4-1-2012

The family name as socio-cultural feature and genetic metaphor: from concepts to methods

Pierre Darlu
UMR7206, CNRS, Muséum National d'Histoire Naturelle, Université Paris 7 Paris

Gerrit Bloothooft
Utrecht University, Utrecht institute of Linguistics

Alessio Boattini
Dipartimento di Biologia E.S., Area di Antropologia, Università di Bologna

Leendert Brouwer
Meertens Institute KNAW, Amsterdam

Matthijs Brouwer
Meertens Institute KNAW, Amsterdam

See next page for additional authors

Authors
Pierre Darlu, Gerrit Bloothooft, Alessio Boattini, Leendert Brouwer, Matthijs Brouwer, Guy Brunet, Pascal Chareille, James Cheshire, Richard Coates, Paul Longley, Kathrin Dräger, Bertrand Desjardins, Patrick Hanks, Kees Mandemakers, Pablo Mateos, Davide Pettener, Antonella Useli, and Franz Manni

The family name as socio-cultural feature and genetic metaphor:

from concepts to methods

Pierre Darlu (1), Gerrit Bloothooft (2,3,4), Alessio Boattini (5), Leendert Brouwer (3), Matthijs Brouwer (3), Guy Brunet (6), Pascal Chareille (7), James Cheshire (12), Richard Coates (8), Paul Longley (12), Kathrin Dräger (9), Bertrand Desjardins (10), Patrick Hanks (8), Kees Mandemakers (4), Pablo Mateos (12) Davide Pettener (5), Antonella Useli (5, 11), and Franz Manni (1)

(1) UMR7206, CNRS, Muséum National d'Histoire Naturelle, Université Paris 7 Paris
(2) Utrecht University, Utrecht institute of Linguistics
(3) Meertens Institute KNAW, Amsterdam
(4) International Institute for Social History KNAW, Amsterdam
(5) Dipartimento di Biologia E.S., Area di Antropologia, Università di Bologna
(6) UMR CNRS 5190 – Université Lyon 2
(7) University of Tours, France, Centre d'Études Supérieures de la Renaissance (CESR)
(8) University of the West of England, Bristol
(9) Deutsches Seminar, Albert-Ludwigs-Universität, Freiburg im Breisgau
(10) Département de Démographie, Université de Montréal
(11) Dipartimento di Zoologia e Genetica Evoluzionistica, Università di Sassari
(12) Department of Geography / Center for Advanced Spatial Analysis, University College London (UCL)

Running title: Family names, from concepts to methods.

ABSTRACT

A recent workshop on “Family name between socio-cultural feature and genetic metaphor – From concepts to methods” was held in Paris on the 9th and 10th December 2010, partly sponsored by the Social Science and Humanity Institute (CNRS), and by Human Biology. This workshop was intended to facilitate exchanges on recent questions related to the names of persons and to confront different multidisciplinary approaches in a field of investigation where geneticists and historians, geographers, sociologists and ethnologists have all an active part. Here are the abstracts of some contributions.

In 1983, Human Biology published a special issue devoted to surnames as tools to evaluate average consanguinity, to assess population isolation and structure, and to estimate the intensity and directionality of migrations. At that time, many population geneticists made major contributions to this field, including Crow, Cavalli-Sforza, Morton, Relethford, Lasker, and Barraï (see review in Lasker, 1985, Colantonio et al., 2011).

Since then, most studies have focused on extending knowledge on population structure, isonymy, and migration. A synthesis was recently published in this journal (Colantonio et al., 2003) showing that surname methodologies have now been applied to about 30 societies all around the world. The geographic scope ranges widely, from the household or village to a whole continent. The authors also underlined the recent methods to analyze Y chromosome DNA polymorphisms which allow the examination of the degree of co-segregation of surnames and Y haplotypes, at least in the occidental naming practice.

The present workshop hoped to go beyond this, even if some presentations were closely allied to classical concerns, and to pinpoint some particularly relevant aspects in current research. There are two main strands. The first rests on the exploitation of databases that are increasing in size and exhaustiveness due to the spread of computerization. In this respect, Pablo Mateos and Paul Longley's UCL Worldnames database (http://worldnames.publicprofiler.org/), which includes about 6 million surnames registered in 26 different countries, constitutes an impressive quantity of information and a wonderful tool for future research (Mateos et al., 2011). However, the data are drawn from diverse sources depending on country, such as national electoral registers or telephone directories, raising problems of homogenization and representativeness that need discussion. Moreover, long distance comparisons between stocks of names with totally different historical and linguistic origins are also a challenge. The corpus of names described by Kathrin Dräger (*Deutscher Familiennamenatlas*) based on the telephone directory of the federal Republic of Germany in

2005 contains a set of one million different types of name for about thirty million telephone lines. These can be organized according to phonology (vowels, consonants, morphology) and to surname type (derived from place names, professions, nicknames, first names). These data allow the exploration of regional variations of names in consideration of lexis, phonology, graphemics, and morphology. Regarding the current distribution of surnames it is possible to trace ancient migratory movements in some cases. In the same vein, Gerrit Bloothooft presented the modern set of 16 million family names of the entire Dutch population collected from the Civil Registration. This includes 314,000 different surnames of which the spatial distribution can be studied online, while etymological and onomastic enrichment is available for 100,000 names. Patrick Hanks and Richard Coates's approach is quite different since they have collected names from various sources, such as ancient or recent dictionaries, primary sources of many kinds, and lists of surnames already published in England, Wales, and Scotland. This approach constitutes the *Family Names of the United Kingdom Project*. It aims to reconstruct the etymology of names and to explain their morphological variations through space and time.

Besides these attempts to draw from modern registers the largest number of surnames in wide geographic areas, the second major research strand involved a focus on historical data. The advantage of surnames over genetic data is that they can be available backward in time for consecutive generations, allowing a more accurate description of population dynamics. Thus Gerrit Bloothooft and Kees Mandemakers included information on collected life cycles of 76,000 persons born between 1811 and 1922; Guy Brunet used the almost exhaustive list of about 400,000 baptisms recorded in Québec from 1600 to 1800; and Pascal Chareille studied the surnames in the Normandy currency tax rolls between 1383 to 1515, and also exploited the household census in Burgundy between 1376 and 1610. Davide Pettener and Alessio

Boattini used the conscription list of individuals born between 1808 and 1987 in Italy's Upper Savio Valley.

The large expansion of the available data, both in time and space, has led to the development of new methods and analytical tools. Among them, and now widely used, are automatic geographic representations of surname diversity, which plot either the variations of frequency of a given name or a set of names sharing some phonetic or grammatical features (see Bloothooft's, Dräger's, and Lisa's figures). Some recent statistical methods, although not entirely new, were also presented, for example a Bayesian approach to infer the origins of migrants (Brunet et al.), Self-Organizing Maps to identify names sharing the same geographic origin (Boattini et al.), or naming network clustering into ethno-cultural groups (Mateos et al, 2011) .

Surnames are efficient markers for tracing the movements of people, and therefore most presentations focus on migration. Gerrit Bloothooft compares the distribution of birth places of current inhabitants of a given town and the corresponding distribution for their great-grandfathers. Guy Brunet discusses the origins of migrants who settled in parts of Québec between the beginning and the end of the 18th century. Pascal Chareille extracts from the household census (14th century, Burgundy) annotations indicating movements of people around Dijon. Patrick Hanks, Richard Coates, and Kathrin Dräger, thanks to their databases providing etymological information on names, can localize the most likely geographic origin of a given name.

One can foresee that the future of surname studies lies probably more in the rich information provided by the set of data preserved through the generations (one of the oldest, which include 8500 names, comes from the 9th century (Chareille, 2011) and in well-defined communities, than in the accumulation of surnames on a wider geographical scale. Moreover, the large amounts of time- and geo-referenced data that will be gathered in the future will

require new statistical methods that take into account the inescapable problems of lemmatization (the grouping together of related surnames) and sampling.

However, names are not just a way to identify individuals that is cheaper and more efficient than by analyzing Y chromosome polymorphisms. They also carry social and economical meanings that merit inclusion in any interdisciplinary approach. Historians, linguists, and geographers, as exemplified during this workshop, can play as active a role as biologists, in surname studies and population analysis. And for the future, the trend should be to expand our traditional western-centered field of investigation, in order to investigate other modes of naming in other countries that have both different cultural traditions and large amounts of available data.

1. The German Surname Atlas Project. Computer-Based Surname Geography [Kathrin Dräger]

German surnames preserve linguistic material which is up to 900 years old, from Middle High German, Middle Low German and Early New High German. This enables us to draw conclusions regarding medieval dialectal variations, writing traditions and cultural life, using the current surname distributions.

The high degree of territorial variation of the German surname system is now being made accessible by the *German Surname Atlas* project (*Deutscher Familiennamenatlas*; begun 2005), a cooperation between the Universities of Freiburg and Mainz under the direction of Prof. Dr. Konrad Kunze and Prof. Dr. Damaris Nübling.

The most frequent and impressive examples are selected from the ~1 million different surnames in Germany to address lexical (e.g., *Schröder*/*Schneider*, both surnames derived from the profession of tailor) as well as phonological (e.g., *Hauser*/*Häuser*/*Heuser*, *Walter*/*Walther*) and morphological (e.g., patronyms such as *Petersen*/*Peters*/*Peter*) questions. The database consists of all of the landline telephone connections in the Federal Republic of Germany in the year 2005 as provided by Deutsche Telekom AG. To estimate the number of people who bear a specific name, one multiplies the number of telephone connections by 2.9. In Germany, telephone connections are the only comprehensive database available. They are arranged by postal code districts comprising five digits each.

The atlas will contain two parts: one grammatical, and one lexical. The first part, comprising phonology, graphemics and morphology, will be published in three volumes: 1) vowels, 2) consonants, 3) morphology and syntax. The second part of the atlas will be divided into three volumes based on the five surname types: 4) provenance and residence names, 5) profession names and nicknames, 6) patronyms. Volume 1 was published in 2009, volume 2

in 2011, and volumes 3 and 4 will follow in 2012. The final two volumes are scheduled for 2015.

Each surname map in the atlas is accompanied by a commentary containing six sections: (i) the topic being illustrated, and why this special case has been chosen. Usually, very frequent names are selected which are preferably etymologically unambiguous; (ii), the quantitative database for the map, with the regular expression applied, the output types and the frequencies of the different types; (iii) etymological information regarding the names; (iv) further details about the map and auxiliary maps, which contain details from the main map or illustrate the same topic with other examples; (v) historical forms of the names. The *German Surname Atlas* is the first linguistic atlas which takes data from both present and past, reaching as far back as the Middle Ages, into consideration; (vi) bibliographical references, cross-references and further information; e.g., the frequency and distribution of names in neighboring countries.

The following case studies are taken from vol. 4 of the *German Surname Atlas*. With surnames derived from the provenance of recently arrived persons, we can illustrate ancient migratory movements because surnames emerged in a time characterized by a large degree of migration within the country.

The example of *Westphal*, which is concentrated in Schleswig-Holstein and Mecklenburg-Vorpommern (see figure 1), illustrates the migration of Westphalian settlers in the context of the German eastward expansion (*mittelalterliche deutsche Ostsiedlung*) of the 9th to the 14th century, in which Germans from modern-day western and central Germany settled less-populated regions of eastern Central and Eastern Europe, formerly inhabited mostly by Slavic and Baltic peoples. As this example shows, Westphalian settlers must have participated in the German eastward expansion to a major extent. This is supported by historical evidence showing that a large part of the population in today's Mecklenburg-Vorpommern has its roots

in the western low German area, as well as by linguistic similarities between dialects in Westphalia and in Mecklenburg-Vorpommern (Schmuck 2009).

Surnames such as *Unger* and *Hunger*, which refer to *Hungary*, are concentrated in Saxony and in the eastern part of Thuringia. The surnames *Böhm* and *Böhme* agglomerate not only in Saxony and Thuringia, but also in northern Bavaria, so that the latter surnames can be found in a curve around Bohemia in today's Czech Republic. According to Walther (1993, 498), the surnames *Unger, Hunger* as well as *Böhm* and *Böhme* reflect the fact that Saxonian miners often moved to Bohemian and Hungarian mining sites. After their return home, they were named after their former places of work.

Figure 2 shows the distribution of the name *Schweizer*. The varieties with *z* exist mainly in Baden-Württemberg, while those with *tz* are largely northern, is attached in the north, mainly in Rheinland-Pfalz and Hessen. These surnames also appear in France in about 3,500 births between 1891 and 1990 (www.notrefamille.com, 28.09.2011), as well as in Switzerland, with about 4,500 telephone connections (www.verwandt.ch, 28.09.2011). The reason why *Schweizer* and its variants appear quite often in Switzerland itself is that during the time when surnames arose, *Schweizer* originally referred to the village *Schwyz* and the surrounding canton. The name of the village and canton *Schwyz* was applied to the entire Swiss confederation only from the 14th century on. Diphthongization led to the standard German form *Schweiz*. Mainly after the Thirty Years War, many people from the village and canton of Schwyz and from the whole Swiss confederation settled in today's southwestern Germany.

Figure 3 gathers surnames which refer to the names of the low mountain ranges *Westerwald*, *Odenwald* and the region of *Bergstraße*, which is part of the Odenwald. The surnames which trace back to the toponym *Westerwald* are located around the corresponding low mountain range: *Westerwald* is concentrated around Frankfurt, *Westerweller*, with assimilation of *ld* to *ll*, in the northeast of Frankfurt and the eastern part of the Ruhrgebiet,

while *Westerwelle* is found in the area of Bielefeld and in the eastern part of the Ruhrgebiet. The surnames which trace back to the toponym *Odenwald* (*Odenwald*, *Odenwälder*, *Odenweller*, *Odenwäller*, *Ottenwälder*, *Ottenweller*) are located in southern Hessen, northwestern Bavaria and northern Baden-Württemberg. Right in the middle, around the homonymous region, *Bergsträßer* and *Bergsträsser* are to be found.

In the Middle Ages, German towns flourished and attracted rural populations, and the newcomers were often named after their place of origin. So with the surnames derived from the provenance of recently arrived persons which relate to single settlements, we can reconstruct where the migrants came from and where they settled down.

Onomasticians such as Grünert (1958, p. 537-553, map 1-9), Hellfritzsch (2007, p. 525-539, maps 1-4), Neumann (1970, p. 182-187, map 2), Neuman (1981, p. 276-283, maps 1-4) collected historical documents regarding surnames related to single settlements and mapped them. Thus they found out that the medieval catchment areas of smaller towns had a radius of barely 100 kilometres.

Conversely, the distribution of surnames can also illustrate where former citizens of a certain town or village moved, because newcomers were often named after their place of origin. In many cases, most persons who bear a specific name based on a small town or village still live within a radius of about 50 kilometres around the eponymous settlement (cf. the contribution of Pascal Chareille in this volume). Figure 4 illustrates this with the example of the surname *Rothenbucher*, with umlaut *Rothenbücher*. Here, the ancestor was named after the small village of *Rothenbuch* in the Spessart.

In addition to the Middle Ages and the early modern period, the database of the *German Surname Atlas* also opens up possibilities to reconstruct migratory movements during the 20th century because it contains not only German but also foreign surnames. This provides a broad

field of research in which linguists, historians, human geographers and geneticists can collaborate.

