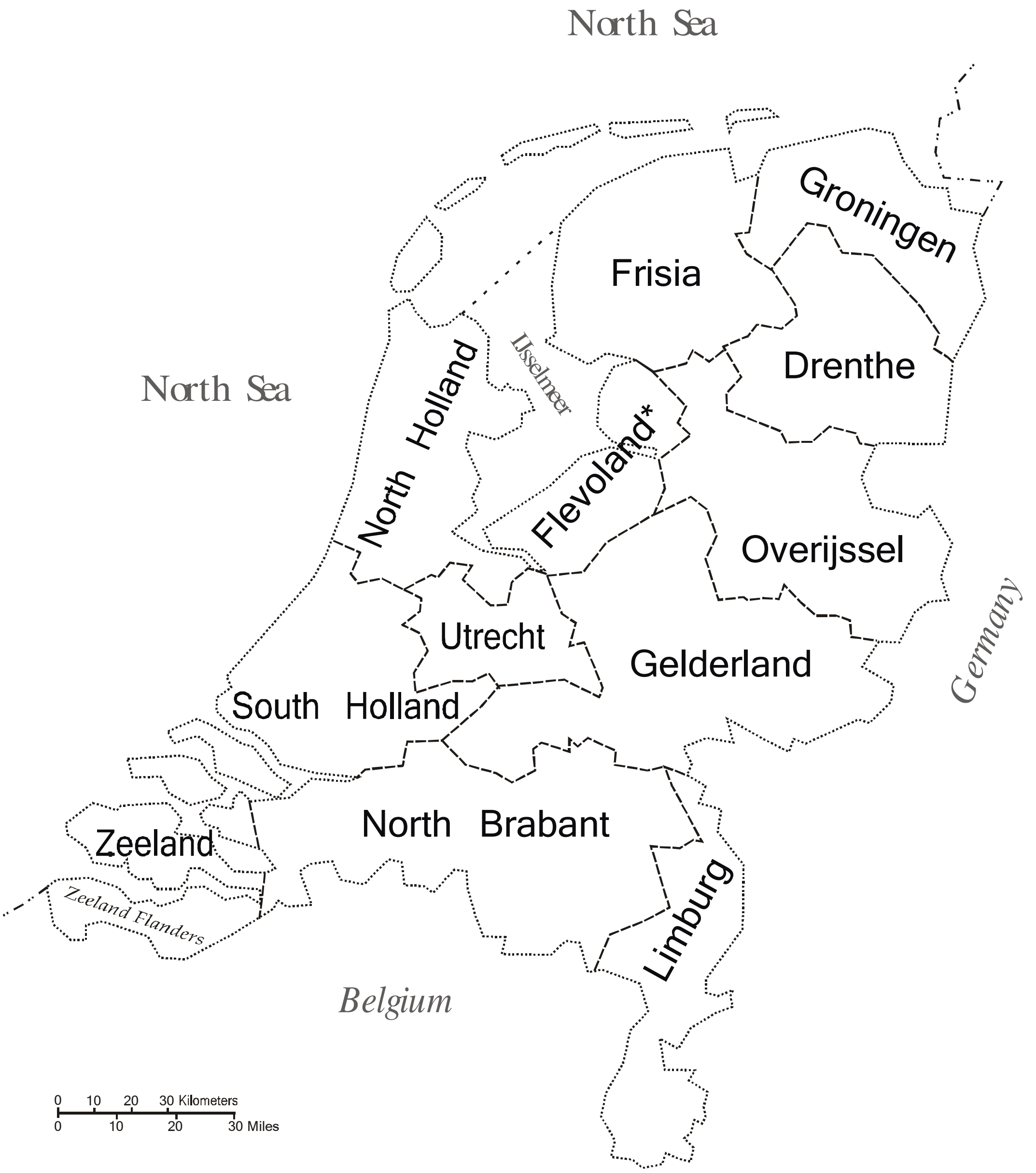
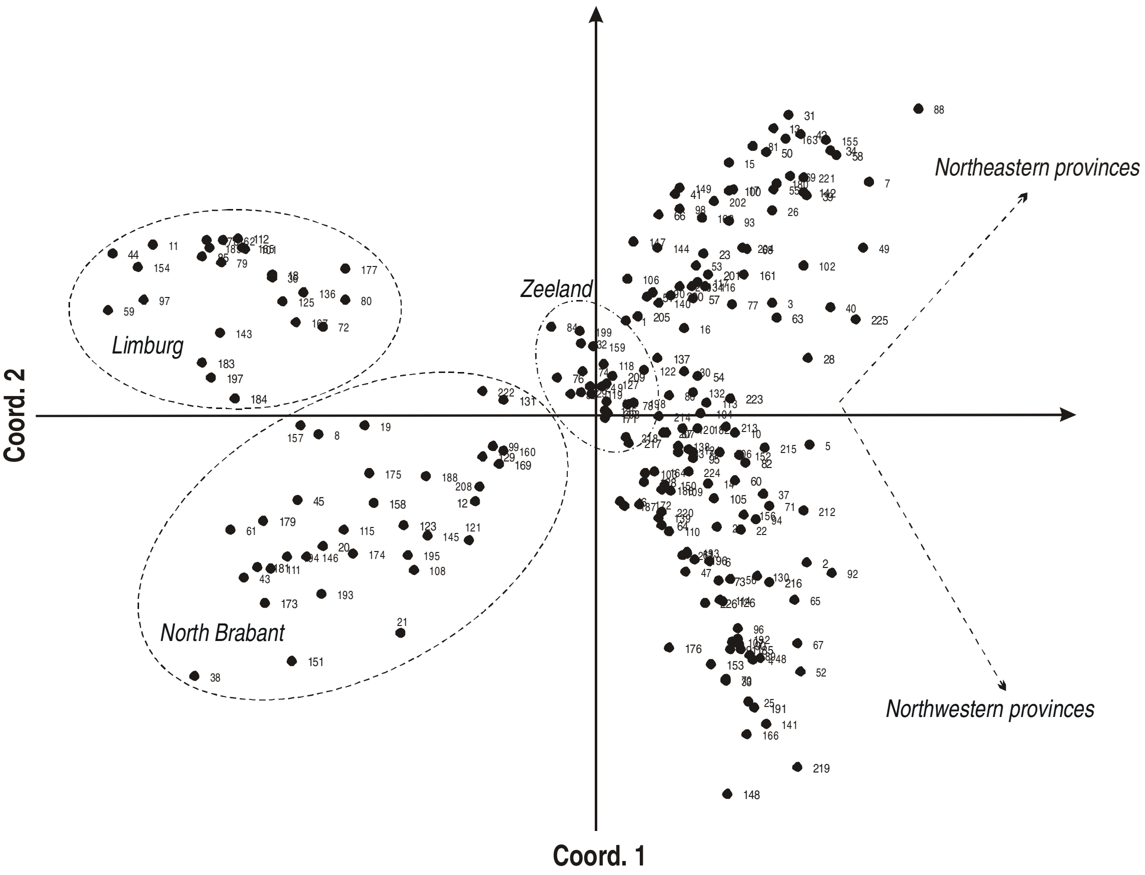
