Franz Manni

Some applications of linguistic research in human population genetics and vice versa.

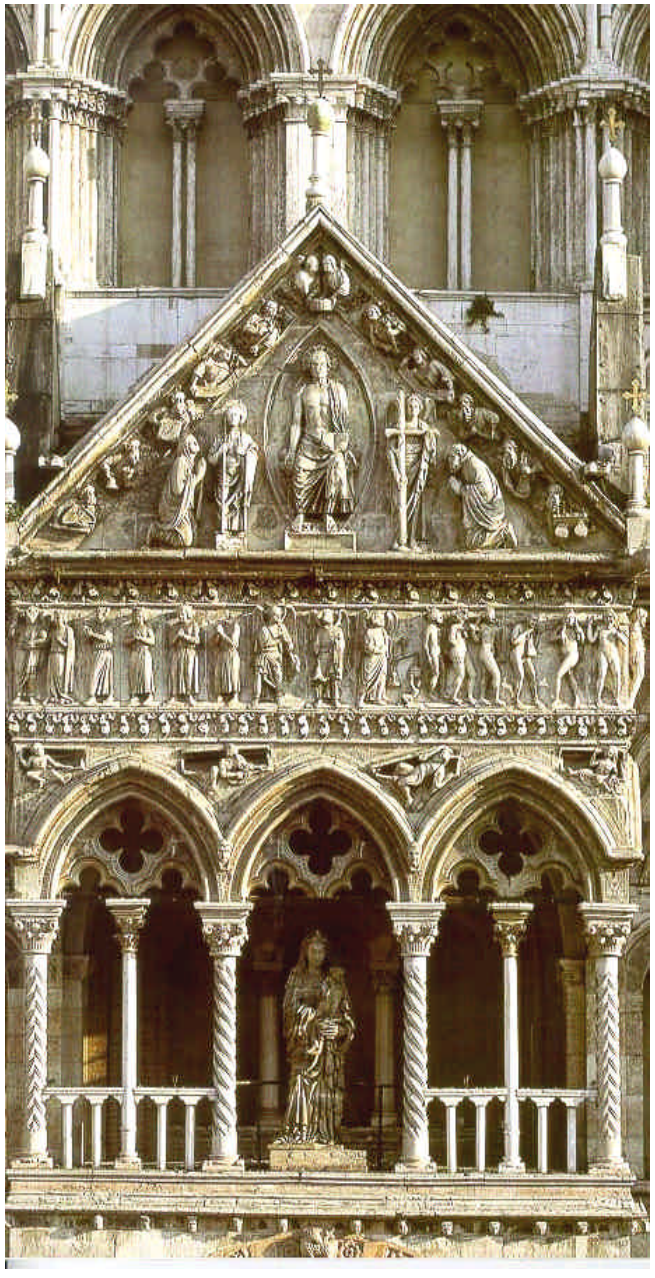
Musée de l'Homme

National Museum of Natural History, Paris, France

manni@mnhn.fr

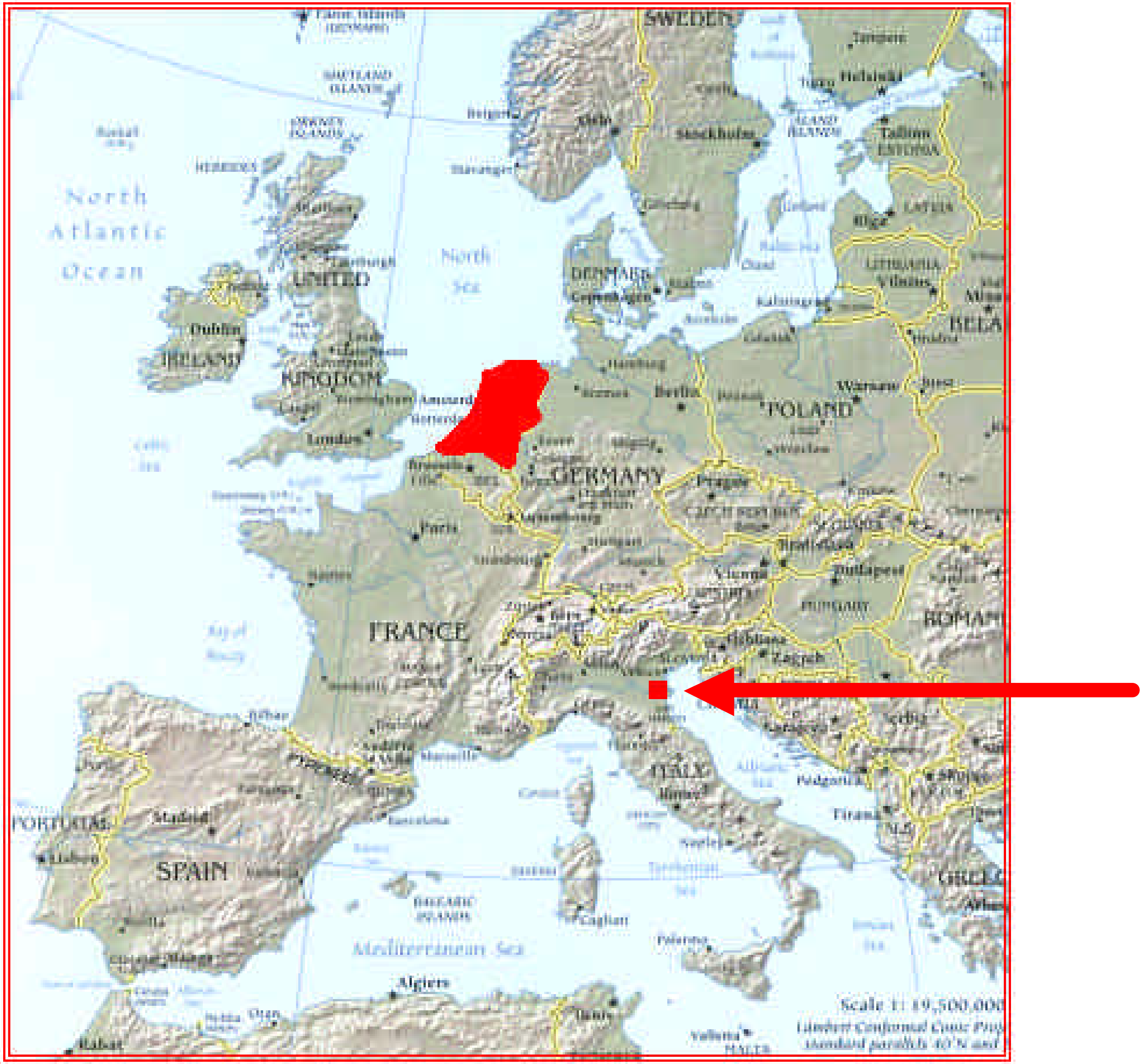
Fifth workshop of the *Netwerk Naamkunde*: family names

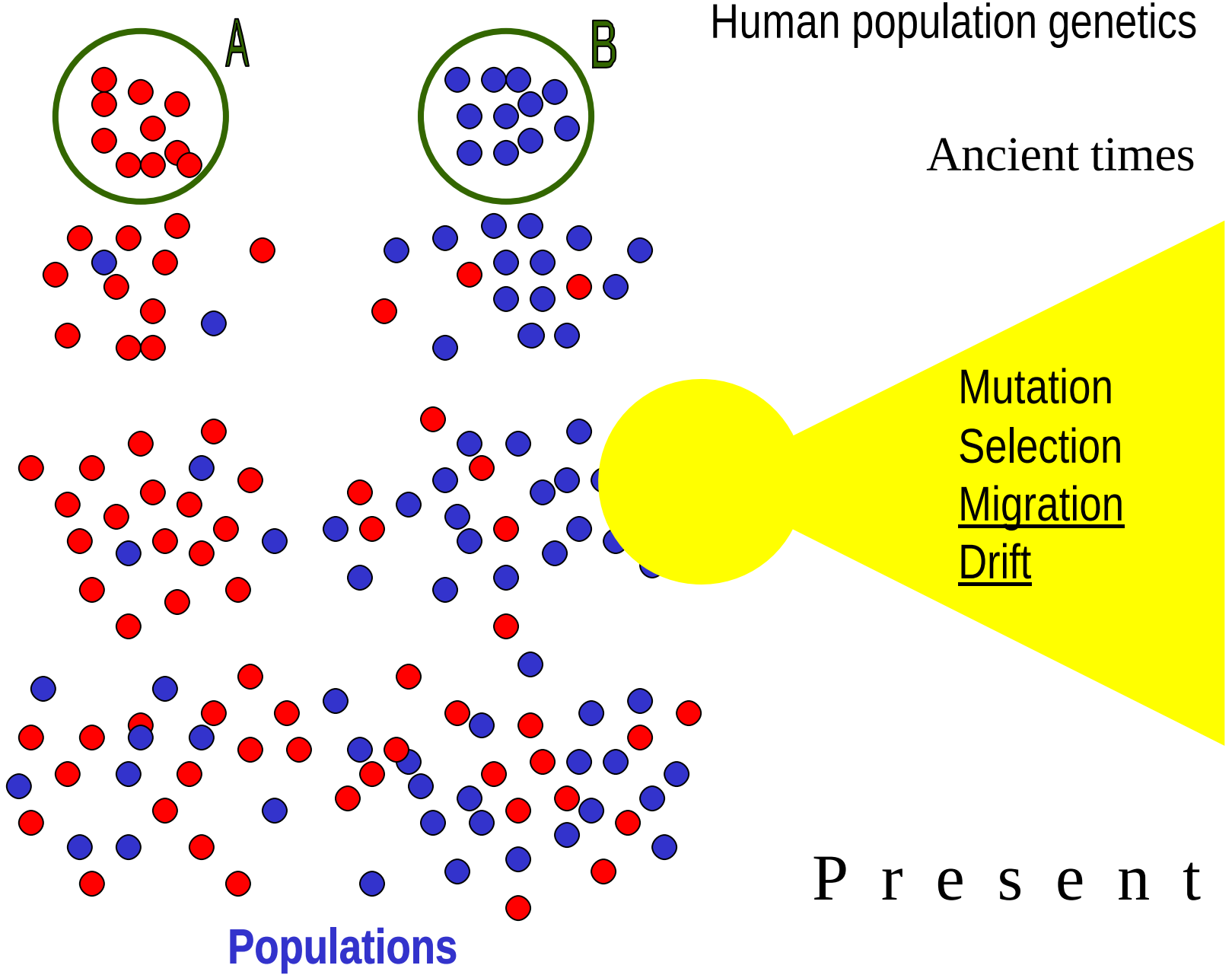
The Hague, Friday 10 October 2008



Ferrara

Cosmé Tura

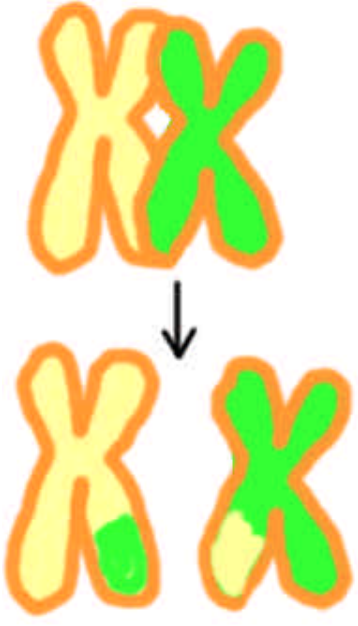

SWEDEN
North Atlantic Ocean
North Sea
Helsinki
Tallinn
ESTONIA
Riga
LATVIA
LITHUANIA
Vilnius
Oslo
Stockholm
Bergen
Stavanger
DENMARK
Copenhagen
UNITED KINGDOM
Edinburgh
Dublin
IRELAND
London
Hamburg
Berlin
POLAND
Warsaw
GERMANY
Prague
Brussels
Paris
FRANCE
Vienna
Bratislava
Budapest
HUNGARY
Ljubljana
Zagreb
Marseille
Barcelona
PORTUGAL
Lisbon
Madrid
SPAIN
ITALY
Naples
Palermo
Cagliari
Tirana
Mediterranean Sea
Algiers
Tunis
Valletta
MALTA
Scale 1: 19,500,000


Human population genetics
A
B
Ancient times
Mutation
Selection
Migration
Drift
Present
Populations

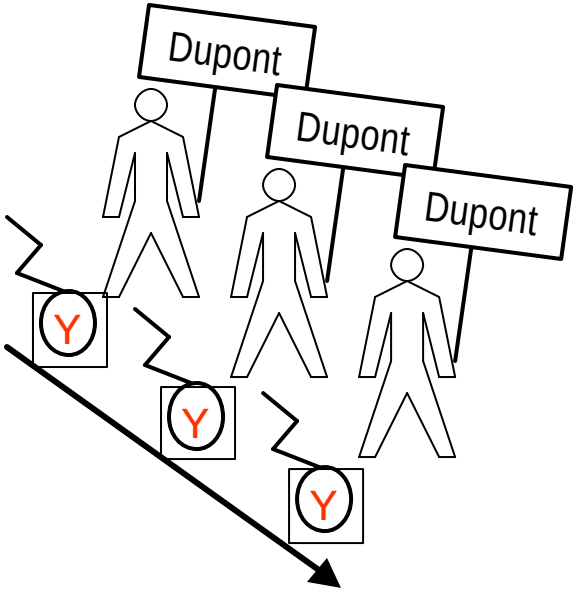
My starting point were the surnames

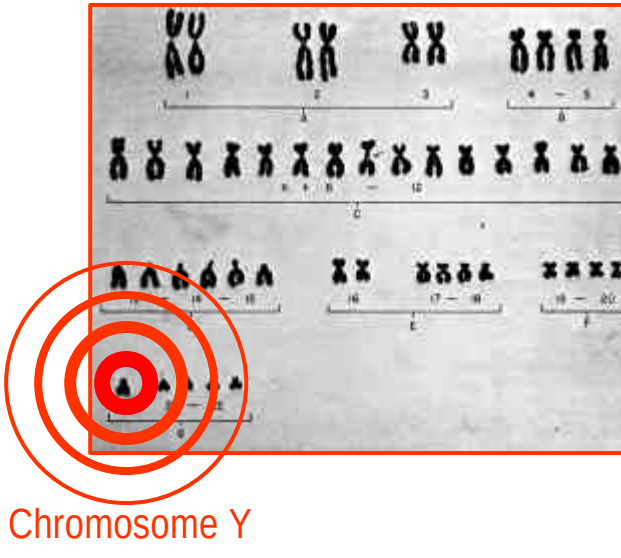
Surnames are a way to look at the variability of the Y-chromosome

Recombinant



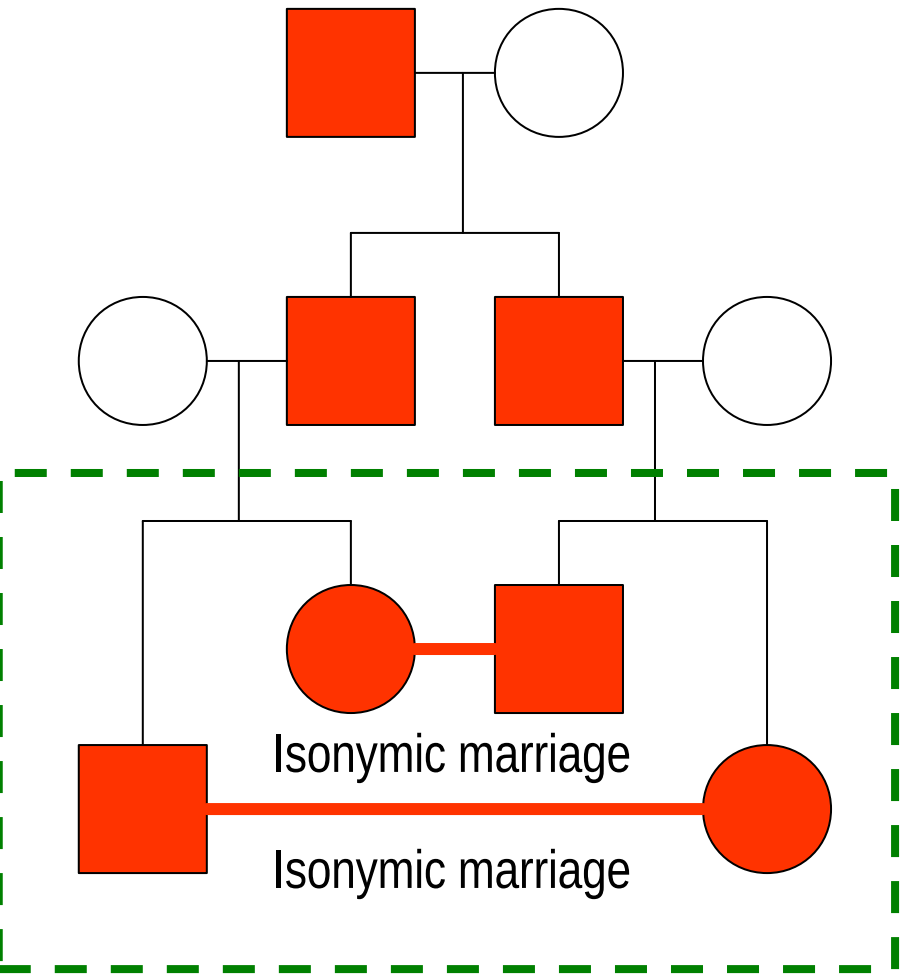
Non Recombinant





Chromosome Y

From surnames to consanguinity: The definition of Isonymy



Parish books

The definition of Isonymy

The real estimation of isonymy can be computed only from real genealogies

When focusing on huge populations it is extremely difficoult to have all the genealogies of the populations and, even when possible, their study requires years.

A WAY-OUT is estimate the levels of isomymy by a probabilistic model, assuming that the husband/wife is not selected according to his surname (Assumption: *I fall in love with someone whatever his/her surname*).

From the distribution of surnames we can estimate the probability of isonymic marriages:

Location **A**: 10 «Nerbonne» over 100 inhabitants

Location **B**: 25 «Nerbonne» over 100 inhabitants

Isonymy $(\mathbf{A})_{\text{Nerbonne}}$ = 10% x 10% = **0.01**

Isonymy $(\mathbf{B})_{\text{Nerbonne}}$ = 25% x 25% = **0.0625**

Isonymy $(\mathbf{AB})_{\text{Nerbonne}}$ = 10% x 25% = **0.025**

Lasker distance:

$$\Sigma n_{si} n_{sj} \,/\, 2 \Sigma n_{si} \Sigma n_{sj}$$

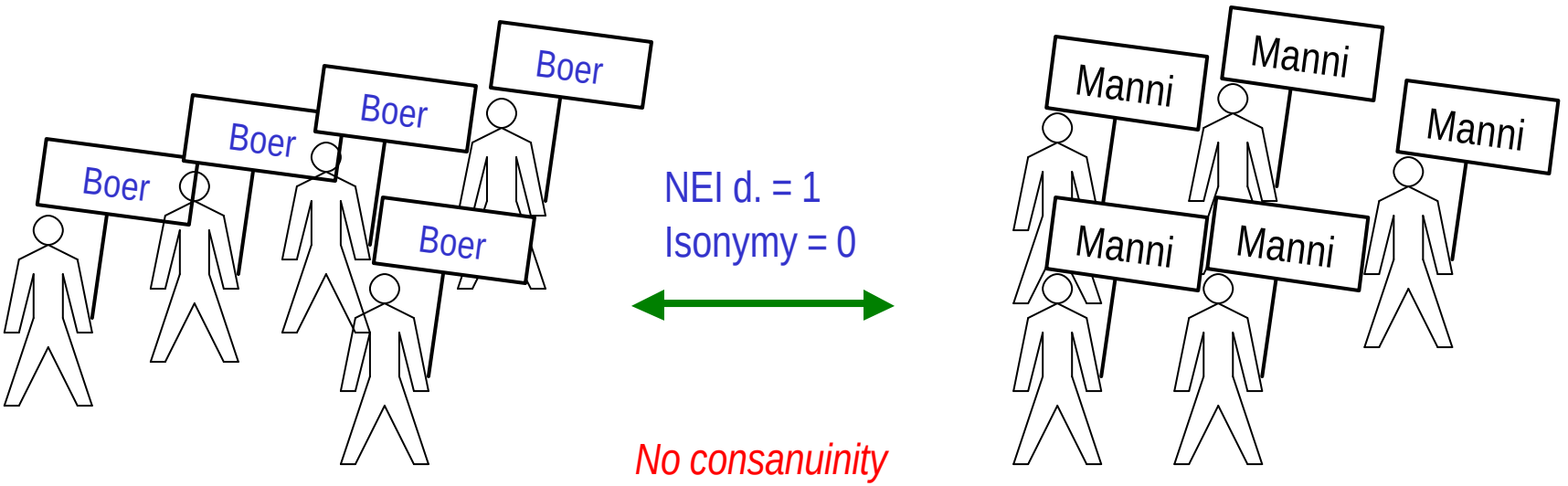
Nei distance:

$$\Sigma n_{si} n_{sj} \,/\, (\Sigma n_{si}^2 \, \Sigma n_{sj}^2)^{1/2}$$

Isonimy and the Netherlands: Nei distance

$$\Sigma n_{si} n_{sj} / (\Sigma n_{si}^2 \Sigma n_{sj}^2)^{1/2}$$

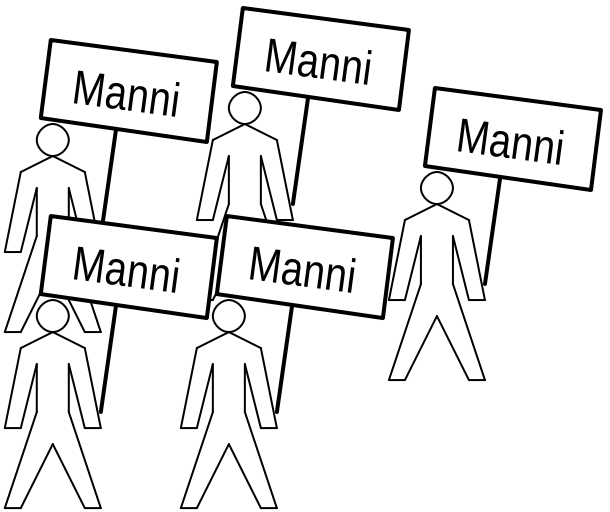
If the two locations have completely different surnames their distance will be **1**, while if they share the same set of surnames with identical relative frequencies their distance will be **null**.



The reverse of isonimy: Nei distance

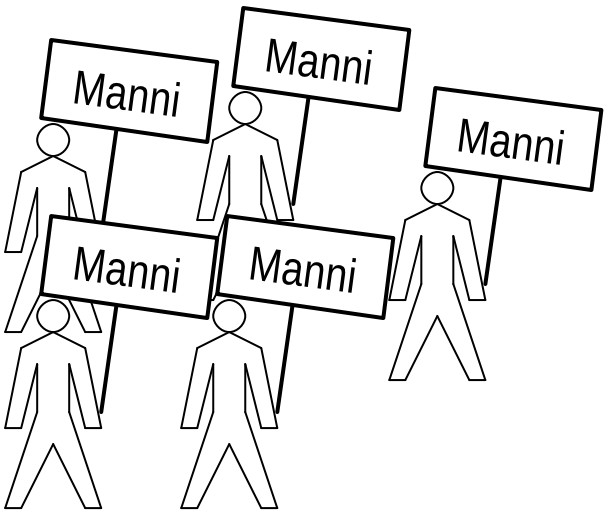
$$\Sigma n_{si} n_{sj} / (\Sigma n_{si}^2 \Sigma n_{sj}^2)^{1/2}$$

If the two locations have completely different surnames their distance will be **1**, while if they share the same set of surnames with identical relative frequencies their distance will be **null**.



NEI d. = 0
Isonymy = 1

HIGH consanuinity



Isonimy : Nei distance

Surname differentiation can be computed, according to NEI as:

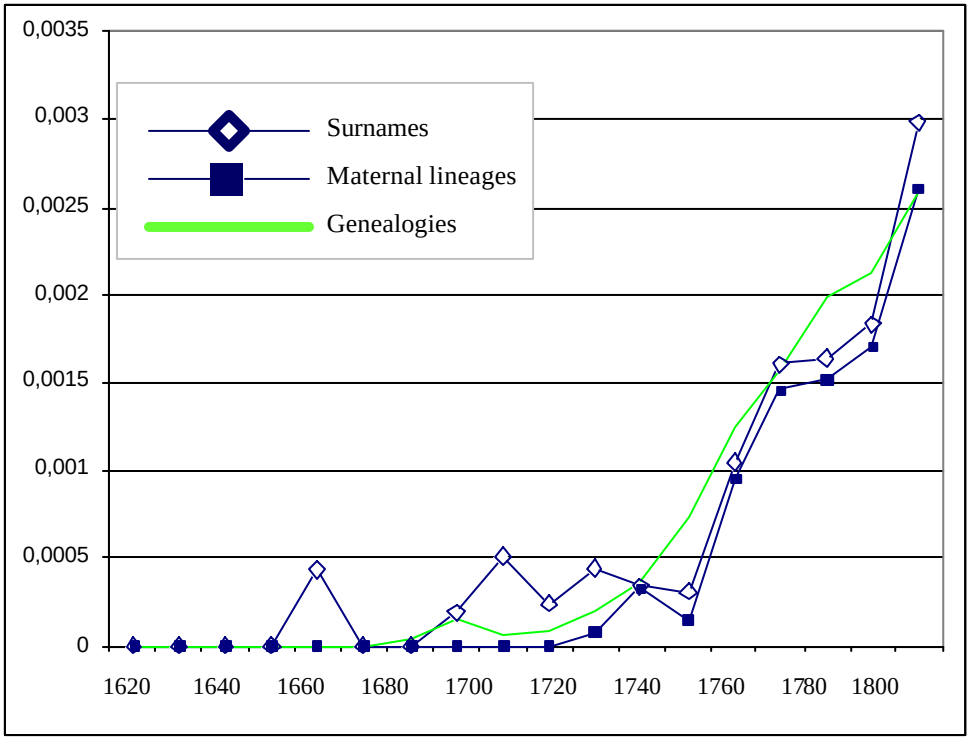
$$\Sigma n_{si} n_{sj} \,/\, (\Sigma n_{si}^2 \, \Sigma n_{sj}^2)^{1/2}$$

where n_{sj} denotes the frequency of a given surname ***s*** in locations ***i*** while n_{sj} denotes the frequency of the same surname in location ***j***.