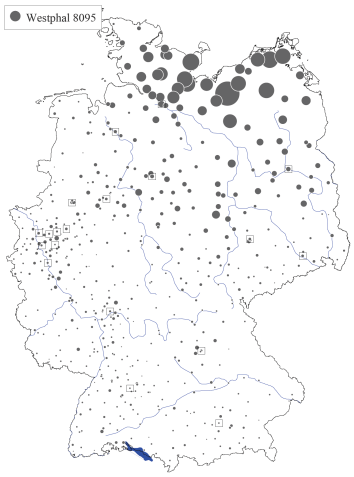

Figure 1: Relative distribution of *Westphal*

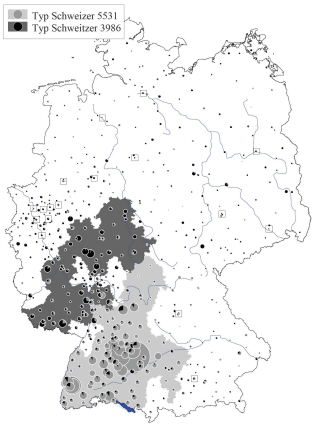

Figure 2: Relative distribution of type *Schweizer* and type *Schweitzer*

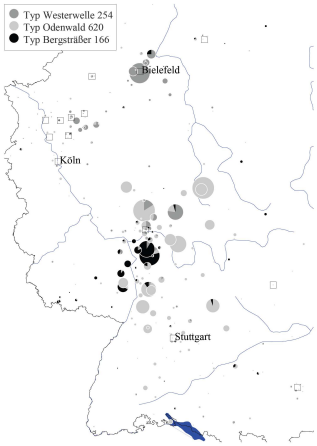

Figure 3: Absolute distribution of type *Westerwelle*, type *Odenwald* and type *Bergsträßer* in Western and Southwestern Germany

Map 4: Absolute distribution of *Rothenbucher* and *Rothenbücher* in Northern Bavaria

2. Data mining in the Dutch (historical) civil registration 1811-present [Gerrit Bloothooft, Kees Mandemakers, Leendert Brouwer, Matthijs Brouwer]

Names identify individual persons. As such, names are central in research dealing with individuals, and groups defined by properties of these individuals – such as families. In the latter, generations also come into play, carrying the dimension of time and historical developments in society. The spatial dimension also influences groups: members migrate and interact. For studies of subjects including genetics, health, demography and sociology, the identification of groups and knowledge of their dispersion in time and space is valuable if not essential information.

In Dutch and other modern civil registrations, people are identified not only by name but also by a persistent ID. By having the parents' IDs in the record of every individual, and a complete and accurate digital registration, all family relations in society are basically known, at least for a couple of generations. In these systems, names are no longer essential to demonstrate relations between people. However, for older registrations, no IDs were used, and reconstructions of relations between people depend strongly on their names and the description of relationships in certificates of birth, marriage and decease. Accuracy of these archives is often problematic, completeness rare, and full digitization a long-term goal only.

II Available data and major ongoing projects in The Netherlands

II.1 Modern Civil Registration

In 2000, a new law on the Civil Registration (CR) opened the possibility to acquire data for scientific research. This opportunity was used by Utrecht University and the Meertens Institute to request two selections of data, one centered around first names, and another around family names. Full population data were acquired for all first names of 21 million persons (5 million deceased). As well as all first names, the (internal) ID, the first names and

IDs of the parents, and the date, place and country of birth of all individuals, were provided. This constitutes a full population genealogy for several generations – but with only the first name known. The data describe the full population born after 1930. They become gradually less complete for earlier years of birth but still provide a 30% sample of all persons born in 1880. All in all, these data entailed 500,000 unique first names which were made public in June 2010 on www.meertens.knaw.nl/nvb. For the family names, full population data were acquired for the 16 million persons alive in 2007 with information about the following attributes: the family name, date, place and country of birth, and the current place of residence (compare Cheshire et al, 2011; Dräger, this paper; Coates and Hanks, this paper). These data were linked to data from the 1947 census. The 16 million persons carried 314,000 unique surnames. The website presenting the surnames was launched in December 2009 on www.meertens.knaw.nl/nfb.

II.2 Historic Civil Registration

Hundreds of volunteers are digitizing historical registers of birth, marriage and decease from the civil registration system that started in 1811, based on Napoleonic law. Currently about half of the job is done. There are now over 16 million registers digitized, containing information on about 70 million (not unique) persons (see www.genlias.nl). Automatic reconstruction of families from these data is now in progress in the LINKS project (*Link*ing *s*ystem for historical family reconstruction). Ideally, the goal of LINKS is to identify all individuals mentioned in the certificates uniquely, and, just like the modern CR, to tag them with a persistent ID and the IDs of their parents. It is possible to link this historical 'population registration' with the modern one, provided privacy reasons do not prevent this.

II.3 Historical Sample of The Netherlands

The Historical Sample of The Netherlands is a project that started in 1991, with the aim to reconstruct life cycles for an unbiased random sample of an eventual 78,000 persons (born

1812-1922) sampled manually from birth certificates. In addition to standard personal data, religious affiliation, occupation, household composition, literacy, social network, and migration history are also collected from the civil certificates and population registers (Mandemakers, 2000). More information can be found on www.iisg.nl/hsn

III. Data mining, considerations, tools and examples

III.1 Geographic spread

Current geographic spread of a family name can be shown immediately on the website of the Dutch Family Name Corpus at the municipality level. By providing an online possibility to search by *regular expression*, properties of all kinds of *sets* of surnames can be shown as well - see the example in Figure 5. These properties may include all kinds of spelling variation, or require the presence of certain morphemic properties which may be typical for some language or dialect. The same options exist for the first names website.

Figure 5 about here.

III. 2 Migration

A complete (historical) civil registration would allow for migration studies by tracing the places of births of subsequent generations. On the basis of our first-name corpus from the modern civil registration, we identified grandparents and their grandchildren, and computed the distance between their places of residence in 2006 (figure 6). When the grandchildren are young they live with their parents at an average distance of a stable 22.5 km. Between the age of 20 and 30, the grandchildren settle themselves and the average distance increases to 34 km, which remains stable again in further life. Distances do not sum over generations since on average grandchildren randomly move in all directions.

Figure 6 about here

Another analysis of geo-distributional nature, and related to migration, can be done for surnames. Given a limited migration some surnames may still be found in the region where the ancestor adopted the name, often many centuries ago. We determined for which surnames 50% of the bearers nowadays live within 30 km of a center municipality. Subsequently we computed per municipality the percentage of the population with such a “regional” name. Results are shown in Figure 7. Rural areas and closed communities such as fishing villages can have up to 43% of the population with a regional name – and a high percentage of consanguinity. Larger towns and newly reclaimed polders are a melting pot of families and obviously have much lower percentages (Bloothooft, 2011).

Figure 7 about here

III.3 Co-variation

An important property of the data in the civil registration (and reconstructed life courses) is that on the basis of known family relations, studies within families and across generations can be performed, thus informing on the social strata of the population. We explored this in a study of modern first names. The assumption was that parents do not chose names for their children at random, but (largely unconsciously) on the basis of what is fashionable or expected in their social environment. This would imply that the names of children in the same family convey some of this fashion. Traditional parents may name their children with old Dutch names like *Willem* and *Dirk*, and this combination of names will appear in such families more frequently than can be expected on the basis of individual probabilities of the names. By analyzing the names of millions of children in families with more than one child, we could cluster the names in such a way that names within a cluster have a higher probability to be found in a single family than across clusters (Bloothooft and de Groot, 2008). For modern naming, fifteen clusters or name groups gave a fair description of the 1,409 most frequent names (naming 75% of all children). These are (1) traditional Latinized names

[*Johannes, Maria*]; (2) Dutch traditional names *Trijntje*]; (3) Hebrew names [*David, Esther*]; (4) Frisian names [*Jelle, Nienke*]; (5) longer premodern Dutch names (popular before 1990) [*Wouter, Suzanne*]; (6) short international names (popular before 2000) [*Mark, Laura*]; (7) English names [*Kevin, Samantha*]; (8) short modern Dutch names [*Tim, Anne*]; (9) other modern names [*Milan, Lara*]; (10) Nordic and French names [*Niels, Anouk*]; (11) elite names [*Floris, Amber*]; (12) French names [*Jules, Dominique*]; (13) Italian and Spanish names [*Lorenzo, Felicia*]; (14) Arabic names [*Mohamed, Samira*]; and (15) Turkish names [*Hakan, Meryem*].

The geographic spread of each name group has significant features across the country, as shown in Figure 8 for traditional Dutch names, which mainly follow the Dutch bible belt – a narrow region of conservative Protestantism from the south-west to the middle of the country – and ends more widely distributed in the Northern provinces, while short English names are preferred in the areas of Catholic dominance, which earlier chose traditional Latinized names.

Figure 8[a and b] about here *COPYEDITOR: please .put them together*

In a subsequent study, we had available diverse socio-economic data from about 281,751 households, including the names of the children in the households. This allowed us to investigate the relation between socio-economic parameters, such as educational level and income of the parents, and the name groups. We also had lifestyle profiles of the households (summarizing all data), and could map the name groups on major lifestyle dimensions related to them (Bloothooft and Onland, 2011). Results are shown in Figure 9, with the horizontal axis related to household income or highest education of the parents (low-high), and the vertical axis related to affinity to tradition versus fashion. Major features are the tendency for well-educated and somewhat traditional parents to choose Dutch, Hebrew or Frisian names, while the medium educated and trendy parents favor foreign or fancy modern names.

Figure 9 about here.

This type of analysis could be done for surnames as well on the basis of known family relations and data from sources external to the civil registration, such as family income, education level, occupation, or ethnicity. This would underpin relationships between surnames and cultural, ethnic and linguistic (CEL) parameters (Mateos et al., 2007).

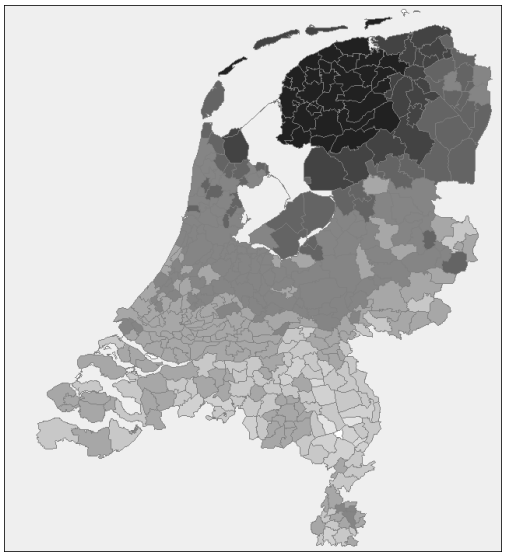

***Figure 5*. Geographic distribution of all surnames that fulfil the regular expression 'stra$', implying 483 names ending with –stra, in percentage per municipality. This is a typical Frisian name ending, expressing 'coming from'. The map shows the province of Friesland with more than 5% -stra names, the circular shape of the decrease of the presence of the name in the North, a relative sharp boundary with the Catholic south of the country - with exceptions in areas of industrial development (in the coal mines of Limburg, around Eindhoven (Philips company) and the textile factories in the eastern part). The 10 gray shades follow a logarithmic scale from over 5% (dark) to less than 0.01% (light).**

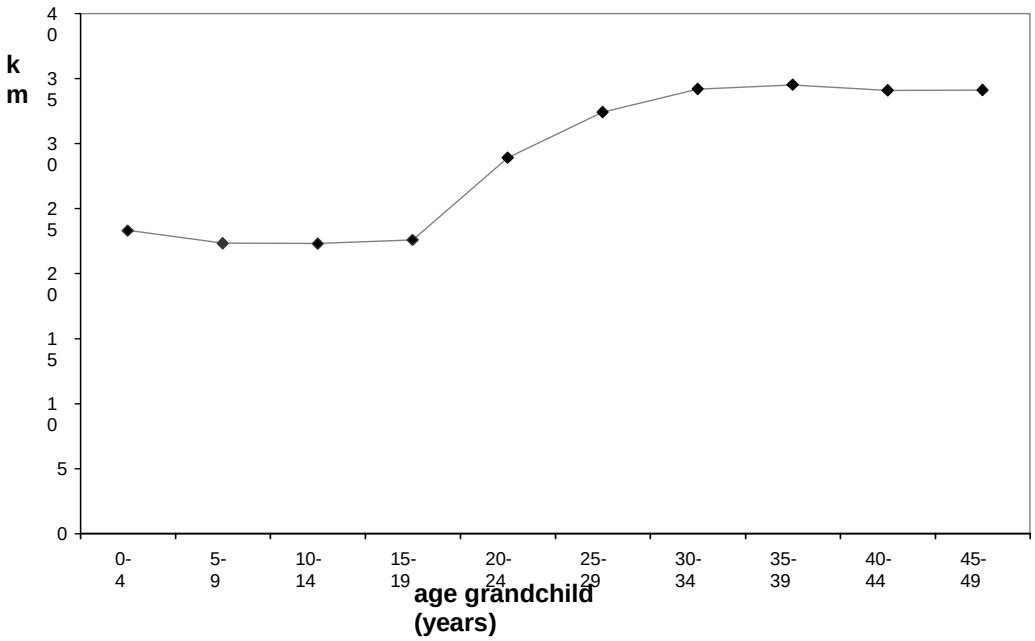

***Figure 6*. Distance between places of living of grandparents and their grandchildren in 2006.**

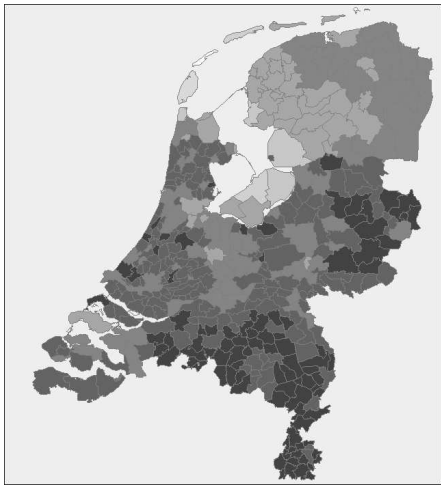

***Figure 7*. Density of regional surnames in The Netherlands. The five gray-shades indicate 1-2%**

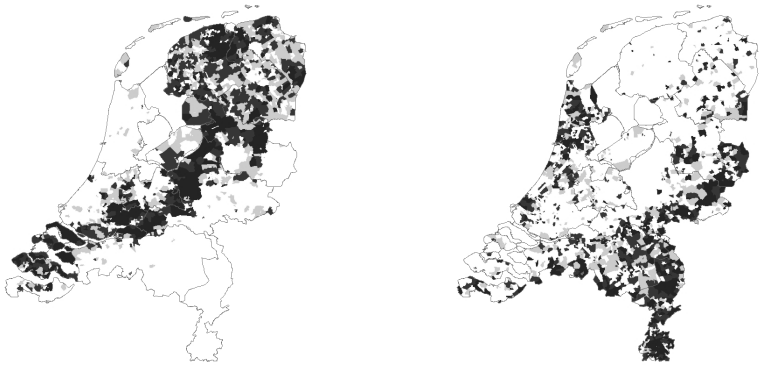

***Figure 8*. Geographic spread of Dutch traditional first names (left) and short English names (right).**

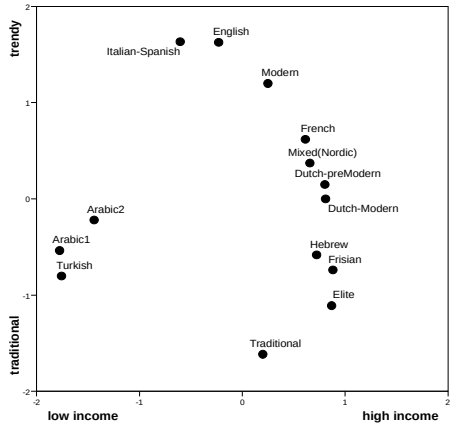

***Figure 9*. Name groups and lifestyle dimensions.**

.

