Inference of migration pattern from space-time surname distribution

Pierre Darlu, CNRS
INSERM U535
Epidémiologie génétique et structure des populations humaines

European Science Foundation
Workshop on Migration
Île de Porquerolles, France,
5-7 september 2007

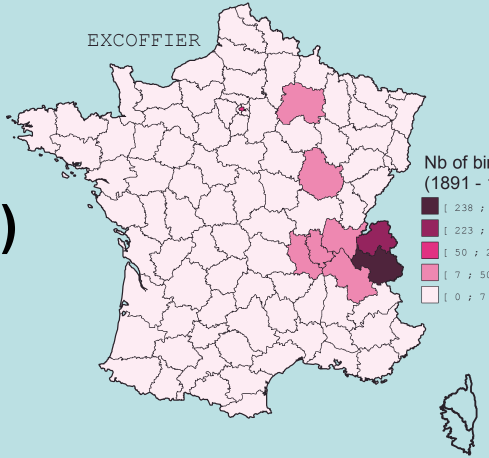
« Le patronyme sous-tend à la fois un champ parental et un champ territorial »
Françoise Zonabend

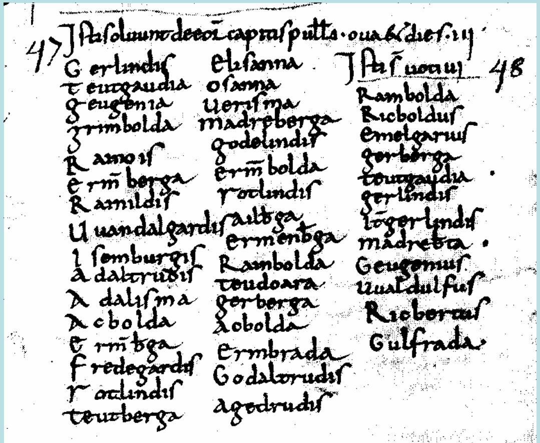
Why surnames ?

Surname as a tool to measure migration flux and origin of migrant in the past

They are :

- **transmitted through male line**
- **highly polymorphic (more than 500000 in France)**
- **peculiar to a given region (example : EXCOFFIER)**
- **diachronic data**




47 Istaſ ſoluunt decem capitaſ pulls · oua & dieſ · iij
Gerlindiſ
Teutgaudia
Geugenia
Grimbolda
Ramoiſ
Ermberga
Ramildiſ
Uuandalgardiſ
Isomburgiſ
Adaltrudiſ
Adaliſma
Acbolda
Ermbga
Fredegardiſ
Rotlindiſ
Teutberga

Eliſanna
Oſanna
Ueriſma
Madreberga
Godelindiſ
Ermbolda
Rotlindiſ
Autbga
Ermenbga
Rambolda
Teudoara
Gerberga
Aobolda
Ermbrada
Godaltrudiſ
Agedrudiſ

Istaſ nouiij 48
Rambolda
Ricbolduſ
Emelgariuſ
Gerberga
Teutgaudia
Gerlindiſ
Rogerlindiſ
Madreberta
Geugeniuſ
Uualdulfuſ
Ricberiuſ
Gulfrada

Polyptique
Saint-Germain-des-Prés
823-8

However…

Some limitations :

▶Only male migration, not female…

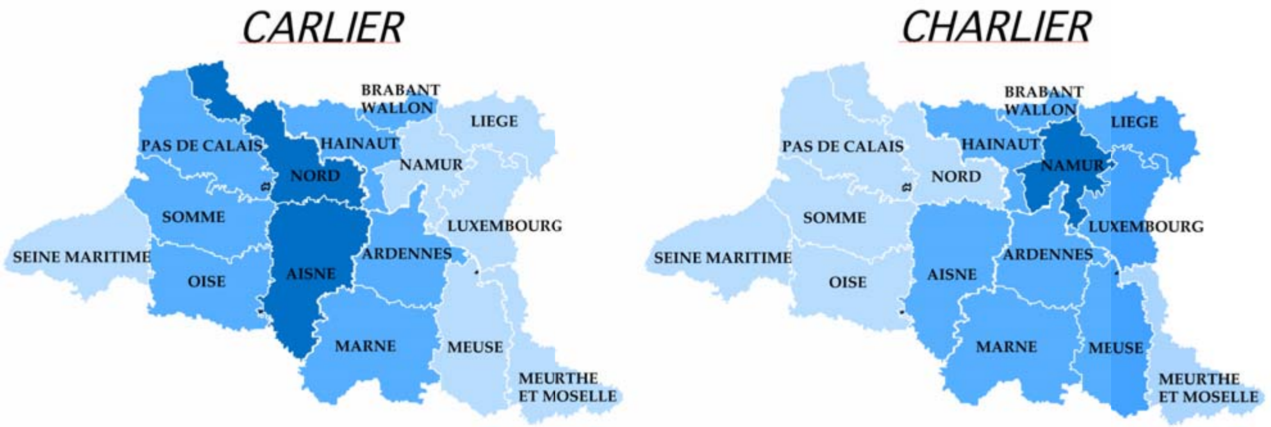
However appropriate records (as marriage registers) afford inferences of female migrtions

▶ Both high polymorphism and diversity in space require an appropriate sampling

Because of the high polymorphism, two distantly areas couldn’t share any surname, then providing no information. So the scale of the investigated areas have to be chosen accordingly

▶ The tricky question of the lemmatization has to be correctly solved

Attempt to distinguish when two orthographical variations are monophyletic (deriving from the same ancestral name) or polyphyletic and showing two different geographical pattern (as “CARLIER” and CHARLIER” in the north of France)



Several methods

- 1) **By deciphering surname distributions**

  immigration at one place – one period

- 2) **Surname distances between places**

  trace of migrations between places, one period

- 3) **Bayesian estimation of geographical origin of migrants**

  migrations between several places – several periods

METHODS

I) Deciphering surname distributions

- **The difference between observed and theoretical surname distribution "at equilibrium", in a given area, provides an estimation of the intensity of *immigration***

Yasuda et al., 1974, Theor. Pop Biol, 5:124-142
Zei et al., 1983, Ann Hum Genet, 47:329-352

Surname distribution and the Fisher's model (1943)