The sums are done for all surnames

By applying the formula to a surname distributions (list of surnames and their relative frequency) of all the places under study it is possible to a pairwise distance matrix accounting for the similarity of surnames in different places.

Surnames enable the estimation of consanguinity (Isonymy)



Alain Gagnon and Bruno Toupance (2002).
Testing isonymy with paternal and maternal lineages in the early Québec population: the impact of polyphyletism and demographic differentials.
Am. J. Phys. Anthrop.

Isonimy and the Netherlands

The Netherlands

Surnames (from 1997 telephone book):

226 sample points (Manni *et al.* 2005)

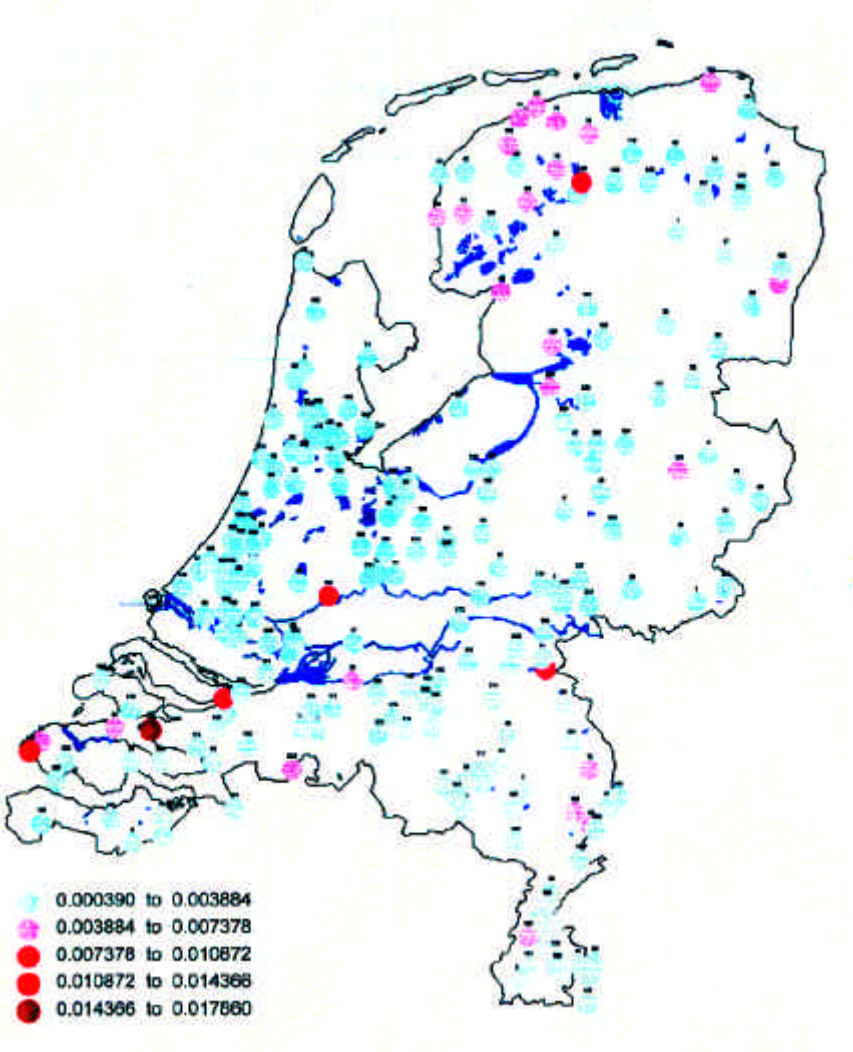
2,400,000 telephone users

The Netherlands

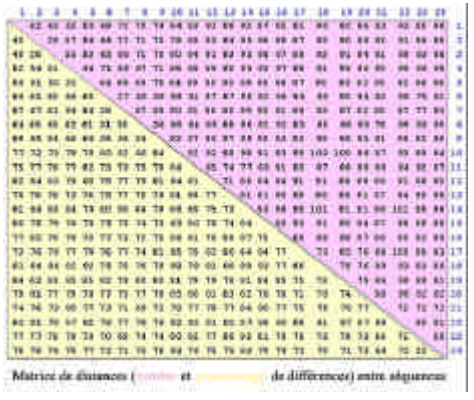
Surnames (from 1997 telephone book):

226 sample points (Manni *et al.* 2005)

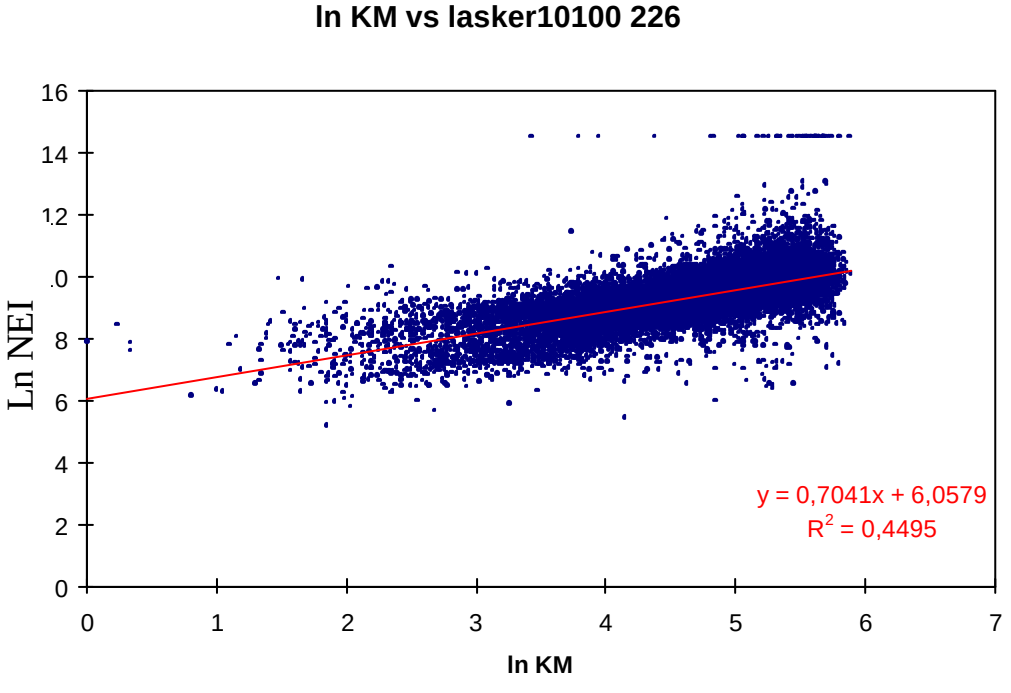
2,400,000 telephone users



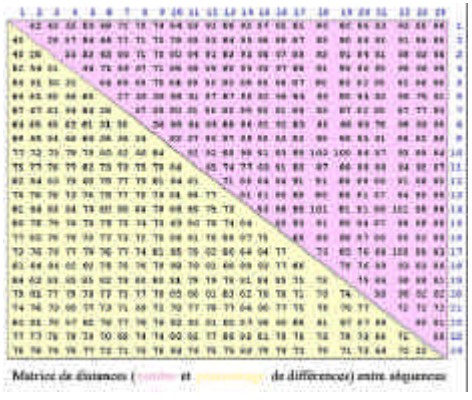
Multidimensional plot of surname pair-wise distances: **Nei**



Matrice de distances (et de différence) entre séquences



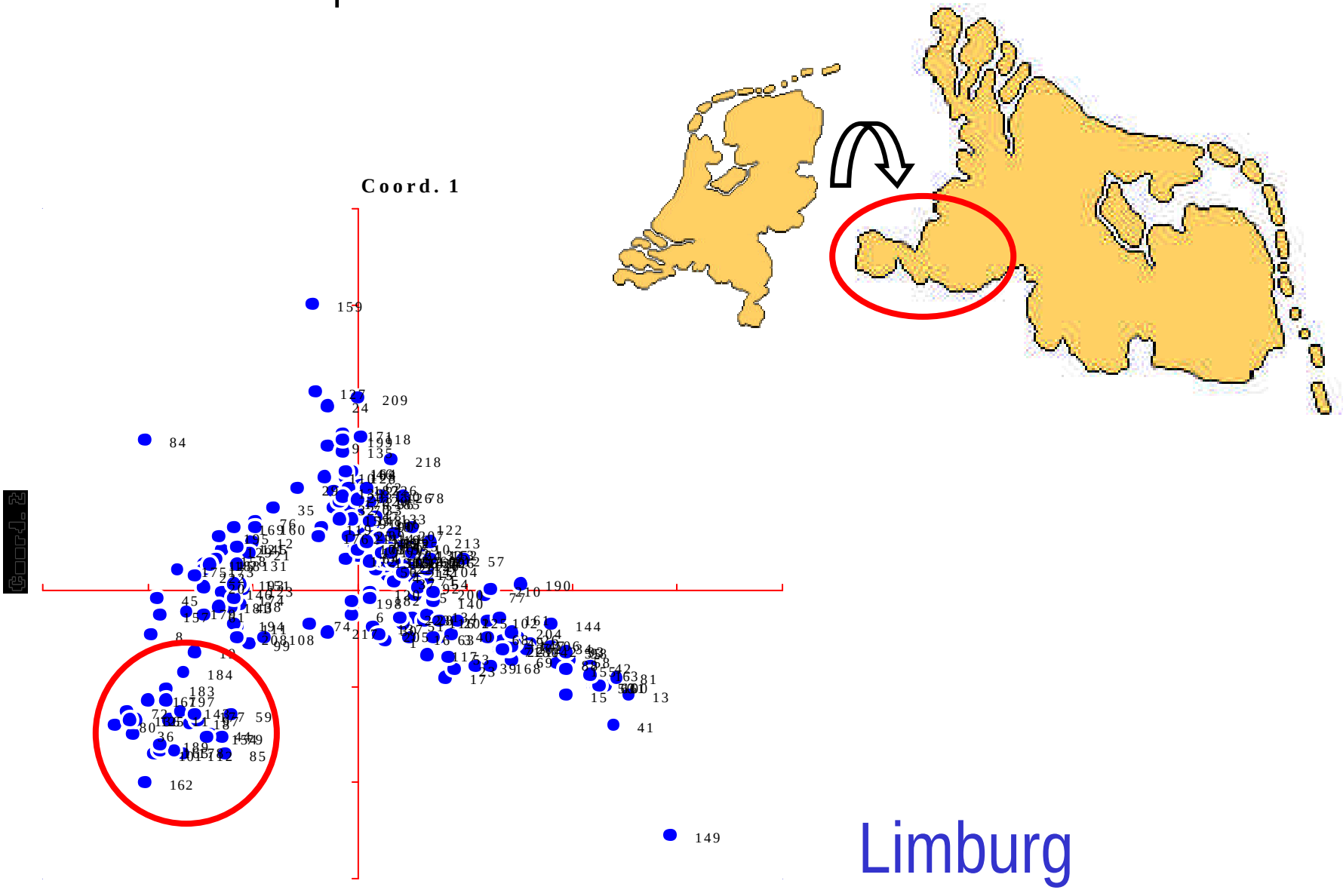
Isonimy and the Netherlands: Nei distance



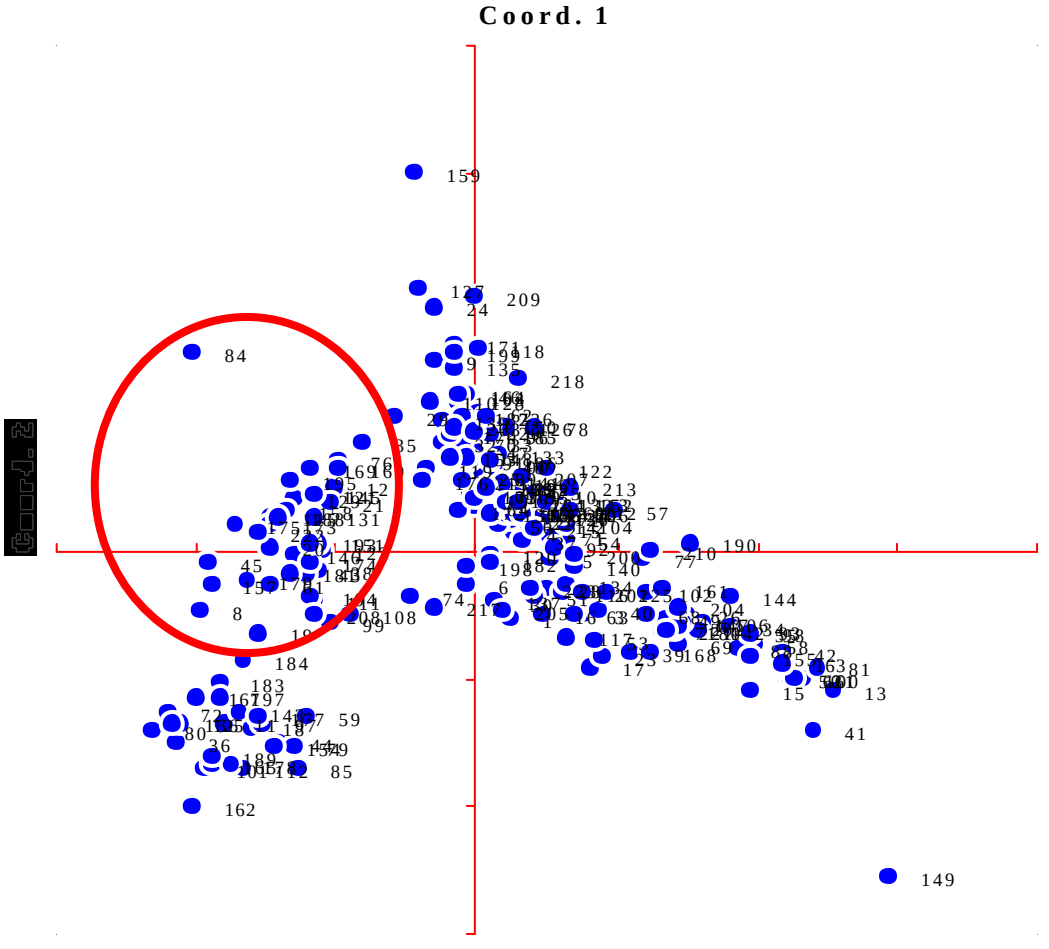
MATRIX

MULTIDIMENSIONAL SCALING

Multidimensional plot os surname distances

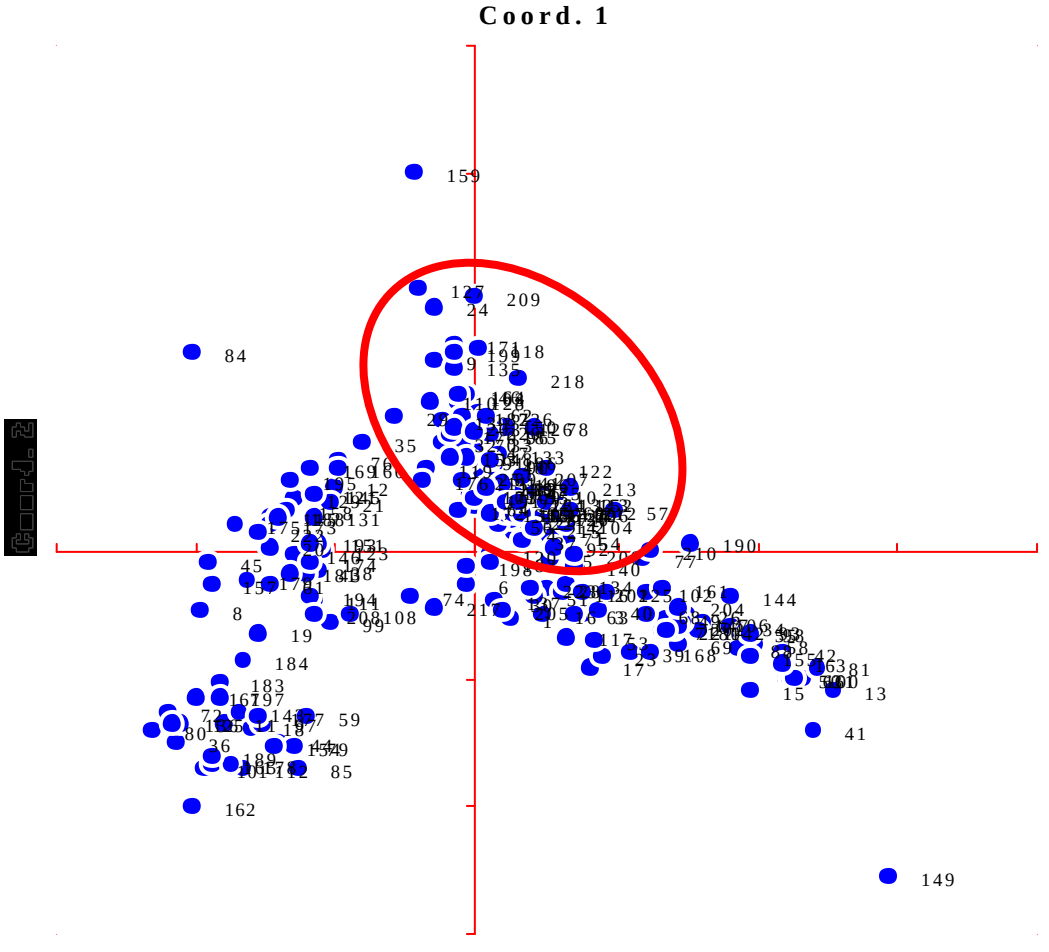


Multidimensional plot os surname distances



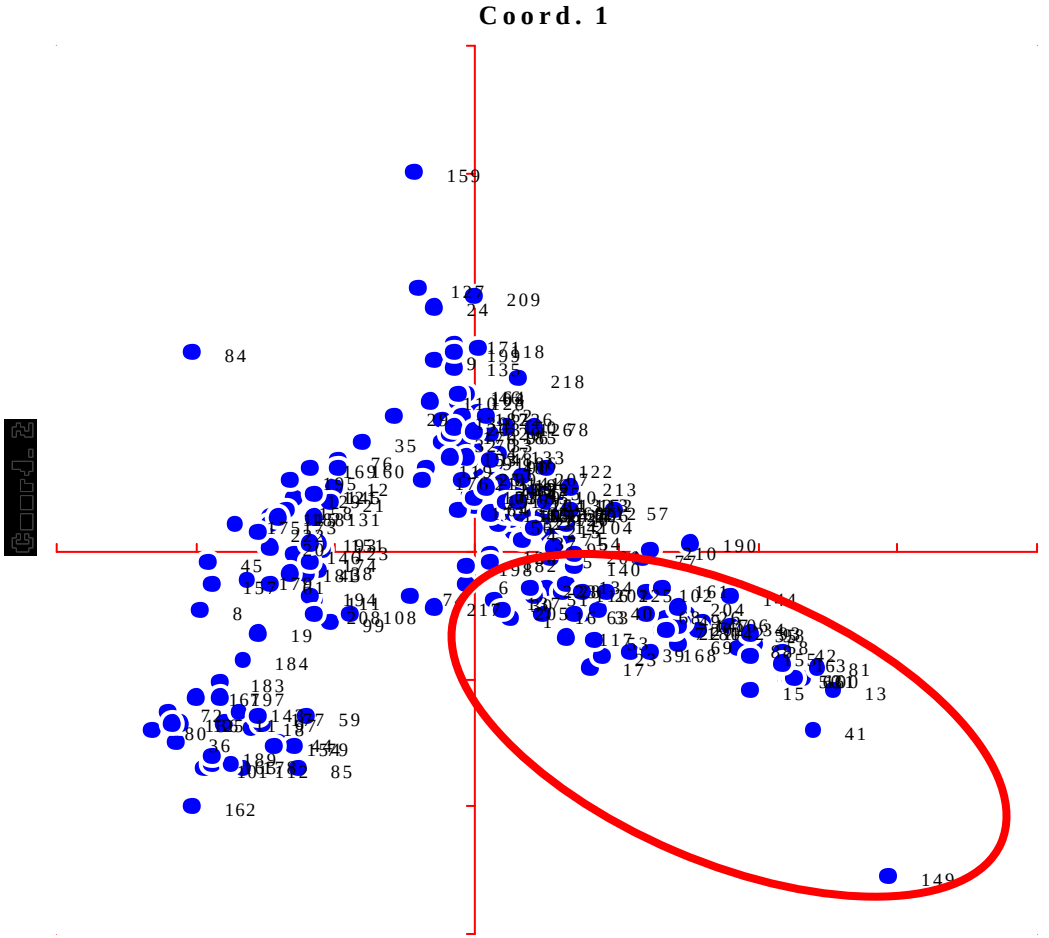
Brabant; Zeeland

Multidimensional plot os surname distances

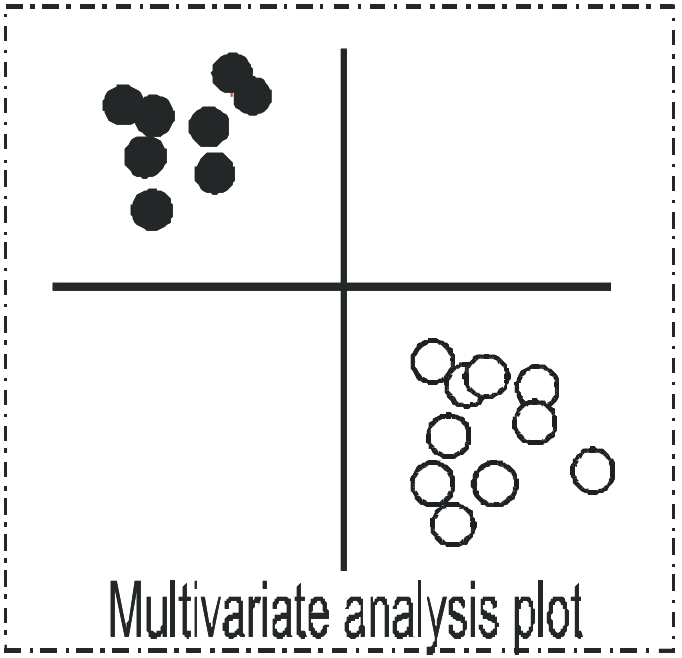


N/S Holland

Multidimensional plot os surname distances



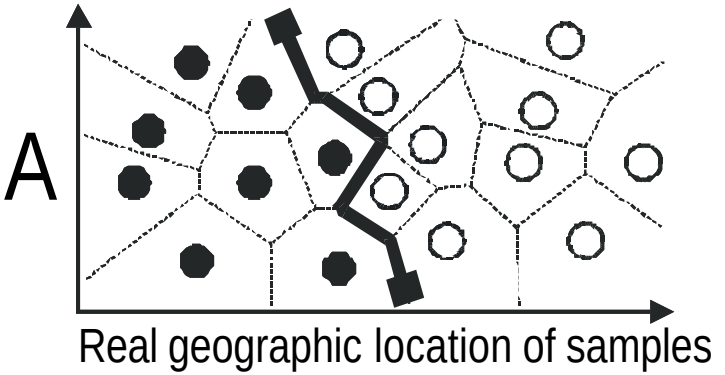
East Netherlands

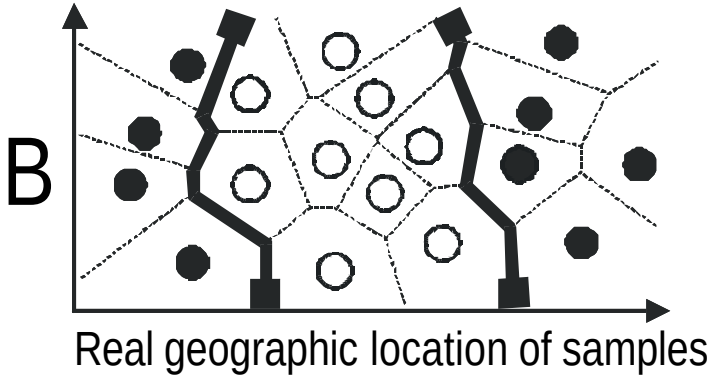


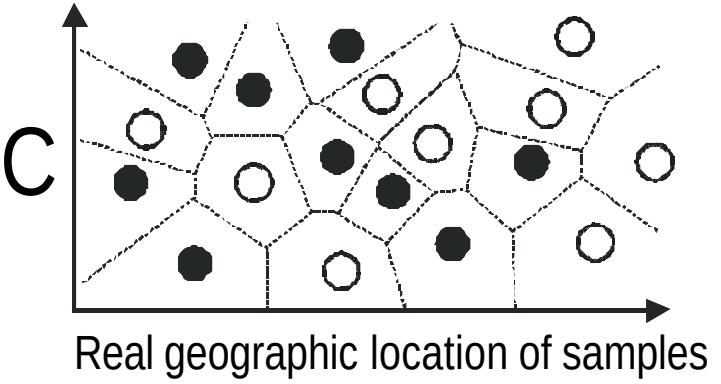
PROBLEM !