3. The new Family Names of the United Kingdom project (FaNUK) [Richard Coates and Patrick Hanks]

The major new research project called *Family Names of the United Kingdom* (*FaNUK*) began on 1 April 2010, and will run for four years, based at the Bristol Centre for Linguistics in the University of the West of England, Bristol. It receives funding from the Arts and Humanities Research Council, and has an attached doctoral studentship. Some 5000 UK family names have no accepted etymological explanation; many others have been wrongly explained. FaNUK's goal is to make good these deficiencies through the creation of a database of family names containing an evidence-based account of the linguistic and geographical origins, history, and demography of at least the 43,000 most frequent extant names.

1. Research context

Public interest in the origins, history, and demography of family names is attested by the vast amount of amateur work and media interest in genealogy. This is poorly served by existing literature, not radically improved since work done in the 1950s (Reaney 1958, 1991). Many seemingly plausible earlier explanations are incompatible with new facts about name history and geographical distribution. Misperceptions have arisen because county-based research by medievalists lacks a national framework. Reliable new resources are needed which are accessible to an increasingly sophisticated public.

Family name research is interdisciplinary. New resources from history, family history, place-name study, official statistics, and genetics include collections and editions of medieval evidence, machine-readable census data, and new statistical methods for correlating family names and locations (cf. the contribution to this article by Pascal Chareille). Geneticists have begun working with local historians on the relationship between distribution of individual family names and their origin. Such work needs bringing together, allowing existing accounts

of family name origins and history to be evaluated, corrected, and supplemented, and allowing a satisfactory multidisciplinary framework to be created. FaNUK will emphasize family names as linguistic and historical entities, rather than focus on genealogy and family history. But it will systematically take account of the work of genealogists and family historians – especially the Guild of One-Name Studies (http://www.one-name.org/) – to ensure maximum credibility for a resource of which they represent the major likely consumers.

Although there is reliable smaller-scale work (e.g. the best one-name studies, and surveys of seven counties dealing with medieval family names), no current resource brings together medieval evidence for comparison with distributional evidence derived from modern online geodemographic tools. FaNUK prepares the ground for detailed genealogical work which will eventually secure the connections across time. When all this material is brought together, critical assessment of previous etymological and historical claims about names and their alleged continuity will be possible, new patterns in their historical demography will appear, and new etymologies for problematic names will be facilitated through direct comparison of the datasets. Research on this scale is entirely new in the UK. The proposed product will be by far the most wide-ranging, complete, and reliable source of relevant information. There is no competing online resource, and FaNUK will counterbalance much misinformation on amateur web-sites (often taken from existing literature).

The standard work on **English** surnames is Reaney (1958, and last revised 1991; R&W). Its defects are now apparent. For example, comparison with 1881 census data reveals no entry for common names such as *Alderson* (northern England), *Blair* (Scotland), and *Critchley* (Lancashire) and over 20,000 other family names with more than 100 modern bearers. Being essentially a dictionary of medieval surnames without declaring this in the title, it includes over 3000 defunct surnames, e.g. some derived from obsolete nicknames (*Ballox*,

Barebone, *Beardless*, etc.) It takes little account of geographical distribution or local sources, explaining *Broadhead* as a nickname and *Gawkrodger* as 'awkward Roger'; both are in fact from minor place-names. Reaney's links between medieval evidence and modern surnames are often demonstrably untenable, and some other etymologies are unreliable or misleading. Other previous **English**-oriented works include: Cottle (1967, 1978, 2009), and the nine counties of the English Surnames Series (ESS), based on McKinley's discontinued programme at Leicester University. A major critique of Reaney's methodology is Redmonds (2002). He and Hey (2000) have shown the need to integrate the study of family history with local history. Hanks and Hodges (1988; H&H), like its successor Hanks (2003; DAFN), is a general resource containing much material relevant to the UK and foreshadowing FaNUK in that its dataset has a broad ethnic and etymological scope, but the etymologies mostly lack medieval evidence.

Despite our reservations about these predecessors, they are usable as a foundation for FaNUK. They offer systematic hypotheses for confirmation or correction, in the light of new evidence. We are therefore grateful to the publishers and copyright owners who have made the material in R&W, H&H, and DAFN available to FaNUK in electronic form.

The best resource for **Welsh** surnames is Morgan and Morgan (1985). However, the headwords are Welsh personal-name forms, not surnames. References are regularly to undated secondary sources, not to dated primary documents. It is therefore not user-friendly for a non-Welsh-speaking public, and potentially misleading for unwary users. For **Scottish** surnames, the standard work is Black (1946), a fine collection of historical data where, as with R&W, names are selected from pre-modern evidence rather than a modern inventory, and the etymologies need systematic revision. The main **Irish** resources (de Woulfe 1923; MacLysaght 1985), are based on old work, though we now have de Bhulbh (2002). Both H&H and DAFN include reliable etymological information on Irish surnames, but none of

these works provides evidence for early bearers of Irish names. Such evidence exists, e.g. in the Tudor Fiants (Nicholls 1994), authorizations to the Court of Chancery in Ireland for the issue of letters patent under the Great Seal of English monarchs in the 16th and 17th centuries, which show surnames in transition from their Irish to their anglicized forms. FaNUK will include, for each Irish family name, evidence from such sources. Whilst the Republic of Ireland is not part of the UK, we cannot omit Irish names, both because of the mass Irish immigration into Britain, and because the north-eastern six counties of Ireland still form part of the UK.

On the basis of such previous work, FaNUK prepares the ground for a history of family names in the UK. Most academic effort will be directed at names of insular origin. However, the UK's multiethnic character will be addressed by including most **immigrant** names (principally **Huguenot** and **Jewish**, and those more recent arrivals having up to 100 current bearers), making FaNUK's range unique. The focus will be on (a) linguistic source (culturally important to those with foreign genealogy), (b) cultural and religious associations, and (c) how and when each name reached the UK, rather than its entire remote history elsewhere. For well-represented cultures, this will lead to projects beyond the end of FaNUK.

Commentaire [MAJ1] : I wonder if this should be 'more than' – can the authors be asked this?

UK surname research lags far behind that in many other European countries. In the Netherlands, two institutions are building large surname databases: Meertens Instituut in Amsterdam (www.meertens.knaw.nl/nfb) and the Central Bureau of Genealogy in The Hague (www.cbg.nl). In Poland, scholars at Pracownia Antroponimiczna (Anthroponymic Research Group, www.ijp-pan.krakow.pl/en/struktura-organizacyjna/zaklad-onomastyki/), Kraków, are researching a comprehensive historical dictionary of Polish surnames whose first volume appeared in 2007. Current UK family name research also compares unfavorably with funded allied areas like English and Scottish place-names. FaNUK seeks to redress the balance.

2. Research methods and project outcomes

We intend to address the research lacunae mentioned above by creating a database using data from the range of sources provided by copyright owners and consultants, gathered by the investigators, and screened, explained, and commented on by the investigators in conjunction with consultants. Machine-readable versions of R&W and H&H were successfully loaded into an experimental prototype database with active collaboration of the publishers. Before the project began, we audited the availability of other reference sources for possible addition to the database. We also have lists of relevant historical resources containing many individuals' surnames, and where such resources exist in e-form, permission is sought for electronic links between these data and the FaNUK database. Where they are not yet available, we are actively exploring with project leaders and copyright owners the potential for digitization to our mutual benefit. As a last resort, FaNUK mines documents conventionally.

The database will also establish the inventory of surnames in the post-1880 UK, accompanied by their geographical distribution and frequency. Surname distributions have been derived computationally by current collaborators from electoral rolls and the 1881 census, both now publicly available online.

FaNUK requires many consultants with various specialisms, philological and computational; we do not have space to mention them all here. As the project has progressed, we have benefited considerably from the cooperation of Steven Archer, who has created mappings of the frequency and geographical distribution of surnames recorded in the 1881 national census. A surname whose association with a particular locality is statistically significant may in many cases have originated there, and this possibility needs to be exhaustively investigated before other possibilities are considered. We say this with confidence, because although people can move around, there is ample evidence that a large number of surnames still cluster around a point of origin. Because of this phenomenon, we

have been able to resolve some issues about the original distribution and source of some surnames deriving from place-names which are recorded from medieval times but wrongly explained in R&W, e.g. that the surname *Harmison* originates in Hermiston in Roxburghshire rather than in Harmston in Lincolnshire. Place-names are comparatively stable, both linguistically and geographically. Surnames are not. Families and individual bearers move around; competing spellings are commonplace; people adopt other surnames; surnames are not necessarily transmitted as counterparts of the Y chromosome; surnames die out. Archer's work (2003, 2011), has confirmed the essential correctness of H.P. Guppy's hypothesis of a significant relation between many surnames and locations, though many such relations remain unexplained. The association between *Fazackerley* and Lancashire is obvious because there is a place in Lancashire called *Fazakerley*; there is no place anywhere else of this name, or with a name remotely like it. Elsewhere there are associations between variants of names and particular places, as with *Pardoe* and *Pardey*, which share a linguistic origin, but have no known genealogical connection or shared source; one may be waiting to be discovered through statistical work on distributions.

3. Summary review of targets and plans for dissemination

FaNUK's primary target is to create reliable explanations for the approximately 43,000 long-established or traditional insular surnames in the UK with more than 100 current bearers. A secondary target is to add explanations for unproblematic names of lower frequency. A tertiary target is to add entries for about 3,000 names of recent immigrant origin, indicating where they came from, what (if anything) is known about their meaning, and giving information relevant to their UK status, such as date of arrival. Data from recent electoral rolls and censuses show that there are over 370,000 different surnames in Britain today, but the vast majority of them are extremely rare, being borne by only a handful of people. Surprisingly, over 300,000 are the names of recent immigrants from a vast number of

countries including, but by no means restricted to, the countries of the former British empire. That leaves the 43,000 surnames referred to above.

The principal output of FaNUK, its publicly accessible database, will be valuable to genealogists, geneticists, local historians, historical demographers, historians of the English and Celtic languages, other philologists, and place-name scholars.

4. Writing the History of the Québec Populations Using Surname Frequencies [Guy Brunet, Pierre Darlu, Bernard Desjardins]

The study of the geographical distributions of surnames obtained from various registers has already demonstrated its efficiency to infer migration of people, either by applying statistical models when surnames are recorded only once at a given time, using Fst statistics (Wright, 1951) or probabilistic models (Karlin and McGregor, 1967; Yasuda et al., 1974, Zei et al.,1983), or by comparing surname frequencies recorded at least twice at the same location (Wijsman et al., 1984; Degioanni and Darlu, 2001; Darlu et al., 2011). This second strategy has been less frequently used because it requires historical records. These are now more abundantly available, thanks to the efforts of historians, as exemplified by several articles in this volume (Bloothooft, 20XX; Chareille, 20XX; Boattini et al., 20XX) and by the present article showing original analysis of migration in Québec.

The arrival of French immigrants in Québec during the 17th century was the starting point for the growth of the French Canadian population, which increased from 18,000 inhabitants in 1700 to 200,000 in 1800 with a corresponding geographic dispersal. On their arrival, the pioneers colonized a strip of land along the Saint-Laurent River expanding first from the two main poles of settlement (Montréal and Québec). During the 18th century, northern and southern parts of the river were progressively occupied, as well as the places between Montréal and Québec.

From the very beginning, baptisms, marriages, and deaths were systematically recorded in parish registers, allowing the reconstruction of the temporal and spatial evolution of the European population. Data on the native Americans were insufficient to allow a similar analysis. The onomastic information drawn from these records were analyzed to infer the demographic growth of this population, its renewal, migration, and geographic expansion.

The present work is based on 392,998 baptism records noted between 1608 and 1799. For each of them, corresponding to a baptized child, the surname, the birth date, and the birth place (in term of parish and County) were noted. Although the question of lemmatization of the surname variants is far less difficult in Québec than in the situation described by Chareille in the case of the 14th and 15th century documentations (Chareille, this volume), surnames had to be first standardized to allow for orthographic variations. Then their frequency was studied by parish and County for four successive periods of time: P1: 1700-1724; P2:1725-1749; P3:1750-1774; P4:1775-1799

Global dynamic of the population

The set of surnames, already largely diversified before 1700 (1349 surnames) was relatively stable in the first part of the 18th century, because of the reduced number of immigrants. The number of baptisms increased fourfold between the first (P1) and the fourth (P4) period. The proportion of surname per baptisms (S/N) was rather high before 1725, and progressively decreased during the rest of the century, indicating that there were new arrivals of migrants with new surnames. This is also stated by the evolution of S' and S'' (See Table 1). Indeed, the number of new surnames arriving at the end of the period (P4, S=4266) was four times higher than those arriving during the previous period (P3, S=923). The turnover between the surnames disappearing (S'') and those arriving (S') leads to a positive although weak balance of 239 surnames in P2, larger in P3 (1947), and in P4 (2679). The burst of growth occurred in the middle of the century, with the arrival of many surnames superimposed upon the maintenance of a core of surnames brought by the first settlers. The proportions of singletons (name occurring only once) confirm this point.

[Table 1]

The two main towns (Québec, Montréal) show a larger diversity of surnames than the parishes or the Counties, obviously following a linear relation with the population, as shown in Figure 10. However, one can also show that Montréal and Québec display an excess of surnames compared to the other places. This excess is larger for the P4 than for the P1 period, meaning that the immigrants are preferentially arriving in these two largest towns, particularly at the end of the century. Actually, the proportion of singletons is respectively 50% and 49% in the parishes of Montréal and Québec (and the weight of the three most frequent surnames is 3.6% and 2.6%) whereas the proportion of singletons is only 29% (and the three most frequent surnames account for 6%) in a typical parish like Saint-Eustache, where 3500 baptisms were recorded.

Such a contrast between large and small populations has long been reported (Zei et al., 1983). The largest towns attract first the immigrants that have heterogeneous origins and consequently have a larger diversity of surnames.

[Figure 10 about here]

Surname resemblance, tree representation, and its geographic projection

To specify the geographic structure of the surname distributions in Québec, we calculated the pairwise surname distances between Counties, using the classical Nei's distance, as used first by Chen and Cavalli-Sforza (1983). The idea is that two Counties sharing close surname frequencies were exchanging people in the past more intensively than two Counties that show a large surname distance.

Once the surname distance matrix was obtained, trees were constructed by the neighbor-joining method (Saitou and Nei 1987), with bootstrap resampling (Felsenstein, 1985) to estimate robustness at nodes of the tree. The consensus tree can be projected on a geographic map, connecting surfaces being clustered together with a given level of bootstrap proportion (Figure 11 about here).

Figure 11 shows that the surname resemblances are clearly high between neighboring Counties, which can exchange individuals readily, and an absence of noticeable division between the two banks of the Saint-Laurent river both near the Montréal and the Québec Counties. Moreover there is no significant structure that distinguishes the area around Montréal from that around Québec. In fact, there are few strong structures except those plotted in Figure 11, suggesting that the dispersion of people (and surnames) was already well advanced on a large scale at the beginning of the 18th century.

Probability of geographic origin (pgo): a Bayesian approach

Since migration of people involves migration of their surnames (or at least the surnames of their children quoted in the birth registers), the movement of people – usually the males because surnames are paternally transmitted – can be reasonably inferred from the movements of their surnames, although with some limitations (Degioanni et al 2001, Darlu and Degioanni, 2007, Darlu et al., 2010, 2011). A Bayesian approach can be applied, as detailed elsewhere (Degioanni and Darlu, 2001, Darlu and Degioanni, 2007, Chareille and Darlu, 2011).