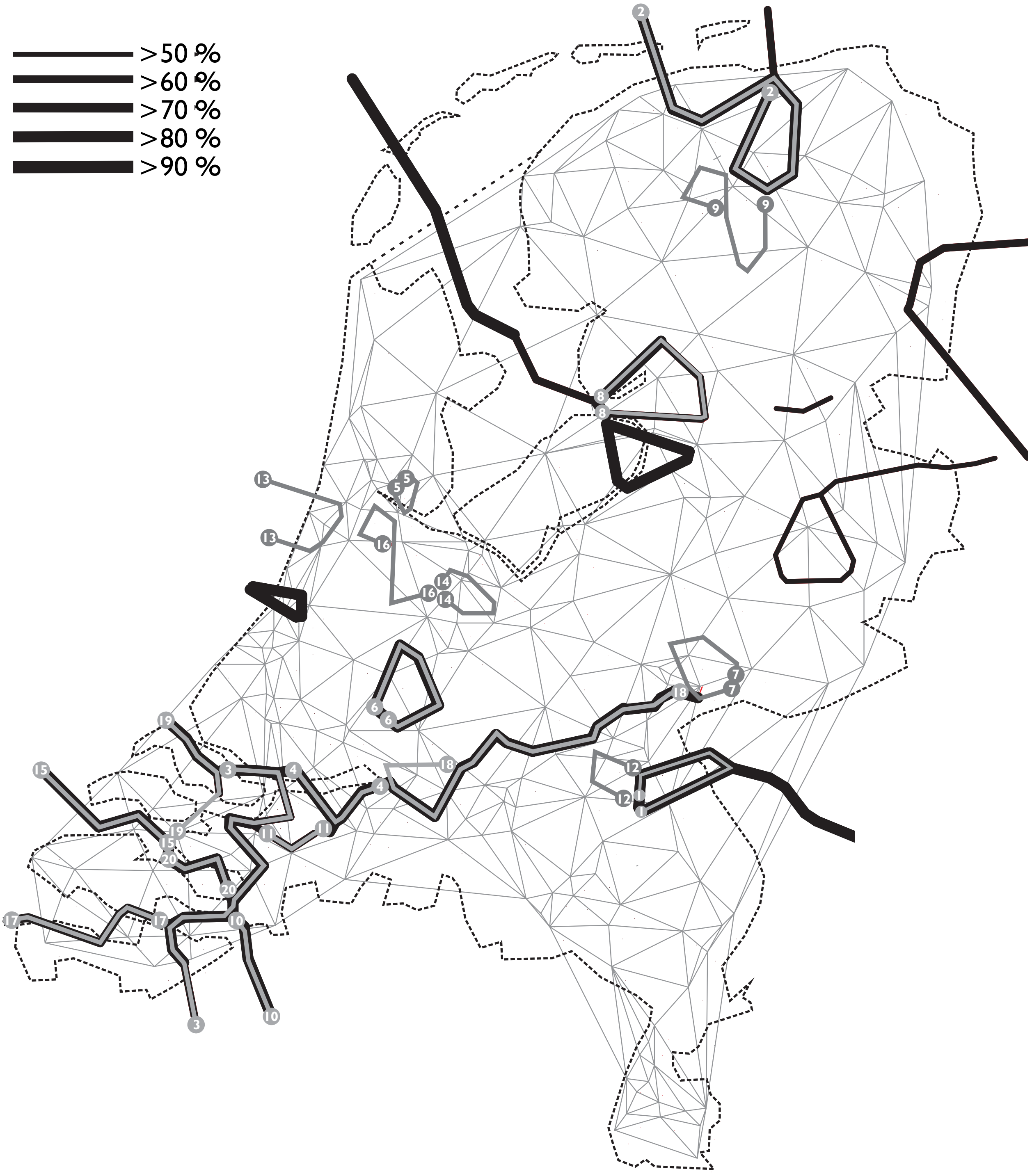