Different scenarios may correspond to a same multivariate analysis

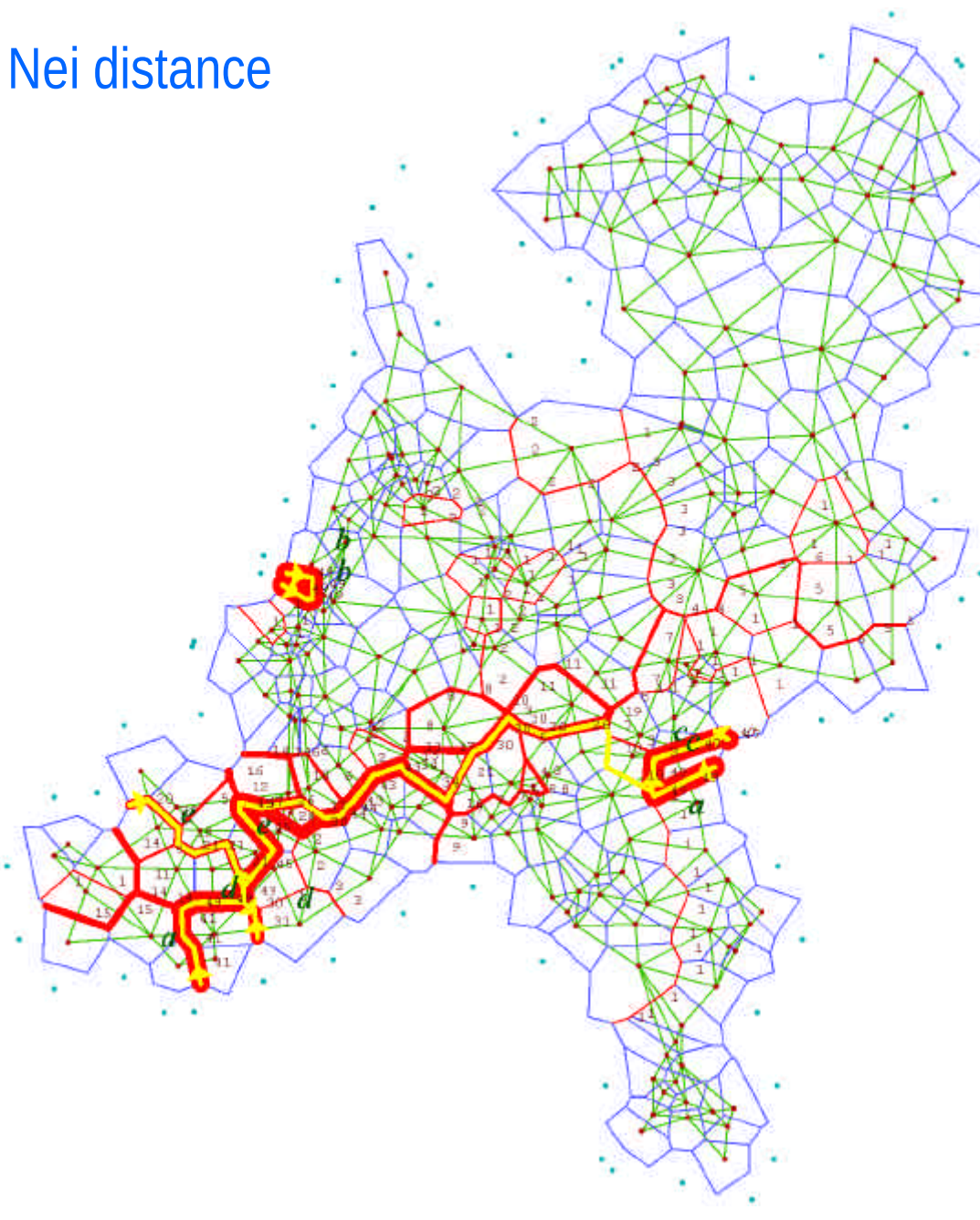
(PCA, MDS, etc.)



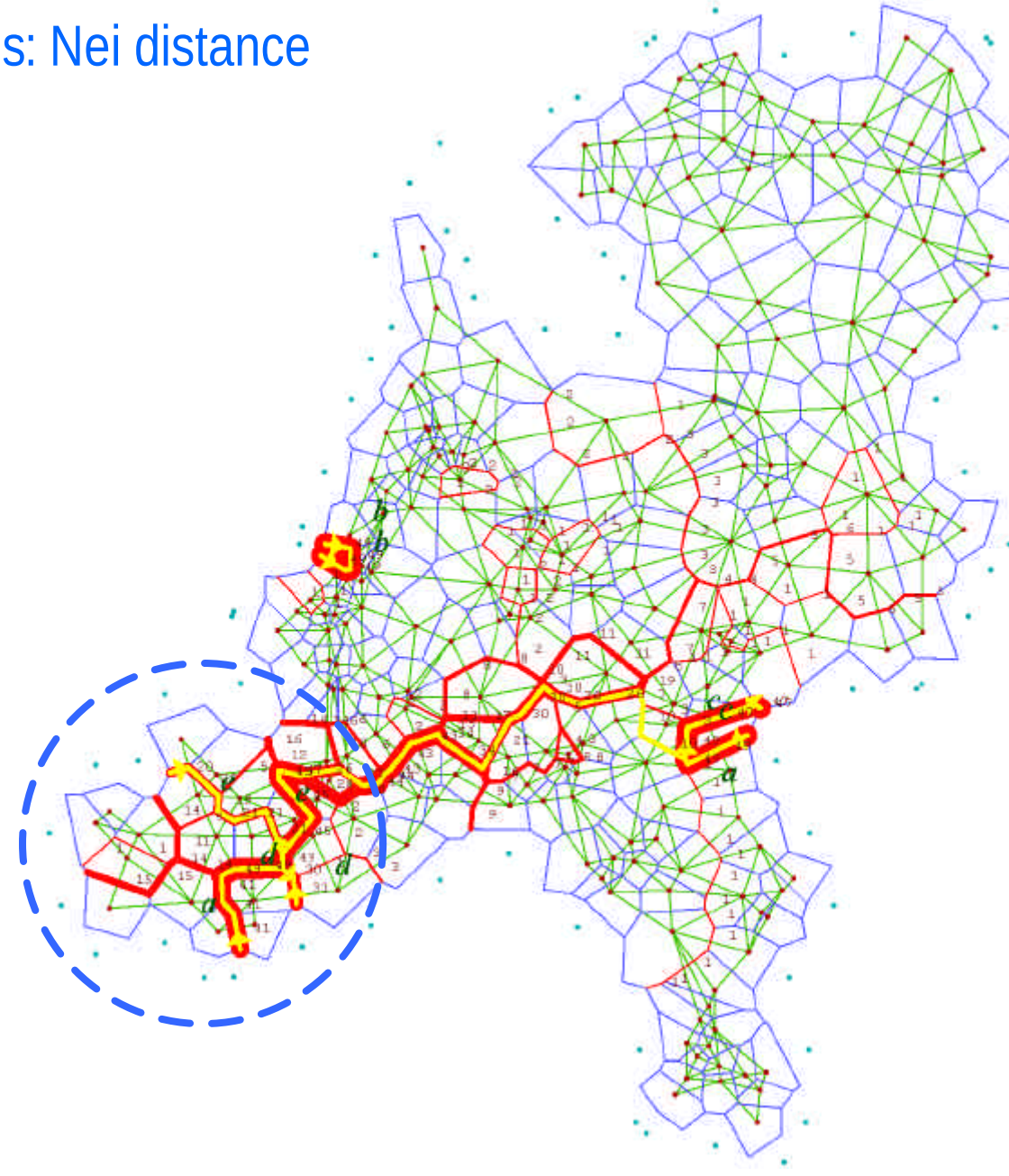




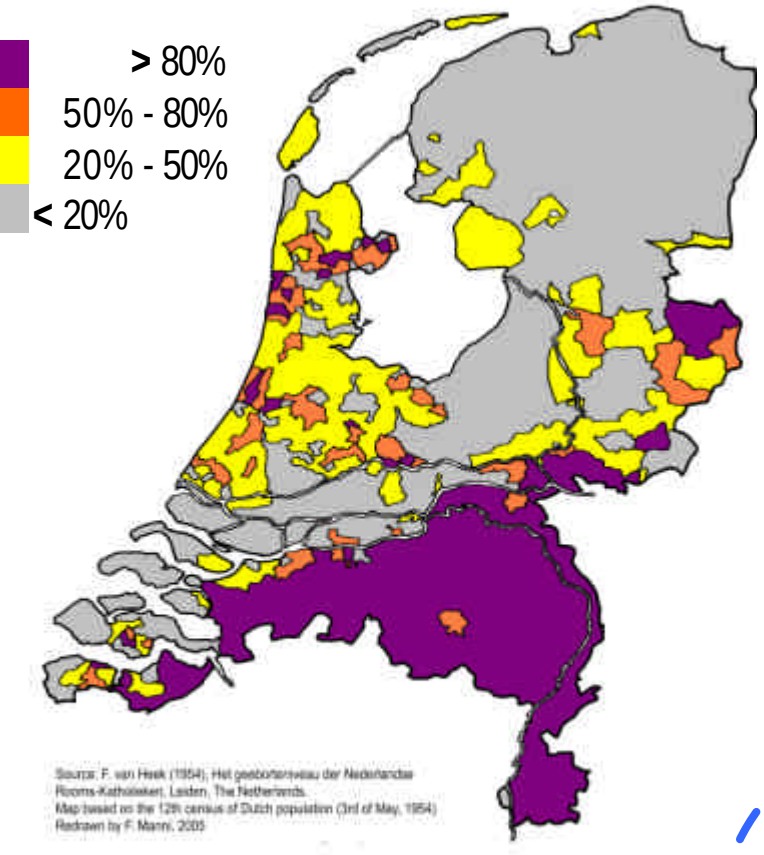
Isonimy and the Netherlands: Nei distance



Isonimy and the Netherlands: Nei distance



Geographic analysis, *islands are another explanation*



Source: F. van Heek (1954), Het geboortenniveau der Nederlandse Rooms-Katholieken, Leiden, The Netherlands.
Map based on the 12th census of Dutch population (3rd of May, 1954)
Redrawn by F. Manni, 2005

Calvin

Benedictus XVI

No intermarriages

End of first part:

« **Surnames cooked with a standard sauce** »

(little pause)

next:

« **Surnames are also words…** »

Is there a link betwween cultural and genetic diversity?

1. **Genetics**
   demographic history of populations, evolution
2. **Linguistics**
   mirrors cultural differences as well as gastronomic traditions, basket technology, etc.
3. **Surnames**

   They are transmitted like genetic traits but they are words…

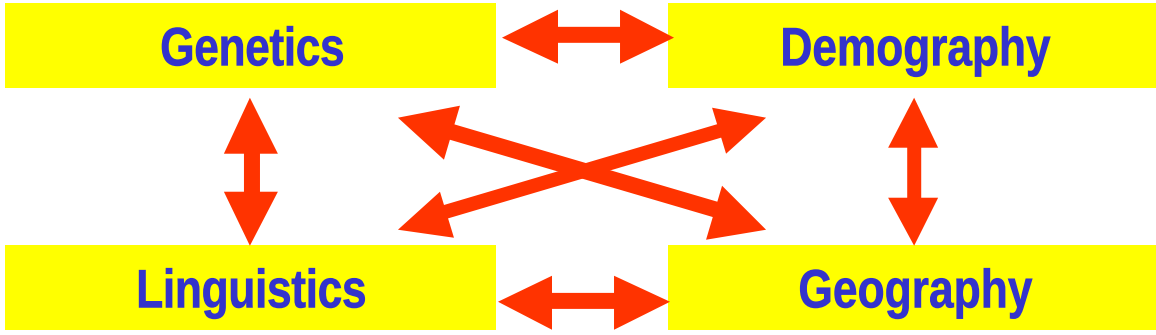
Is there a link betwween cultural and genetic diversity?

Literary and Linguistic Computing Advance Access published September 15, 2006

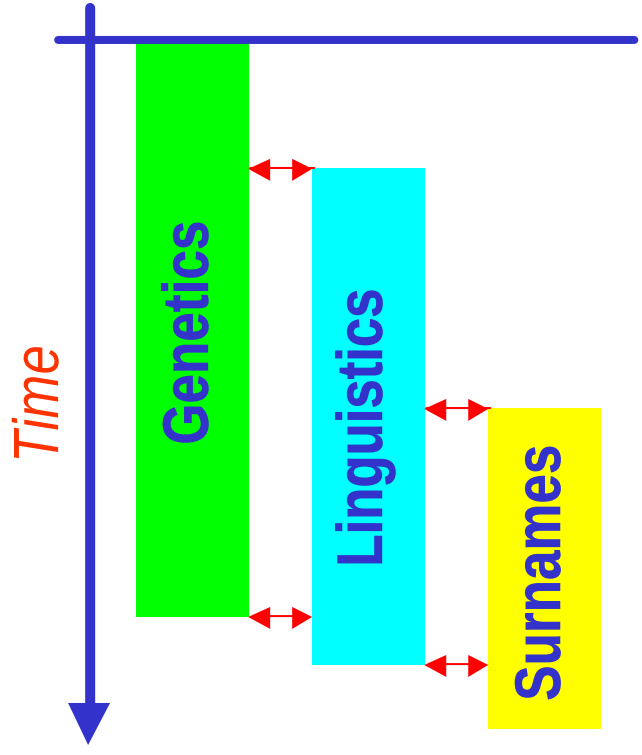
To What Extent are Surnames Words? Comparing Geographic Patterns of Surname and Dialect Variation in the Netherlands

Franz Manni
UMR 5145 CNRS, Musée de l'Homme MNHN, Paris, France

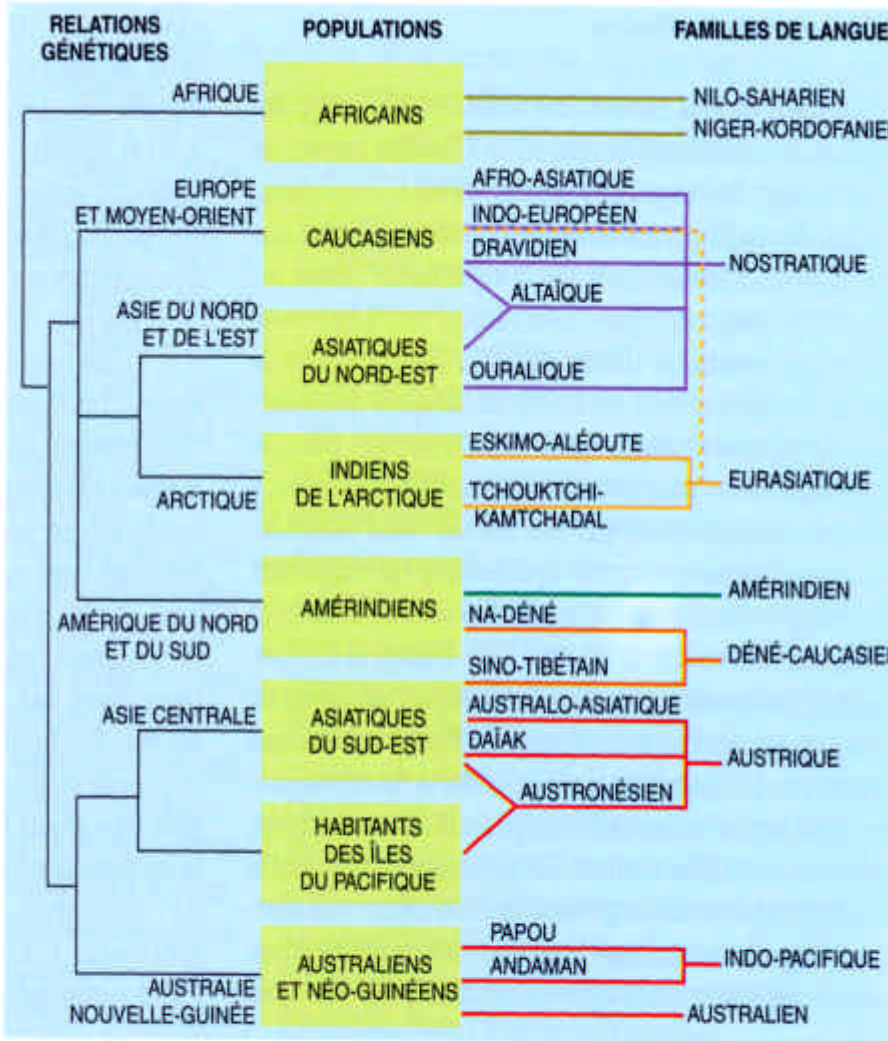
Wilbert Heeringa and John Nerbonne
Alfa-Informatica, Faculty of Arts, University of Groningen, The Netherlands



The three markers have different time dephts,
Therefore each one of them represents a variability that originated over a different time frame.



Genetics vs linguistics



Cavalli-Sforza et al. 1989

Comparing genetic and linguistic variability

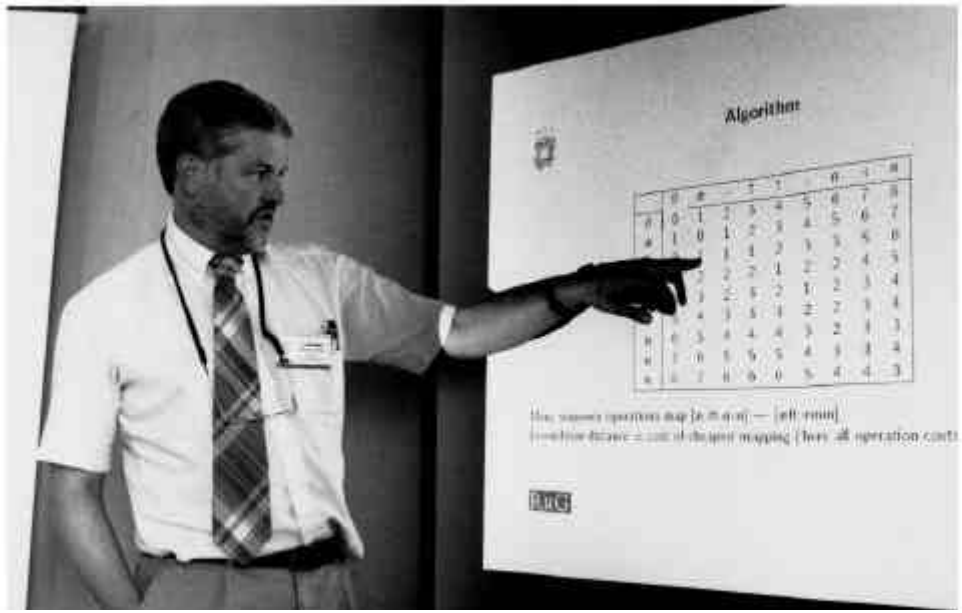
Worldwide analyses are quite controversial.

Better to focus on smaller geographic scales

Nowadays it is possible to computationally analyze dialects and similar languages. It is safer and probably more testable, to date.

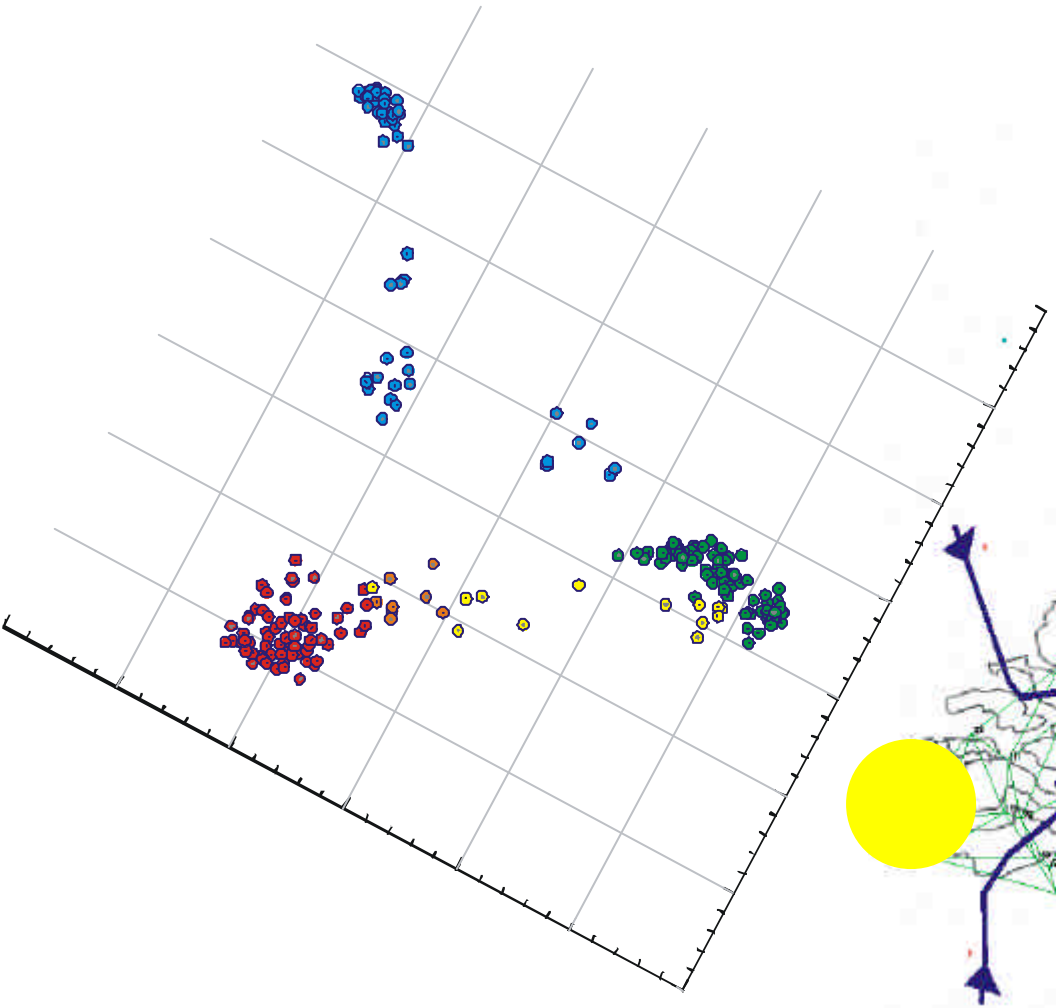
DUTCH **DIALECTS**: credits

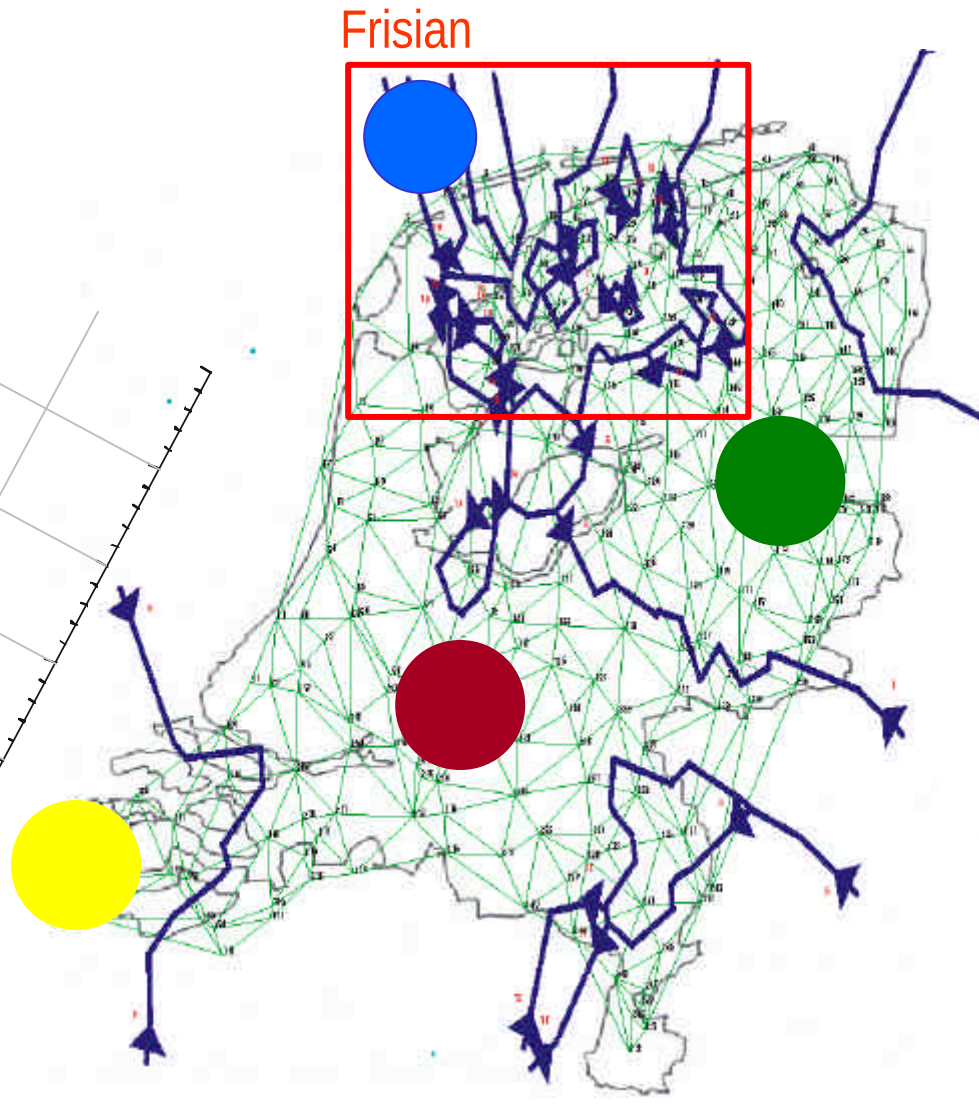
J. Nerbonne



W. Heeringa

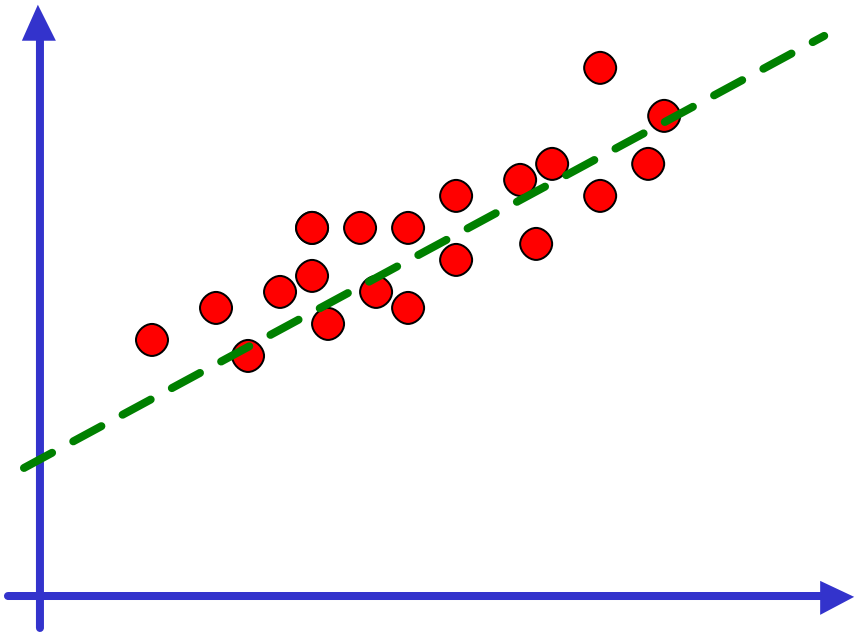
DUTCH **DIALECTS**: MDS and geographic analysis with Monmonier alg.