For the area under investigation (here a County), called the "recipient area", the probability that the surname s_k which is present at time $t+1$ and absent at time t originated from another area, a_i called the "source area" i, is, according to Bayes' Theorem:

$$p(a_i|s_k) = \frac{\pi(a_i)p(s_k|a_i)}{\sum_i \pi(a_i)p(s_k|a_i)}$$

Where $p(s_k|a_i)$ is the probability of observing the surname s_k within the a_i-th area. This probability can be estimated by the observed frequency of the kth surname in the a_i-th area. $\pi(a_i)$ is the *a priori* probability of emigration from the geographic area a_i to any other area, whatever the surname. The sum is over all considered geographic areas.

As this probability of origin of surnames is estimated for each surname s_k, one obtains a more accurate estimate by summing all surnames and then by calculating the weighted mean probability of geographic origin, pgo_i, of any surname newly arriving between two periods in a given recipient area i as:

$$pgo_i = \frac{1}{\sum_k \omega_k} \sum_k \omega_k \, p(a_i | s_k)$$

where ω_k is a weight taking into account the fact that several persons could share the same surname. Once these probabilities are obtained, they are used as a new estimate of the *a priori* probability $\pi(a_i)$ and are replaced into the Bayesian formula which is recalculated. This iterative process is carried on until a convergence criterion is met (for extensive discussion, see Degioanni and Darlu, 2001).

Figure 12 shows the probability of geographic origin of newly arriving immigrants at Rimouski. Most of them did not come from the 43 Counties (outside: 23%). The most part came from the neighboring Counties (Kamouraska, 30% ; Montgagny, 17%). Clearly, the settlement in this part of Québec was done from place to place at short distance. A large town like Québec did not participate much in this process of migration.

The same method was applied to the migrations between the three main towns, Montréal, Trois-Rivières, and Québec. Table 2 shows the probability of geographic origin for each town. Most of the immigrants were coming from “outside”, p=0.44 and 0.57 for Montréal and Québec respectively, much more than for Trois-Rivières (p=20). If some migrants to Montréal came from Québec (p=0.18) the reverse is not true (p=0.06). Trois-Rivières received its immigrants mainly from Québec.

[Table 2]

Conclusion

As demonstrated by the example of the Province of Québec for which accurate and exhaustive data are available for a long period of time, the use of surname frequencies in a geographic and historical context allows inferences on the peopling and on the spatial population structuring. The few methods used in this paper (analysis of surname distribution, calculation of the surname distances between places, use of agglomerative procedures to estimate robustness of surname proximities and their geographic representation, estimation of the probabilities of origin of migrants) allow us to conclude that the various Canadian parishes in Québec were, at the end of the 18th century, not very strongly structured, reflecting the dispersal of the previous generations, but nevertheless maintaining exchanges and migrations at short distances between neighboring places, and retaining the Saint-Laurent River and the two main centers of population (Montréal and Québec) as the most important delineating geographical elements.

		Number N of Baptisms	Number S of Surnames	Proportion (%) of Surnames among the Baptisms	Proportion (H/S %) of Singletons among the Surnames	Number S' of newly arriving Surnames	Number S'' of Surnames disappearing next period
Before 1700		41759	1349				
1700-1724	P1	44857	2709	6.0	8.7	1704	544
1725-1749	P2	56246	2768	4.9	1.9	411	348
1750-1774	P3	107919	4798	4.4	13.4	923	1587
1775-1800	P4	183961	7571	4.1	19.2	4266	

Table 1 Distribution of the numbers of baptisms (N), surnames (S) and of their ratio according to periods. H/S is the proportion of Singletons (Hapax), S' is the number of new surname arriving at a given period and still found in all next periods, S'' is the number of surnames already present or arriving at a given period and disappearing at the next periods.

		From Montreal	Trois Rivieres	Quebec	Outside
	Montreal		0.01	0.19	0.50
To	Trois-Rivieres	0.04		0.28	0.21
	Quebec	0.06	0.00		0.66

Table 2 : Probabilities of geographic origins of migrants coming from Montréal, Trois-Rivières, Québec, and from outside to these cities, between 1725-1775 (P2+P3) and 1775-1799 (P4)

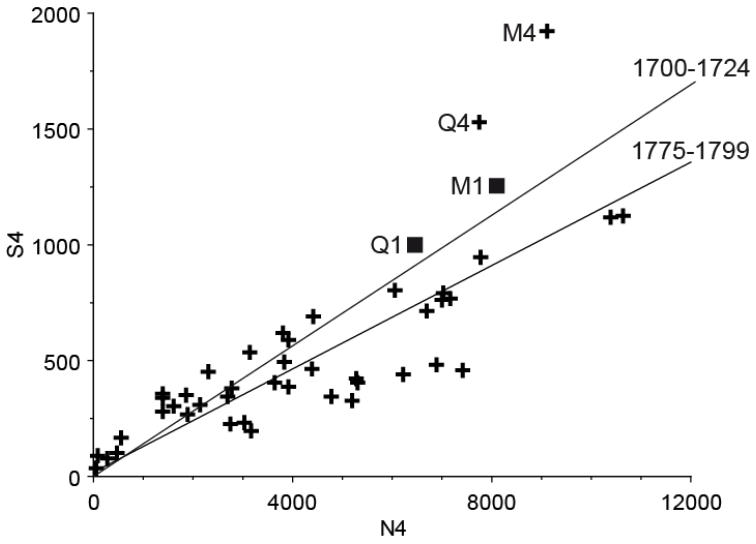

Figure 10. Regression of the number of surnames (S4) on the number of baptisms (N4) observed in 43 Counties for the period 1775-1799 (P4). The line of regression for the period 1700-1724 periods (P1), is also drawn for comparison, and is identical for the 1725-1749 (P2) and 1750-1774 (P3). Montréal and Québec are plotted for the P1 and P4 periods (M1, M4, and Q1,Q4 respectively), to show the larger than expected increase of the number of surnames between these two periods of time (P4 versus P1) whereas the trend is stable or even inverted for the other towns.

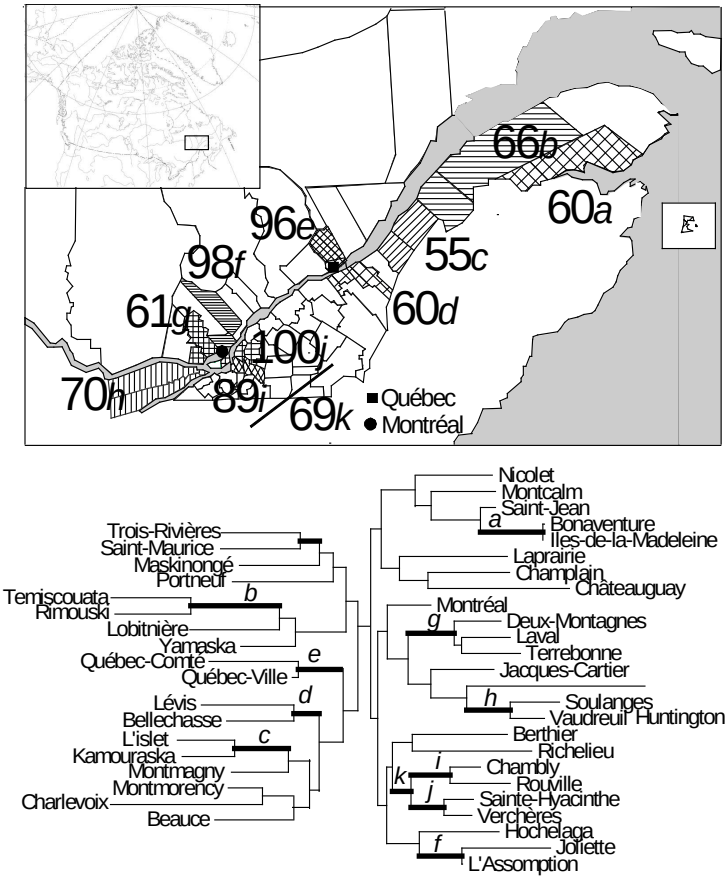

Figure 11 - Projection of the clusters defined by bootstrap proportion larger than 55% in the unrooted tree reconstructed by Neighbor-Joining from the Nei's pairwise surname distances between the 43 Counties (P3 and P4 pooled). Numbers in the map are the bootstrap proportions (%) attached to the branches labeled with the corresponding italic letters.

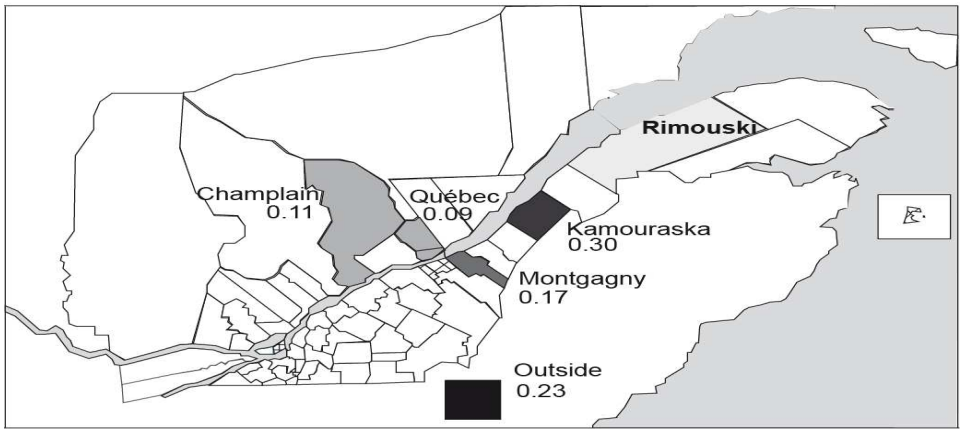

Figure 12. Probabilities of geographic origin of migrants newly arriving at P4 (1775-1799) in the Rimouski county from other Counties of the Province of Québec, or from elsewhere (Outside) (e.g. 30% of the migrants arriving in Rimouski at P4 came from Kamouraska)

5. A long-term perspective on anthroponymic corpora [Pascal Chareille]

It was the 11th century that saw the emergence of the two-element naming system still in use in France today. While this system was certainly not initially patronymic, the transmission of the surname—although not systematic before the 18th century—probably became "usual" as early as the 13th century. In the written sources used by historians, names provide abundant material for study. In France, since the Revolution and the establishment of a civil status register, potentially exhaustive nominative data for the whole territory are available, strengthening a system of registration which had existed since the early 16th century. The vicissitudes of archival conservation, however, are such that not all these documents have come down to us. Indeed, they are even relatively rare for the 16th century. And the further back in time one goes, the less the data are spatially exhaustive. The documents which predate the parish registers never contain the whole population. Thus the tax rolls from the 14th and 15th centuries, some admirable regional series of which have survived, only name the head of the household, and almost never the other members. In these documents, in which men are over-represented, the mode of designation of individuals already very broadly associates a name (or forename) with a surname (either individual, family or patronymic), and hence it is possible to envisage a study of anthroponymic stocks, in particular stocks of surnames, over a long duration (15th to 20th centuries).

The exploitation of medieval sources in this perspective, however, remains a perilous exercise: identifying individuals, and hence anthroponyms, may be uncertain: is the *hug[ue]s boy laigue* thus designated in a census of households in Dijon in 1376 the same person as the *hug[ue]s boilleaux* identified a year later in the same street? Examples of this type are legion, and it is often not simple to decide, since the transcription of names was largely phonetic at a period when writing was not yet in general use and spelling still inconsistent. Numerous

criteria (orthographic, linguistic, phonetic, etc.) can be involved in the differentiation of variants, and the choice whether to group the latter together or treat them separately is obviously decisive for the constitution of such historical corpora. The differentiation of names such as *Fabre, Favre, Febvre, Fèvre, Lefebvre, Lefèvre, Lefébure,* etc., or *Gauthier, Gautier, Galtier, Vautier, Vaultier,* etc., which goes uncontested in present-day lists of patronyms, is not necessarily pertinent for the Middle Ages. Lemmatization is therefore a necessary and unavoidable stage in the anthroponymist's task. In practice, it leads to the establishment of separate corpora depending on the level of lemmatization adopted, either only grouping together the minor spellings and/or variants ("weak lemmatization"), or else associating, in a common "root form", all the related forms ("strong lemmatization").

Patronymic stability: Normandy 1383 to 1515...

Normandy is one of the regions for which we have at our disposal a considerable historical corpus of 64,000 anthroponymic occurrences, concerning more than 55,000 individuals, drawn from the perusal of some 1,400 *rôles du monnéage* [rolls of a currency stabilization tax], dating from 1383 to 1515, concerning nearly 550 parishes scattered over five viscountcies (Bayeux, Caen, Falaise, Vire and Orbec) (Angers and Chareille 2010). Nearly 13,000 different patronyms have been identified, a number which was reduced to 7,600 after "strong lemmatization".

Despite this high level of lemmatization, nearly three out of every four patronyms is only attested in a single viscountcy, and less than 3.3% are present in all five. In 15th-century Normandy, then, the monophyletic character of patronyms is marked, suggesting an essentially local distribution of patronymic homonymy and a rooting of populations. It is, however, difficult to determine whether the high degree of micro-regional specificity in the 15th century is ascribable to low population mobility or to the relatively recent adoption of patronyms, as the spatial dispersion of the hypothetical original corpora proves to be a slow

process. Furthermore, the linguistic dimension of the problem, which is indisputable, still needs to be evaluated.

Despite these specificities proper to the above viscountcies, the most frequent patronyms are those which are also to be found in various places all over Normandy. None of the 100 most frequent patronyms in the whole set of corpora from 14th- and 15th-century Normandy is absent from more than two viscountcies.

The division of this corpus into four periods (P1=1383-1413; P2=1416-1449; P3=1452-1479; P4=1482-1515) makes it possible to examine its evolution over a long duration: *Lefebvre*, *Jehan*, *Hue*, *Martin* and *Hebert* are the five most frequent patronyms and, with the exception of *Hebert*, they always occupy one of the eight leading positions. The stability of these results over a very long duration is remarkable. The 25 most frequent patronyms in the 15th-century corpus are all, with the exceptions of *Regnault* and *Gueroult*, among the 150 most frequent today in the department of Calvados. This stability, however, only concerns the most frequent patronyms. In those parishes for which the documentation is continuous, less than 15% of these patronyms are attested over the total period (1383-1515), one which admittedly was particularly troubled. It is not an easy task to interpret this renewal, but the latter does not appear to be specific to either the period or the chosen analytical scale (see Darlu et al. 1997, for the period 1891-1940).

The question of migrations: the example of the Dijonnais region, 1376-1610

Historians, following the example of geneticists, use anthroponymy as one of the ways of tracing population movements, whether it be a matter of studying "long-distance" migrations within a vast territory or between one linguistic area and another, or of intra-regional migrations.

A few rare documents allow a systematic count of instances of explicit extra-urban mobility. This is the case with a household census carried out in Dijon during 1376-1377 (see

the extract in Figure 13): the origin (parish and street) and destination of the known migrants are often clearly mentioned (Beck and Chareille 1998).

In the absence of direct information, the study of migrations can also be envisaged on the basis of the count of surnames corresponding to place-names. We are aware that the method is imperfect and questionable (Emery 1952, 1955; Kedar 1973), but its application to the above enumeration concerning the Dijonnais region allows the construction of a map (Figure 14) which is perhaps less indicative of the main axes of migration toward Dijon than of a perception of the surrounding space (Beck and Chareille 1997).