North Sea
North Sea
Groningen
Frisia
Drenthe
IJsselmeer
North Holland
Flevoland*
Overijssel
Utrecht
Gelderland
Germany
South Holland
North Brabant
Zeeland
Zeeland Flanders
Limburg
Belgium
0 10 20 30 Kilometers
0 10 20 30 Miles

Fig. 2



Northeastern provinces
Northwestern provinces
Limburg
North Brabant
Zeeland
Coord. 2
Coord. 1
88
31
13
163
155
34
58
81
50
15
7
149
41
100
202
221
142
66
98
93
26
147
144
23
204
49
102
53
201
161
106
140
57
77
3
63
205
1
16
40
225
28
137
122
30
54
132
223
214
213
10
215
5
217
224
60
105
37
71
212
110
22
2
92
47
130
216
65
96
67
176
153
52
25
191
141
166
219
148
11
44
154
112
79
177
59
97
136
125
80
72
143
183
197
184
84
199
159
118
209
127
76
222
131
157
8
19
99
160
169
175
188
208
12
158
45
179
61
115
123
145
121
20
174
195
108
43
173
193
21
151
38

Fig. 3



>50 %
>60 %
>70 %
>80 %
>90 %
2
2
9
9
8
8
13
5
5
13
16
14
16
14
7
18
7
6
6
19
12
15
3
4
18
4
12
1
1
11
19
15
11
20
20
17
17
10
3
10

