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
Human population genetics

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.
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 difficult to have all the genealogies of the populations and, even when possible, their study requires years.

A WAY-OUT is to estimate the levels of isonymy 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:

<table>
<thead>
<tr>
<th>Location</th>
<th>Surname Distribution</th>
<th>Isonymy (A)_{Nerbonne}</th>
<th>Isonymy (B)_{Nerbonne}</th>
<th>Isonymy (AB)_{Nerbonne}</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>10 «Nerbonne» over 100 inhabitants</td>
<td>10% x 10% = 0.01</td>
<td>25% x 25% = 0.0625</td>
<td>10% x 25% = 0.025</td>
</tr>
<tr>
<td>B</td>
<td>25 «Nerbonne» over 100 inhabitants</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Lasquer distance:**
$$\frac{\sum n_{si}n_{sj}}{2\sum n_{si}\sum n_{sj}}$$

**Nei distance:**
$$\frac{\sum n_{si}n_{sj}}{\left(\sum n_{si}^2 \sum n_{sj}^2\right)^{\frac{1}{2}}}$$
Isonymy and the Netherlands: Nei distance

\[ \frac{\sum n_{si} n_{sj}}{\left( \sum n_{si}^2 \sum n_{sj}^2 \right)^{\frac{1}{2}}} \]

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 isonymy: Nei distance

\[ \frac{\sum n_{si} n_{sj}}{\left( \sum n_{si}^2 \sum n_{sj}^2 \right)^{\frac{1}{2}}} \]

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.

HIGH consanuinity
Surname differentiation can be computed, according to NEI as:

\[
\frac{\sum n_{si} n_{sj}}{\left( \sum n_{si}^2 \sum n_{sj}^2 \right)^{\frac{1}{2}}}
\]

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.
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

\[ y = 0.7041x + 6.0579 \]

\[ R^2 = 0.4495 \]
Isonimy and the Netherlands: Nei distance
Multidimensional plot of surname distances

Coord. 1

Without polyphyletic surnames

Limburg
Multidimensional plot of surname distances

Brabant; Zeeland
Multidimensional plot of surname distances

Coord. 1

N/S Holland
Multidimensional plot of surname distances

Coord. 1

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*
No intermarriages
End of first part:

« Surnames cooked with a standard sauce »

(little pause)

next:

« Surnames are also words... »
Is there a link between 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 between cultural and genetic diversity?

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 depths, 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.
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 explanation*
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

<table>
<thead>
<tr>
<th></th>
<th>SURNAMES</th>
<th>DIALECTS (72)</th>
<th>GEOGRAPH. DISTANCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>SURNAMES</td>
<td>1</td>
<td>0.298</td>
<td></td>
</tr>
<tr>
<td>Nei</td>
<td>0.298</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>DIALECTS</td>
<td>0.298</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>72</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GEOGRAPH.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DISTANCE</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Merging dialect and surname matrices: 72 sampling points
Are surnames words?
Correlation between surname and dialect dissimilarity matrices for the NL

<table>
<thead>
<tr>
<th></th>
<th>SURNAMES Nei</th>
<th>DIALECTS (72)</th>
<th>GEOGRAPH. DISTANCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>SURNAMES Nei</td>
<td>1</td>
<td>0.298</td>
<td>0.565</td>
</tr>
<tr>
<td>DIALECTS 72</td>
<td>0.298</td>
<td>1</td>
<td>0.632</td>
</tr>
<tr>
<td>GEOGRAPH. DISTANCE</td>
<td>0.565</td>
<td>0.632</td>
<td>1</td>
</tr>
</tbody>
</table>

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 [127]. The data set is the same as in Fig. 1.5 and Fig. 2.4. 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 [127]. When computing the distance between a pair of data items, only the (squared) differences between components having values that are available are computed. The rest of the differences are thus set to the average of the computed differences.

Fig. 1.6. Sommon’s mapping of the same data set that was projected using the PCA in Fig. 1.5 [127].

Fig. 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 Problem 7, 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 missed when forming the projections. A key to the abbreviated country names is given in the Appendix.