The application to historical corpora of tools developed for the study of population genetics is not impossible and, moreover, enables an approach to the question of mobilities (Darlu et al. 2010; Bourin and Sopena 2010). Their use can, however, be difficult, constrained as it is by the limitations of the documentation: the absence of exhaustivity in the corpus, and the relative uncertainty as to both the hereditary nature of surnames and the extent to which they were fixed, which was certainly the norm in the 14th century but was by no means an exclusive rule. Nominative lists do not, except in exceptional cases, make it possible to identify a migrant who might have given up his former surname in favor of another recording his provenance or, on the contrary, sealing his adhesion to a new community through the use of local sound patterns in place of “exotic sonorities”. And we do not know the possible extent of this phenomenon, which is attested in various places.

Despite these difficulties, the diachronic analysis of spatial distributions allows the—however fragmentary—reconstitution of the histories of certain patronyms, and hence possibly of families, and thereby makes it possible to formulate hypotheses on migrations.

Phylogenetic methods make it possible to evaluate the more or less close proximity between the corpora on the sole basis of the presence/absence of a patronym in various places

without taking into account the variability of patronymic frequencies, the latter data being potentially unreliable as far as the medieval period is concerned.

The exhaustive reading of the household census of the bailiwick of Dijon for the years 1376, 1424, 1470 and 1610 makes it possible to construct a corpus of more than 35,500 occurrences distributed over 288 continuously documented localities grouped together by canton (on the basis of present-day administrative divisions). The anthroponymic structure of the populations thus observed highlights four groups within each of which the patronymic proximity suggests more intense exchanges. The relationships between the cantons can be represented in the form of a tree constructed by neighbor-joining (Saitou and Nei 1987) with bootstrap values (Felsenstein 1985) (Figure 15). The comparison with 20th-century data, taken from the *Registre français des noms patronymiques* [French register of patronymic names] for the period 1891-1940, reveals an astonishing stability: the present-day anthroponymic structure was already in place, with few differences, in the Middle Ages. This result needs to be further refined, but it does seem to suggest that the most recent migrations have not, at this scale of analysis, had a destructuring effect on micro-regional patronymic corpora, and hence that the privileged axes of population interchanges have not undergone any fundamental changes.

The (re)constitution of patronymic corpora for past periods is a difficult exercise, but the problems inherent in historical documents are not insurmountable. It is surprising to discover, as far as the regions which it has been possible to investigate are concerned, that many of the points that seem to characterize contemporary corpora (diversity of corpora, a high degree of local specificity for most patronyms, renewal of the overall corpus, yet stability of the most frequent names in the results, etc.) already seem to be in place in 14th- and 15th-century France.

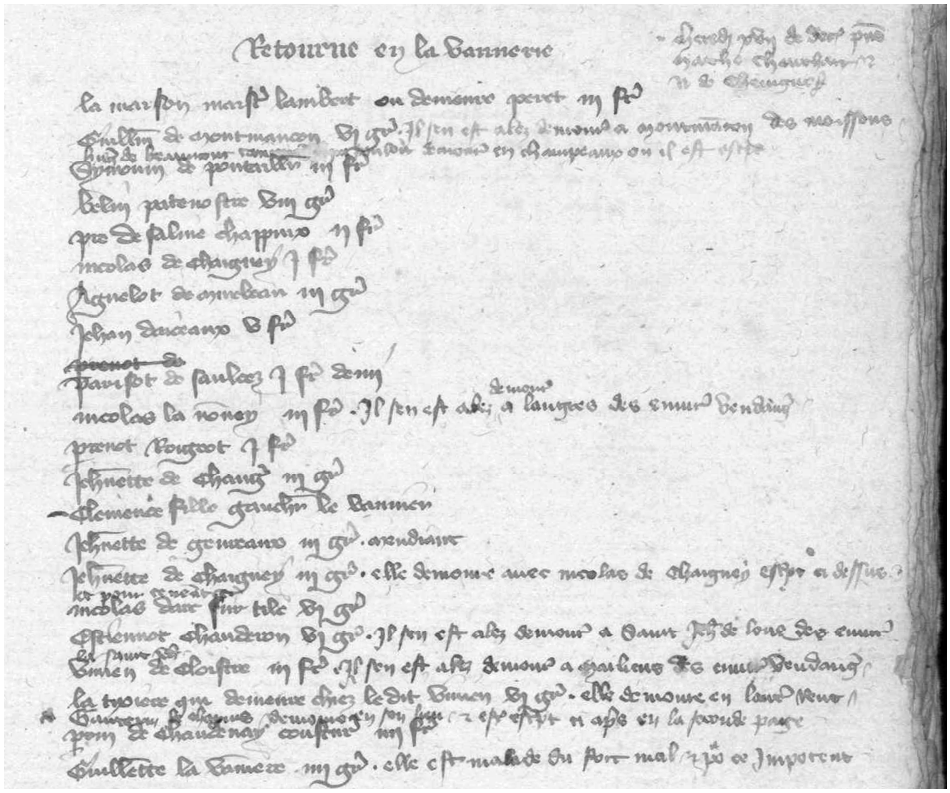

Figure 13. Annotated household census (1376-1377) [*dénombrement des feux*] for Dijon, available at:

http://archivesenligne.cotedor.fr/console/ir_ead_visu_lien.php?ir=630&id=73969140

(FRAD_021_B_11574_0109, Chambre des Comptes de Bourgogne Dijonnais).

In this extract, concerning a street known as "Retourne en la Vannerie", the annotations mention that, for instance, "Guill[em]in de Montmancon" (entry 2) "left to live in Montmançon at harvest-time" [*Guill[em]in de Montmancon sen est alez demour[é] à montma[n]con des moissons*], and that "Nicolas la Monney" (entry 12) "left to live in Langres around the time of the grape harvest" [*nicolas la mon[n]ey sen est alez demour[é] a langres des envir[ons] vendang[es]*], etc.

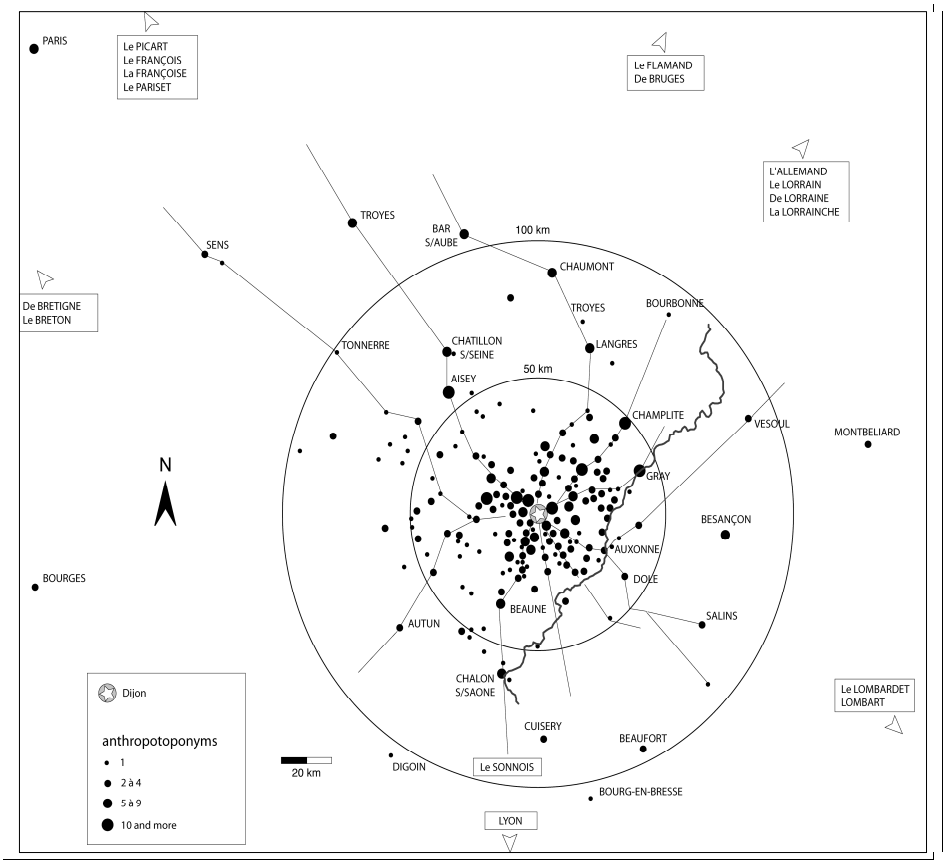

Figure 14. Surnames with place-name elements (or 'anthropotoponyms') at Dijon in 1376-1377.

This map, which is visibly articulated along the main routes from or towards Dijon (the strategic, political and economic routes of Burgundy in the period of the Valois dukes), is probably a fair reflection of both a large proportion of the migratory realities of the time and also, indirectly, of the perception of their surrounding space by late 14th-century inhabitants of Dijon.

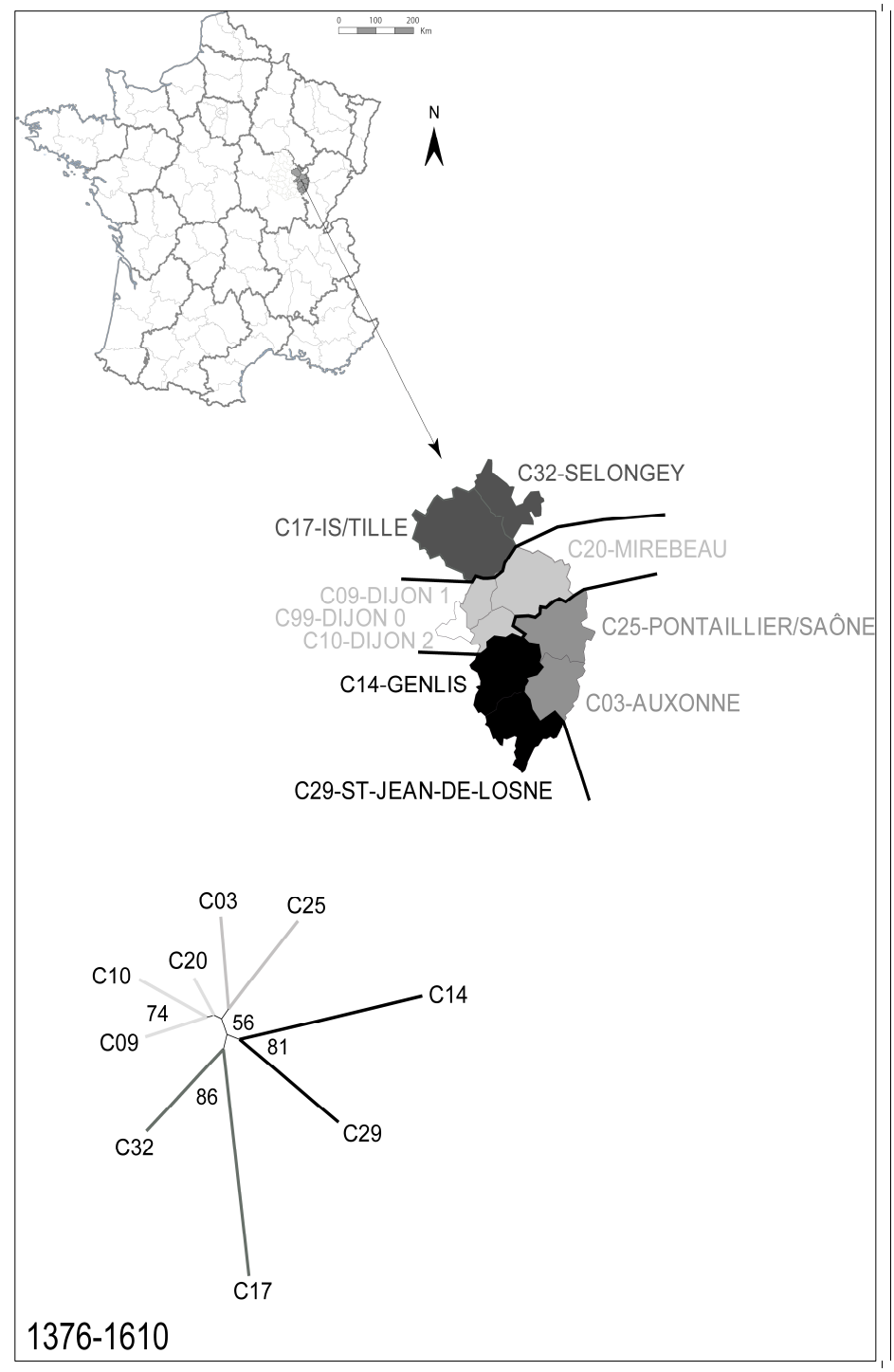

Figure 15. Division of cantons based upon the presence/absence of (sur)names.

The data for 1376-1610 make it possible to identify, from a surname perspective, four groups: 1) Selongey and Is-sur-Tille, which correspond to the enclaved, afforested land of "La Montagne"; 2) Mirebeau and the cantons lying to the east of Dijon on the Côte and near the capital; 3) and 4) the low-lying land on the plain of the Saône, divided by the Tille and its marshes, which were later drained and were long an almost impassable barrier and thereby a *de facto* limit on people's movements: Pontarlier and Auxonne are on the left bank (to the east) of the river; and Genlis and Saint-Jean-de-Losne on the right bank (to the west).

6. Reconstructing past genetic structures in recently transformed populations: Surnames and Y-chromosomes in the Upper Savio Valley (Central Apennines, Italy).

[Alessio Boattini, Antonela Useli, Davide Pettener]

Many of the preceding contributors (Bloothooft et al., Brunet et al., Chareille, Coates & Hanks, Dräger) focused on the efficacy of surnames in tracing movements of people as well as in reconstructing historical changes in migration patterns and/or similarity/dissimilarity coefficients between populations. These features make surnames an interesting tool for human population genetics inferences *per se*.

Recently, in the context of molecular anthropology studies focused on the variability of the Y-chromosome – with which surnames share a patrilineal ancestry (King and Jobling, 2009) – the study of surnames found a new field of application. Most frequently, surnames have been advocated to design more careful sampling strategies (Manni et al., 2005, Boattini et al., 2010a). Surnames have been used to increase the 'archaeogenetic' power of genetic studies through the analysis of historical records and pedigrees (Bowden et al., 2008; Boattini et al., 2011). In this way, researchers were able to infer 'past' genetic structures of populations by selecting those individuals who carry surnames that were proved to be present in a certain area at the time of surname introduction. In particular, Manni et al. (2005) introduced a 'general' surname method, based on Self-Organizing Maps (SOMs), that provides an efficient identification of groups of surnames that share a geographic origin and migration history. The method was first tested in the case study of the Netherlands (Manni et al., 2005, Manni et al., 2008), then successfully replicated in microgeographic contexts (Boattini et al., 2010a, 2010b; Rodriguez Diaz & Blanco-Villegas, 2010).

Here we apply the SOMs methodology in order to unravel the genetic structure of a population that was subjected to radical transformations during the last century. The Upper

Savio Valley – a mountain population located in Italian Central Apennines – experienced a series of demographic phenomena that were common to great part of Italian mountain communities: major depopulation and migrations towards the most important urban centers. In this study, we will compare surname clusters identified by SOMs with Y-chromosome variability in the Upper Savio Valley. Our main purposes are: 1) to test the power of the SOMs method to discover 'real' (biologically significant) clusters, and, if this condition is met, 2) to search for historical changes in surname structure of the population and 3) to identify remnants of historic genetic structures within the investigated area.

The data and methods

Surname analysis is based on 10,202 records from conscription lists for the years 1828-2005, corresponding to individuals born between 1808 and 1987. Following historic/geographic criteria, the Upper Savio Valley was subdivided into five areas (A, B, C, D, E), of which A and B correspond to the main urban centers of the valley – where the great part of the population is currently settled – while C, D and E are very rural areas, that nowadays are largely deserted (Figure 16).