Fig. 4

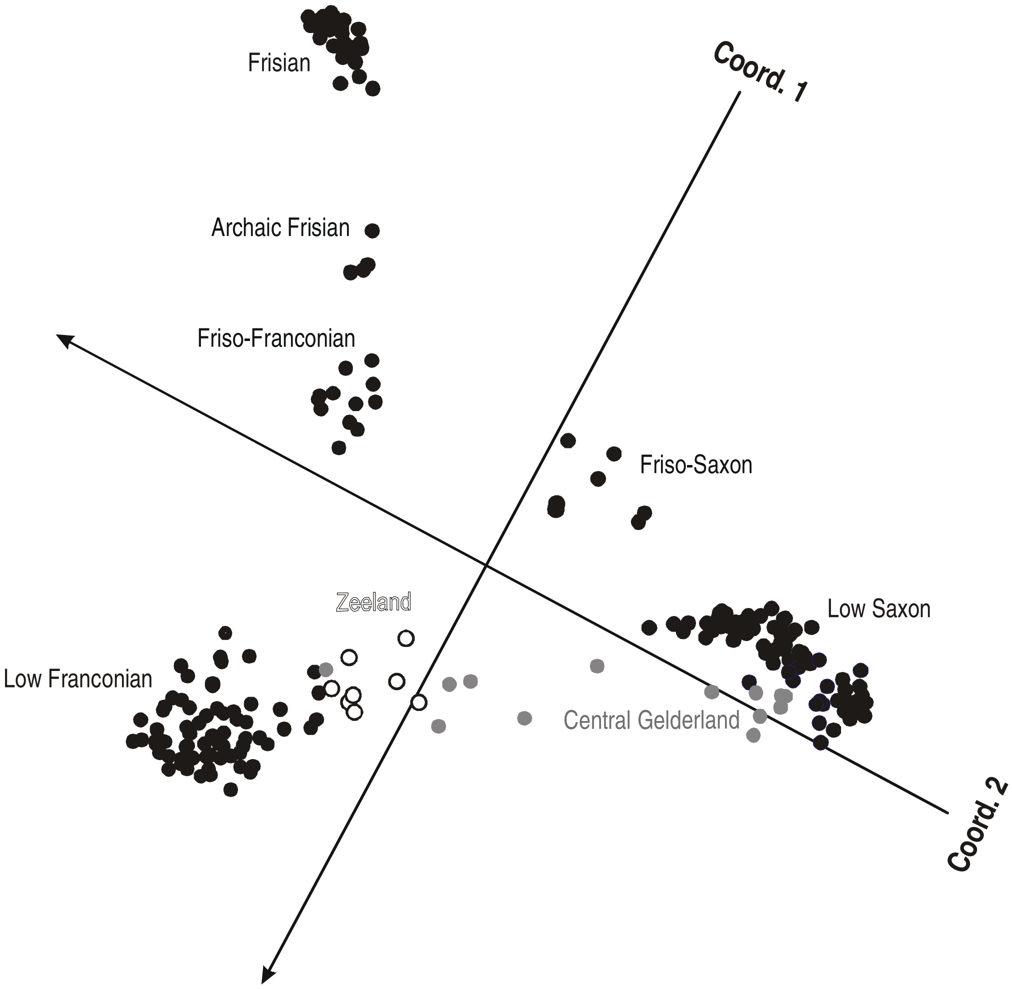

Frisian
Coord. 1
Archaic Frisian
Friso-Franconian
Friso-Saxon
Zeeland
Low Saxon
Low Franconian
Central Gelderland
Coord. 2