☺ N size of the constant population

☺ S number of surnames

☺ Number of children per individual : Poisson distribution, m=2

☺ surname transmitted without selection

☺ Each individual is replaced by an other

- ☞ with identical surname : p = 1 - v
- ☞ with an other surname : q = v

☺ k = number of individuals sharing the same surname

If $N >> k$

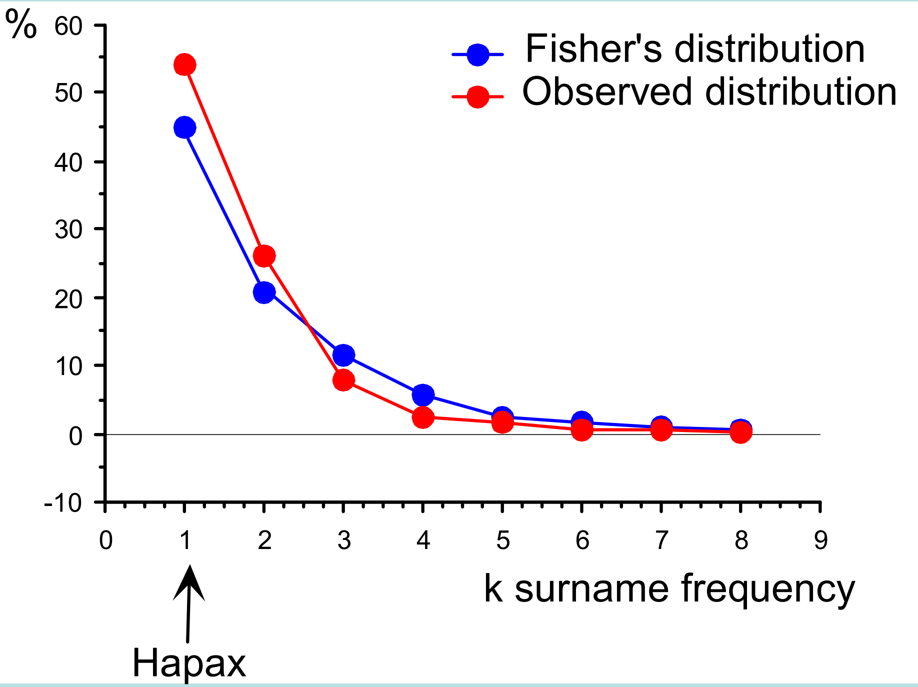
$$E(n_k) = \frac{N\nu}{k(1-\nu)}(1-\nu)^k$$

$$S = \sum_k E(n_k) = \frac{-N\nu}{(1-\nu)} \ln \nu$$

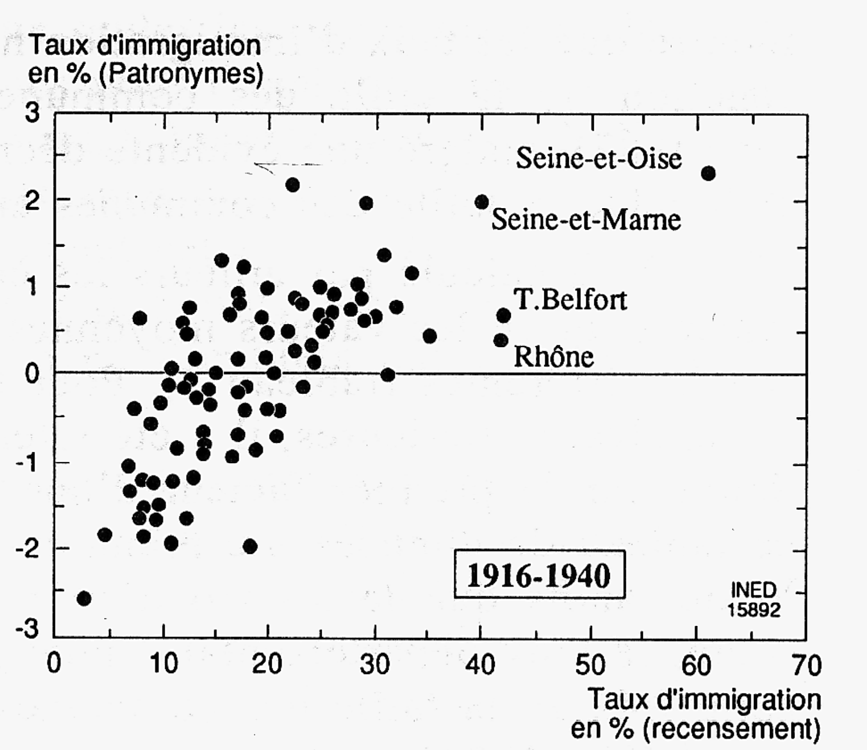
$$\frac{S}{N} = \frac{-\nu}{(1-\nu)} \ln \nu \Rightarrow \text{estimation of } \nu$$

Test of adjustment: $\chi^2 = \sum_k (n_k - E(n_k))^2$

Immigration and surname distribution



The excess of hapax (k=1) provides a rough estimation of the immigration rate



Comparison between immigration rates
estimated from surname distribution and from census
In: Darlu P., Ruffié J., Population, 1992

The relationship between the immigration rate estimated from surname distribution and from census is rather strong. This result provides a kind of validation of the Fisher’s method. It was obtained from surname data recorded in births register and averaged by communes belonging to the same French department. Identical result was also found from Italian regions (Piazza et al., Nature, 1987)

METHODS

II) Surname distance between places

- The “surname distance” between two places gives an estimate of the strength of the long-term spatial exchange of migrants

$$\varphi_{ij} = \frac{\sum_k p_{ik} p_{jk}}{\left(\sum_k p_{ik}^2 \sum_k p_{jk}^2 \right)^{\frac{1}{2}}},$$

$$\Rightarrow d_{ij} = -a \ln(\varphi_{ij} - b)$$

This distance was proposed by Chen and Cavalli-Sforza, *Human Biology*, 1983
p_{ik} : frequency of the kth surname in the ith area. a and b are constant adjusted to the data

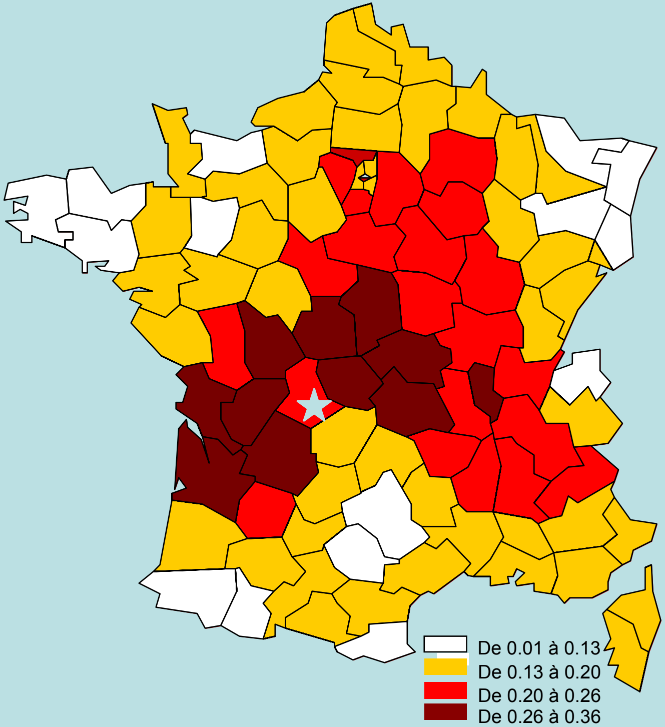
“surname distance” and long-term exchange

nom	Area		
	A	B	C
Durand	0.6	0.5	0.2
Marie	0.3	0.4	0.1
Da Silva	0.1	0.1	0.7

d(AB)=weak distance ⟷

d(AC)= strong distance ⟷

Example : A and B are areas showing close profile of surname frequencies and, consequently, short “surname distance”. This short distance could result from long-term exchanges of people between the A and B areas. However one cannot specify the directionality of migrations, from A to B or from B to A. On the other hand, surname distance between B or A and C is large so that exchanges must be rare between these areas



Geographical distribution of surname distances between 8 counties from Limousin and 90 departments 1890-1915

Clearly, the long-term exchanges of populations between the “star” location (Limousin) and the other parts of France were preferentially done south-westward and north-eastward

METHODS

III) Probability of origin of migrants

- **To estimate the origin of migrant newly arriving in a given area**

This method is not intended to estimate the quantitative flux of migrant, but solely their geographical origin in term of probability.