Surname distributions were analyzed with SOMs. The SOMs method is a clustering technique through neural networks based on “competitive learning”, an adaptive process in which the cells (“neurons”) simulating a neural network (“map”) gradually become sensitive to different input categories (Kohonen, 1984). The main idea is that different neurons specialize to represent different types of input vectors; in doing so they interact with the neighboring neurons by means of a “neighborhood function”. This procedure will result in the differentiation of the whole map-space: a) identical vectors will be mapped at the same neuron, b) slightly different ones at close neurons, while c) very different vectors will be mapped at far neurons. The shape (rectangular or square) and size (number of cells) of the

SOMs are defined by the user. The size of the map determines the maximum number of different clusters; therefore, larger maps will classify items (surnames, in this study) more accurately than smaller ones. Nevertheless, it may happen that some cells remain empty, while others collect many items. Manni et al. (2005) demonstrated that the SOMs method can be considered a "blind" automated approach to identify the geographic origin of surnames. For the study of Y-chromosome variability, we collected peripheral blood samples from 59 individuals who were selected on the basis of a) pertinence of their surname to one of the main SOMs clusters (see below), b) ascertained patrilineal residence in the Upper Savio Valley for the last three generations. For each sample, 31 binary polymorphisms (M213, M9, 92R7, M173, SRY1532, P25, TAT, M22, M70, 12f2, M170, M62, M172, M26, M201, M34, M81, M78, M35, M96, M123, M167, M17, M153, M18, M37, M126, M73, M65, M160) and 12 short tandem repeats [STRs] (DYS391, DYS389I, DYS439, DYS393, DYS390, DYS385a/b, DYS438, DYS437, DYS19, DYS392, DYS389II) were typed.

Results and Discussion

The geographic distribution of surnames was analyzed using SOMs. This revealed four main surname clusters: clusters I (33 items) and II (99 items) are mainly represented in areas C, D and E, thus these groups of surnames may be considered as indigenous to rural areas, while clusters III (72 items) and IV (125 items) are mostly found in areas A and B, thus the corresponding surnames very likely had their origin in the urban centers of the Upper Savio Valley (Figure 17). For some of these, we were able to confirm their inferred place of origin based on 16th-century surname information for two Upper Savio Valley parishes from previous research (Boattini & Pettener, 2005). As a second step, we explored diachronic changes in SOMs cluster frequencies by subdividing our data according to six 30-year

intervals (referring to the year of birth: 1808-1837, 1838-1867, 1868-1897, 1898-1927, 1928-1957, 1958-1987).

All the considered areas show a temporal increase in the degree of within-area surname diversity (Figure 16), particularly for the two more recent periods. These results were confirmed by continuous descending Fst patterns for the Upper Savio Valley for the whole historic interval considered (results not shown) and suggest that our population was characterized by considerable internal mobility (in particular towards the urban areas).

These results suggest strongly that social-cultural factors gave rise to a reproductive barrier between inhabitants of the chief towns and those of the surrounding areas, despite their sharing the very same environment. Nevertheless, historical changes in SOMs cluster frequencies and Fst show a shift towards a higher degree of surname homogeneity between areas, meaning that the reproductive barrier has been disappearing, especially during the last two periods (i.e. the second half of the 20^{th} century). Unfortunately, our study was not able to discriminate between monophyletic and polyphyletic surnames, as was the case for Manni et al. (2005), but this was expected given the microgeographic setting of this research; regarding this last point, analogous results were obtained for the Alpine isolate Val di Scalve (Boattini et al., 2010a).

The next step of our research was to verify if SOMs results were confirmed by Y-chromosome analyses. The 59 total samples were divided into two groups corresponding to: 29 individuals whose surnames are included in clusters I and II (rural), and 30 individuals whose surnames are included in clusters III and IV (urban). While haplogroup frequencies between the two sub-populations were not significantly different (with the exception of haplogroup G, that was found almost exclusively in the urban sub-population) (Figure 17), Fst calculations based on STR haplotypes revealed a slight but significant differentiation (Fst = 0.022, p = 0.02). This means that these differences lay mainly within haplogroups, as is

clearly demonstrated by a network representation of haplogroup R1b1-P25 (Figure 2), the most widespread in the Upper Savio Valley, to which corresponds Fst = 0.074, p = 0.02. "Urban" haplotypes mostly cluster in the same branch of the network, while "rural" ones form different branches (stemming from the same "urban" haplotype). Summing up, it seems very likely that the two sub-populations evolved from the same ancestral population, a process that – for historical reasons – probably had its origins during the late middle ages.

In conclusion, we can affirm that surname results, as obtained with the SOMs, are confirmed and enhanced by Y-chromosome data. Furthermore, the combined use of cultural markers (surnames) and molecular markers (Y-chromosomes), enabled us to bring to light a 'fossil' reproductive barrier between two different groups of individuals – urban and rural ones – within the same population and environment. The demographic changes that intervened during the studied period and in particular in the second half of the 20^{th} century (increased population mobility, depopulation of the rural areas), caused that barrier to disappear. At a more general level, this study underlines the contribution that surname analysis can bring to molecular anthropology studies and in particular to those aimed at the reconstruction of genetic histories of populations.

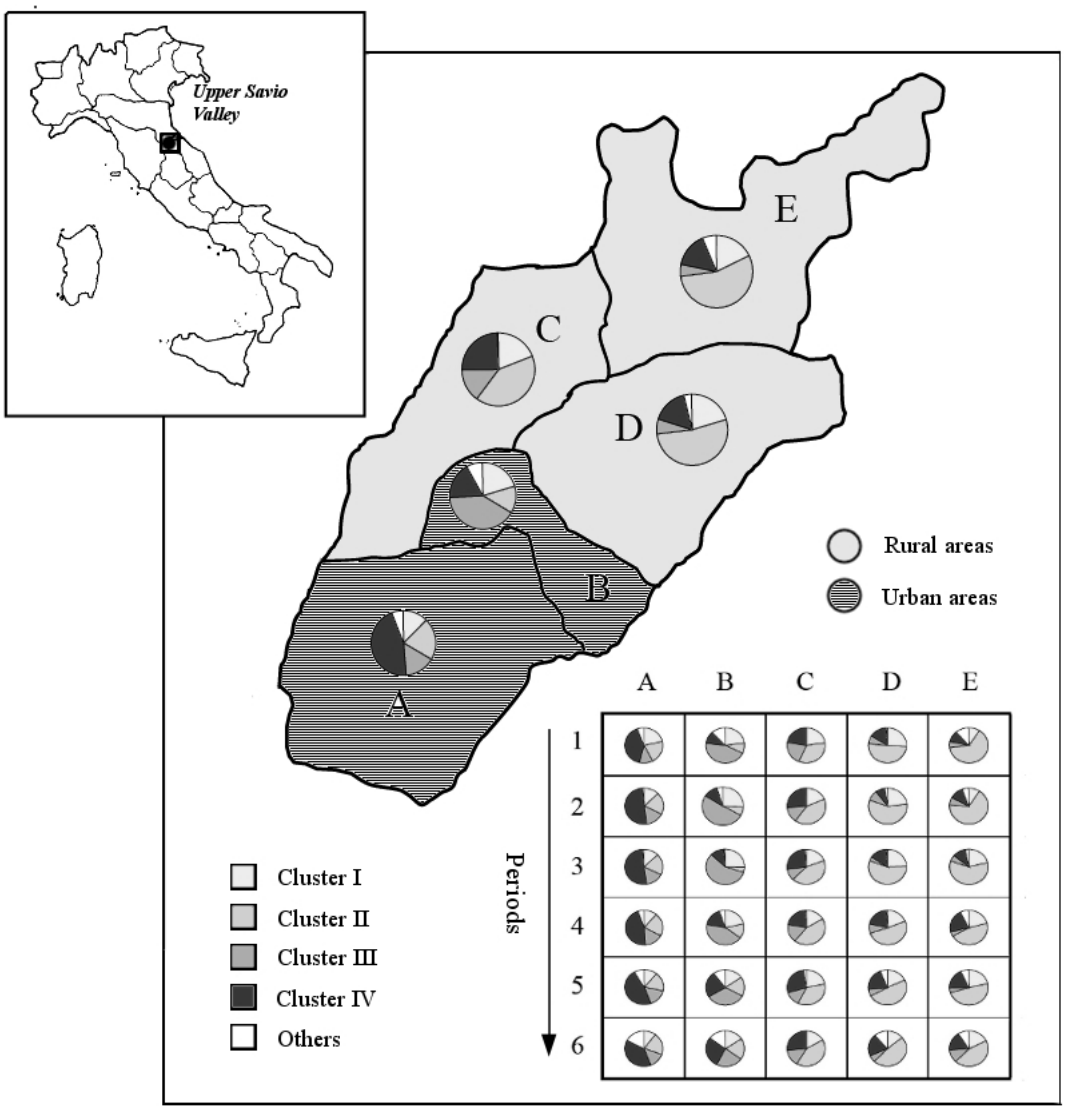

Figure 16. Geographic location and frequencies of the main surname clusters from SOMs with their temporal changes (right, below) in the Upper Savio Valley.

	Frequencies	
Hg	*Rural areas*	*Urban areas*
E1b1b1a-M78	0.07	0.00
E1b1b1b-M81	0.03	0.03
G-M201	0.03	0.23
I*(xI2a2)-M170	0.00	0.07
J*(xJ2)-12f2	0.03	0.00
J2-M172	0.21	0.20
R1*(xR1a,R1b1)-M173	0.03	0.00
R1a1-M17	0.03	0.00
R1b1-P25	0.55	0.43
T-M70	0.00	0.03
Sample size	29	30

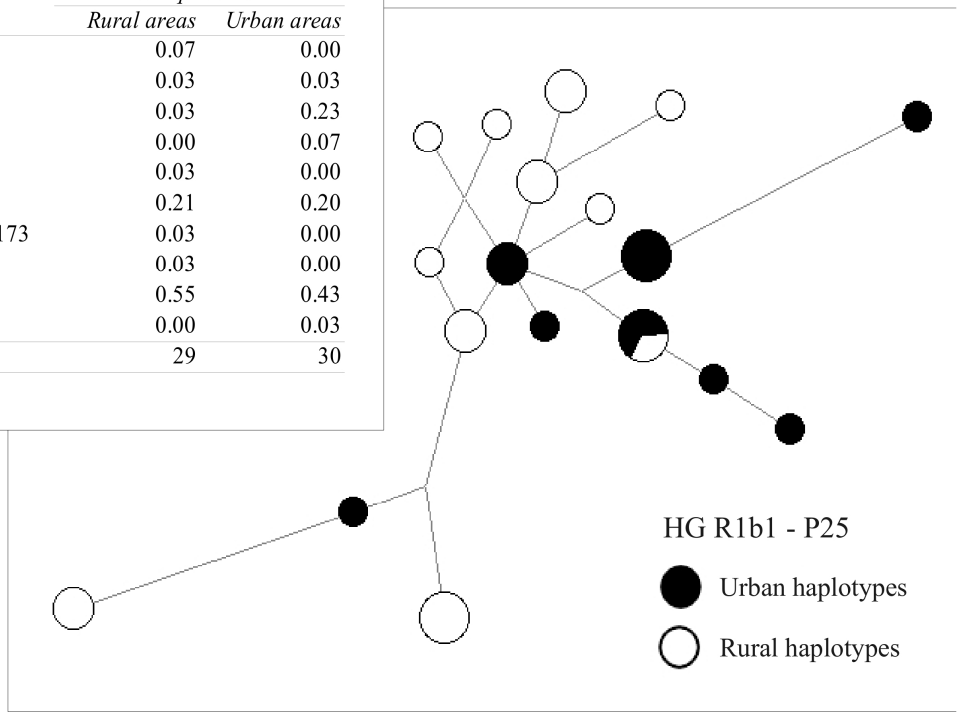

Figure 17. Haplogroup frequencies and network of the R1b1-P25 haplogroup in the rural and urban sub-populations.

7. Cross-disciplinary perspectives in the spatial analysis of names [Pablo Mateos, Paul Longley, James Cheshire]

Supprimé : Title of the contribution of Pablo Mateos et al.

Looking at the recent past, surnames have provided a reliable indicator of population structure (Lasker, 1985), migration flows (Piazza et al 1987, Darlu and Ruffié, 1992), intermarriage (Bugelski, 1961), endogamy and genetic inheritance (Jobling, 2001). Moreover, focusing on disentangling contemporary migration and population diversity, surnames have also proved very useful to classify populations by ethno-cultural origin in health registers (Lakha et al 2011; Lauderdale and Kestebaum 2000), electoral candidates (Dancygier, 2010; Wood et al 2011), inventors (Natham, 2010), and even in social communities on the web such as MySpace and Facebook (Chang et al 2010) or Wikipedia (Ambekar et al 2009). Some of these studies (Bloothooft, this paper) also use forenames as well as surnames frequencies, since the former are also intergenerationally assigned, following cultural norms, social networks and persistent group practices (Alford, 1998). Taken together both forenames and surnames can provide even more refined understandings of population structure and relatedness (Mateos et al 2011).

Supprimé : let's

Moving from the age of migration to the age of DNA research, we will illustrate how such cross-disciplinary understandings of people's names can help to disentangle some of the overlapping and recent episodes that conform the origins of our common past at much finer geographical scales. We will do so through a few examples also presented at the Paris workshop by a team of geographers at University College London (UCL) (Longley, Mateos and Cheshire).

In a recent study on the genetic structure of Great Britain (Winney et al 2011) UCL geographers provided a set of methodologies that were key in enhancing the sampling of thousands of DNA donors as well as the statistical analysis of population sub-structures in

genotyped information. The contributions by geographers were twofold. First, they geocoded (assigning a pair of geographic coordinates) the places of birth of 3865 subjects and those of their parents and grandparents across Great Britain (circa 27000 of places of birth). This then facilitated the task of calculating mean distances between grandparental places of birth, determining distance and population size thresholds to filter out persons with non-local or "urban" grandparents contributing to their DNA makeup. 75% of the sample turned to have a mean distance between grandparental places of birth of 37.3 km, while 70% of grandparental places of birth were classified as "rural". Second, they established whether each subject's surname was actually originated close to the area where his/her grandparents were born. This was mainly successful for monophyletic surnames, or those with no more than two or three distinct areas of origin, but the task of determining such "areas of origin" was eased through the use of a historical population register, as opposed to a contemporary one. The 1881 Census with the names of 29 million respondents were geocoded at a small area level for which relative surname frequency distributions were computed (through a collaboration between geographers and historians). This allowed us to divide the sample between DNA donors with "local" vs "non-local" surnames to seek further genetic patterns. Given the relatively old ages of the donors, the fact that only grandparental places of births were used, and that surname areas of origin were taken from 1881 frequency distributions, the study substantially improved the knowledge about a sample taken in the 2000s by "bringing it back" to a population closer to the late 19th century, regardless of contemporary locations of DNA donors. For detailed results about substrucutres found in Great Britain and some of the historic explanations see Winney et al (2011). Tyler-Smith and Xue (2011) have praised this detailed approach to the geographic, historical and onomastic analysis of genetic data and highlighted its value in improving disease association studies. As they put it, this "*microcosmic survey of genetic variation in a set of small islands off the western coast of the*

Supprimé : that

Supprimé : obviously

Supprimé : worked

Supprimé : by

Supprimé : ing

Supprimé : dividing

Supprimé : subjects

Supprimé : look for

Supprimé : subjects

Supprimé : Xiu

Eurasian continent is revealing the level of differentiation that builds up over millennia via events well documented by archaeology and history, so these alternative data sets can be compared to address questions about the initial peopling of the area, and its subsequent reshaping by internal and external forces" (Tyler-Smith and Xue 2011: 130).