Fig. 5

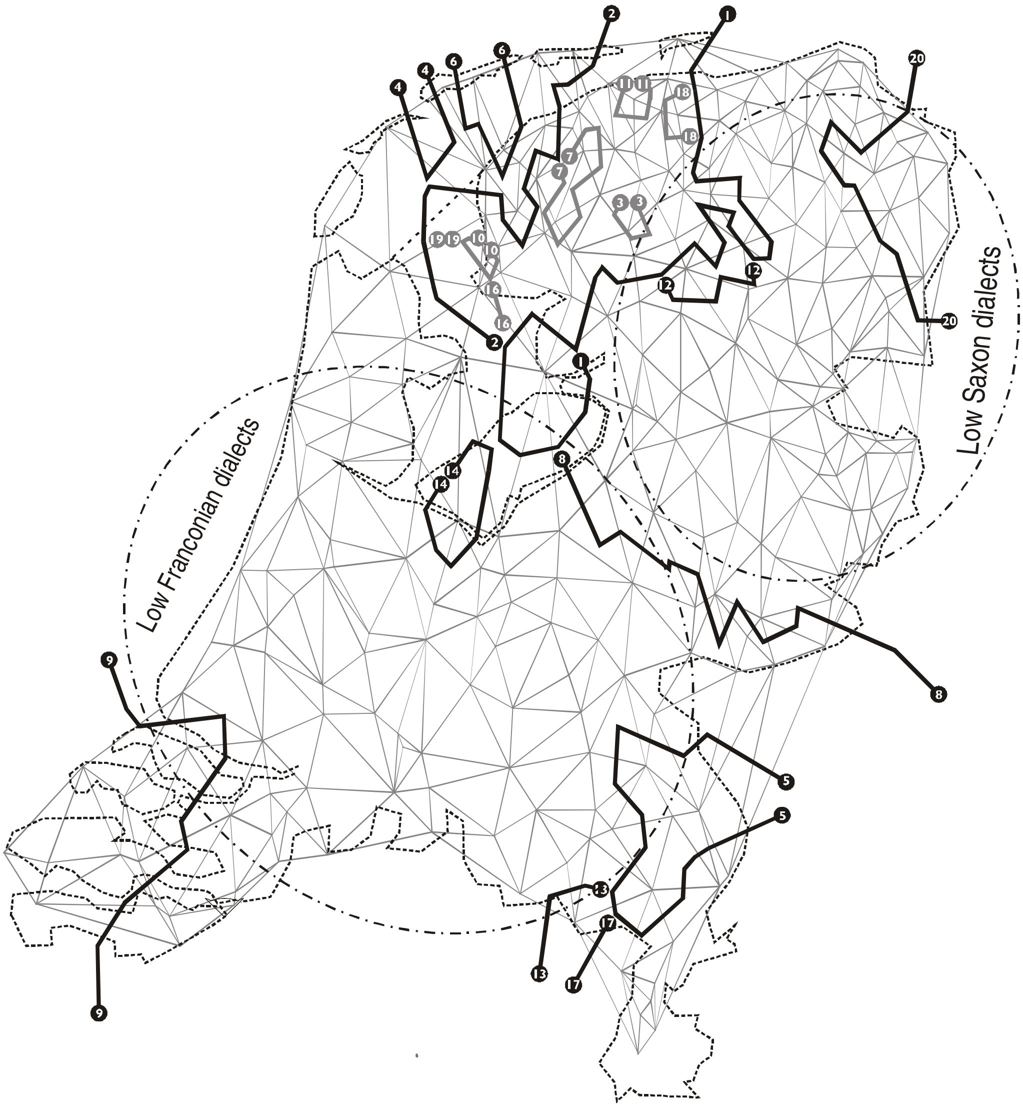

Low Franconian dialects
Low Saxon dialects
1
2
3
4
5
6
7
8
9
10
11
12
13
14
16
17
18
19
20