255 samples; 20 barriers

DUTCH **DIALECTS**: excellent correlation with geography

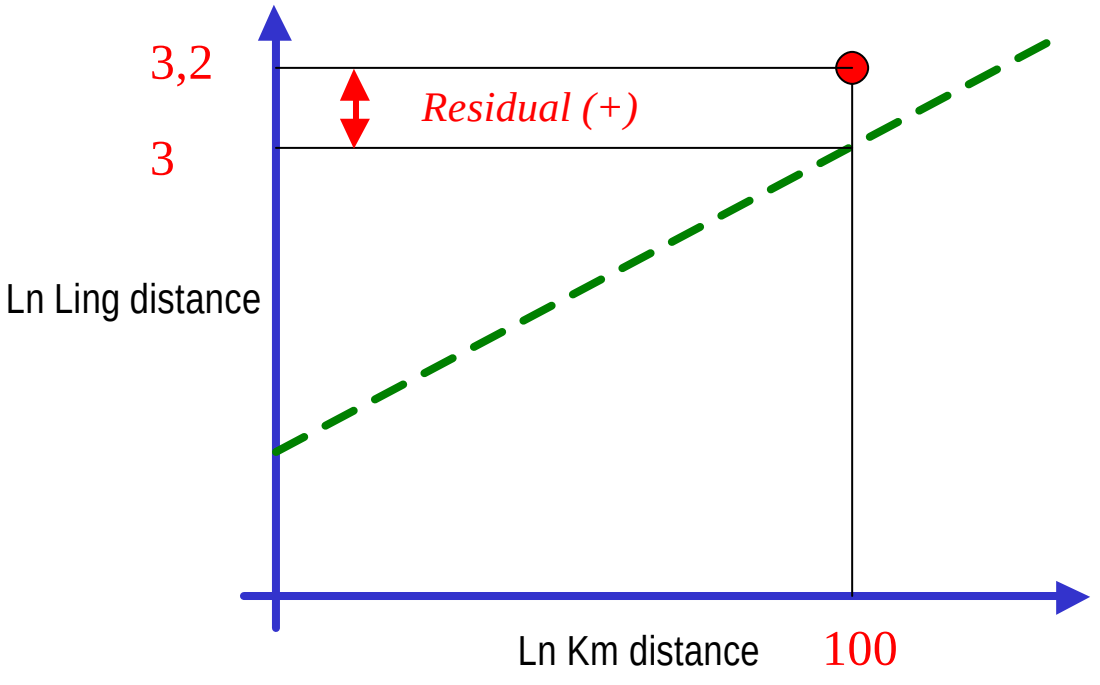


If the regression is good it means that there are some rules

We can accept the model and use it to make inferences of **EXPECTED** values of difference.

The difference between the **EXPECTED** value and the **OBSERVED** value is the **RESIDUAL**

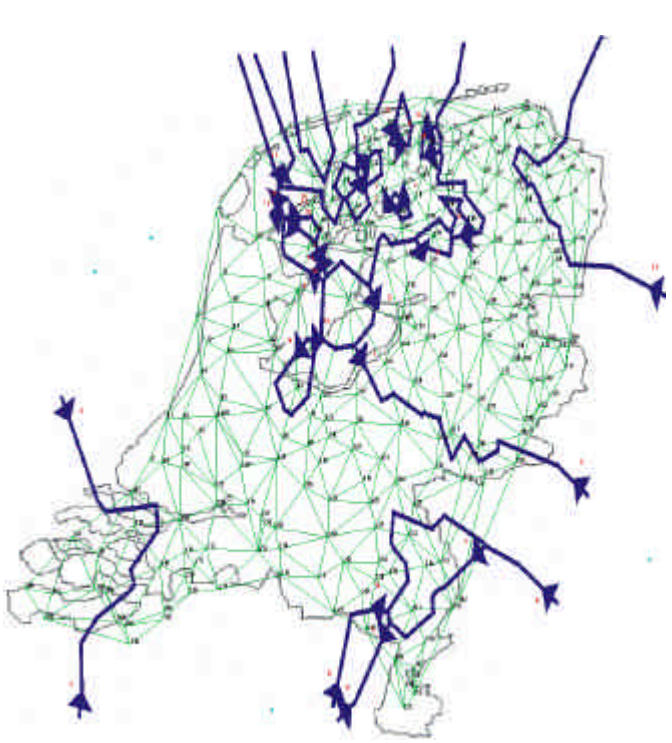
DUTCH **DIALECTS**: excellent correlation with geography

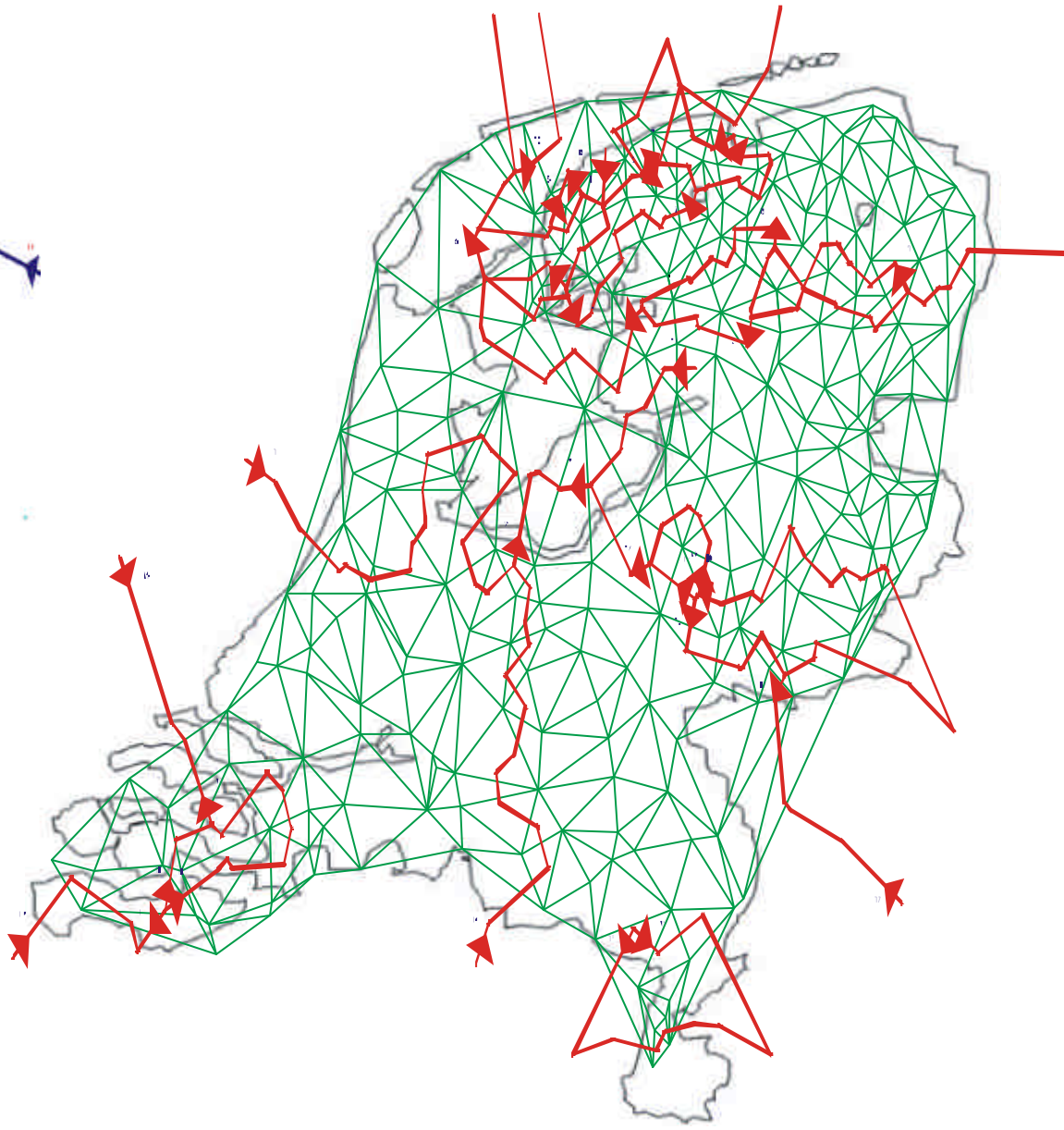


By computing the residuals for all the pairwise measures of linguistic distance we can compile a matrix of residuals and do the same kind of analysis we did before,

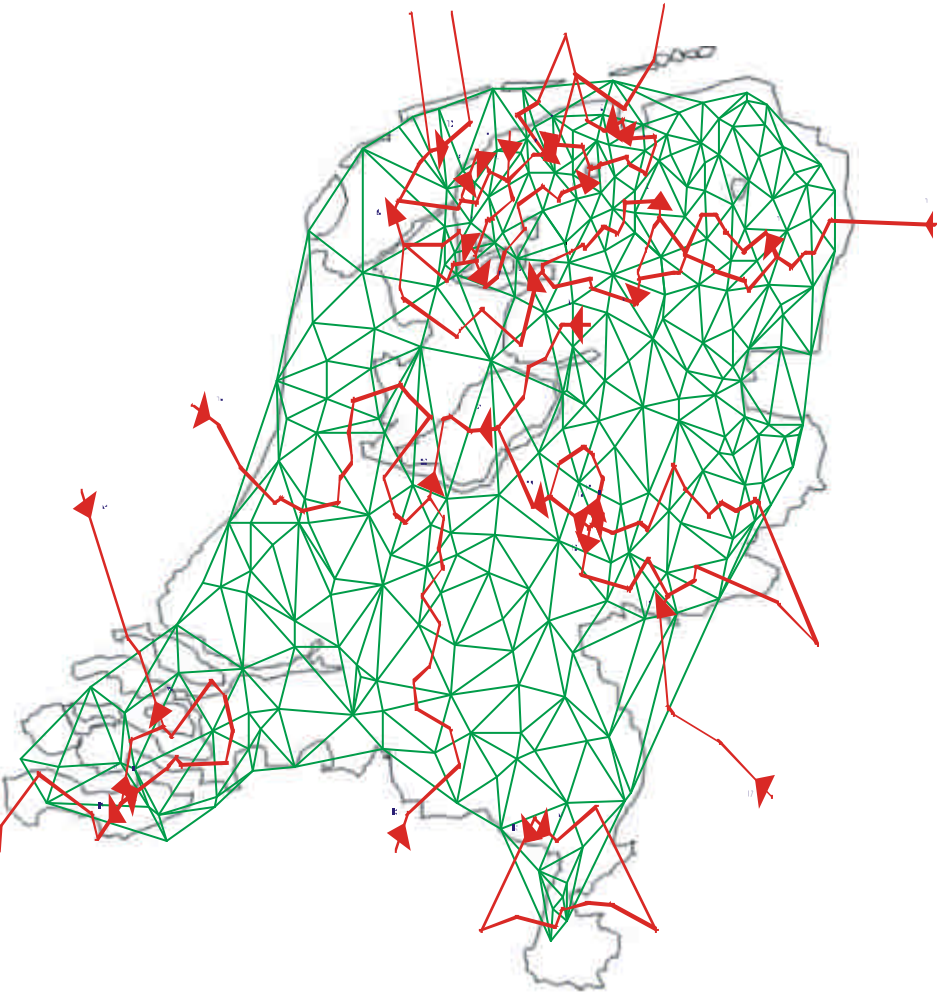
The new results are expected **NOT** to be conditioned by geography.

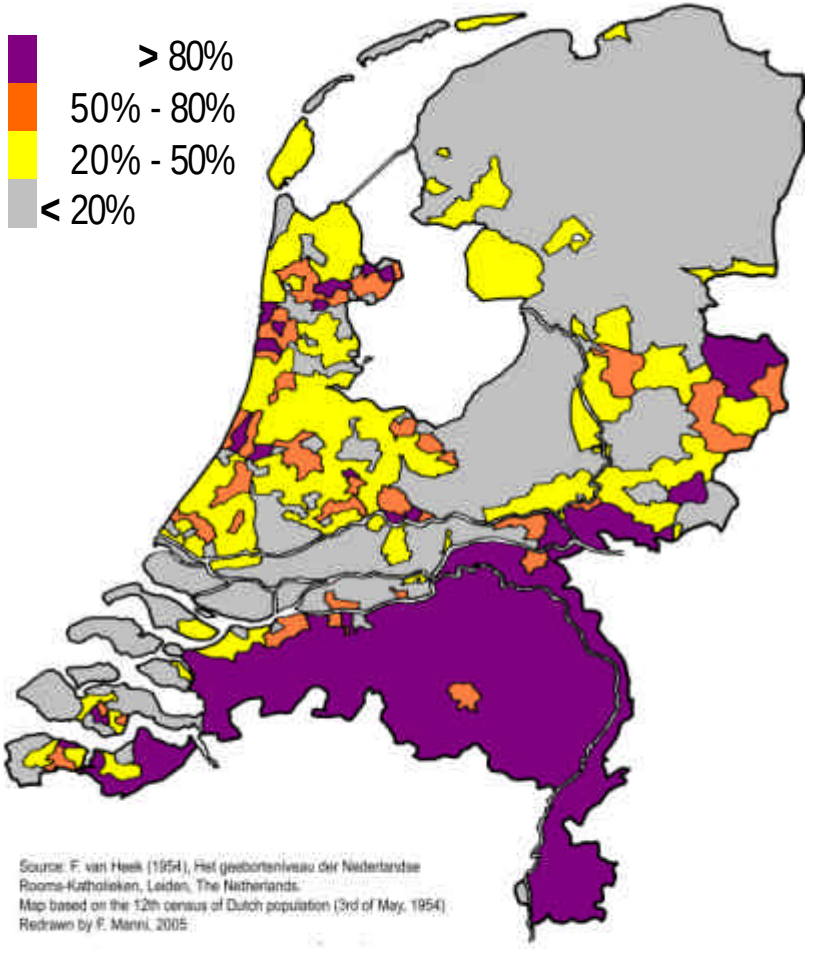
Geographic analysis with Monmonier algorithm, *«normal» vs. residuals*



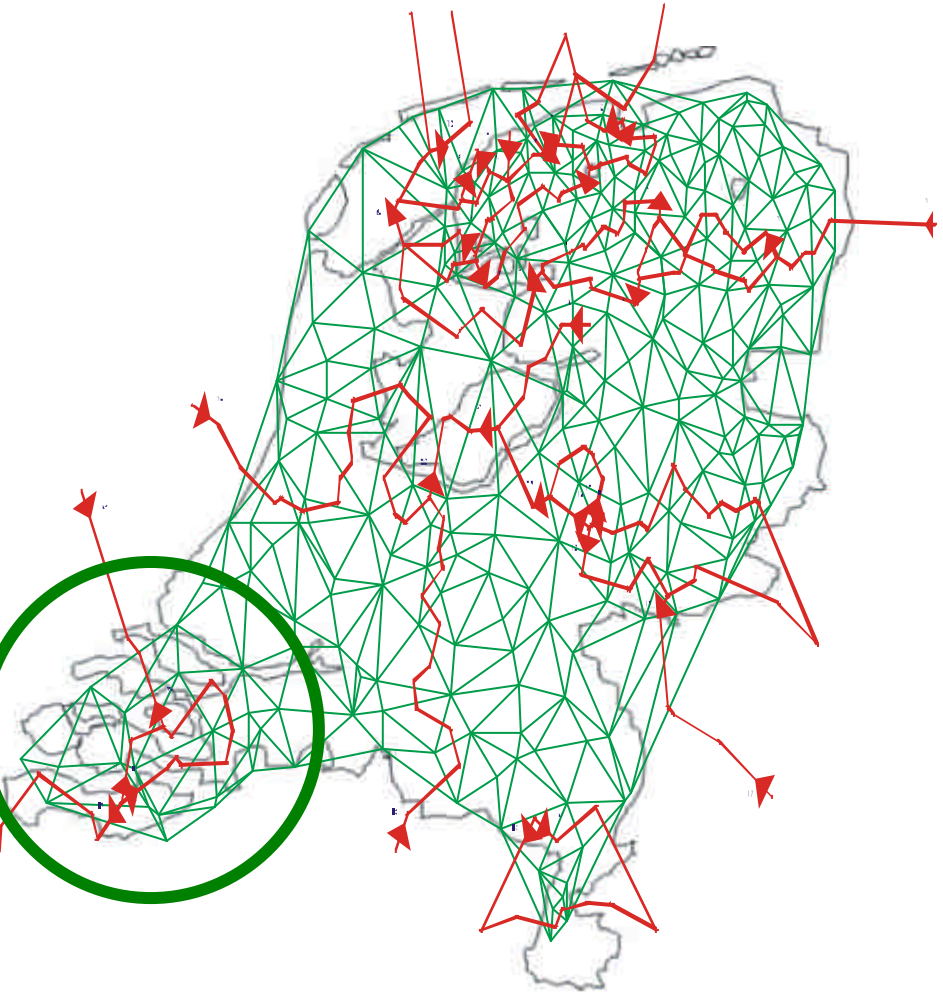


Geographic analysis, *religion is **NOT** an explantion*

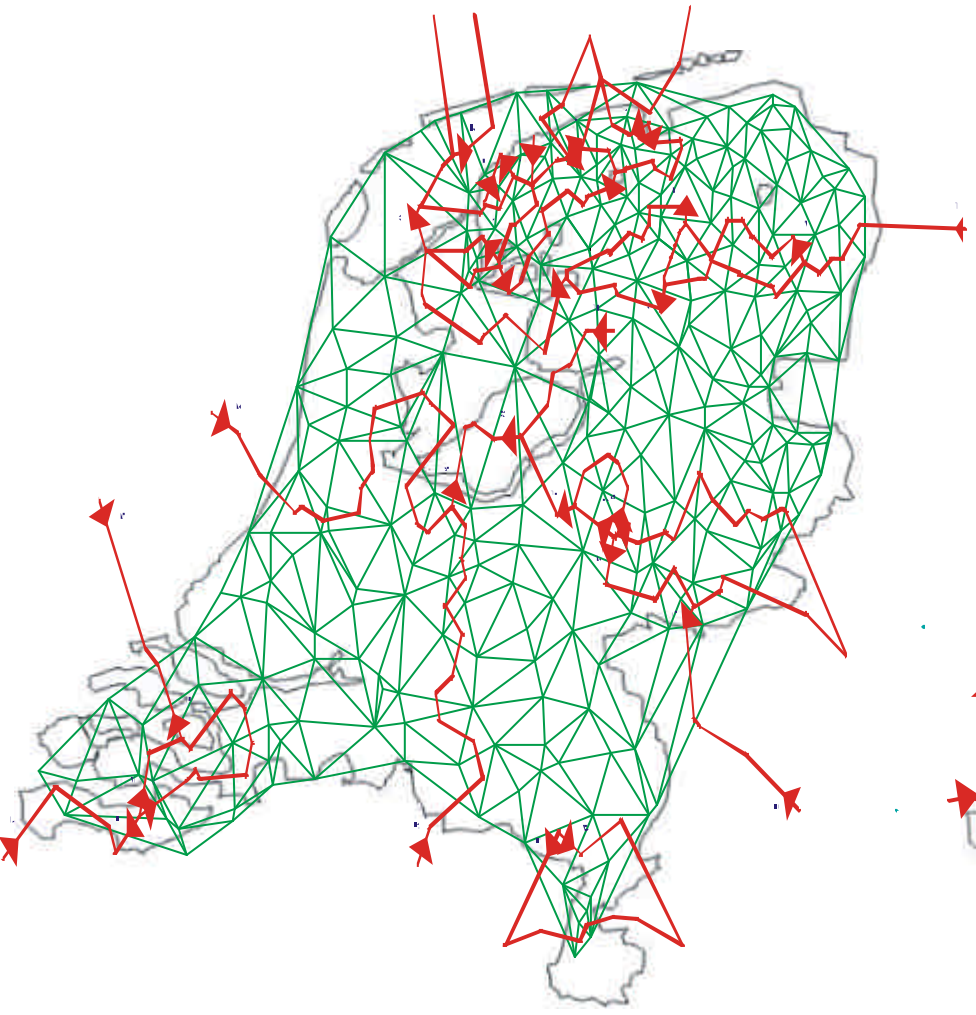


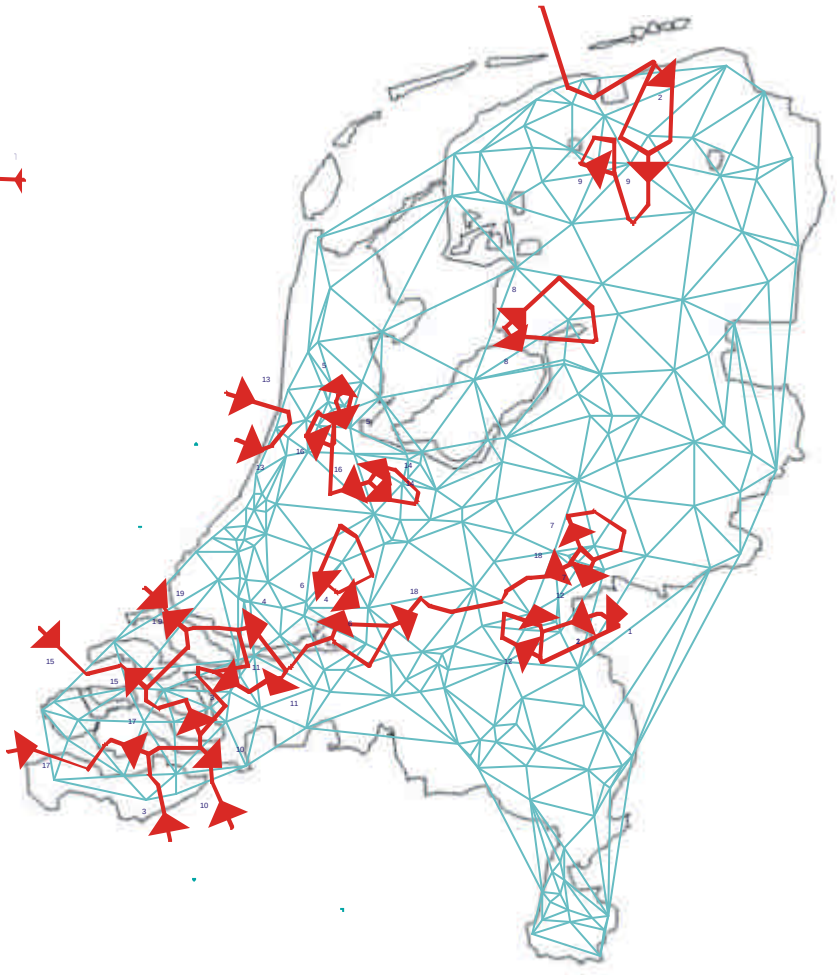


Geographic analysis, *islands can be an explanation*



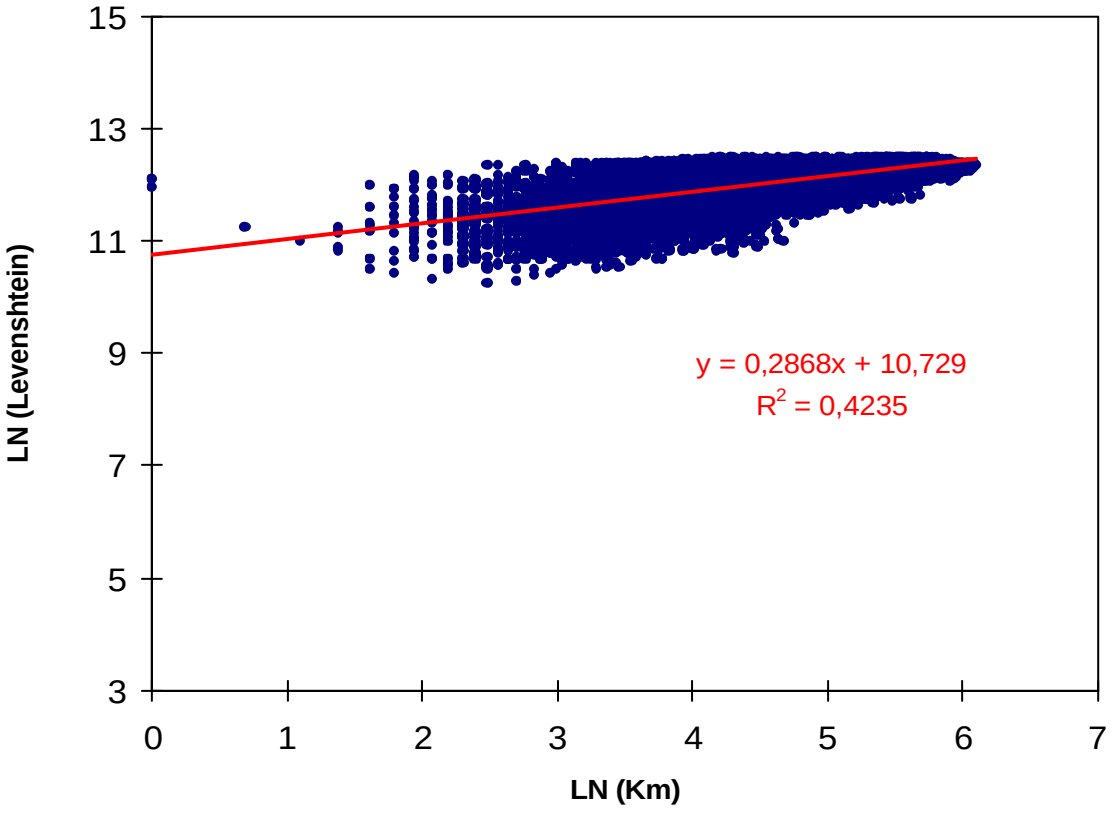
DIALECTS vs. SURNAMES





After regressions

DUTCH **DIALECTS**: excellent correlation with geography



We have seen something similar with the surnames…

Are surnames words?