$$p(g_k|s_i) = \frac{\pi(g_k)p(s_i|g_k)}{\sum_k \pi(g_k)p(s_i|g_k)}$$

$\pi(g_k) \quad = a\ priori$ probability of migration from g_k

$p(s_i|g_k) =$ probability of the surname s_i in the area k

$$pgo_k = \frac{1}{\sum_i \omega_i} \sum_i \omega_i p(g_k|s_i)$$

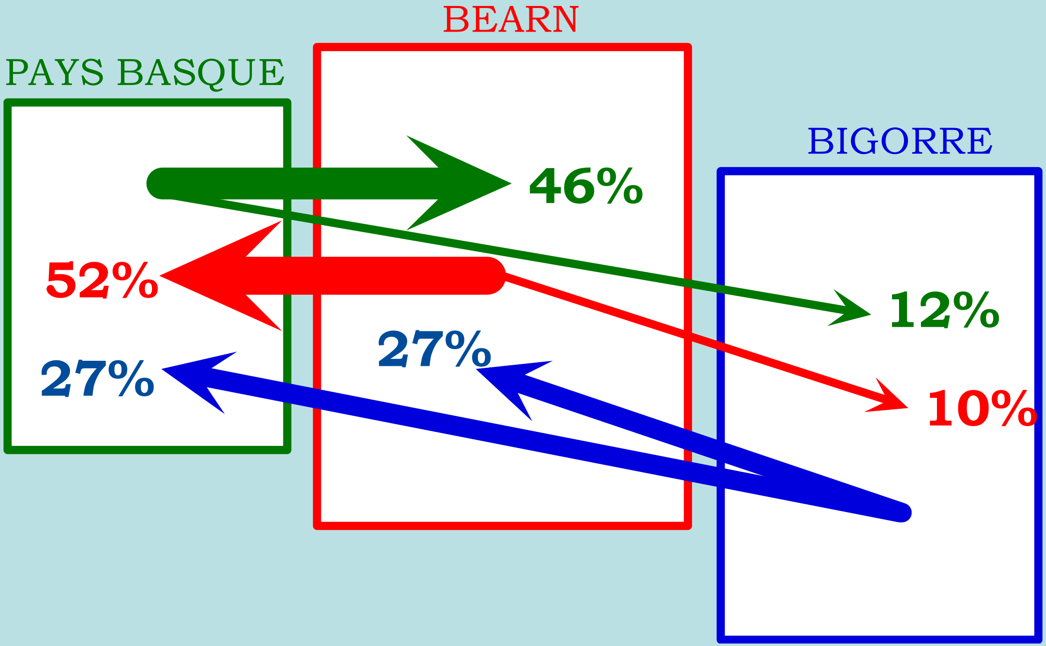
pgo$_k$: probability of *k*th geographical origin

Degioanni A. Darlu P., *Ann Hum Biol*, 2001

Origin of surnames

Name	Time	Area						
		A	B	C	D	E	F	out
Fabre	T1	0.6	0.5	0.2	0	0	0.1	?
	T2	0.3	0.4	0.1	0.1	0.2	0.3	?
Martinez	T1	0	0	0	0	0	0	?
	T2	0.3	0	0	0	0	0	?

Example: the frequency of « Fabre » and « Martinez » were recorded twice, at two successive periods T1 and T2, in 6 different areas [A..F].
The “Fabre” surnames observed at time T2 in the D and E areas, where they were not observed before (T1), have a high probability of coming from the A or B areas (the place where the frequencies of “Fabre” were high at T1), and a low probability from the F area
The “Martinez” surnames arriving in A at time T2 come from “outside”, since this surname was absent in all the areas at time T1.



Proportion of surname attested in only one area between 1891-1915
and found in an other area between 1916-1940

Darlu P., Degioanni A., Jakobi L., In: *Le Patronyme, Histoire, anthropologie, société*, CNRS Editions, 2001