Fig. 6

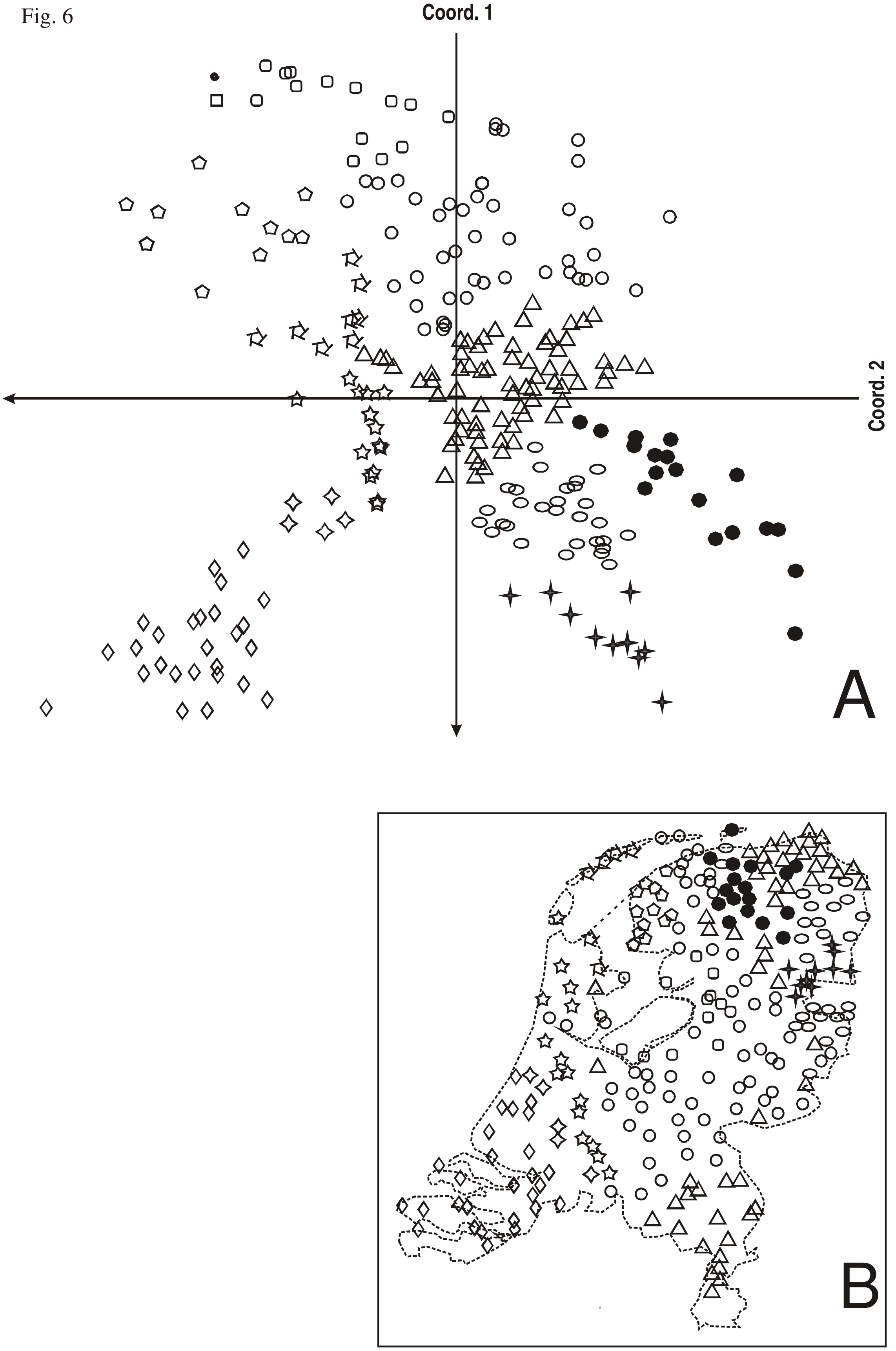


Fig. 7

