# Exploring co-variation in the (historical) Dutch civil registration

*Gerrit Bloothooft, Kees Mandemakers*

DOI: 10.2436/15.8040.01.32

## Abstract

Civil registration (CR) contains a wealth of information on inhabitants of a country; their names, their dates-, places- and countries of birth, marriage and decease, with links to partners and children. From the onomastic point of view this is one of the best sources one can have for very many research questions. Availability of the data from civil registration is a problem, however, for privacy reasons for modern CR, and for lack of digitized data from historic CR.

In the Netherlands, much progress has been made in recent years: full population selections (for 16 million inhabitants) from CR were made available for scientific research on the basis of a new law on CR, separately for first names and surnames. In 2010, websites were launched, based on these data: the Dutch Corpus of First Names (www.meertens.knaw.nl/nvb) with 500,000 names and the Dutch Family Names Corpus (www.meertens.knaw/nl/nfb) with 314,000 names. The sites show first name popularity per gender from 1880 onwards, family name figures for 1947 and 2007, name distribution maps at the municipality level, name explanations (partly) and other documentation.

For historical CR, certificates of birth, marriage and death from 1811 onwards are being indexed and digitized by hundreds of volunteers, half of the job being currently done (16 million certificates). A project on record linkage aims to reconstruct families from these data to build a historic CR.

A typical property of a source like CR is the information network it consists of. The relations among people are explicitly present, as partners, as a family, as inhabitants of some village, as born in some year, and so on. This opens new ways for onomastic research based on co-variation. The names of children in the same family inform us on specific parental preferences, the spatial distribution of names on regional influences in naming and migration, the first names in different generations in a family on dynamics in naming over time, family reconstruction processes (needed to create historical CR) can learn us about spelling and name variation for the same person, and so on. All these analyses start at the very detailed personal level, but can be aggregated to demonstrate societal processes. Several examples of this kind of studies are presented.

*****

## I. Introduction

Names typically 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, also generations come into play, carrying the dimension of time and historical developments in society. The dimension of space equally influences groups: members migrate and interact. For studies of, among other things, genetics, health, demography and sociology, the identification of groups and knowledge of their dispersion in time and space is valuable if not essential information.

Identifying individuals need not to be difficult in contemporary digital systems of civil registration, if access to this information is not severely restricted for privacy reasons. For several countries, telephone directories may provide a significant sample and a useful snapshot, but family relations among people (including generations) remain unknown. For older registrations the privacy limitation does not hold, but (digital) availability and data quality constitute a serious issue.

In Dutch and other modern civil registrations, people are identified not only by name but also by a persistent ID. By having 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 not essential anymore to demonstrate

relations between people. However, for older registrations, no IDs were used, and reconstruction of relations between people highly depends on their names and the description of roles in certificates of birth, marriage and decease. The accuracy of these archives is often problematic, completeness rare, and full digitization a long term goal only.

This paper reviews the current status of availability of data from modern and historical civil registration in The Netherlands. The role of names in projects is discussed and some analytic approaches in the studies of names – under the availability of full population data – are demonstrated.

## II. Available data and major ongoing projects in The Netherlands

### *II.1. Modern Civil Registration*

In 2000, a new law on 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 focused on first names, and another on family names. These full population selections were provided in 2006 (update on 2011-01-01) and 2007, respectively.

#### II.1.a. FIRST NAMES

Full population data were acquired for all first names of 16 million persons alive in 2006 and of 1.5 million persons deceased since the digitization of CR in 1994, and of 3.5 million persons deceased before 1994 but who were mentioned as parents in the records of their children. The latter required an extensive reconstruction process since a parent could appear in the records of several children, while especially the deceased parent's information – not essential anymore for the current administration – contained serious numbers of errors. Besides all first names, also the (internal) ID, the first names and IDs of the parents, the date, place and country of birth of all, were provided. This basically constitutes a full population genealogy for several generations – but with only the first name known. The data are largely complete from 1930 onwards, but still provide a 30% sample in 1880. All in all, these 21 million persons entailed 500,000 unique first names which were made public in June 2010 on www.meertens.knaw.nl/nvb. The website provides for each first name the number of namesakes alive in 2006. Information on the first names is differentiated according to first or subsequent name positions and gender. The names are presented by way of a distribution per year from 1880 until now, which shows the popularity of names through the ages, a geographic distribution of places of birth of namesakes alive in 2006 (468 municipalities), and an etymological description of the name (limited to 20,000 names only, but many variants still to be linked), see an example screenshot figure 1. All presentations are available in absolute and relative figures.