Correlation between surname and dialect dissimilarity matrices for the NL

	SURNAMES Nei	DIALECTS (72)	GEOGRAPH. DISTANCE
SURNAMES Nei	*1*	**0.298**	
DIALECTS 72	**0.298**	*1*	
GEOGRAPH. DISTANCE			

Merging dialect and surname matrices: 72 sampling points

Are surnames words?

Correlation between surname and dialect dissimilarity matrices for the NL

	SURNAMES Nei	DIALECTS (72)	GEOGRAPH. DISTANCE
SURNAMES Nei	*1*	**0.298**	**0.565**
DIALECTS 72	**0.298**	*1*	**0.632**
GEOGRAPH. DISTANCE	**0.565**	**0.632**	*1*

Merging dialect and surname matrices: 72 sampling points

End of second part:

« **Are Surnames words? (They are not)** »

(little pause)

next:

« **Surnames cooked with a spicy sauce …** »

A New Method for Surname Studies of Ancient Patrilineal Population Structures and its Possible Application to the Improvement of Y-Chromosome Sampling

Franz MANNI, B. TOUPANCE & E. HEYER

Unité de Génétique des population - Musée de l'Homme MNHN
17, Place du Trocadéro - 75016 Paris (manni@mnhn.fr)

The maps of Kohonen

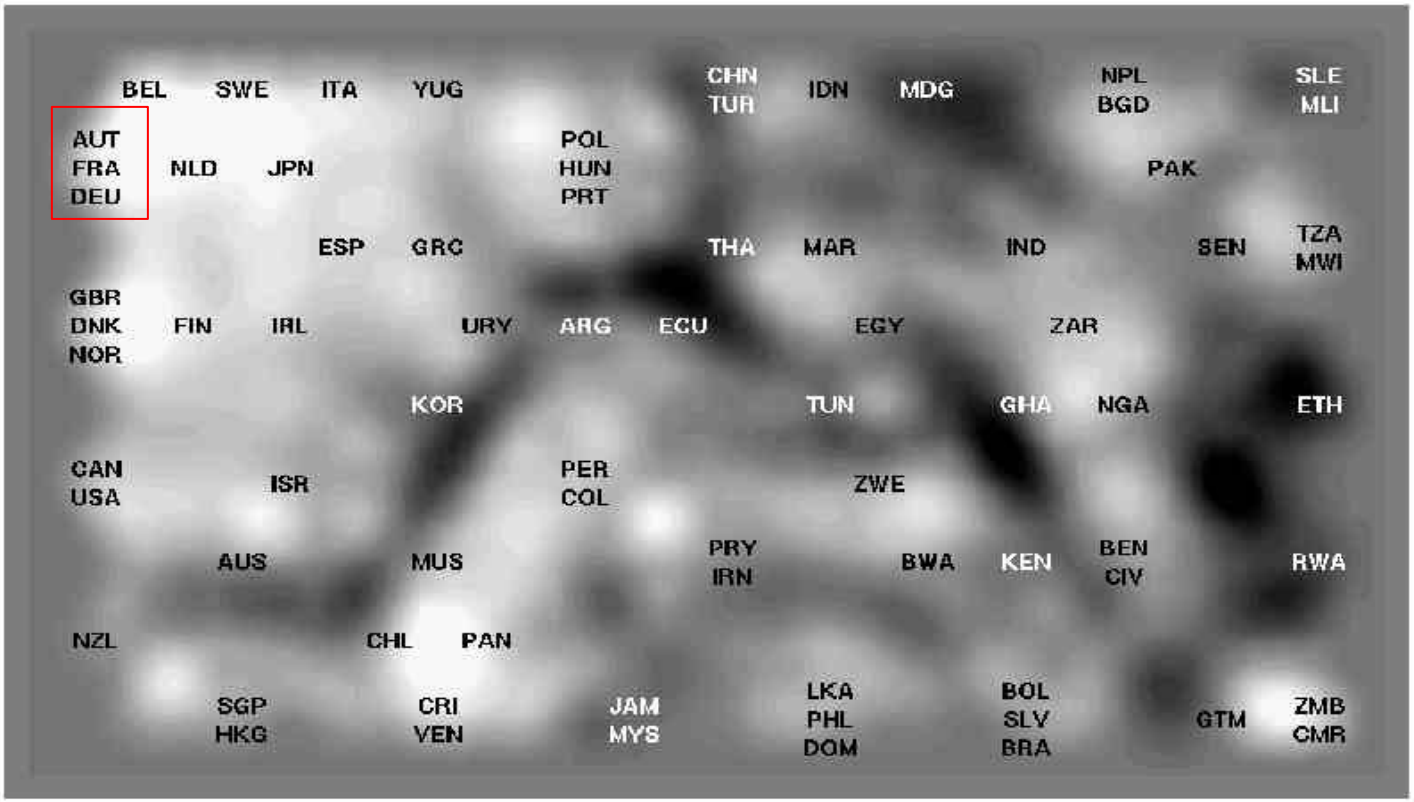


Figure 5: A map display constructed using the SOM algorithm. The overall order of the countries seems to correspond fairly closely to the Sammon's mapping of the same data set (Fig. 4). The most prominent clustering structures are also visible in both displays. Details on how the map was constructed are presented in Publication 2. The size of the map was 13 by 9 units.

Example : poverty in the world

Comparing different methods

Example : poverty in the world

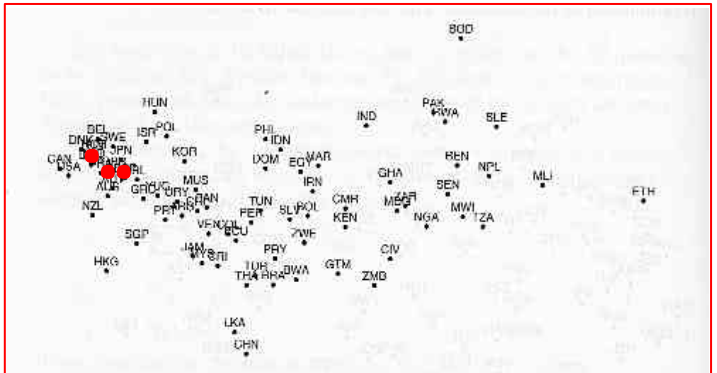


Fig. 1.7. A nonlinear projection constructed using nonmetric MDS [1.27] The data set is the same as in Fig. 1.5 and Fig. 1.6 Missing data values were treated by the following simple method which has been demonstrated to produce good results at least in the pattern recognition context [1.37]. When computing the distance between a pair of data items, only the (squared) differences between component values that are available are computed. The rest of the differences are then set to the average of the computed differences

MDS

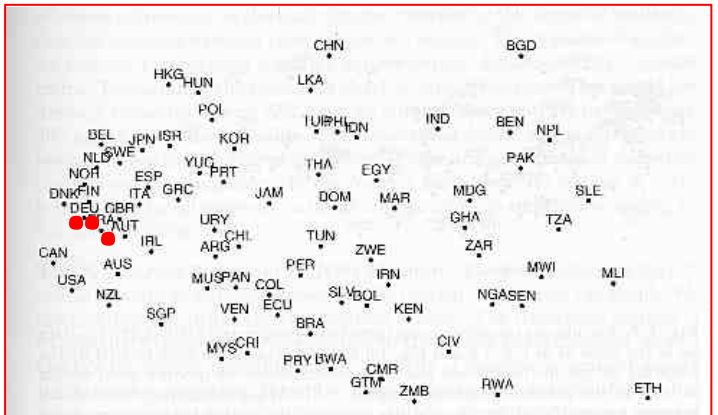


Fig. 1.6. Sammon's mapping of the same data set that was projected using the PCA in Fig. 1.5 [1.27]

Sommon's mapping

Figure 2: A dataset projected linearly onto the two-dimensional subspace obtained with PCA. Each 39-dimensional data item describes different aspects of the welfare and poverty of one country. The data set consisting of 77 countries, used also in Publication 2, was picked up from the World Development Report published by the World Bank (1992). Missing data values were neglected when computing the principal components, and scaled when forming the projections. A key to the abbreviated country names is given in the Appendix.

ACP

A discrete classification

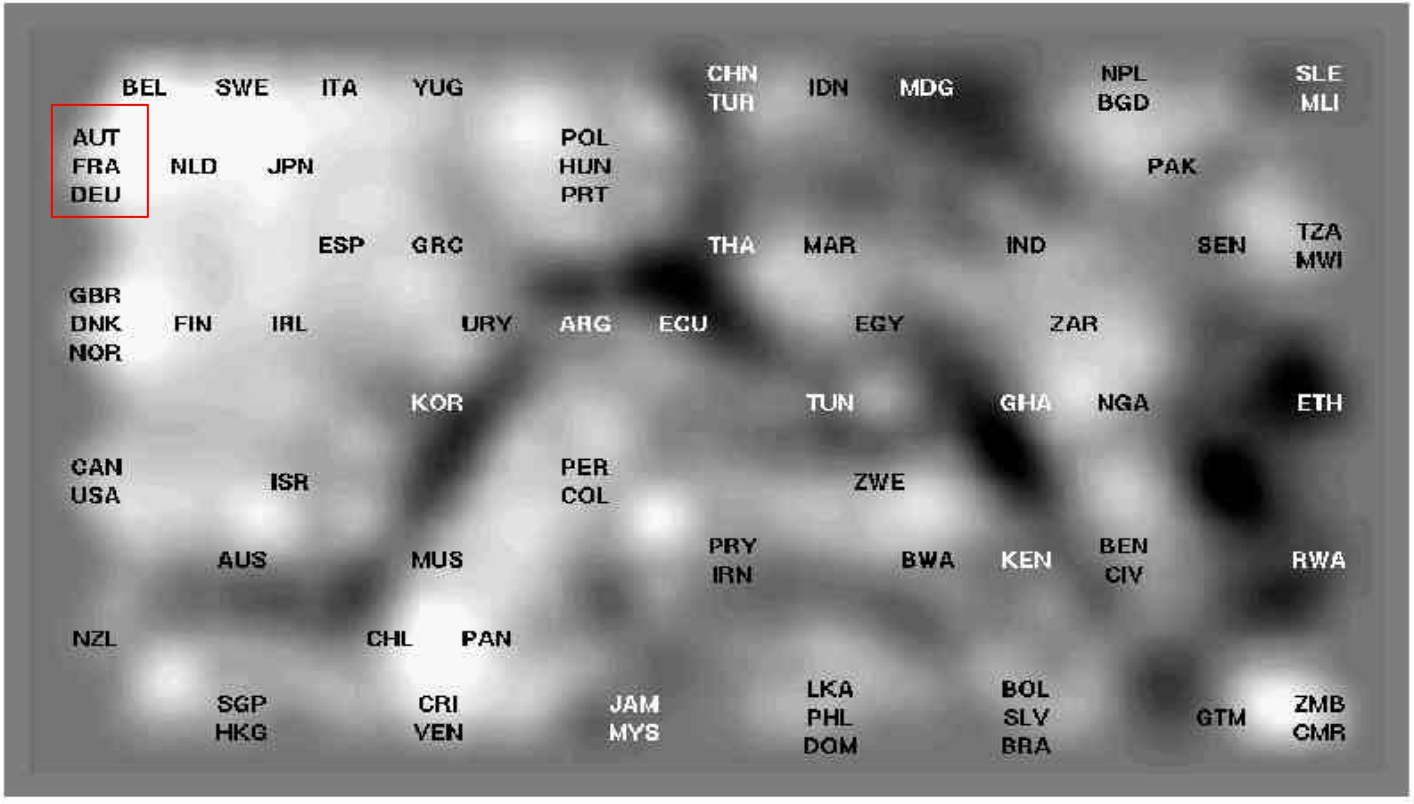
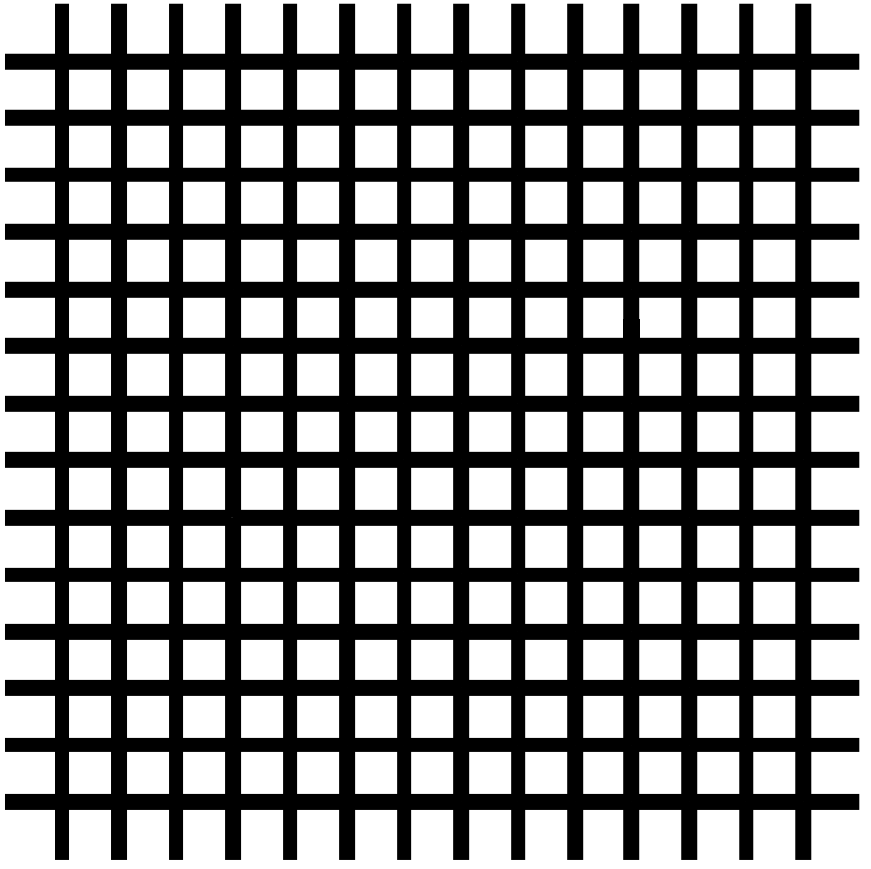


Figure 5: A map display constructed using the SOM algorithm. The overall order of the countries seems to correspond fairly closely to the Sammon's mapping of the same data set (Fig. 4). The most prominent clustering structures are also visible in both displays. Details on how the map was constructed are presented in Publication 2. The size of the map was 13 by 9 units.

Example : poverty in the world

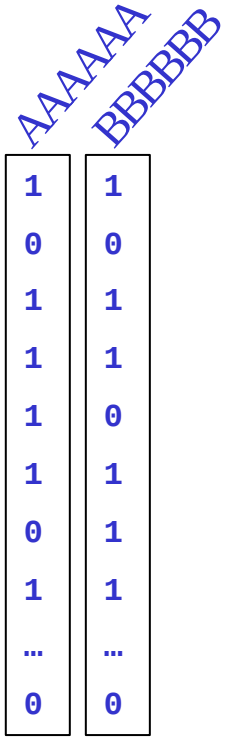
Surnames: *Cluster* a*nalysis*

	Surname 1	Surname 2	Surname 3	Surname 4	
Town1	1	1	1	1	…
Town2	0	0	0	0	
Town3	1	1	1	1	
Town4	0	0	0	0	
…	…	…	…	…	
Town5	0	0	0	0	