This example is a simplified application of the previous described method

EXAMPLES

Saint-Germain-des-Prés - IXth°

Savoie – XVIII-XXth°

Cévennes – XIX-XXth°

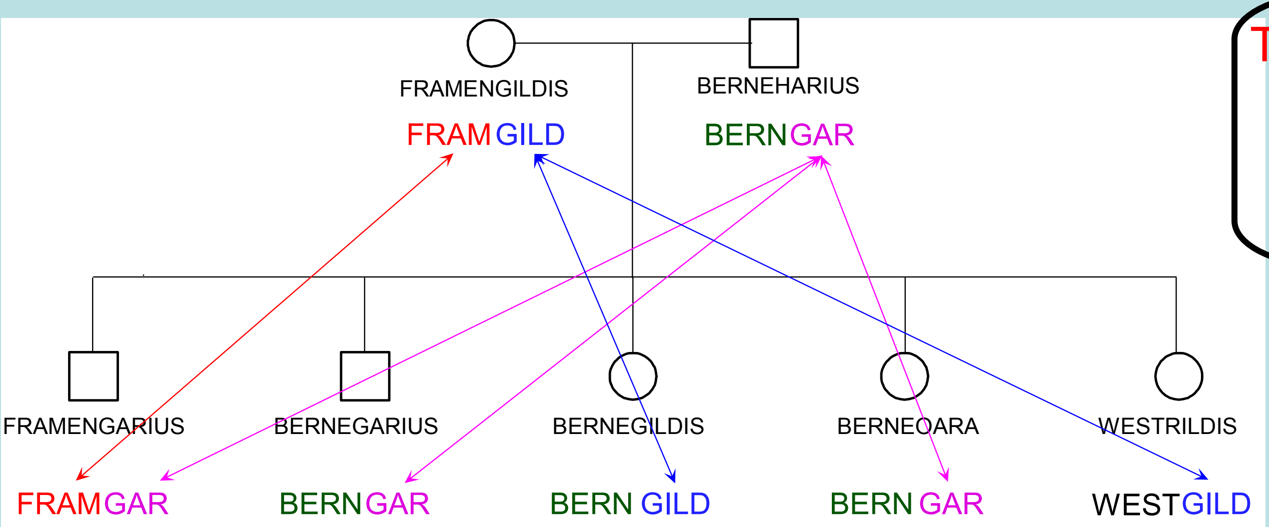
France – XIX-XXth

SAINT-GERMAIN-DES-PRES

IXth

Collaborations : Monique Bourin, Pascal Chareille, Jean Pierre Devroey

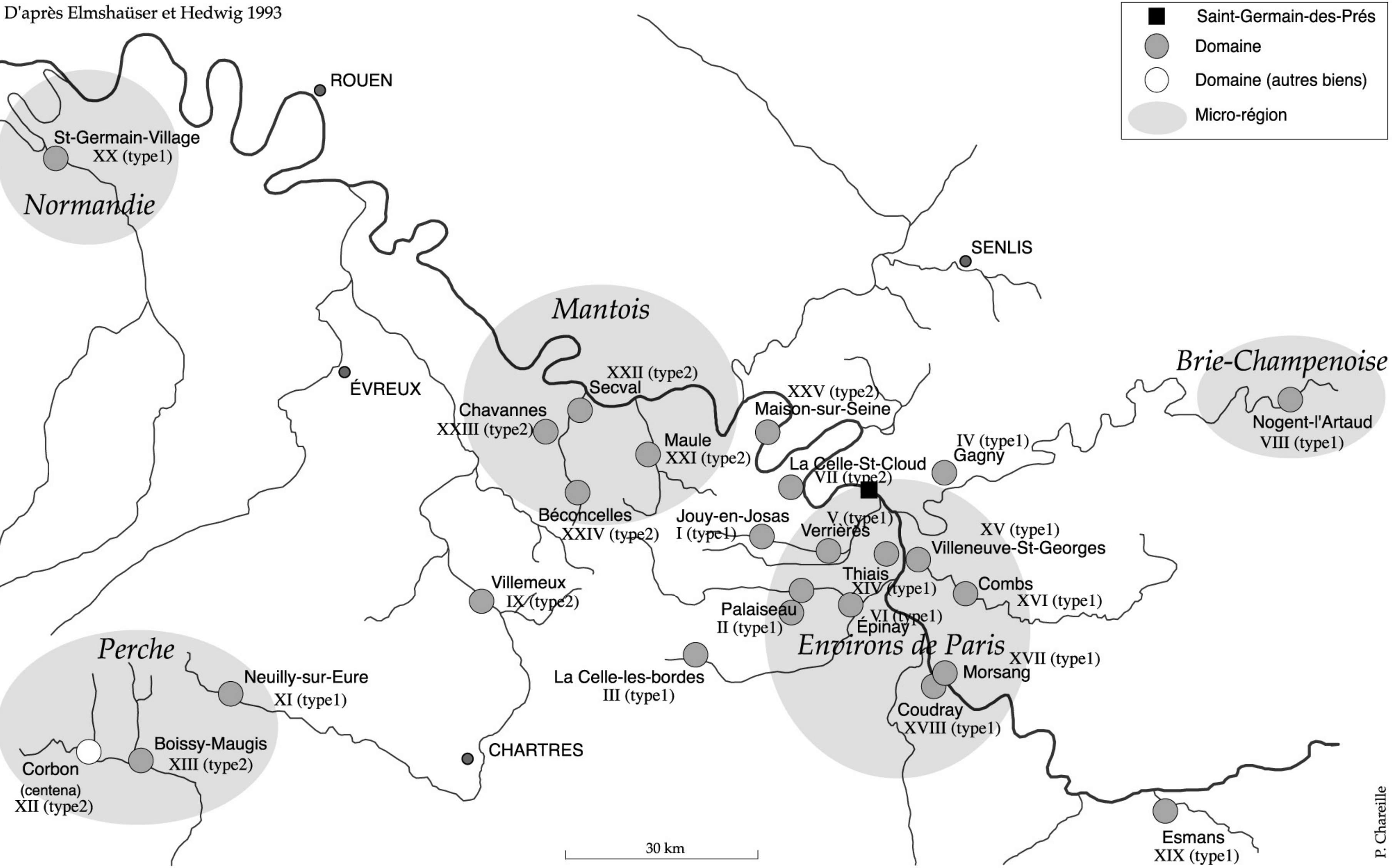
Transmission of lexemes looking like genetic transmission



Germanic system used at Saint-Germain-des-Près (IXème)
Each child's name is composed of two lexems herited from his/her parents .
Sex is attested by the ius or is/a ending.
cf. M. Bourin, *Pour la Science*, 1996

	Garçons			Filles		
	n	% pur	%altéré	n	% pur	%altéré
1er nommé	834	54.7	4.3	213	52.6	4.7
……						
5ème nommé	31	45,2	0	124	41,9	4,8

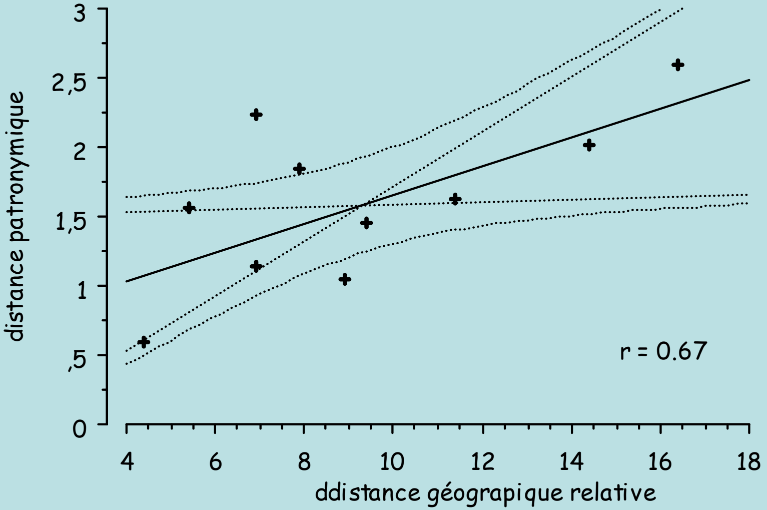
Proportion of transmitted names which follow the “Germanic” rule, according to sex and rank
For other investigated factors : heir/non heir, social status, see :
Bourin et Chareille, 2002


D'après Elmshaüser et Hedwig 1993

P. Chareille

Domaines de Saint-Germain-des-Prés, IXème siècle

Relation between surname distance and geographic distance



Relationships between surname distance and geographical distance between domains. (Saint-Germain-des-Près (Normandie, Perche, Mantois, Paris, Brie-Champenoise ; IXe siècle, see map)

(From Chareille, Thèse, 2003; Darlu P., 2004, *Ann Demo Hist*)

From

to

	I	II	III	IV	IX	V	VI	VII	VIII	XI	XII	XIII	XIV	XIX	XV	XVI	XVII	XVIII	XX	XXI	XXII	XXIII	XXIV	XXV	out
I	18	65	35	12	172	50	32	36	16	6	16	73	55	34	54	42	18	5	12	42	49	8	90	29	30
II	11	62	28	16	174	46	28	32	18	20	27	83	45	40	63	47	17	6	25	35	48	8	69	22	31
III	12	54	55	13	181	47	20	38	14	11	25	103	47	39	49	39	18	11	16	31	43	5	81	20	28
IV	15	56	23	14	211	48	24	34	16	7	18	84	52	42	47	37	17	5	16	33	54	10	98	19	20
IX	10	52	31	7	228	37	19	26	15	7	27	101	37	46	37	37	13	4	15	25	49	12	80	15	69
VI	15	55	25	16	196	57	27	36	17	12	18	84	56	39	44	50	20	5	21	30	45	10	79	23	18
VIII	11	53	24	14	178	40	22	34	24	12	26	95	44	49	49	42	22	5	21	33	48	7	77	25	44
XI	13	82	35	15	167	44	22	26	10	9	25	100	50	32	43	32	15	5	16	30	42	7	103	20	56
XII	9	60	36	12	188	51	21	32	14	9	23	115	37	33	40	44	16	5	17	34	39	17	87	26	34
XIII	8	55	35	13	190	40	19	31	13	9	23	124	36	45	44	44	18	4	24	28	42	9	78	26	44
XIV	9	54	20	13	172	55	34	34	17	5	24	92	75	35	49	41	16	5	16	31	49	9	97	22	26
XIX	12	57	24	12	188	43	27	32	22	5	23	86	41	120	40	36	21	4	19	29	54	8	77	22	0
XV	12	65	29	12	195	43	20	30	21	9	21	78	52	65	58	42	15	6	17	37	52	9	76	24	12
XVI	15	61	32	15	183	46	22	41	18	8	17	78	49	35	52	57	25	4	25	32	41	6	71	20	48
XVII	11	59	32	11	212	41	22	29	17	7	18	75	52	44	45	33	35	3	21	28	59	7	98	20	20
XVIII	11	73	24	22	176	54	12	43	18	18	17	54	45	39	52	42	20	6	23	40	41	11	82	22	56
XX	11	65	24	12	165	41	34	31	20	11	22	75	48	42	37	40	17	5	58	27	43	9	100	20	43
XXI	12	50	32	13	185	36	27	33	14	6	33	94	45	46	62	42	16	4	26	42	61	8	78	24	10
XXII	17	67	27	14	207	35	24	30	14	7	21	92	40	47	46	37	12	4	32	37	52	15	76	20	25
XXIII	11	53	25	15	188	53	32	25	21	6	19	57	53	46	44	67	13	6	43	36	67	15	78	27	0
XXIV	8	55	22	14	180	44	26	24	22	9	21	85	40	46	39	40	18	6	26	24	55	6	114	17	60
XXV	18	43	26	13	163	51	16	53	12	0	23	81	88	54	48	59	18	7	15	49	46	7	63	46	0