Supprimé :

Supprimé : Xiu

Other more advanced techniques to determine a surname's "core region" (or regions) of origin have been also developed by the UCL Geography team using a probability approach to create "surname surfaces" with which to identify spatial concentrations of surnames and compare them over time (Cheshire and Longley 2012). This is done through Kernel Density Estimation (KDE), conforming a backdrop against which a DNA sampling strategy could be designed. Furthermore, geographical analysis methods can also complement the existing range of approaches to classify continuous space into discrete cultural regions identifying barriers to population interaction through surname frequencies and densities. In this vein two recent studies published by this team were also presented at this workshop using a range of clustering, areal classification and spatial analysis techniques (Cheshire et al 2011, and Longley et al 2011).

Supprimé : s

Supprimé : already

Finally, a different approach to surname origin classification, also presented by the UCL team at the Paris workshop, draws from recent advances in physics to cluster large and complex networks (Mateos et al 2011). Naming networks were constructed linking surnames through the forenames they share in 17 countries at the individual person level drawn from the aforementioned UCL Worldnames database (a network comprising 118 million people, 4.6 million unique surnames and 1.5 million unique forenames, linked through 46.3 million unique forename-surname pairs). Algorithms to search for community structure in very large networks were used to identify clusters of cultural ethnic and linguistic origin. Clusters of surnames were automatically classified and identified by their cultural origin checking them against surname dictionaries to validate the methodology (with results varying between 0.71

to 1 in terms of sensitivity and specificity values). Such an approach permits the automatic classification of large numbers of surnames into clusters of cultural commonality that can then be further analyzed for linguistic, historic or geographical patterns. This methodology automatically identified clusters of these surnames originated outside Europe, which given that most of the data were drawn from Europe shows its value to disentangle recent migration episodes as well as population structure in the non-Western World.

In a final example drawn from the same paper (Mateos et al 2011) the population of a single city; Auckland, New Zealand was also clustered using a naming networks approach (see Figure 18 for a visualization of such a network). It identified clear clusters of Pacific populations that retain intra-marriage practices in contemporary New Zealand, in particular Tongan, Samoan, and to a lesser extent Maori surnames as well as other Pacific Islanders. The fact that this was achieved automatically is particularly striking, especially when there is a lack of knowledge of the Pacific name corpora, their frequencies and origins in the literature. The surname classification methods used in this example, combined with those proposed in Winney et al (2011) could be used to design a much more efficient DNA sampling strategy in Auckland without the expense of conducting research in each of the Pacific Islands, magnified by the fact that there are probably more Samoans and Tongans in Auckland than in those islands. As such, surnames can be used to enhance sampling strategies in urban areas for populations that may have perhaps become diluted compared to their areas of origin as a result of mass migration. This application has already been proposed in the population genetics literature for rural-urban migration in western Europe (Manni et al, 2003). The aforementioned methods brought in by geographers to conventional population genetics research, should be complemented by those presented in this paper by historians, linguists and anthropologists (for example see those led by Darlu, Chareille, Coates and Dräger). Such cross-disciplinary research approaches will be invaluable to improve methods to classify the

Supprimé : the
Supprimé : then
Supprimé : analysed
Supprimé : was
Supprimé : visualisation
Supprimé : done
Supprimé : from
Supprimé : Brunet

ethno-linguistic origin of surnames, pinpoint them to historical areas of origin and trace subsequent waves of migration and specific population dynamics over space. We hope to have made a small contribution to the wealth of cross-disciplinary approaches presented in this paper, and instilled some curiosity amongst researchers in population genetics in enhancing their research findings through the incorporation of methods of spatial analysis of names.

Supprimé : .¶

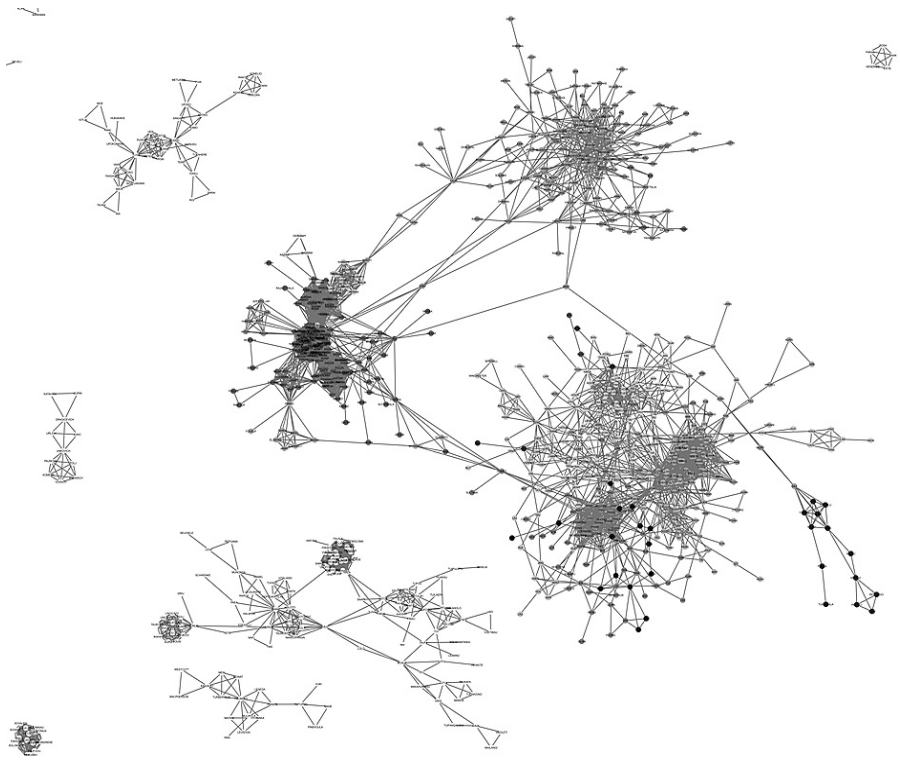

Figure 18: A naming network of Auckland, New Zealand (adapted from Mateos et al 2011).

Supprimé : caption is too short and it is difficult to understand the image without reading the text, please write a self explanatory caption

<subcaption>

The network shows a selection of surnames in the city of Auckland (NZ) using data from the 2006 Electoral Register. Each node represents a unique surname and the links are common forenames shared between a surname pair. The network's topological structure reveals clear clustering in Auckland's naming practices, reflecting closely-knit social networks and ethno-cultural customs that prevents cross-cultural (fore-)naming. For example the bottom third of the figure contains a sub-network of names that are Tongan, Samoan and other Pacific Islanders. The full version of this network including the actual surnames can be visualized in an on-line version available at http://www.onomap.org/naming-networks/fig2.aspx: this Figure can be navigated with full panning and zooming capabilities for flexible exploration. For more details on how the network was built see Mateos et al 2011.

Final remarks

Here, we have provided an overview of some ongoing research about surnames to understand population dynamics in Western Europe and Canada. The rather narrow geographical area addressed here is explained only by the venue of the workshop (Paris). While similar studies are conducted in other regions and continents (see Colantonio et al. 2003; and Mateos *forthcoming*), we think that the examples presented here are representative of the kind of research questions that surnames allow; questions that often go beyond the simplistic use of surnames as a proxy to Y-chromosome diversity. In any case, let us start our discussion from this traditional use of surnames, since this contribution will mainly address a readership of anthropologists and population geneticists that are directly involved in the description of human genetic variation on a world scale. From this perspective, our discussion will then expand to other disciplines and applications highlighting the clear need for increased cross-disciplinary study of population dynamics across space and time in order to better understand human diversity.

Continuous technological improvements have made possible the analysis of very large portions of our DNA. Full genome sequencing will soon become an easy and widespread technique allowing very deep inference about regional and microregional genetic differences that can be explained by demographic factors that, in turn, can rely on historical and cultural processes. Family names of patrilineal descent have proved to mirror a single locus on the Y-chromosome (King and Jobling, 2009). However, they have a temporal depth that is very limited (between 4 and ±30 generations) when compared to the scale of demographic processes inferred with molecular markers, and in any case variations in the Y-chromosome represents an extremely small amount of genetic information. *In this context, why should*

anthropologists take into consideration surname information that, albeit easier to collect than DNA data, is sometimes tricky to interpret, as it is suggested in this summary paper?

The easiest answer is that surnames allow a retrospective look at human variation. They permit comparisons between recent and ancient surname corpora, as historical documents often report surname information over several successive generations, and with a degree of polymorphism that (for the moment) is larger than that available with DNA. *Is this not similar to the scientific interest in ancient DNA technology, which is now being applied to past populations?* Once extant human diversity has been satisfactorily described (and at large geographical scales this is not too far away), one of the major questions will be to explain when and how it arose. If nowadays there are already several clues based on statistical analysis of genetic markers, direct evidence is seldom available, and ancient DNA extraction and typing will remain difficult as the molecule inevitably degrades and appropriate bones cannot be found. This is why Boattini (this paper) most appropriately uses the expression of *archaeogenetic power* to define the interest of surnames in anthropology and biodemography.

Today, in an age of global migration (Castles and Miller, 2009), surnames have indeed the potential to allow researchers an intermediate level of access to the recent past and to small geographical scales that are difficult to obtain otherwise. Interest in surname research is ultimately related to their hereditary character in most societies, but also to their group identity function (Alford, 1998), making them very useful to classify populations according to ancestral proximity. Studies in this area are all based on one simple assumption: the distribution of people's names over space and time is far from random, even in today's highly mobile societies. Therefore, surnames have already proved very useful to provide evidence of migration phenomena in different periods making it possible to identify past genetic isolates and population structures that have been modified or disappeared altogether.

This is where the potential of surnames in population studies goes well beyond the traditional paternal lineage demonstrated in Y-chromosome research.

In order to seize the opportunities lying ahead in such surname studies, more cross-disciplinary research is required that addresses the following key research challenges: a) determine the most probable geographical, temporal and cultural origin of surnames; b) distinguish polyphyletic from monophyletic surnames; c) identify common surname linages in variations of spellings; d) establish finely detailed surname frequency distribution across space and time, e) delineate areas of surname origin and barriers to cultural and population interaction, and f) combine the above advances to tease out the different population episodes that have been overlaid across space and over time. It is obvious that such scientific endeavor will only be possible through the close collaboration with disciplines outside population genetics, as most of this paper's contributions clearly show. We encourage researchers is such cognate fields to participate in the exciting challenge of improve understandings of our shared past through future contributions in Human Biology in this direction.

Acknowledgements:

Pascal Chareille describes a work in progress undertaken jointly with Denise Angers on Normandy and Patrice Beck on the Dijonnais region. He is especially indebted to Pierre Darlu for help with the analysis of the data. Alessio Boattini and Davide Pettener wish to thank the Municipality of Bagno di Romagna and the local section of AVIS (Associazione Italiana Volontari del Sangue) for their kind collaboration. The study was partly funded by the Comunità Montana dell'Appennino Cesenate.

Literature cited

Alford R (1988) *Naming and Identity: A Cross-Cultural Study of Personal Naming Practices*. New Haven, CT: Hraf Press.

Angers, D. and P. Chareille. 2010. Patronymes et migrations en Normandie de la fin du XIVe à la fin du XVe siècle : premiers résultats. In *Anthroponymie et migrations dans la Chrétienté médiévale*, M. Bourin and P. Martinez Sopena, eds. Madrid, Spain: Casa de Velázquez (Colección de la Casa de Velázquez 116), 275-316.

Ambekar A, Ward C, Mohammed J, Male S, Skiena S (2009) Name-ethnicity classification from open sources. Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; June 28– July 1; Paris, France. pp 49–58. Available: http://delivery.acm.org/10.1145/ 1560000/1557032/p49-ambekar.pdf?key1 = 1557032&key2 = 1502083521&- coll = GUIDE&dl = GUIDE&CFID = 53350992&CFTOKEN=96858509 Accessed 2010 Dec 18.

Archer, Steven (2003) *The British 19th-century surname atlas.* CD-ROM. Dartford: Archer Software. [New edition (2011).]

Beck, P., and P. Chareille. 1997. Espaces migratoires et aire d'influence de la ville de Dijon à la fin du XIVe siècle. *Cahiers de Recherches Médiévales (XIIIe-XVe s.)* 3:17-32.

Beck, P., and P. Chareille. 1998. Sédentarité et mobilité à Dijon à la fin du XIVe siècle. In *La ville au moyen Âge*, vol. 2: *Sociétés et pouvoirs dans la ville*, N. Coulet and O. Guyotjeannin, eds. Paris, France: Éditions du CTHS, 95-104.

Black, George F. (1946) *The surnames of Scotland: their origin, meaning and history.* New York: New York Public Library. [Reprinted Edinburgh: Birlinn (1993).]

Bloothooft et al. (this article)

Bloothooft, G. (2011), 'Linguistics and geography, the surname case', in: W. Zonneveld, H. Quené, and W. Heeren (Eds.), *Sound and Sounds, studies presented to M.E.H. (Bert) Schouten*, Utrecht, UiL-OTS, 9-20.

Bloothooft, G. and D. Onland (2011), 'Socioeconomic determinants of first names', Names 59:1, 25-41.

Bloothooft, G. and L. Groot (2008), 'Name clustering on the basis of parental preferences', Names 56:3, 111-163

Boattini et al. (this article)

Boattini A, Griso C, Pettener D. 2010b. Are ethnic minorities synonymous for genetic isolates? Comparing Walser and Romance populations in the Upper Lys Valley (Western Alps). J Anthropol Sci, 89:161-173.

Boattini A, Luiselli D, Sazzini M, Useli A, Tagarelli G, Pettener D. 2011. Linking Italy and the Balkans. A Y-chromosome perspective form the Arbereshe of Calabria. Ann Hum Biol, 38:59-68.

Boattini A, Pedrosi ME, Luiselli D, Pettener D. 2010. Dissecting a human isolate: Novel sampling criteria for analysis of the genetic structure of the Val di Scalve (Italian Pre-Alps). Ann Hum Biol, 37:604-609.

Boattini A, Pedrosi ME, Luiselli D, Pettener D. 2010a. Dissecting a human isolate: Novel sampling criteria for analysis of the genetic structure of the Val di Scalve (Italian Pre-Alps). Ann Hum Biol, 37:604-609.

Boattini A, Pettener D. 2005. Tra crinali e confini: mobilità matrimoniale e barriere riproduttive in Romagna Toscana (Bagno di Romagna, 1572-1930), in Breschi M, Fornasin A (editors), Il matrimonio in situazioni estreme: isole e isolati demografici, Forum, Udine, pp. 127-142.

Bourin, M., and P. Martínez Sopena, eds. 2010. *Anthroponymie et déplacements dans la Chrétienté médiévale*. Madrid, Spain: Casa de Velázquez (Collection de la Casa de Velázquez, 116).

Bowden GR, Balaresque P, King TE, Hansen Z, Lee AC, Pergl-Wilson G, Hurley E, Roberts SJ, Waite P, Jesch J, Jones AL, Thomas MG, Harding SE, Jobling MA. 2008. Excavating past population structures by surname-based sampling: the genetic legacy of the Vikings in northwest England. Mol Biol Evol. 25:301-9.

Bugelski BR (1961) Assimilation through intermarriage. *Social Forces* 40: 148

Castles, S and Miller, M (2009) *The Age of Migration*, Palgrave Macmillan: London

Cavalli Sforza L.L., Menozzi P., Piazza A. 1994 The History and geography of human genes. Princeton University Press, Princeton, New Jersey, USA.