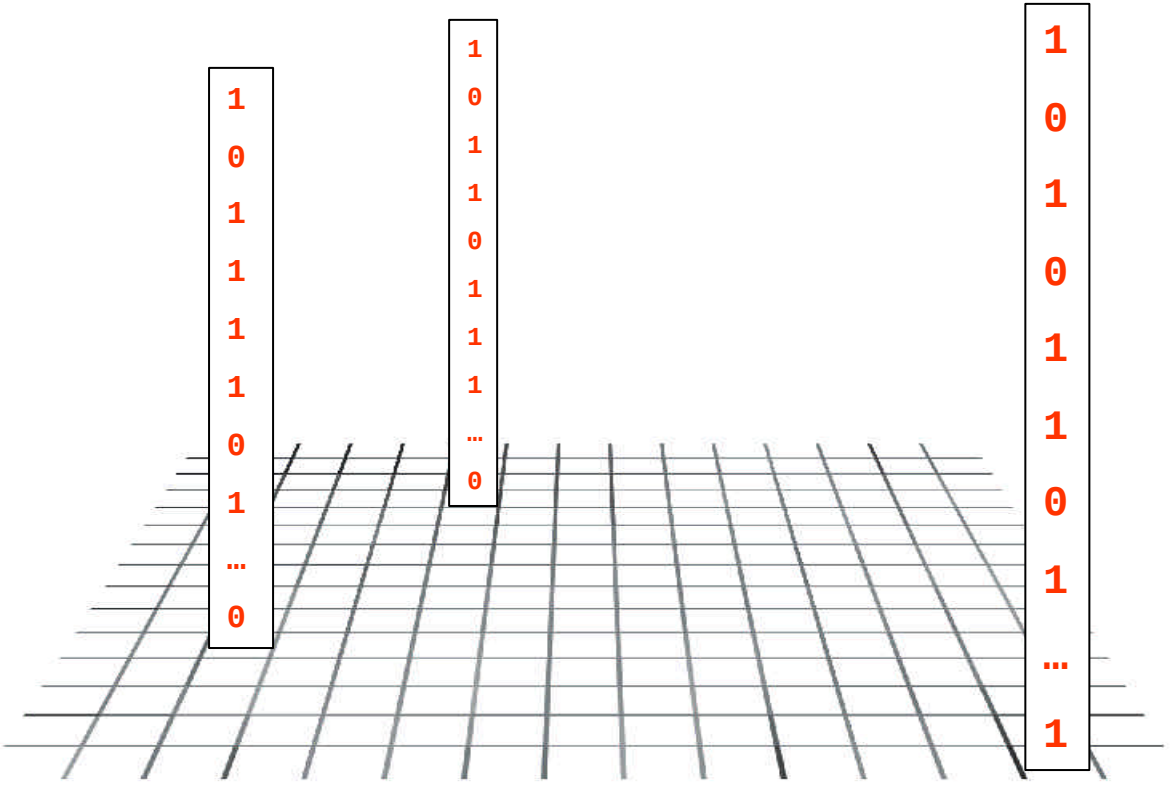


15 x 15

Surnames : *How it works* 1

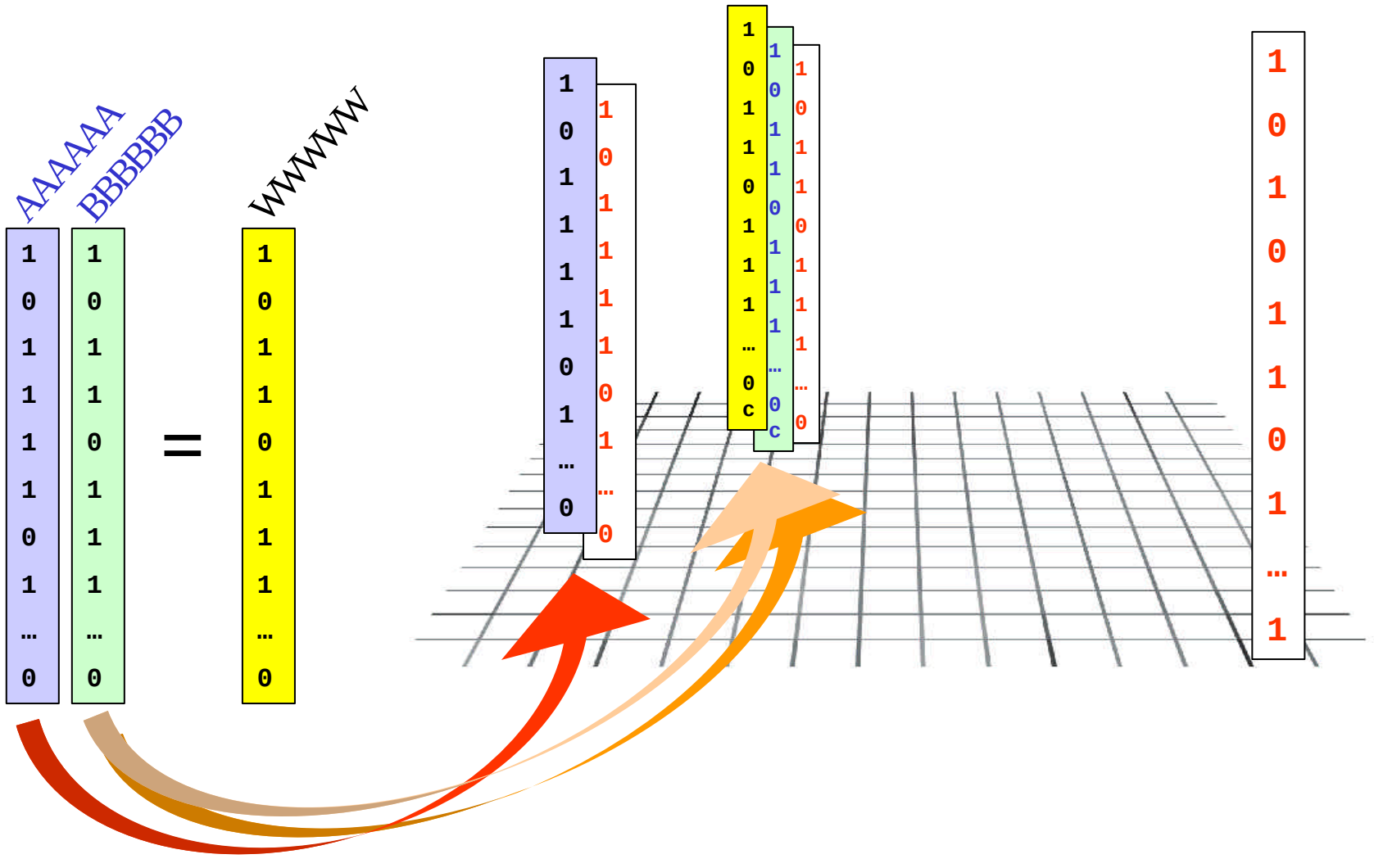


Surnames



Reference vectors on the map

Surnames : *How it works* 2

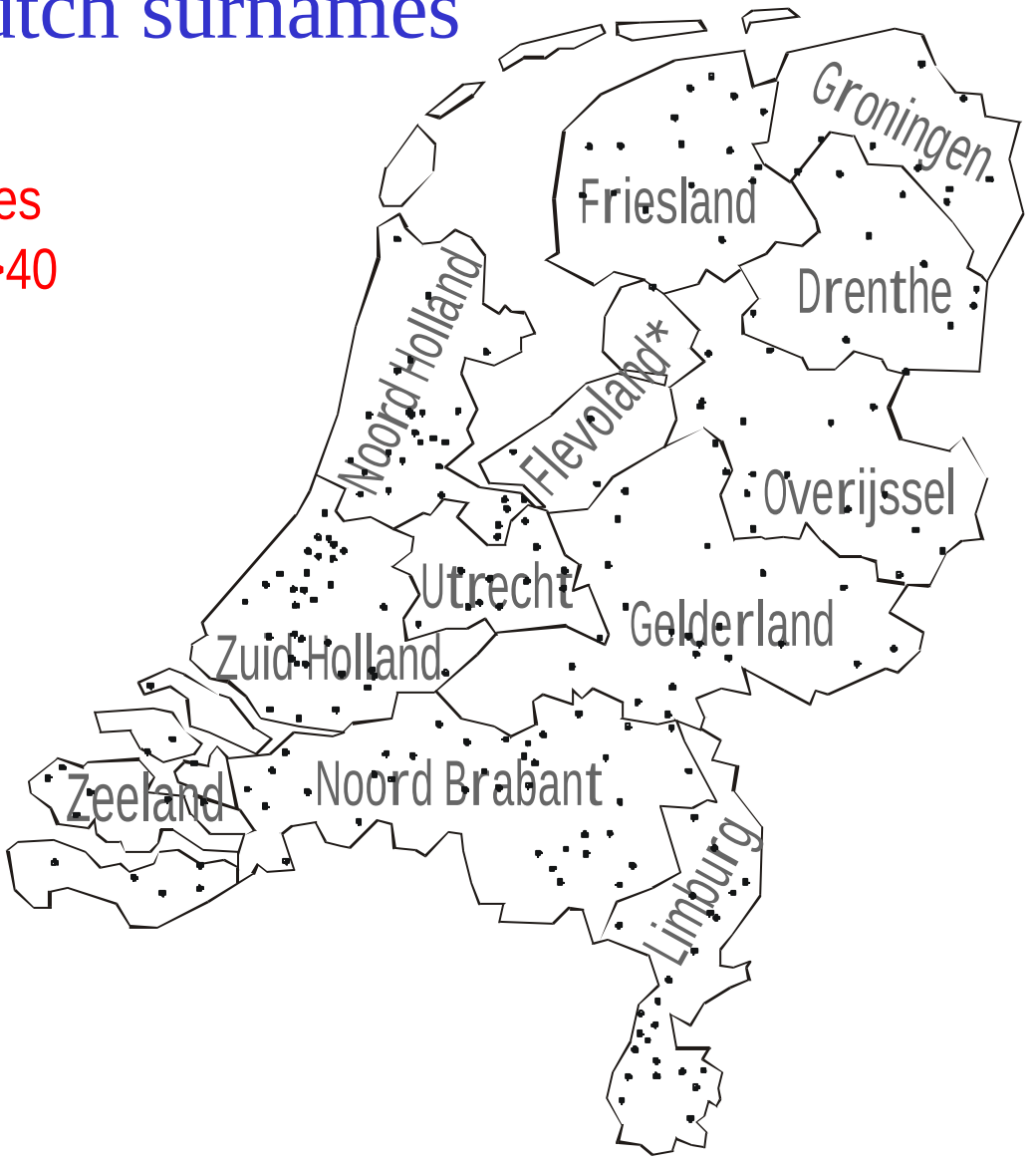


Kohonen maps: *advantages*

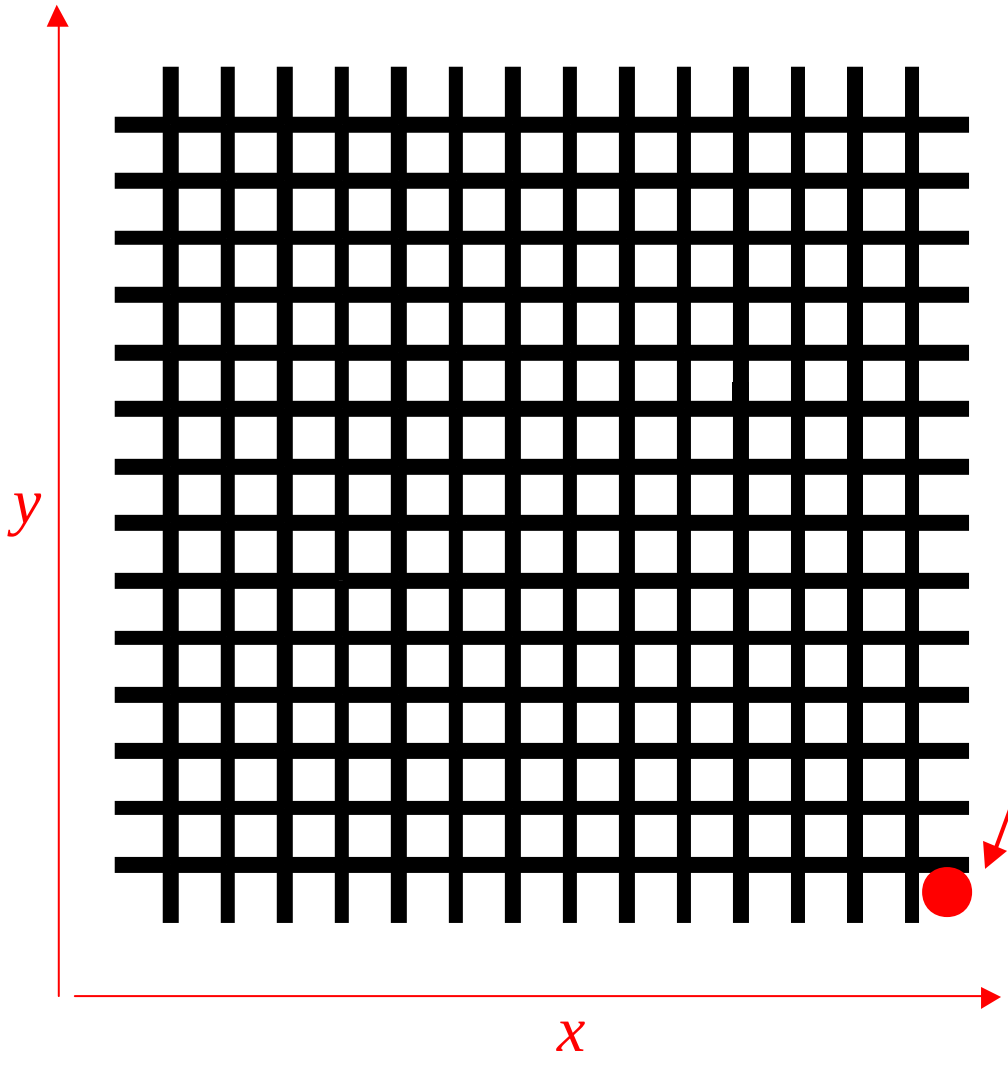
1. Can handle missing values (vectors)
2. It has been shown that the topology in more exact than in MDS or PCA representation.
3. As a consequence, the relations between weakly differenciated populations are more clear.
4. Softwares are very stable and can handle up to 10.000 vectors (surnames) in 226 dimensions (towns) (Penthium II, 1000 MHz, 256 Mo RAM).

Application to Dutch surnames

- 9,929 different surnames (1,642,354 families) f >40
- 226 towns and cities
- 15 x 15 cells map (225 clusters)



How results look like…

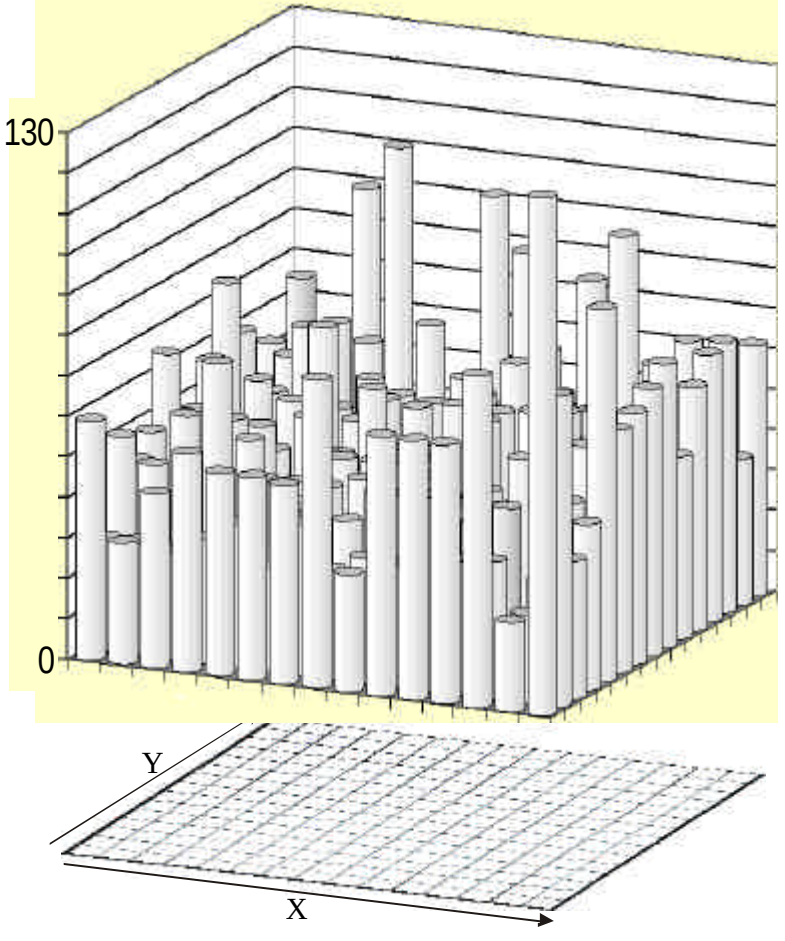


36 surnames:

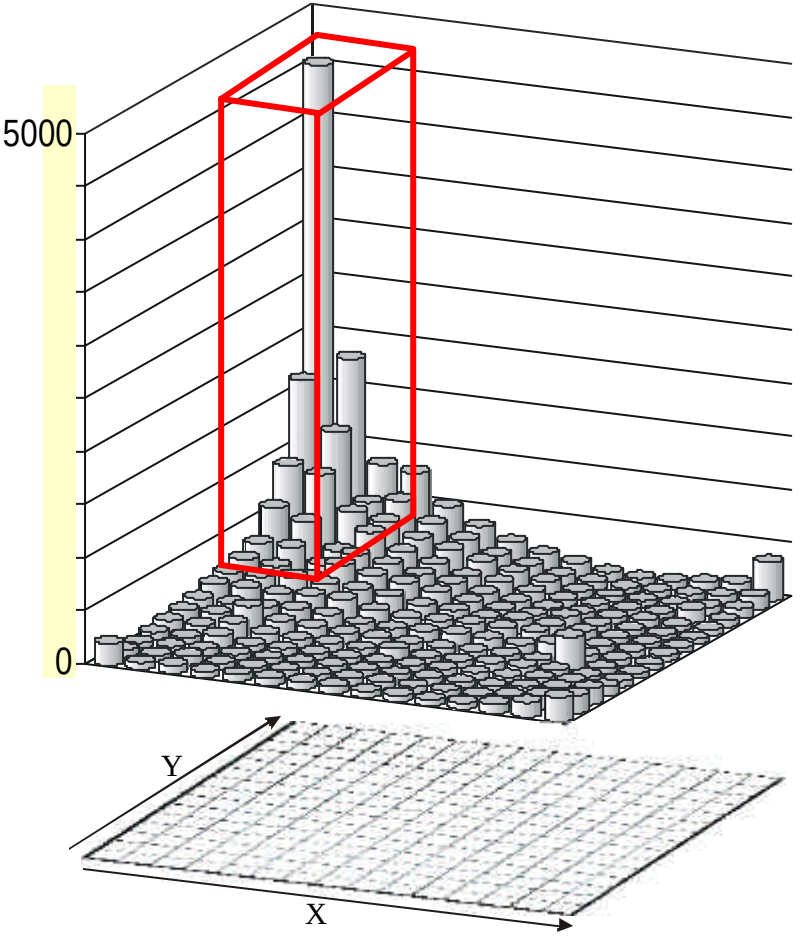
Abma, Algra, Anema, Attema, Baarda, Betten, Bonnema, Bontekoe, Bottema, Brinksma, Cnossen, Cuperus, Damstra, Deelstra, Duiker, Haitsma, Hengst, Hoeksma, Huitema, Hylkema, Iedema, Jelsma, Kammen, Kooiker, Kuiken, Minnema, Mollema, Monsma, Numan, Piersma, Popma, Rienstra, Schaper, Sinnema, Steensma, Vlietstra

Surnames grouped in a same cell will be considered as if they were a single SUPER – SURNAME (GSSGD) (families having a similar migration history)

Absolute frequency of surnames per cell (cluster) ...

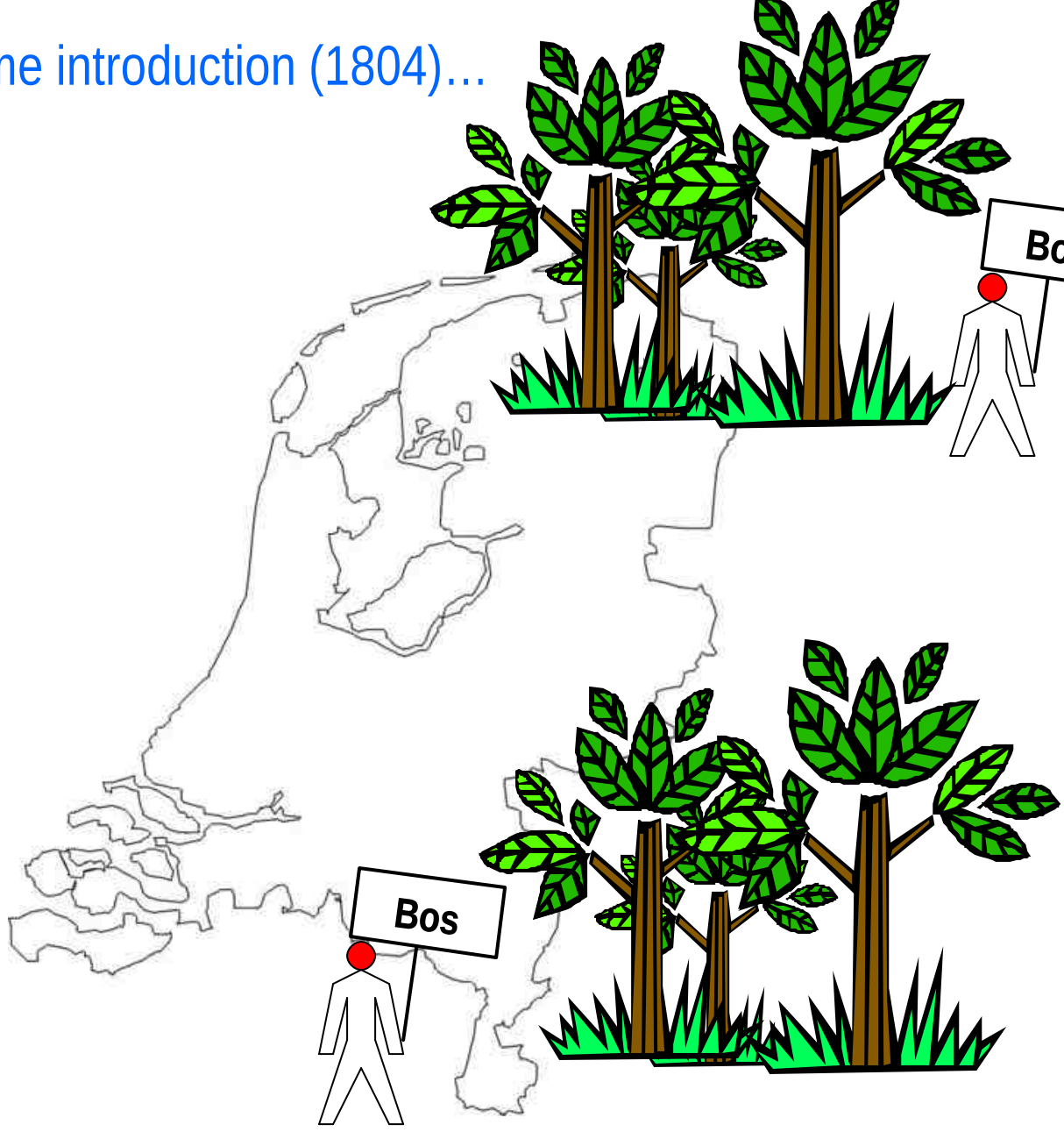


Average number of families sharing such surnames…



At the time of surname introduction (1804)…

Polyphyletism

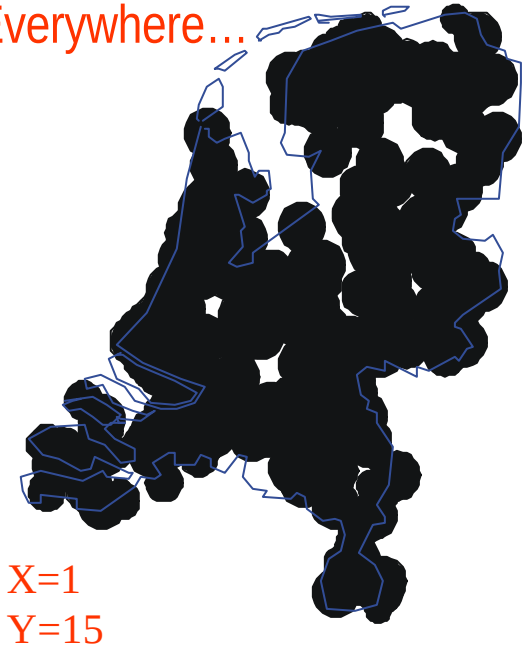


Surnames' frequency vectors undergo a correction by the size of the town/city…

Surname	City 1 *(200,000 hbt.)*	City 2 *(5,000 hbt.)*	
Johnsson	***1000***	***56***	***1556***
	1000 / 200000 = 0.5	***30 / 5000 = 0.65***	***1.15***
	0.50 / 1.15 = 0.43	***0.65 / 1.15 = 0.56***	***1.00***
	43 %	56 %	

Polyphyletism: *its signature…* *(24% of individuals)*

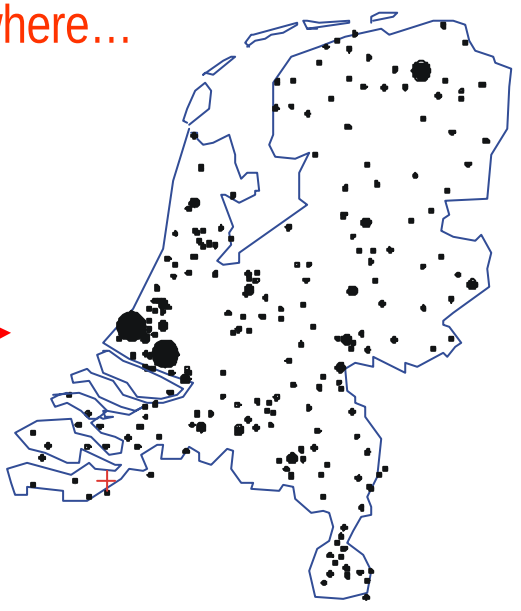
Everywhere…



X=1
Y=15

correction

No where…

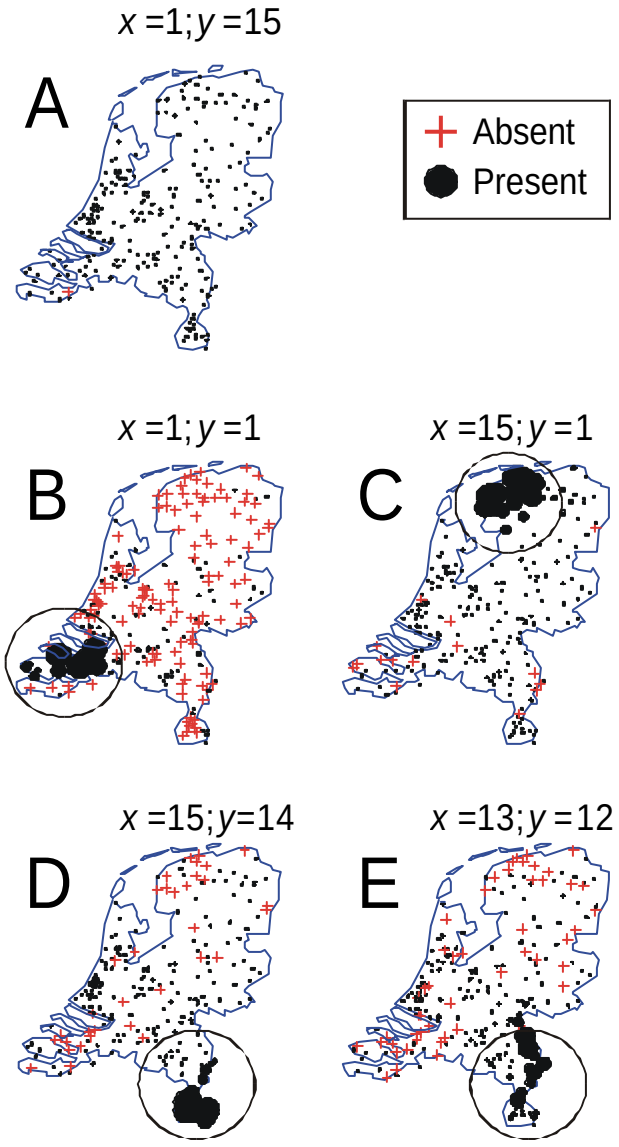


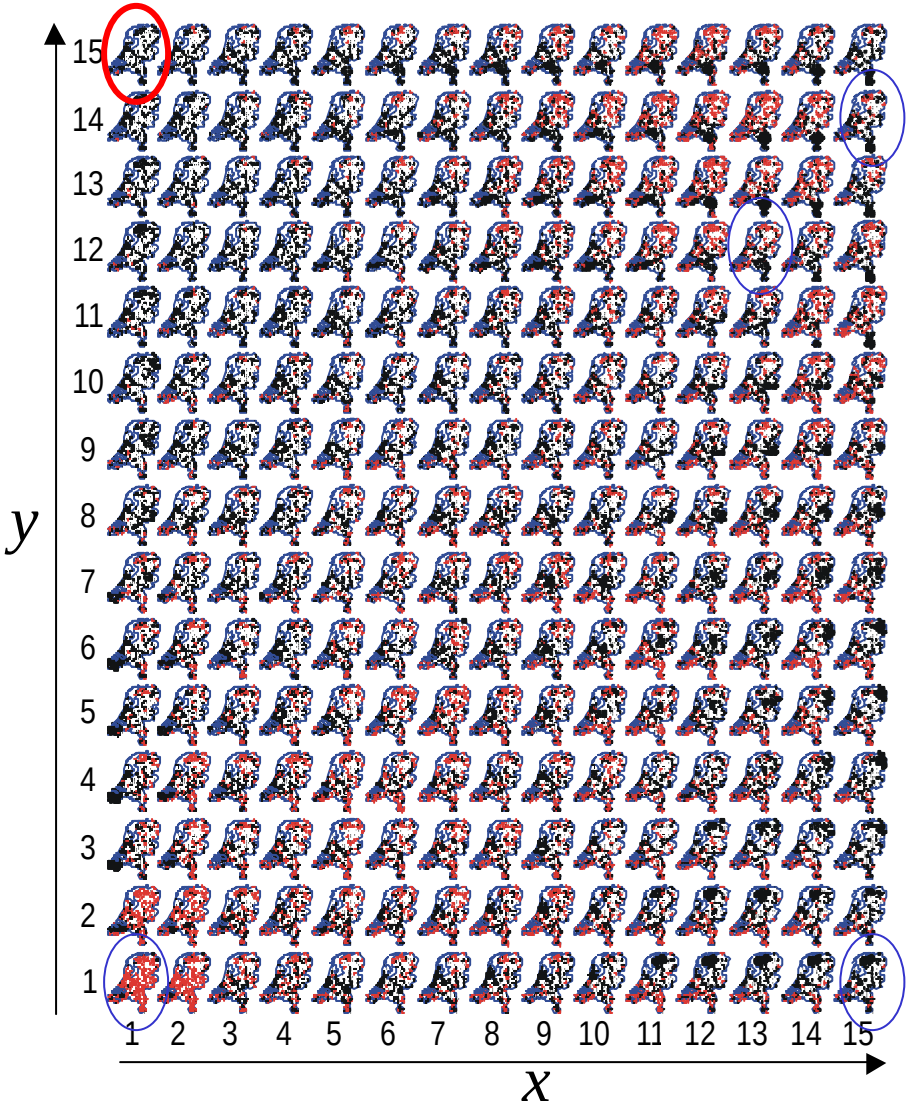
68 surnames:

Beek, Berg, Bijl, Blom, Boer, Bos, Bosch, Bosman, Brink, Broek, Brouwer, Dam, Dekker, Dijk, Dijkstra, Graaf, Groot, Haan, Hendriks, Hoek, Horst, Huisman, Jager, Jansen, Janssen, Jong, Jonge, Kamp, Kok, Koning, Koster, Kramer, Kroon, Kuipers, Laan, Lange, Leeuw, Leeuwen, Linden, Meer, Meijer, Mulder, Peters, Post, Roos, Ruiter, Smits, Valk, Veen, Velde, Vermeulen, Visser, Vliet, Vonk, Vos, Vries, Wal, Wijk.