Probability (‰) of origin of lexemes (E1 or E2), second generation

How to read this table :

Example: 22.8% of lexemes given to the children were also found in the previous generation inside the IXth domain (the most "isolate" domain in a sense), while 6.9% were not present in any domains (thus coming from "outside").

One can see that some domains (XIX, XXIII, XXV) named their children only with lexemes already attested in the domains [I to XXV].

Knowing the mode of transmission of lexemes, this table could help to figure out how the names were originated and could provide a way to estimate possible migrations between domains and from outside.

Some (cautious) conclusions

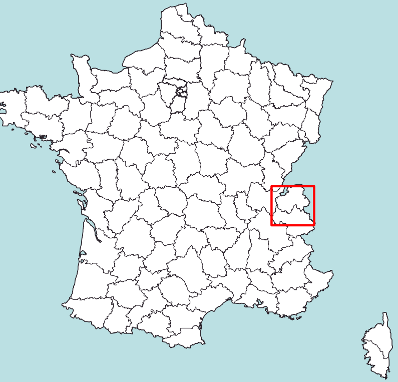
- Weak surname diversity among domains compared to diversity between manses within domains
- Differentiation with geographic distance
- Some domains had their own lexemes (hapax) (IX,XXIV, XIII…)
- Few lexemes were becoming from « outside »:
  « closed area/ weak immigration ?» ?
- lexemes given to a child were frequently chosen from
  an other domain: exchanges among domains ?

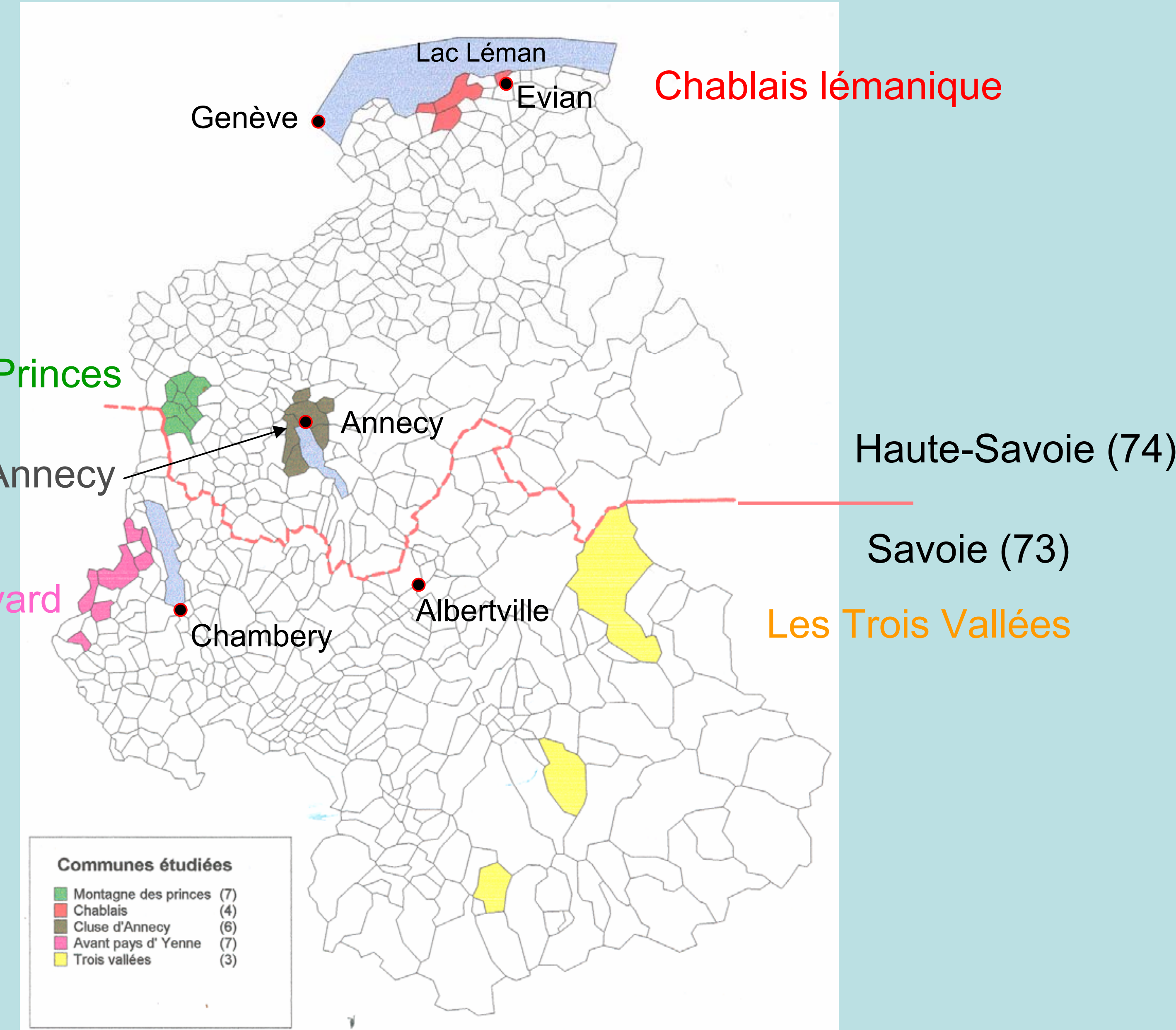
SAVOIE

From XVIIIth to XXth

From Darlu P., Brunet G., Barbero D
First presented at the International Conference on Space and Time in Historical Demographic Studies, new methods and models IUSPP, 2006.
(supposed to be published soon)

« SPACE »