Cavalli-Sforza L.L., Moroni A. and G. Zei 2004 Consanguinity, Inbreeding and Genetic Dfift in Italy. Princeton University Press, Princeton, New Jersey, USA, pages 90-148.

Chang J, Rosenn I, Backstrom L, Marlow C (2010) ePluribus: Ethnicity on Social Networks. Procceedings of the 4th International AAAI Conference on Weblogs and Social Media23–26 May; Washington. Association for the Advancement of Artificial Intelligence (AAAI). pp 18–25. Available: http:// www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/viewFile/1534/ 1828 Accessed 2011 Feb 03.

Cieślikowa, Aleksandra, ed. (2007-) *Antroponimia polski od XVI do XVIII wieku* [Polish anthroponymy from the XVI to the XVIIIth century], vol. 1 (A-G); vol. 2 (H-Mą). Kraków: Lexis (2007 and 2009 respectively). [Vol. 3 is at press and vol. 4 in preparation.]

Cottle, Basil (1967) *The Penguin dictionary of surnames.* Harmondsworth: Penguin. [3rd, fully revised, edn by John Titford (2009).]

Mis en forme : Anglais (Royaume-Uni)

Chareille (this volume)

Chareille P., and P. Darlu, 2010. Anthroponymie et migration: quelques outils d'analyse et leur application à l'étude des déplacements dans les domaines de Saint-Germain-des-Près au IXe siècle. In *Anthroponymie et migrations dans la chrétienté médiévale*. M. Bourin and P. Martinez Sopena, eds. Madrid, Spain: Casa de Velázquez (Collection de la Casa de Velázquez 116), 41-73.

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

Cheshire, J.A., P. Mateos, P.A. Longley (2011). 'Delineating Europe's Cultural Regions: Population Structure and Surname Clustering', Human Biology 83(5):573-598

Cheshire, JA and Longley, PA (2012) Identifying spatial concentrations of surnames, *International Journal of Geographical Information Science*, 26 (2) 309-325

Dancygier, R.M. (2010) *Immigration and Conflict in Europe*, Cambridge University Press: New York.

Dammel, Antje, and Mirjam Schmuck. 2008. Der Deutsche Familiennamenatlas (DFA): Relevanz computergestützter Familiennamengeographie für die Dialektgeographie. In: Elspaß, Stephan, and Werner König (eds.), *Sprachgeographie digital: die neue Generation der Sprachatlanten*, 73–104; 254–260. Hildesheim et al.: Olms.

Darlu P. and Ruffié J., 1992. L'immigration dans les départements français étudiée par la méthode des patronymes. *Population*, 3 : 719-734.

Darlu P., Brunet G., Barbero D., 2011. Spatial and temporal analyses of surname distributions to estimate mobility and changes in historical demography: the example of Savoy (France) from the XVIIIth to XXth century. In: *Navigating Time and space in Population studies*.Gutmann, M.P.; Deane, G.D.; Merchant, E.R.; Sylvester, K.M.

(Eds.)Series. International Studies in Population, vol 9, Springer, 1st Edition, 2011, XII, 245 p.

Darlu, P. and A. Degioanni. 2007. Localisation de l'origine géographique de migrants par la méthode patronymique: exemple de quelques villes de France au début du XXème siècle. *Espace géographique* 3: 251-265.

Darlu, P., Degioanni, A., Ruffié, J. 1997. Quelques statistiques sur la distribution des patronymes en France. *Population*, 3:607-634.

de Bhulbh, Seán (*c.* 1997) *Sloinnte na h-Éireann = Irish surnames.* Faing, Co. Luimnigh: Comhar-Chumann Íde Naofa. [2nd edn titled *Sloinnte uile Éireann = All Ireland surnames* (2002).]

De Felice E. 1978 Dizionario dei cognomi italiani, Mondadori, Milano, Italy.

de Woulfe, Patrick (1906) *Sloinnte Gaedheal is Gall = Irish names and surnames.* Dublin: M. H. Gill [2nd edn (1922/3).]

Degioanni, A. and P. Darlu. 2001. A Bayesian approach to infer geographical origins of migrants through surnames. *Annals of Human Biology* 28: 537-545.

Dräger, Kathrin, and Mirjam Schmuck. 2009. The German Surname Atlas Project - Computer-based surname geography. In: Ahrens, Wolfgang, Embleton, Sheila and Lapierre, André (eds.): *Names in multi-lingual, multi-cultural and multi-ethnic contact. Proceedings of the 23rd International Congress of Onomastic Sciences, August 17-22, 2008, York University, Toronto, Canada*. [CD-Rom].

Emery, R. 1952. The use of the surname in the study of medieval economic history. *Medievalia et Humanistica* 7:43-50.

Emery, R. 1955. A further note on medieval surnames. *Medievalia et Humanistica* 9:104-106.

Felsenstein, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, 39:783-791.

Grünert, Horst. 1958, *Die Altenburgischen Personennamen. Ein Beitrag zur mitteldeutschen Namenforschung.* Tübingen: Niemeyer.

Hanks, Patrick, and Flavia Hodges, eds (1988) *A dictionary of surnames.* Oxford: Oxford University Press.

Hanks, Patrick, ed. (1990) *Dictionary of American family names*, 3 vols. Oxford: Oxford University Press.

Hanks, Patrick, Peter McClure, and Richard Coates (forthcoming 2012) Family Names of the United Kingdom: a new research project in British anthroponomastics. *Proceedings of the 24th International Congress of Onomastic Sciences, Barcelona, Spain, 5-9 September 2011*.

Hellfritzsch, Volkmar. 2007, *Personennamen Südwestsachsens. Die Personennamen der Städte Zwickau und Chemnitz bis zum Jahre 1500 und ihre sprachgeschichtliche Bedeutung*. Leipzig: Leipziger Universitätsverlag.

Hey, David G. (2000) *Family names and family history.* London: Hambledon and London.

ICOS = *Proceedings* of the International Congresses of Onomastic Sciences.

INSEE (1985) *Registre français des noms patronymiques*

Jobling MA (2001) In the name of the father: surnames and genetics. *Trends in Genetics* 17: 353–357.

Karlin, S. and J. McGregor (1967). "The number of mutant forms maintained in a population." Proceedings of the 5th Berkeley Symposium on Mathematics, Statistics,and Probability 4: 415-438.

Kedar, B. 1973. Toponymic surnames as evidence of origin : Some Medieval Views. *Viator* 4:123-129.

King TE, Jobling MA. 2009. What's in a name? Y chromosomes and the genetic genealogy revolution. Trends Genet., 25:351-360.

Kohonen T. 1982. Self-organized formation of topologically correct feature maps. Biol Cybern 43:59-69.

Kohonen T. 1984. Self-organization and associative memory. Berlin: Springer.

Kunze, Konrad, and Damaris Nübling. 2007. Der Deutsche Familiennamenatlas (DFA): Konzept, Konturen, Kartenbeispiele. *Beiträge zur Namenforschung (N.F.)* 42/2, 125–172.

Kunze, Konrad. 2004. *dtv-Atlas Namenkunde: Vor- und Familiennamen im deutschen Sprachgebiet*. 4th edition. München: dtv.

Kunze, Konrad; Damaris Nübling (eds.). Deutscher Familiennamenatlas. Berlin, New York: de Gruyter. Bd. 1: Graphematik/Phonologie der Familiennamen I: Vokalismus. Von Christian Bochenek, Kathrin Dräger (2009). Bd. 2: Graphematik/Phonologie der Familiennamen II: Konsonantismus. Von Antje Dammel, Kathrin Dräger, Rita Heuser, Mirjam Schmuck (2011).

Lakha, F., Gorman D, Mateos, P. (2011) Name analysis to classify populations by ethnicity in public health: Validation of Onomap in Scotland, *Public Health* 125 (10) 688-696

Lauderdale, D.S. & Kestenbaum, B., 2000. Asian American ethnic identification by surname. *Population Research and Policy Review*, 19(3) 283-300

Longley, PA and Cheshire, JA and Mateos, P (2011) Creating a regional geography of Britain through the spatial analysis of surnames. *Geoforum* , 42 (4) 506 – 516

MacLysaght, Edward (1985) *The surnames of Ireland*, 6th edn. Blackrock, Co. Dublin: Irish Academic Press.

Mandemakers, K. (2000), 'The Netherlands. Historical Sample of the Netherlands', in: P. Kelly Hall, R. McCaa & G. Thorvaldsen (ed.), *Handbook of International Historical*

Microdata for Population Research (Minnesota Population Center Minneapolis 2000), 149-177.

Manni F, Heeringa W, Toupance B, Nerbonne J. 2008. Do surname differences mirror dialect variation? Hum Biol 81:41-64.

Manni F, Toupance B, Sabbagh A, Heyer EDA-F. (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(2):214-228

Manni F., Toupance B., Sabbagh A., Heyer E. 2005 A new method for surname studies of ancient patrilinear population structures, and possible application to improvement of Y-chromosome sampling. Am. J. Phys. Anthropol., 126:214-228

Mateos, P. (forthcoming) *Ethnicity, geography and populations: Tracing diversity and migration through people's names*, Springer: Heidelberg

Mateos, P. (2007) A review of name-based ethnicity classification methods and their potential in population studies, *Population Space and Place*, 13 (4): 243-263.

Mateos, P., Longley, P.A. and O'Sullivan, D. (2011) Ethnicity and Population Structure in Personal Naming Networks. *PloS ONE* 6 (9) e22943

Mateos, P., R. Webber, P. Longley (2007). 'The Cultural, Ethnic, and Linguistic Classification of Populations and Neighborhoods Using Personal Names', CASA working paper 116, UCL, London http://discovery.ucl.ac.uk/3472/

McKinley, Richard A. (1975) *The surnames of Norfolk & Suffolk.* Oxford: Leopard's Head Press.

McKinley, Richard A. (1977) *The surnames of Oxfordshire.* Oxford: Leopard's Head Press.

McKinley, Richard A. (1981) *The surnames of Lancashire.* Oxford: Leopard's Head Press.

McKinley, Richard A. (1988) *The surnames of Sussex.* Oxford: Leopard's Head Press.

McKinley, Richard A. (1990) *A history of British surnames.* Harlow: Longman.

McKinley, Richard A., George Redmonds, and David Postles (1973-98) A series of county-based surname studies. Oxford: Leopard's Head.

Morgan, T.J., and Prys Morgan (1985) *Welsh surnames.* Cardiff: University of Wales Press.

Nathan, M. (2011) The economics of super-diversity: findings from British cities, 2001-2006. SERC Discussion Paper Series, SERCDP0068, London School of Economics: London.

Nei, M. 1973. Genetic distance between populations. *American Naturalist* 106: 283-292.

Neumann, Isolde. 1970. *Die bäuerlichen Familiennamen des Landkreises Oschatz.* Berlin: Akademie-Verlag.

Neumann, Isolde. 1981. *Die Familiennamen der Stadtbewohner in den Kreisen Oschatz, Riesa und Großenhain bis 1600*. Berlin: Akademie-Verlag.

Nicholls, Kenneth, ed. (1994) *The Irish Fiants of the Tudor Sovereign during the reigns of Henry VIII, Edward VI, Philip and Mary and Elizabeth I*, 4 vols. Dublin: Burke.

Nübling, Damaris, and Konrad Kunze. 2005. Familiennamenforschung morgen: der deutsche Familiennamenatlas (DFA). In: Brendler, Andrea; Brendler, Silvio (eds.), *Namenforschung morgen: Ideen, Perspektiven, Visionen*, 141–151. Hamburg: Baar.

Nübling, Damaris, and Konrad Kunze. 2006. New perspectives on Müller, Meyer, Schmidt: computer-based surname geography and the German Surname Atlas project. *Studia anthroponymica scandinavica* 24, 53–85.

Piazza A, Rendine S, Zei G, Moroni A, Cavalli-Sforza LL (1987) Migration rates of human populations from surname distribution. *Nature* 329: 714–716.

Postles, David (1995) *The surnames of Devon.* Oxford: Leopard's Head Press.

Postles, David (1998) *The surnames of Leicestershire & Rutland.* Oxford: Leopard's Head Press.

Reaney, Percy H. (1958, 1976) *A dictionary of British surnames.* London: Routledge and Kegan Paul. [Third edn by Reaney and Richard M. Wilson, *A dictionary of English surnames* (1991). R&W.]

Redmonds, George (1973) *The surnames of Yorkshire, West Riding.* Oxford: Leopard's Head Press.

Redmonds, George (2002) *Surnames and genealogy: a new approach.* Bury: Federation of Family History Societies.

Rodriguez Diaz R, Blanco Villegas MJ. 2010. Genetic structure of a rural region in Spain: distribution of surnames and gene flow. Hum Biol. 82:301-314.

Rohlfs G. 1997 Studio e ricerche su lingue e dialetti di Italia, Sansoni Editore, Fireze, Italia.

Saitou, N. and Nei, M. 1987. The Neighbor-Joining Method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4): 406-425.

Schmuck, Mirjam. 2009. Personennamen als Quelle der Grammatikalisierung. Der *ing*-Diminutiv in Mecklenburg-Vorpommern, *Beiträge zur Namenforschung (N.F.)* 44, 35-65.

Tooth, Edgar (2000) *The distinctive surnames of North Staffordshire*, 2 vols. Leek: Churnet Valley Books.

Tyler-Smith, C. & Xue, Y., 2012. A British approach to sampling. European journal of human genetics *European Journal of Human Genetics*, 20(2), p.129-130.

Walther, Hans. 1993. *Zur Namenkunde und Siedlungsgeschichte Sachsens und Thüringens. Ausgewählte Beiträge 1953-1991*. Leipzig: Reprint-Verlag.

Supprimé :

Wijsman, E., G. Zei, Moroni A., Cavalli-Sforza L.L.. (1984). "Surnames in Sardinia. II. Computation of migration matrices from surname distribution in different periods." Annals of Human Genetics 48: 65-78.

Winney B, Boumertit A, Day T, Davison D, Echeta C, Evseeva I, Hutnik K, Leslie S, Nicodemus K, Royrvik EC, Tonks S, Yang X, Cheshire J, Longley P, Mateos P, Groom A, Relton C, Bishop DT, Black K, Northwood E, Parkinson L, Frayling TM, Steele A, Sampson JR, King T, Dixon R, Middleton D, Jennings B, Bowden R, Donnelly P, Bodmer W.(2012). People of the British Isles: preliminary analysis of genotypes and surnames in a UK-control population. *European Journal of Human Genetics*, 20(2), p.203-10.

Wood, J.; Badawood, D.; Dykes, J.; Slingsby, A.(2011) BallotMaps: Detecting Name Bias in Alphabetically Ordered Ballot Papers, *IEEE Transactions on Visualization and Computer Graphics*, 17 (12) 2384 - 2391 doi: 10.1109/TVCG.2011.174

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

Zei G., Barbujani G., Lisa A., Fiorani O., Menozzi P., Siri E. and L. Cavalli-Sforza 1993 Barrier to gene flow estimated by surname distribution in Italy. Ann Hum Genet, 57:123-140.

Zei G., Lisa A., Fiorani O., Magri C., Quintana-Murci L., Semino O. and S. Santachiara-Benerecetti 2003 From surnames to History of Y-chromosome: the Sardinian population as a paradigm. Europ J Hum Genet, 11(10):802-7.

Zei, G., R. Guglielmino, et al. (1983). "Surname in Sardinia. I. Fit od frequency istributions for neutral alleles and genetic population structure." *Annals of Human Genetics* **47**: 329-352