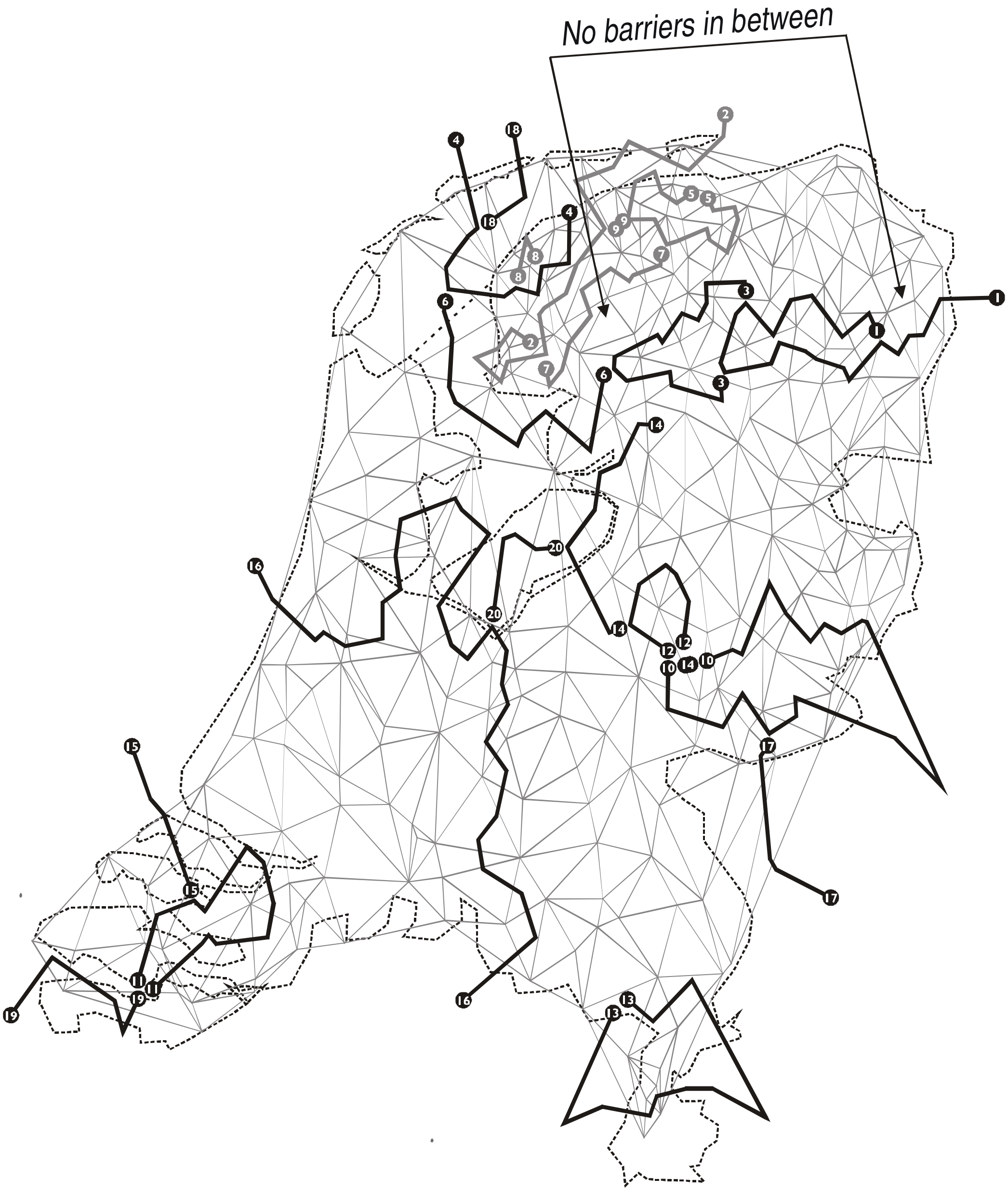
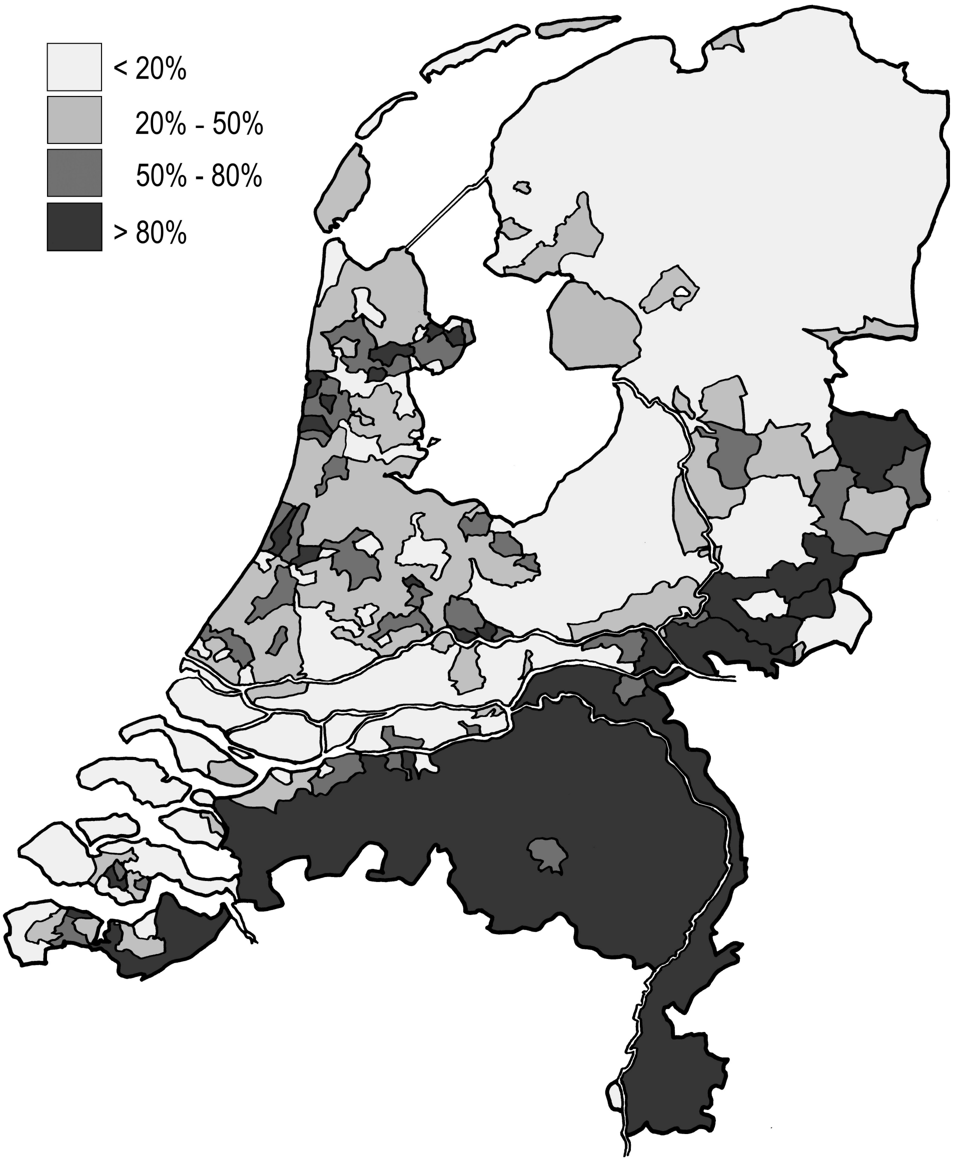