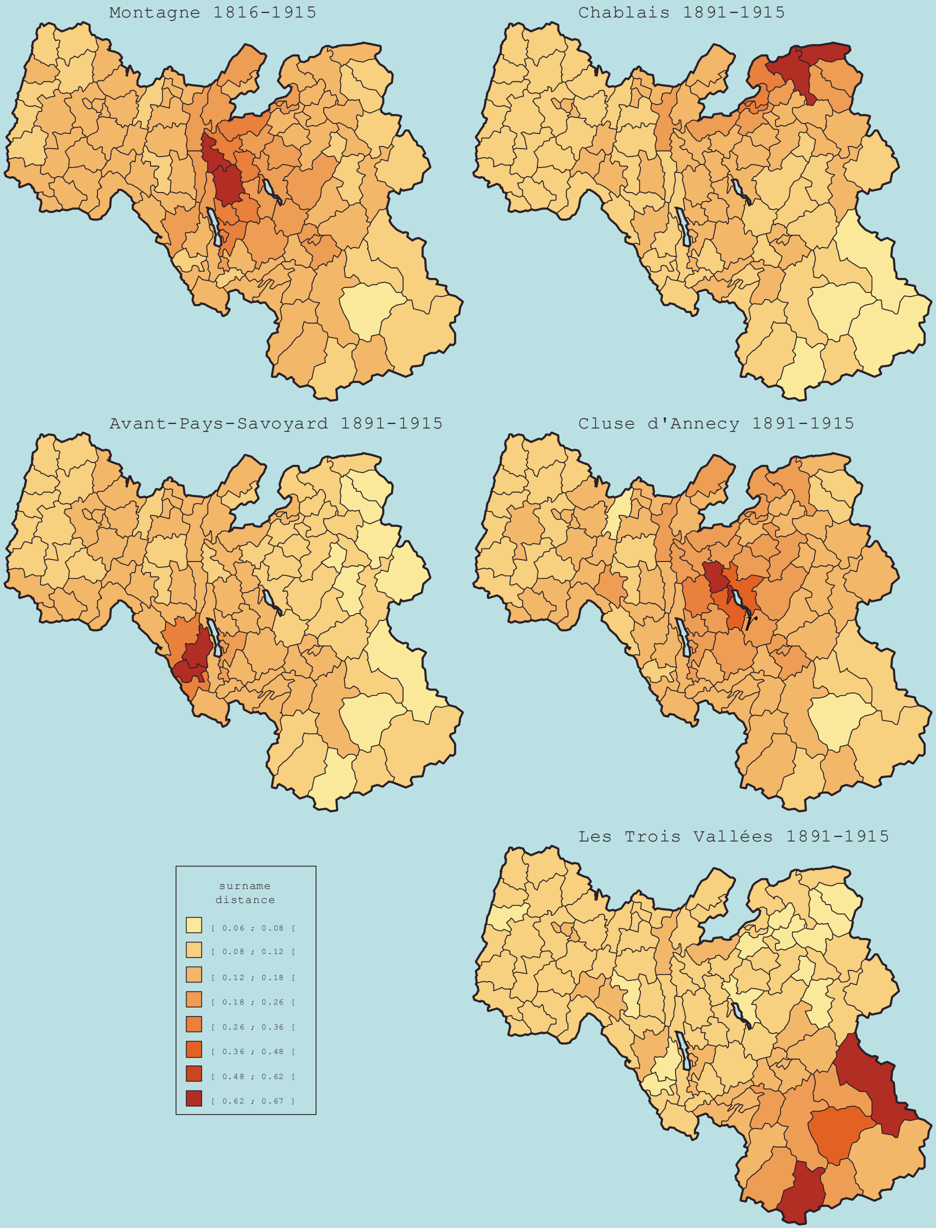



Lac Léman
Evian
Genève
Chablais lémanique
Montagne des Princes
Annecy
Cluse d'Annecy
Haute-Savoie (74)
Savoie (73)
Avant Pays savoyard
Albertville
Chambery
Les Trois Vallées
Communes étudiées
🟩 Montagne des princes (7)
🟥 Chablais (4)
🟫 Cluse d'Annecy (6)
🟪 Avant pays d' Yenne (7)
🟨 Trois vallées (3)

« TIME »

Five periods	Sources
▶ 1710-1729	Parish birth records
▶ 1810-1829	Parish birth records
▶ 1891-1915	INSEE birth registers
▶ 1916-1940	INSEE birth registers

9414 surnames



Surname distances between the five regions and the other counties

Each map indicates how the resemblances between the surname profile of a given region and those of its surrounding counties decrease in space

	To ->	Montagne	Chablais	Cluse d'Annecy	Avant Pays Savoyard	3 Vallées
a) Proportion of migrants between communes of the same region						
Among communes Within regions	*T1->T2*	0.152	0.279	0.409	0.241	0.024
	T2->T3	0.196	0.256	0.320	0.283	0.070
	T3->T4	0.098	0.293	0.402	0.192	0.007
b) Proportion of extra-regional migrants						
From outside	*T1->T2*	0.409	0.562	0.456	0.503	0.686
	T2->T3	0.412	0.544	0.507	0.571	0.782
	T3->T4	0.456	0.538	0.522	0.468	0.800

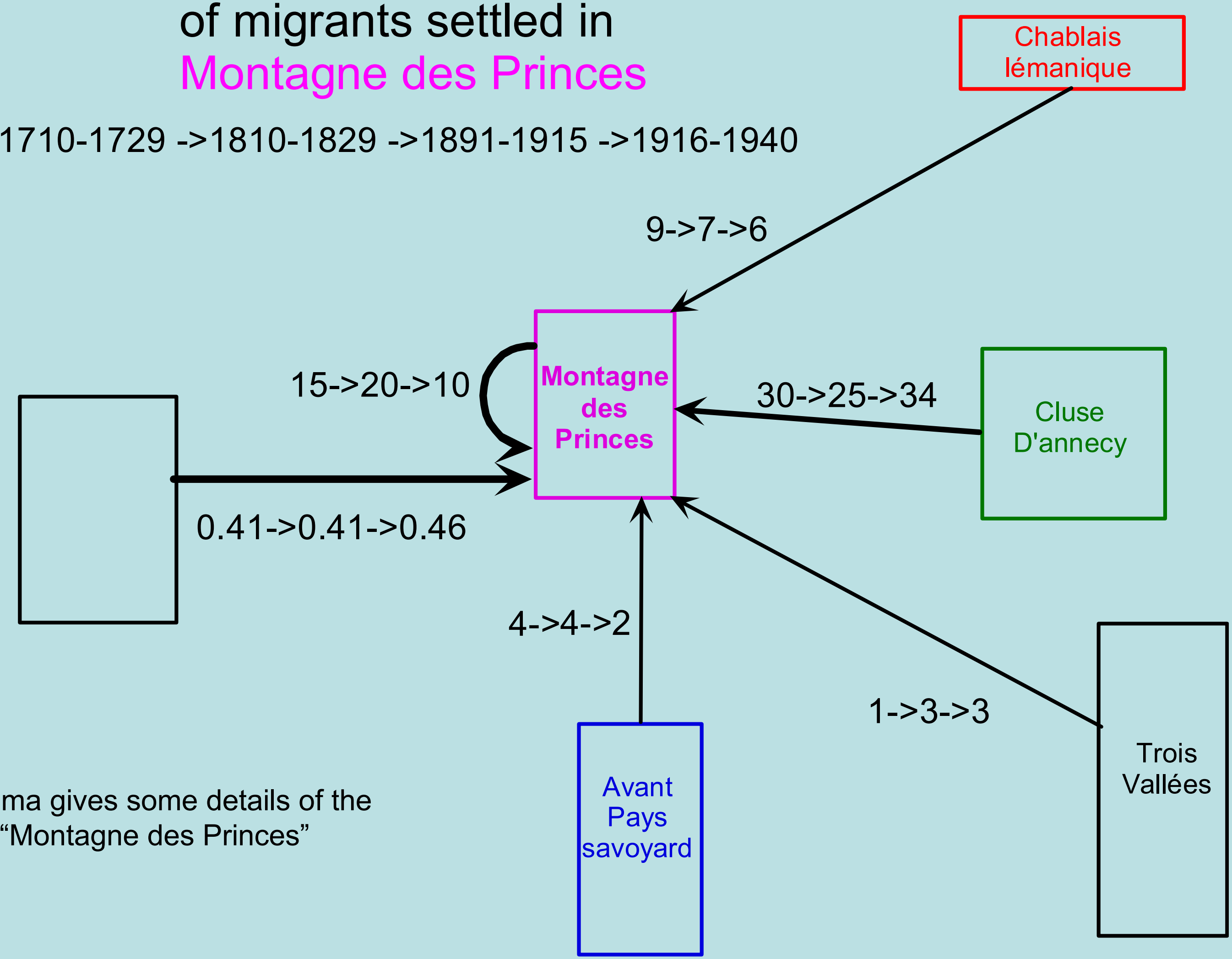
How to read it:

a) The probability that migrants settled in a **commune** located in the Cluse d’Annecy region, during the periods between T1 (1710-1729) and T2 (1810-1829), were **coming from other communes belonging to the same region** is 0.409 (quite high, meaning that people were migrating very locally). This probability is lower in the case of the “Montagne” region (meaning that migrants came from more distant regions); and almost null for the “3 vallées” region which includes only three communes largely separated by mountain ranges.

b) Most of the migrants who settled in the “3 vallées” region came from a place non included in this study (p between 0.686 and 0.800 depending on the periods). On the other hand, Montagne welcomed the smallest proportion of migrants coming from outside (p betwen 0.409 and 0.456)

probability of geographic origin (%)
of migrants settled in
Montagne des Princes

1710-1729 ->1810-1829 ->1891-1915 ->1916-1940



Chablais lémanique

9->7->6

Montagne des Princes

15->20->10

30->25->34

Cluse D'annecy

0.41->0.41->0.46

4->4->2

1->3->3

Avant Pays savoyard

Trois Vallées

This schema gives some details of the
pog from “Montagne des Princes”

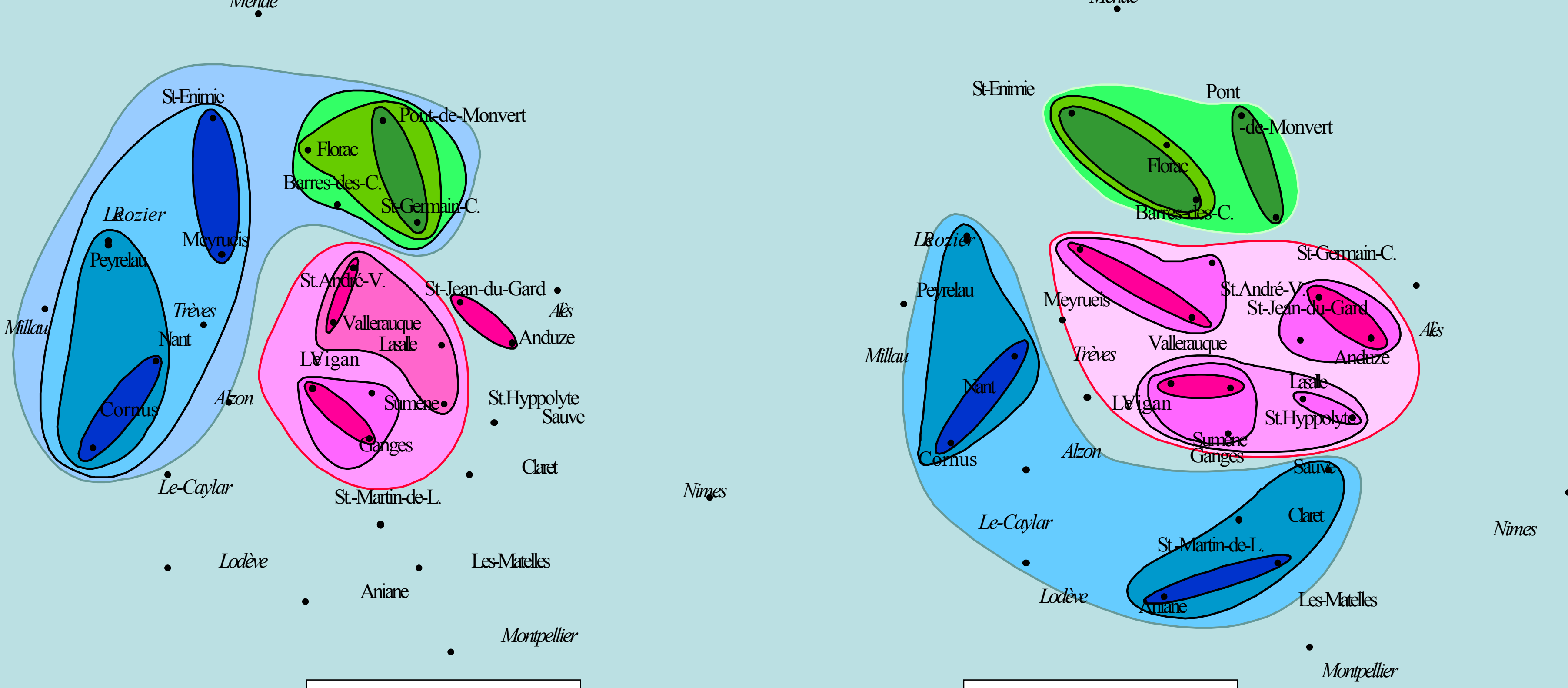
(short) conclusions

- High stability of the populations since the XVIIIth
- Most immigrants were originated :
- from neighbor communes (e.g. among Chablais, Cluse)
- from a particular region (e.g. from Cluse to Montagne)
- from outside (e.g. Trois vallées)

CÉVENNES

XIX-XXth

Collaborations: Anna Degioanni, Josef Smets



1843-1862

1890-1915

Clustering diagram of the surname distances between counties, Cévennes (Neighbor-Joigning and %bootstrap)

Darlu et al., In: *Spatial analysis of biodemographic data*, INED,1996

How to read it:
From the surname distance matrix between pairs of counties, NJ clustering algorithm was performed. The confidence of each cluster was estimated by bootstrap and plotted with light/dark colored shapes depending on the bootstrap values.
Between the two periods, the relationships between places drastically changed, following ways of rural exodus toward labor market area, and made easier by development of roads through the Mont Aigoual

Conclusions

Path of the rural exodus due to:

- Loss of influence of Millau
- Improvement of major routes between north and east
- Attractiveness of Montpellier *vs* Nimes-Alès

Urban Centres, France

XIX-XXth

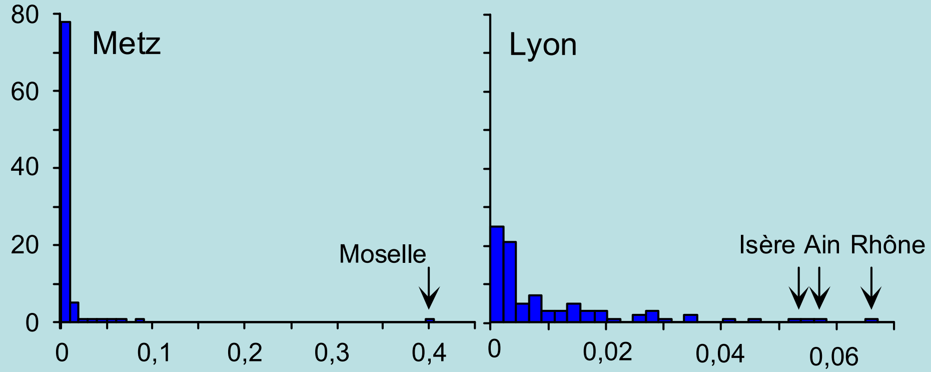
Darlu P, Degioanni A., *Espace Géographique,* 2007