MDS

Sommon’s mapping

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 analysis**

```
<table>
<thead>
<tr>
<th></th>
<th>Surname 1</th>
<th>Surname 2</th>
<th>Surname 3</th>
<th>Surname 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Town1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Town2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Town3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Town4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Town5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
```

15 x 15
Surnames: *How it works*
Surnames: How it works
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) (Pentium 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, Brinksm, 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…

<table>
<thead>
<tr>
<th>Surname</th>
<th>City 1 (200,000 hbt.)</th>
<th>City 2 (5,000 hbt.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Johnsson</td>
<td>1000</td>
<td>56</td>
</tr>
<tr>
<td></td>
<td>1000 / 200000 = 0.5</td>
<td>30 / 5000 = 0.65</td>
</tr>
<tr>
<td></td>
<td>0.50 / 1.15 = 0.43</td>
<td>0.65 / 1.15 = 0.56</td>
</tr>
<tr>
<td></td>
<td>43 %</td>
<td>56 %</td>
</tr>
</tbody>
</table>
Polyphyletism: *its signature*…

(24% of individuals)

68 surnames:

→ 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**  
![Distribution Map]

**Frenquence**  
![Frenquence Map]

**Origin**  
![Origin Map]

**36 surnames:**
- Abma, Algra, Anema, Attema, Baarda, Betten, Bonnema, Bontekoe, Bottema, Brinkema, Chossen, Cuperus, Damstra, Deelstra, Duiker, Haitsma, Hengst, Hoekema, Hylkema, Iedema, Jelsma, Kammen, Kooiker, Kuiken, Minnema, Mollem, Monsma, Numan, Piersma, Popma, Rienstra, Schaper, Sinnema, Steensma, Vlietstra

**65 surnames:**
A well identified population:

- 675 families in Friesland
- 273 families in Drenthe

- Extinction of surnames
- Size of families: clans?
- We only considered those surnames \( f > 40 \) individuals
## Correlations with census data

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

<table>
<thead>
<tr>
<th>Census</th>
<th>1830</th>
<th>1840</th>
<th>1849</th>
<th>1859</th>
<th>1869</th>
<th>1879</th>
<th>1889</th>
<th>1899</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correlation</td>
<td>0,57</td>
<td>0,56</td>
<td>0,55</td>
<td>0,51</td>
<td>0,49</td>
<td>0,46</td>
<td>0,42</td>
<td>0,4</td>
</tr>
<tr>
<td>p&lt;0.05</td>
<td>p&lt;0.05</td>
<td>p&lt;0.05</td>
<td>p&lt;0.05</td>
<td>p&lt;0.05</td>
<td>p&lt;0.05</td>
<td>p&lt;0.05</td>
<td>p&lt;0.05</td>
<td>p&lt;0.05</td>
</tr>
</tbody>
</table>
How many didn’t move?
(with reference to the time of surnames’ origin)


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

1880/1930

Grandparents

Surnames criterion

Present
Fig. 5

Area “α”
Area “β”
Area “χ”
Area “δ”

Generations

A B C D E F

Present

- 1 - 2 - 3 - 4 - 5 - 6 - 7 - 8
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
Prof. dr. P. de Knijff (Peter)

Prof. dr. Manfred Kayser

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.
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
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**

<table>
<thead>
<tr>
<th>Province</th>
<th>Locations</th>
<th>Individuals</th>
</tr>
</thead>
<tbody>
<tr>
<td>N. Holland</td>
<td>60</td>
<td>2007</td>
</tr>
<tr>
<td>Z Holland</td>
<td>104</td>
<td>10,582</td>
</tr>
<tr>
<td>Utrecht</td>
<td>22</td>
<td>301</td>
</tr>
<tr>
<td>Zeeland</td>
<td>24</td>
<td>468</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Province</th>
<th>Locations</th>
<th>Individuals</th>
</tr>
</thead>
<tbody>
<tr>
<td>N. Holland</td>
<td>60</td>
<td>2007</td>
</tr>
<tr>
<td>Z Holland</td>
<td>104</td>
<td>10,582</td>
</tr>
<tr>
<td>Utrecht</td>
<td>22</td>
<td>301</td>
</tr>
<tr>
<td>Zeeland</td>
<td>24</td>
<td>468</td>
</tr>
</tbody>
</table>

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…
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...
The matrix is partially used

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>54</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>80</td>
<td>38</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>33</td>
<td>61</td>
<td>78</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>40</td>
<td>28</td>
<td>74</td>
<td>33</td>
<td>0</td>
</tr>
</tbody>
</table>

Geographic analysis
1) Search for the higher distance value
1) Search for the higher
2) Trace a segment perpendicular to the edge
1) Search for the highest distance value
2) Trace a segment perpendicular to the edge of triangle
3) To the left or to the right?

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

Monmonier algorithm
Barrier vs. 2.2 for MS Windows

a software to compute geographic barriers from a distance matrix

By F. Manni and E. Guérard

Google: “barrier” “2.2”