No barriers in between
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20

Fig. 8



NOTES

[i] Isonymy, a measure of surnames' overlap in a population, estimates the degree to which the population is related, i.e. its consanguinity. Real isonymy is obtained by counting the number of marriages where partners have the same surname (isonymic marriages). Isonymy can by estimated by computing the probabilities of isonymic marriages for all surnames. The probability depends on the relative frequency and the number of all the different surnames. This latter measure is called 'random isonymy' and assumes that the choice of the partner is not influenced by his family-name, being—in this respect—completely random. In a village where all the inhabitants have different surnames isonymy is 0, in another village where all the inhabitants have the same surname isonymy is 1.

[ii] A specific software, "Barrier 2.2", is available at **`http://www.mnhn.fr/mnhn/ecoanthropologie/software/barrier.html`**

[iii] To avoid a frequent misunderstanding we note that the edge associated with the maximum distance do not need to be on the borders of the triangulation, being such case more an exception than the rule. If this is the case, the extension of the boundary occurs in one direction only; otherwise, it takes place in two directions simultaneously.

[iv] Here we adopt the standard scientific notation where (*) means a significance level of 5%, while (**) and (***) respectively indicate significance levels of at 1% and 0.1%.

[v] Population genetic differences heavily depend on demographic phenomena.

[vi] This claim specially applies to the rural Netherlands. It must be emphasised that in middle of the 17th century a large proportion of the Dutch population was already living in towns and cities.

[vii] By patrilocality we mean a residential pattern in which a married couple settles in the husband's home or community.

[viii] By the way, this is the reason why the scientific definition of races does not apply to humans; they are too similar to be partitioned into separate, biologically meaningful groups.