The bayesian method is applied to infer the origin of migrants leaving rural or small districts to settle in large urban centres at the time of the World War I

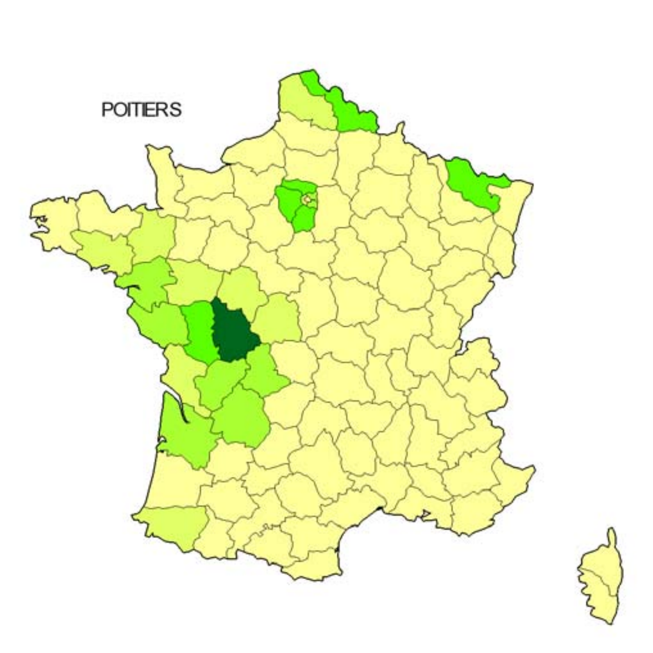
▶Two periods : P1: 1891-1915
P2: 1916-1940

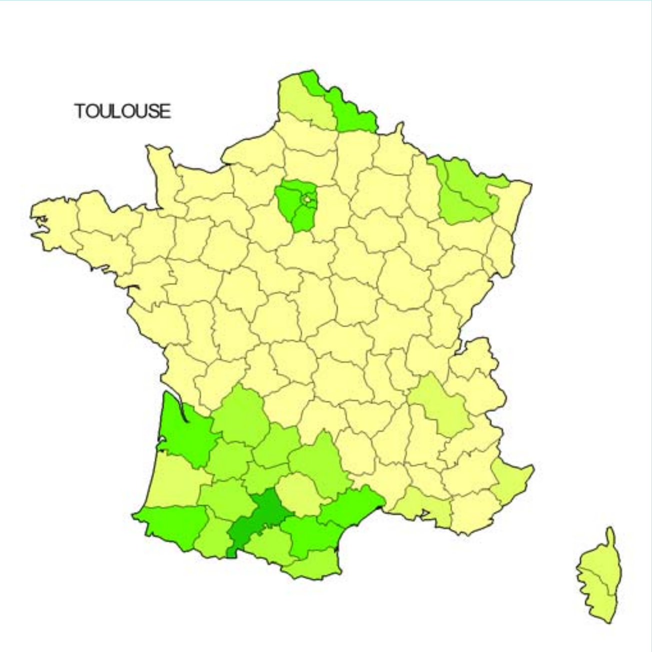
▶29 urban centres

▶More than 30000 rural districts

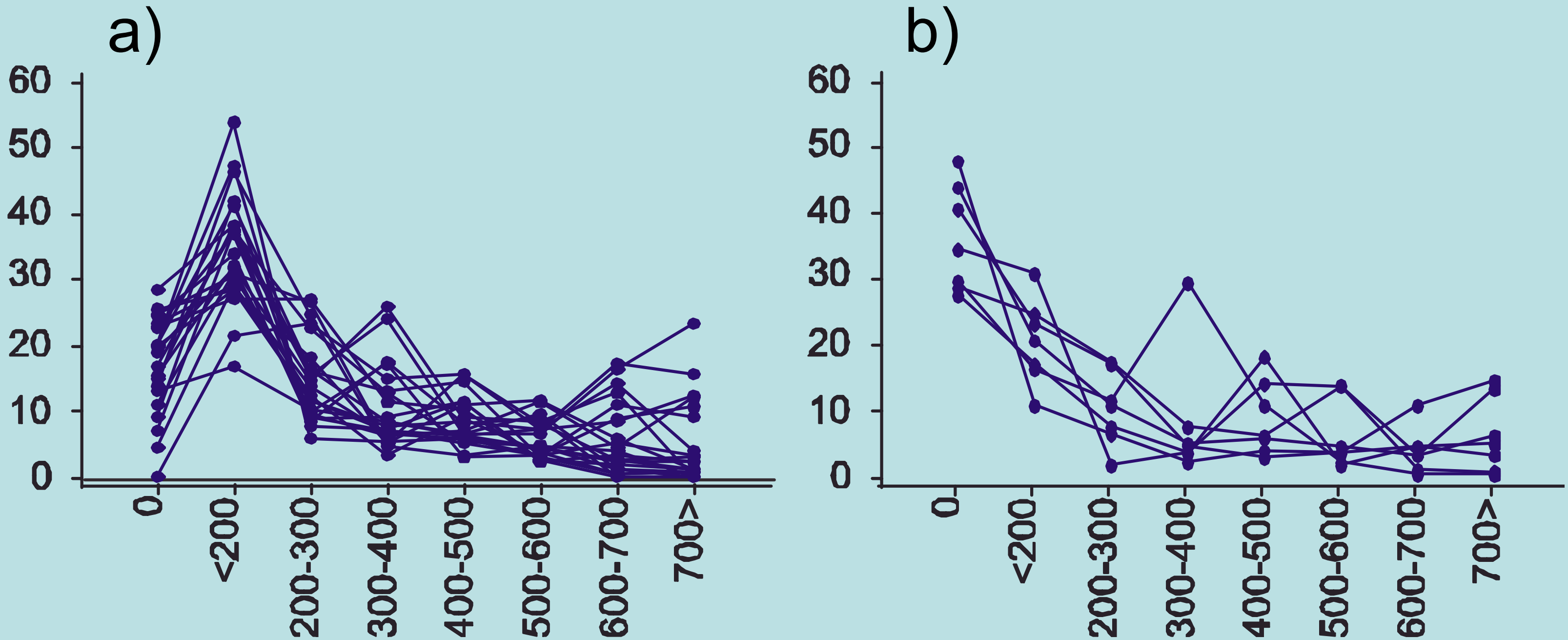


Departmental distribution of the probability of geographical origin of migrants (pog, abscissa) between P1 and P2, in Metz and Lyon. Lyon shows a large heterogeneity of origin compared to Metz





Two examples showing the departmental distribution of the probability of geographical origin (pog):
The migrants settling in Poitiers were coming particularly from rural communes located in the same department, and less from the neighbor departments. The origins of people migrating to Toulouse were more diversified, since they came from the largest South-West part of France



The probability of geographical origin (pog, %,) is directly linked to the geographic distance (abcissa) between the « rural source » and the "urban recipient" area. Each curve corresponds to a urban centre showing an intra-departmental pog smaller (**a**, on the left) or larger (**b**, on the right) than the pog of the most neighboring department,