® Paesant; Wood; «From the wood»; King; Brewer; Hunter; «From the dam»; Chevalier; Fisherman; Big; Young; The young…

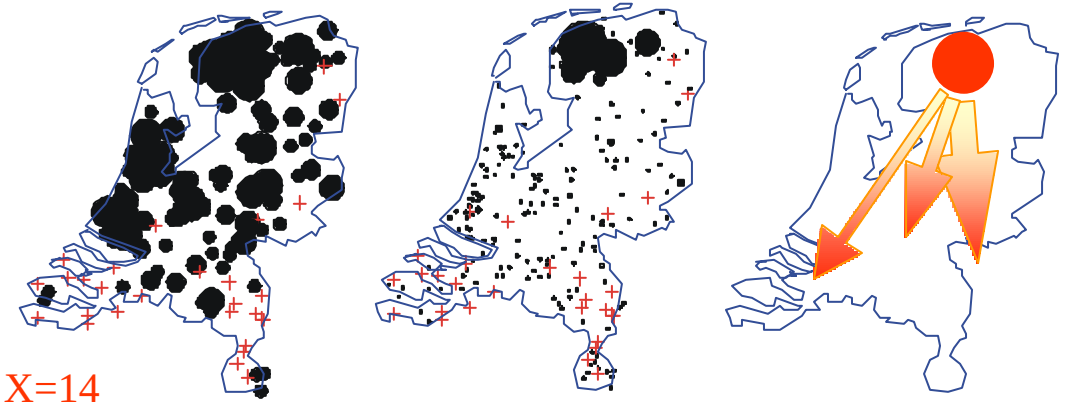
Geographic origin…



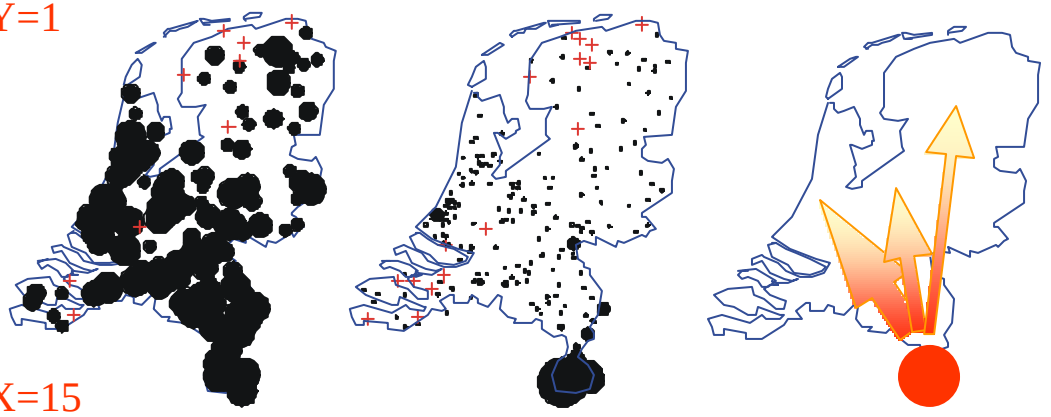


Migrations: *How many? Where from? Where to?*

Distribution Frequence Origin



X=14
Y=1



X=15
Y=15

2

36 surnames:

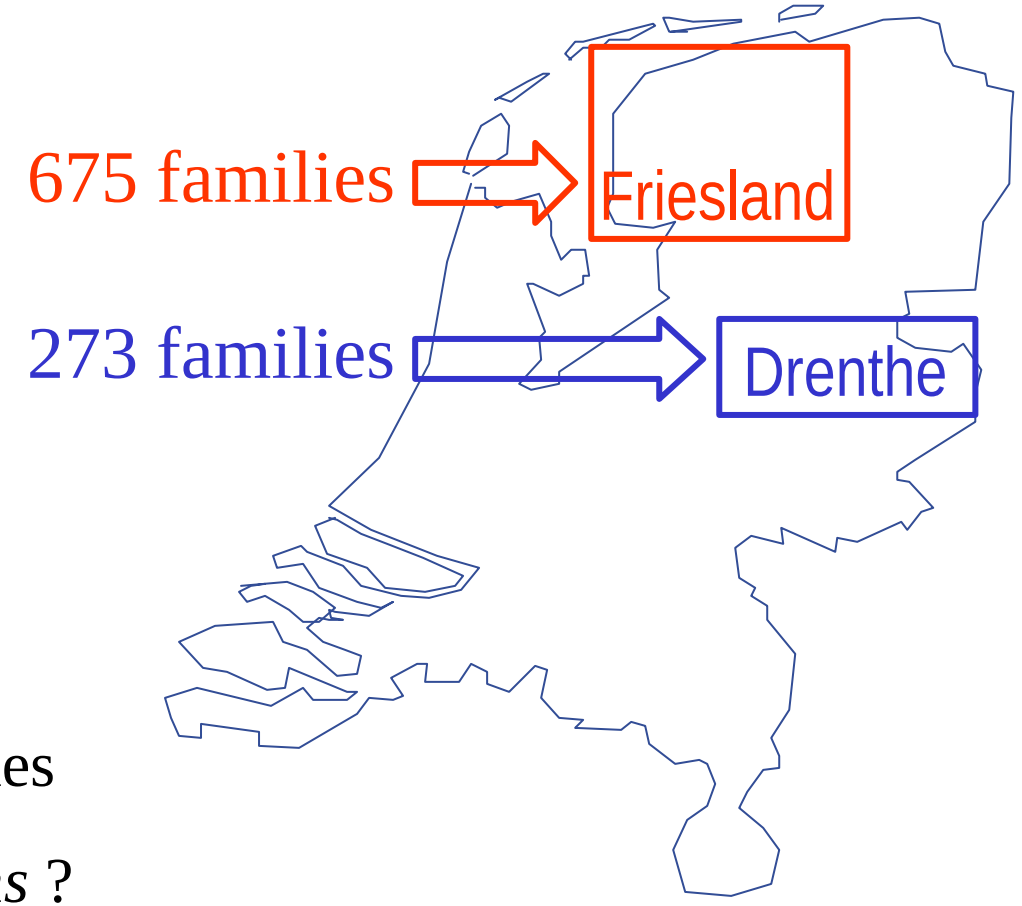
Abma, Algra, Anema, Attema, Baarda, Betten, Bonnema, Bontekoe, Bottema, Brinksma, Cnossen, Cuperus, Damstra, Deelstra, Duiker, Haitsma, Hengst, Hoeksma, Huitema, Hylkema, Iedema, Jelsma, Kammen, Kooiker, Kuiken, Minnema, Mollema, Monsma, Numan, Piersma, Popma, Rienstra, Schaper, Sinnema, Steensma, Vlietstra

3

65 surnames:

Beurskens, Bisschops, Boonen, Breuer, Brouns, **Claassens**, **Claessen**, Cleef, Coenders, Coumans, Creemers, **Cuijpers**, Custers, **Cuypers**, Daemen, Daniels, Dassen, Dohmen, Eyck, Frenken, Gerards, Gijsen, Godschalk, Gubbels, Habets, Haenen, Hermens, Heynen, Hillen, Houwen, Jetten, Jeurissen, Knippenberg, Knops, Kurvers, Lenssen, Leurs, Maessen, Mans, Megen, Meuwissen, Michiels, Mommers, Palmen, Paulissen, Paulussen, Puts, Ramaekers, Ramakers, Reynders, Rijks, Rongen, Sassen, Schendel, Schols, Seuren, Sieben, Theelen, Thissen, Tummers, Verheggen, Verlinden, Vinken, Vroomen, Weerts

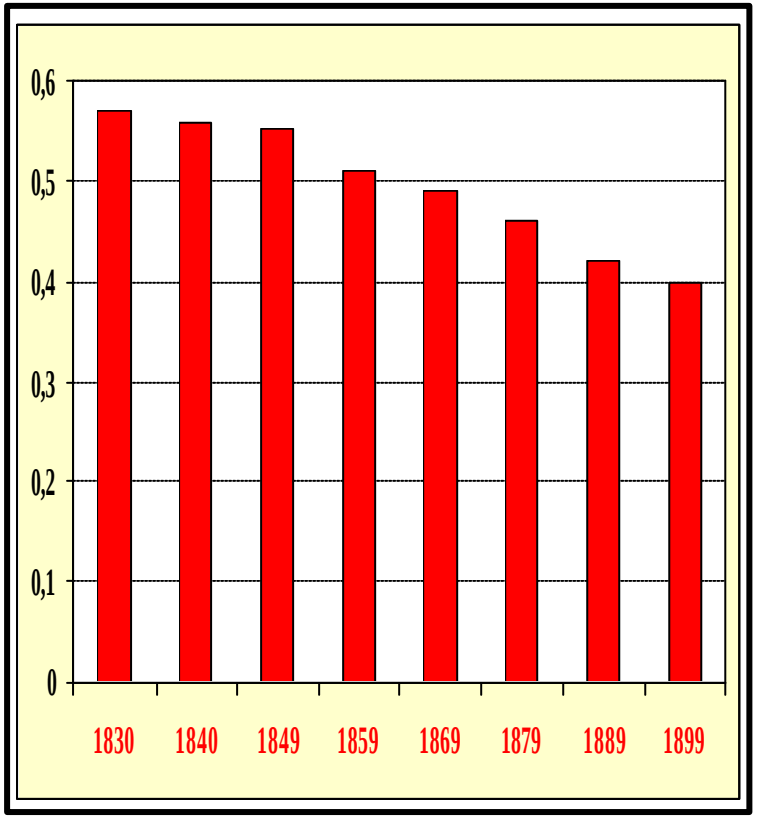
A well identified population:



- Extinction of surnames
- Size of families: *clans* ?
- We only considered those surnames $f > 40$ individuals

Correlations with census data

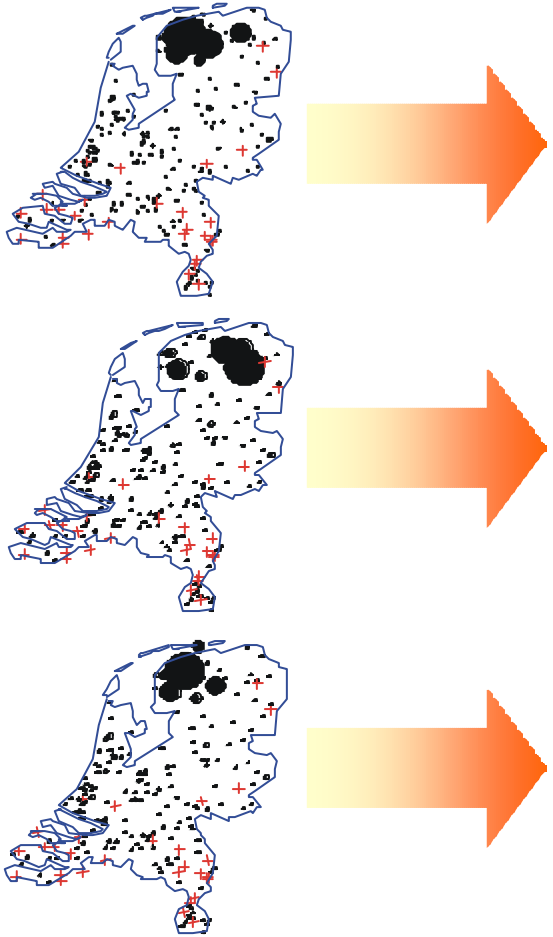
	1830 census	SOMs
Drenthe	21,328	273
Friesland	64,841	675
Gelderland	101,594	716
Groningen	50,161	345
Limburg	60,527	1,025
Noord Brabant	116,244	1,193
Noord Holland	117,971	610
Overijssel	58,344	699
Utrecht	41,814	399
Zeeland	42,824	864
Zuid Holland	141,41	1,035
Total	*817,058*	*7,834*



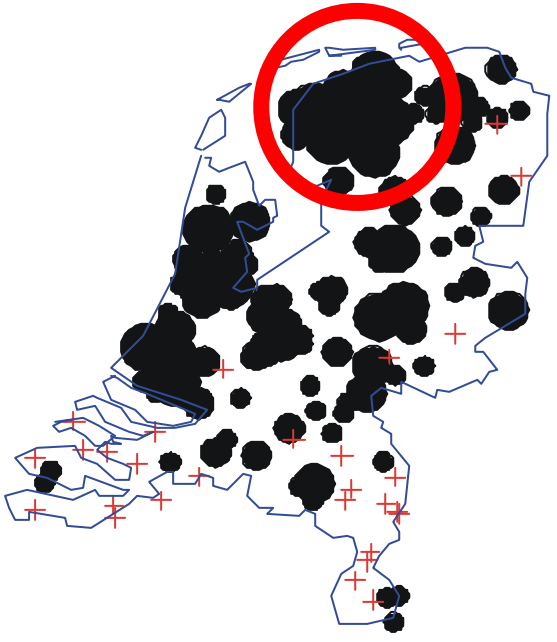
Census	1830	1840	1849	1859	1869	1879	1889	1899
Correlation	0,57	0,56	0,55	0,51	0,49	0,46	0,42	0,4
	p<0.05	p<0.05	p<0.05	p<0.05	p<0.05	p<0.05	p<0.05	p<0.05

How many didn't move?

(with reference to the time of surnames' origin)

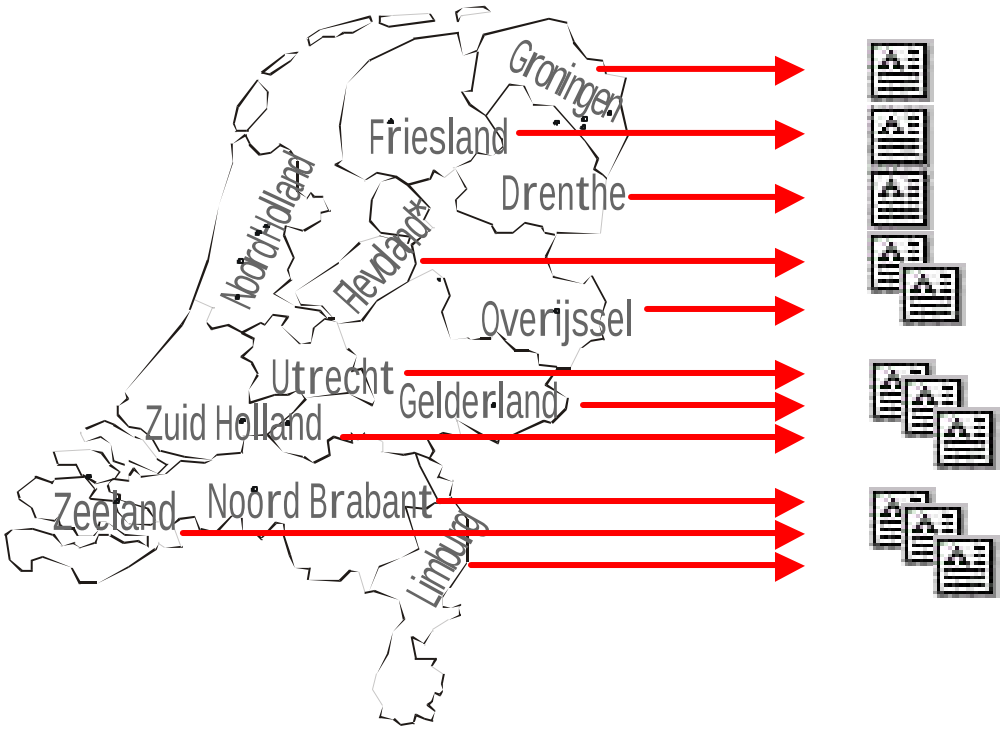


Abma, Algra, Anema, Attema, Baarda, Betten, Bonnema, Bontekoe, Bottema, Brinksma, Cnossen, Cuperus, Damstra, Deelstra, Duiker, Haitsma, Hengst, Hoeksma, Huitema, Hylkema, Iedema, Jelsma, Kammen, Kooiker, Kuiken, Minnema, Mollema, Monsma, Numan, Piersma, Popma, Rienstra, Schaper, Sinnema, Steensma, Vlietstra Beurskens, Bisschops, Boonen, Breuer, Brouns, Claassens, Claessen, Cleef, Coenders, Coumans, Creemers, Cuijpers, Custers, Cuypers, Daemen, Daniels, Dassen, Dohmen, Eyck, Frenken, Gerards, Gijsen, Godschalk, Gubbels, Habets, Haenen, Hermens, Heynen, Hillen, Houwen, Jetten, Jeurissen, Knippenberg, Knops, Kurvers, Lenssen, Leurs, Maessen, Mans, Megen, Meuwissen, Michiels, Mommers, Palmen, Paulissen, Paulussen, Puts, Ramaekers, Ramakers, Reynders, Rijks, Rongen, Sassen, Schendel, Schols, Seuren, Sieben, Theelen, Thissen, Tummers, Verheggen, Verlinden, Vinken, Vroomen, Weerts…

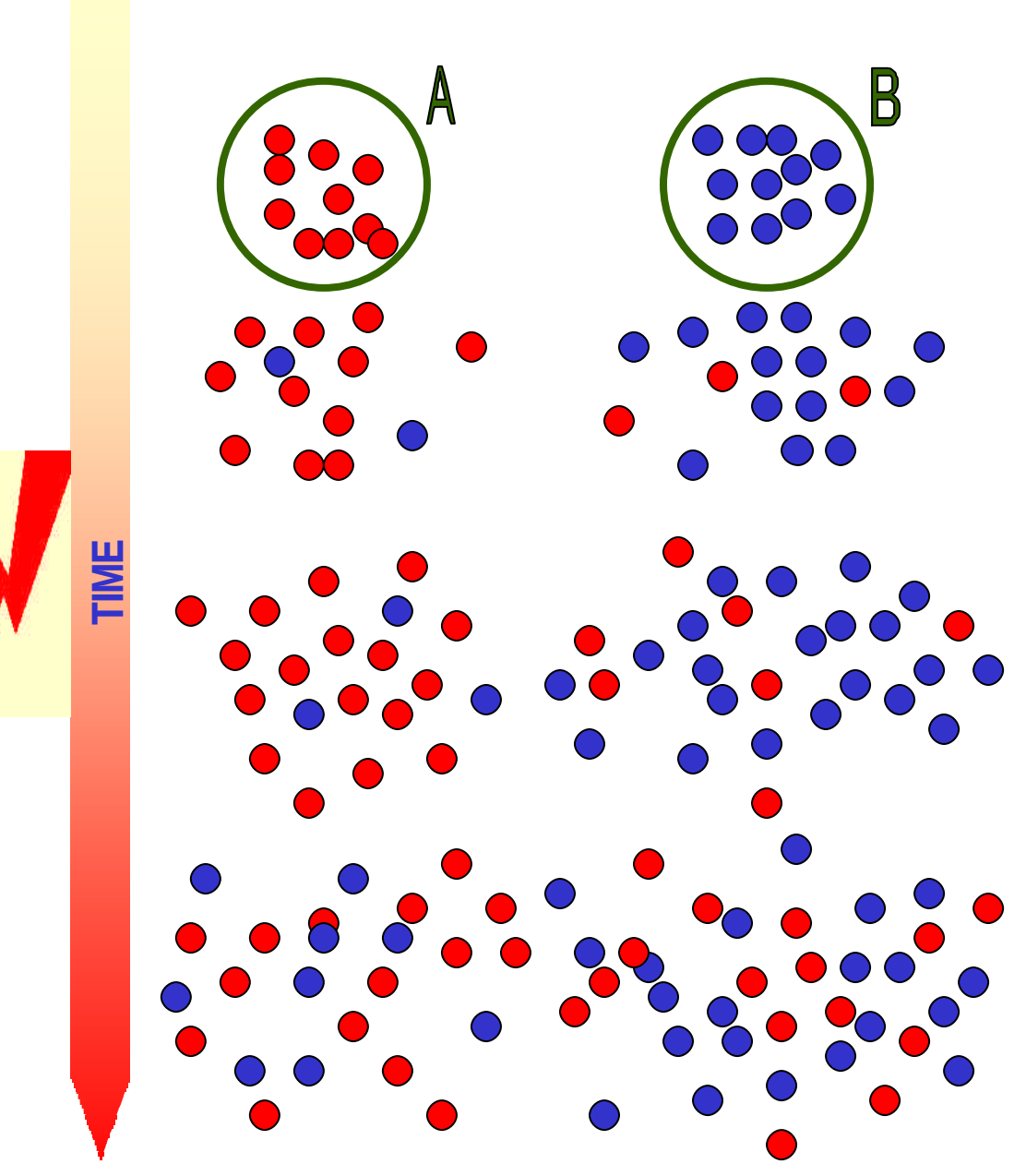


20 %

Improving the quality of Y chromosome samplings …



Since we know which individuals live where their ancestors lived two centuries ago, then we can sample **only** corresponding families (surnames) !!!



Ancient times (?)

1800 – Netherlands

XVI / XVII century Europe

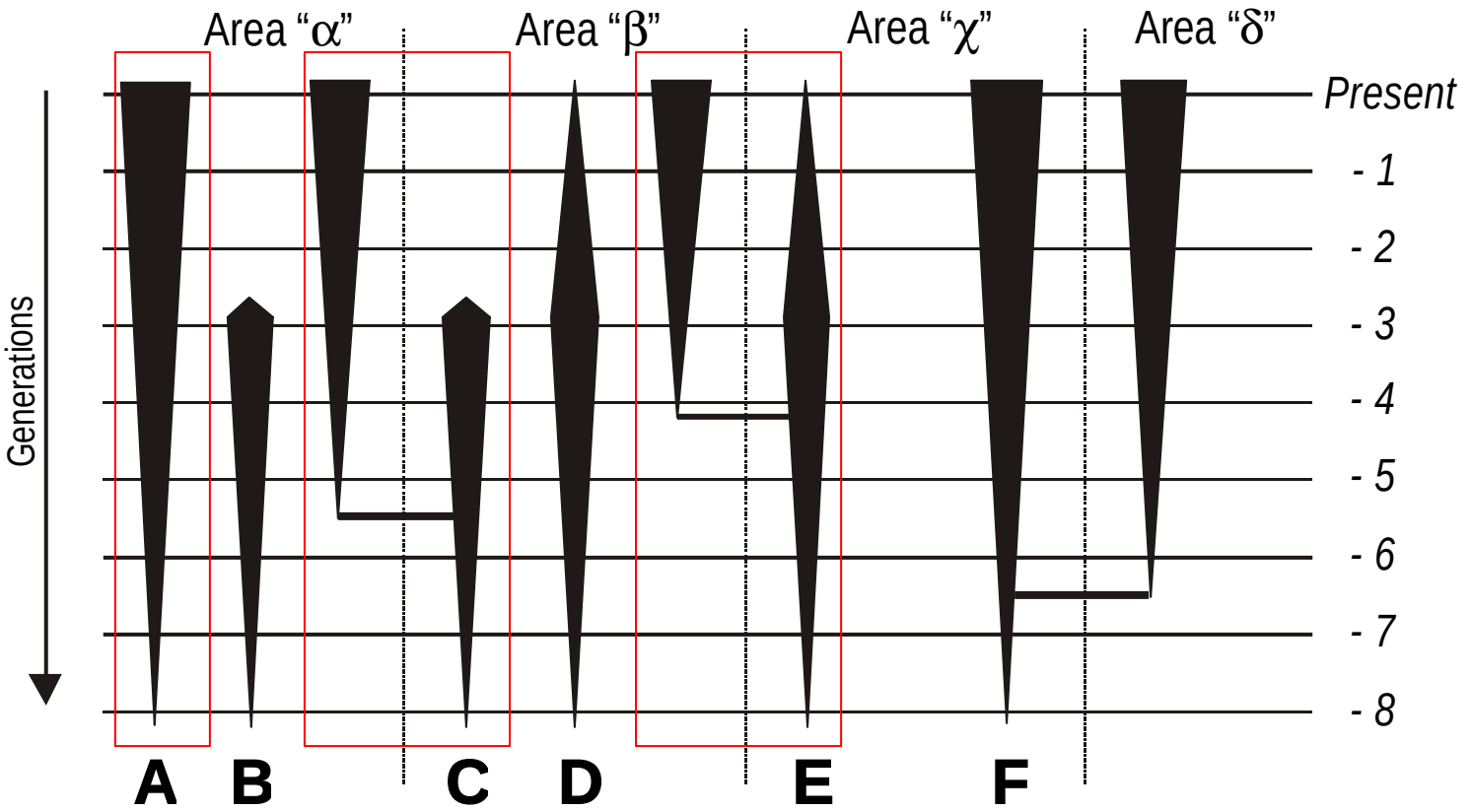
Surnames criterion

1880/1930

Grandparents

P r e s e n t

Fig. 5


Area “α”
Area “β”
Area “χ”
Area “δ”
Generations
Present
- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
A
B
C
D
E
F

AMERICAN JOURNAL OF PHYSICAL ANTHROPOLOGY 126:214–228 (2005)

New Method for Surname Studies of Ancient Patrilineal Population Structures, and Possible Application to Improvement of Y-Chromosome Sampling

Franz Manni,* Bruno Toupance, Audrey Sabbagh, and Evelyne Heyer

Leidse Wetenschappers
professoren

English Version
ontact

Prof.dr. P. de Knijff (Peter)

Prof. dr. Manfred Kayser

Erasmus Universiteit Rotterdam
Erasmus University (Internationaal)

Presentatie | Kwaliteit | Opleidingen

Welkom bij de Erasmus Universiteit Rotterdam | Uitstekende studiefaciliteiten | Rotterdam St

De Universiteit voor ambitieuze studenten

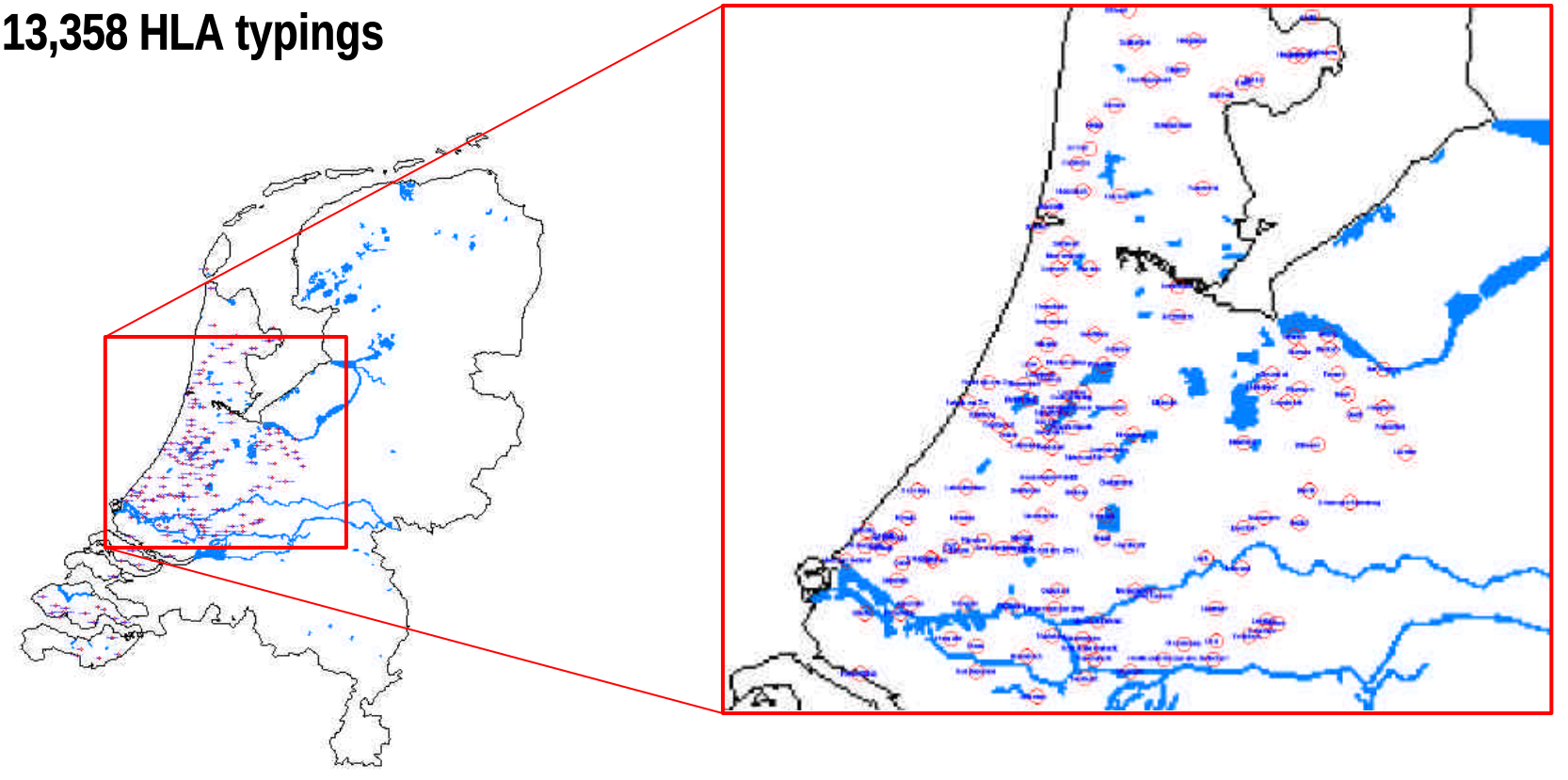
De Erasmus Universiteit Rotterdam (EUR) is een middelgrote universiteit met zo'n 24.000 studenten. De EUR biedt opleidingen op de gebieden **Economie en Management, Geneeskunde en Gezondheid**, en **Recht, Cultuur en Maatschappij**. De universiteit kenmerkt zich door de combinatie van academische vorming, internationale oriëntatie en maatschappelijke relevantie. Studeren aan de EUR betekent persoonlijke begeleiding, netwerken bouwen en natuurlijk uitstekende faciliteiten.

The Wenner-Gren Foundation

supporting worldwide research in all branches of anthropology

We would like to compare a **random sample** of the Dutch population with a sample selected according to surnames specific of given locations.

13,358 HLA typings



The Wenner-Gren Foundation

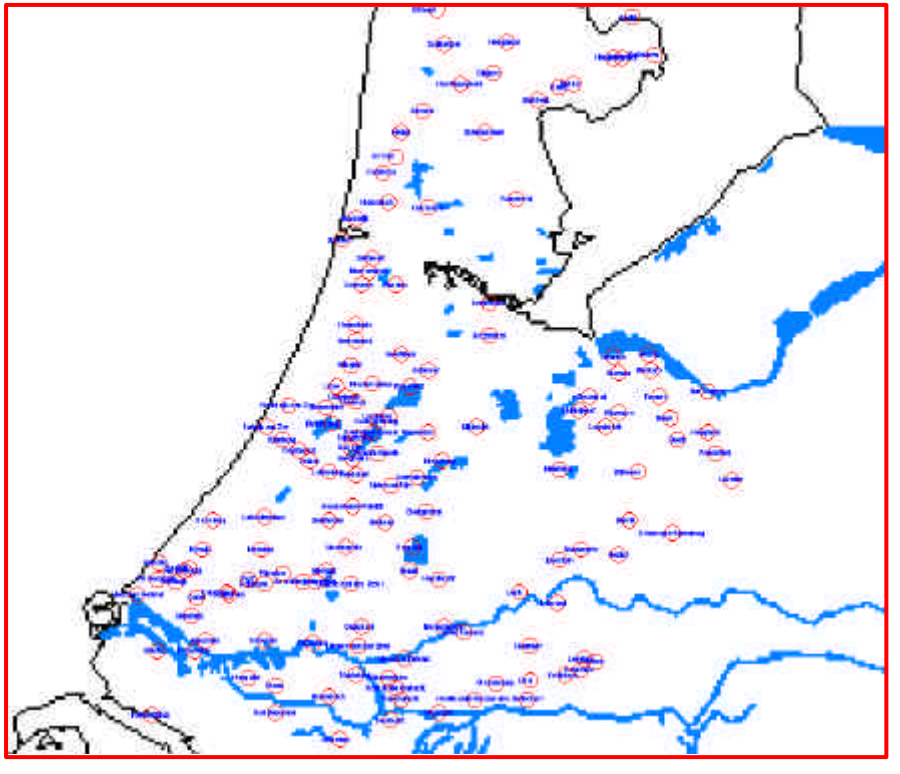
supporting worldwide research in all branches of anthropology

We would like to compare a **random sample** of the Dutch population with a sample selected according to surnames specific of given locations.

13,358 HLA typings

210 locations

Province	Locations	Individuals
N. Holland	60	2007
Z Holland	104	10,582
Utrecht	22	301
Zeeland	24	468



The Wenner-Gren Foundation

supporting worldwide research in all branches of anthropology

We would like to compare a **random sample** of the Dutch population with a sample selected according to surnames specific of given locations.

13,358 HLA typings

210 locations

Province	Locations	Individuals
N. Holland	60	2007
Z Holland	104	10,582
Utrecht	22	301
Zeeland	24	468

People with a surname really from Noord Holland, Zuid Holland, Utrecht and Zeeland: **1310**

We compared such 1310 individuals with the resting 12,048 ones

There is a statistically significant difference

Wait a second, please…

WAYNE STATE UNIVERSITY PRESS

http://www.humbiol.com

New editorship (october 2008)

Evelyne Heyer

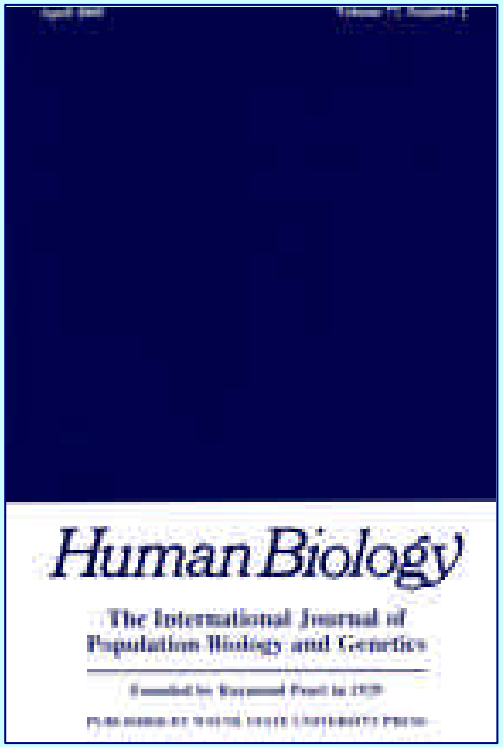
Editor-in-Chief

Franz Manni

Executive Editor

Guido Barbujani

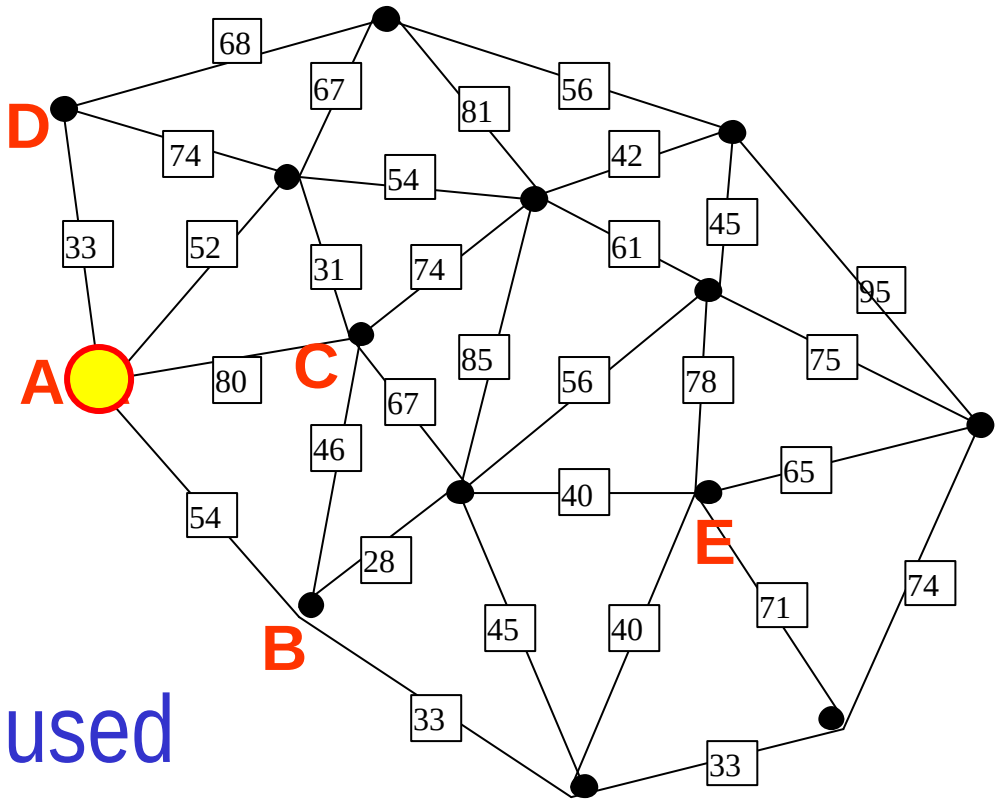
Associate Editor



Interdisciplinary papers : population genetics, cultural evolution, demography, anthropology...

	A	B	C	D	E
A	0				
B	54	0			
C	80	38	0		
D	33	61	78	0	
E	40	28	74	33	0

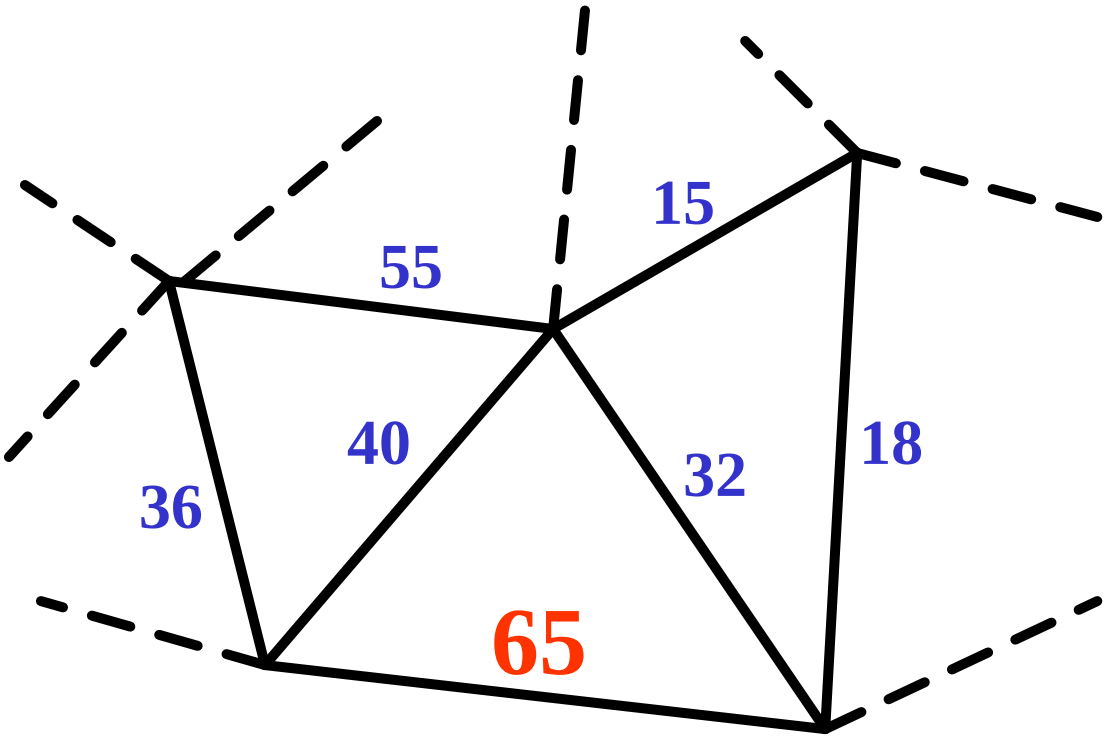
...



The matrix is partially used

Monmonier algorythm

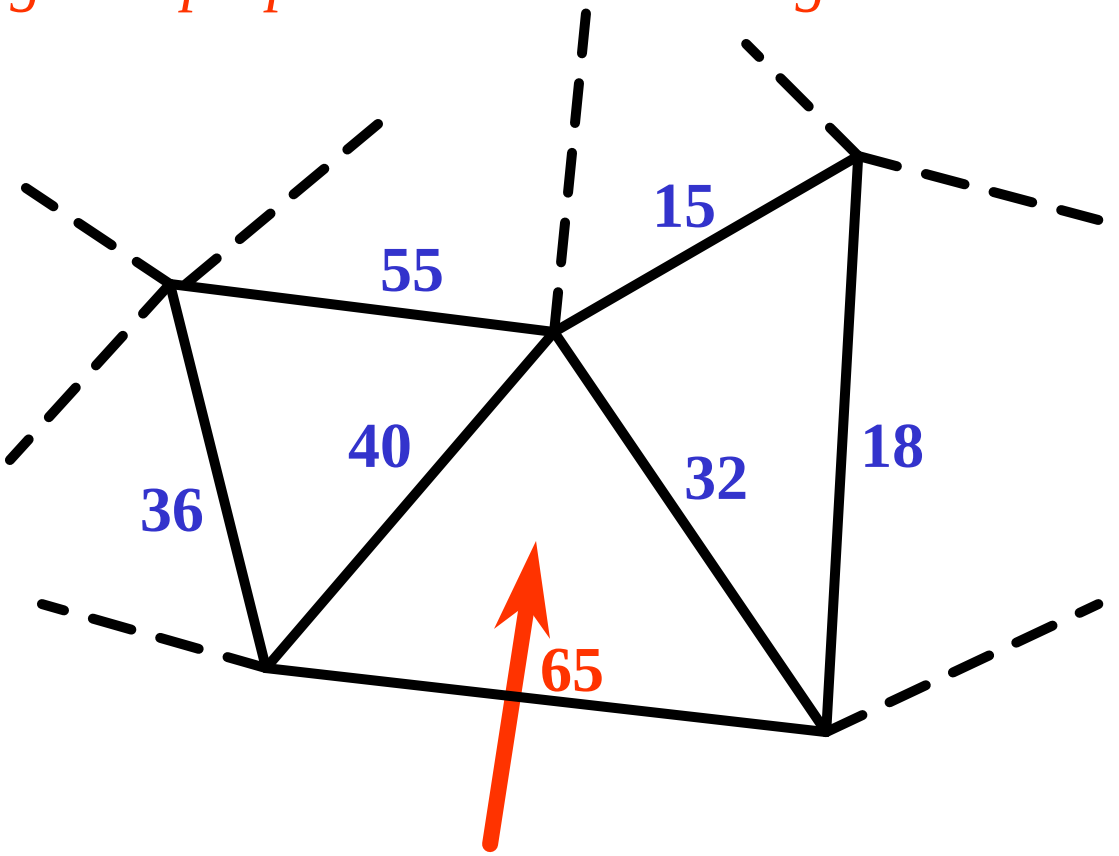
1) Search for the higher distance value



Monmonier algorythm

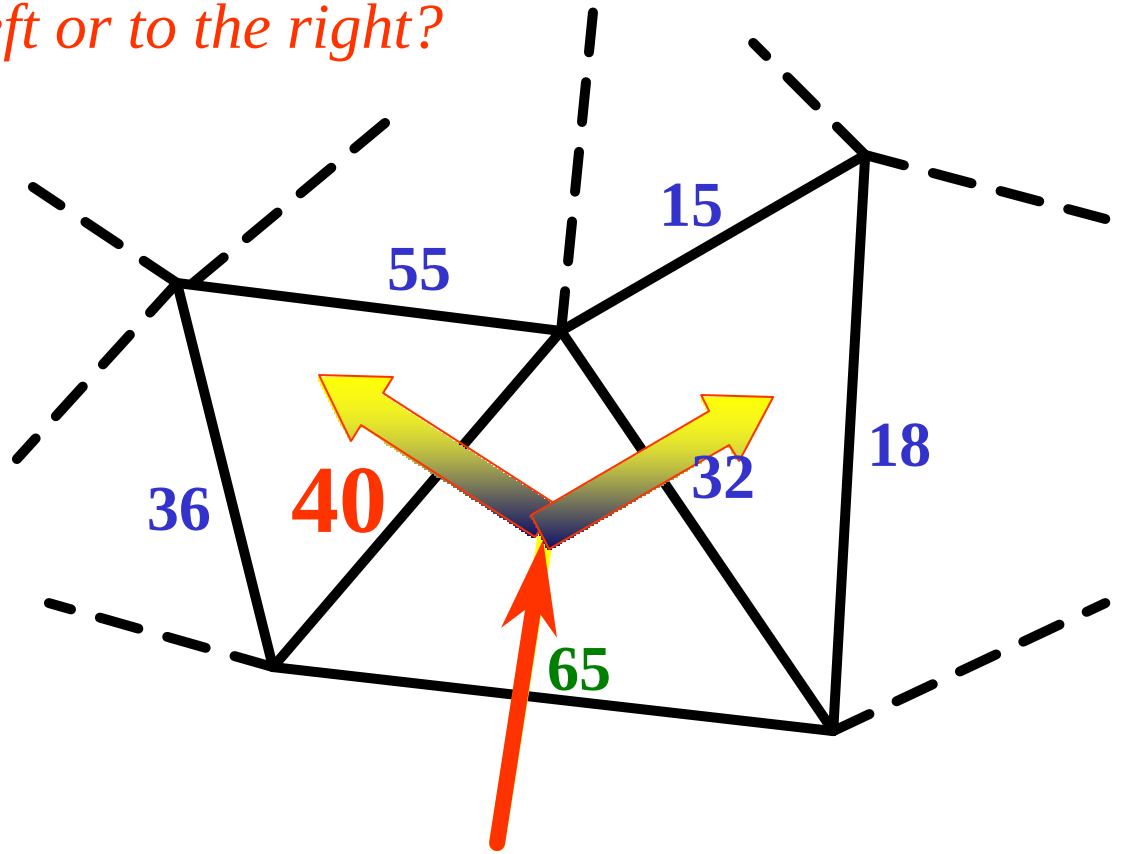
1) Search for the higher

2) Trace a segment perpendicoular o the edge



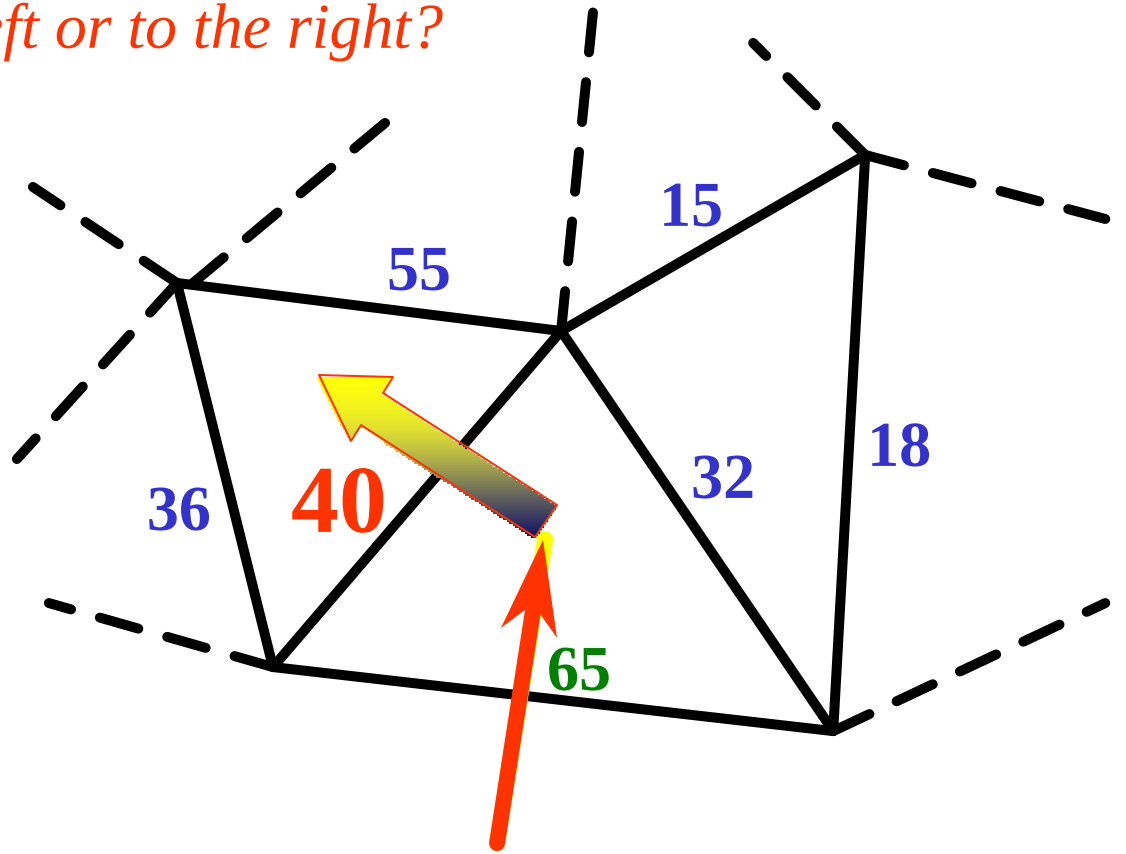
Monmonier algorythm

1) *Search for the highest distance value*
2) *Trace a segment perpendicoular to the edge of triangle*
3) *To the left or to the right?*



Monmonier algorythm

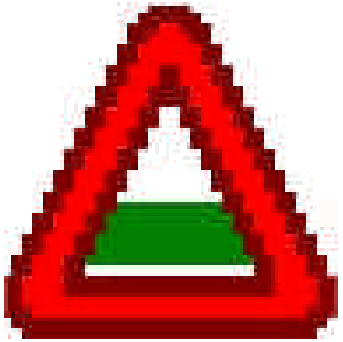
1) *Search for the highest distance value*
2) *Trace a segment perpendicoular to the edge*
3) *To the left or to the right?*



Barrier vs. 2.2 *for MS Windows*

a software to compute geographic barriers from a distance matrix

By F. Manni and E. Guérard



Google Nederland : **“barrier” “2.2”**