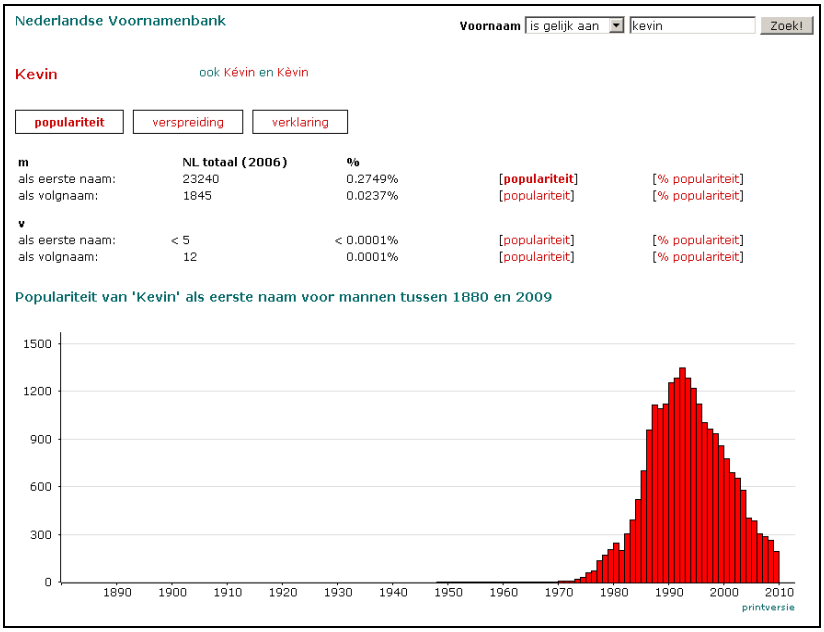
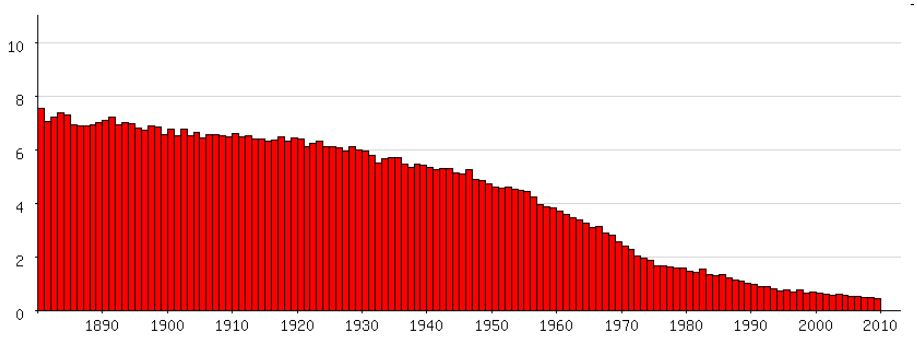
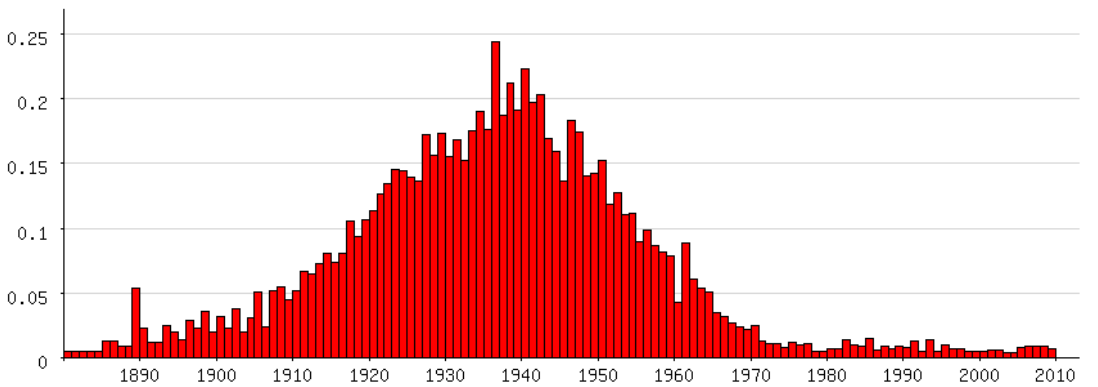



**Figure 1**. Screenshot of the website of Dutch first names, showing the popularity of Kevin.

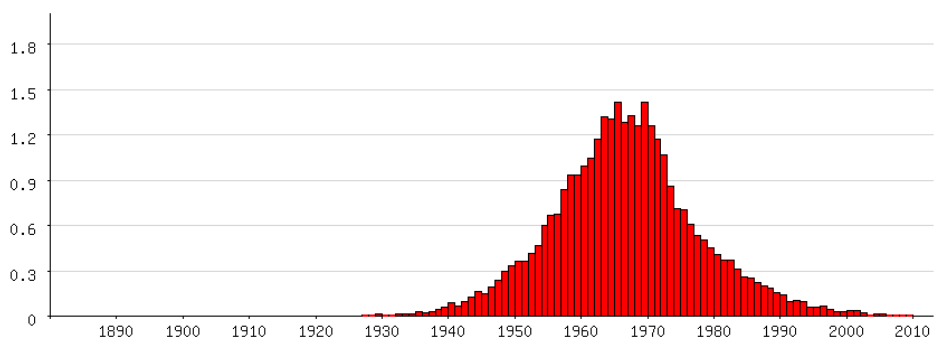
The popularity of names over time demonstrates the decline of naming after grandparents (with traditional names, like *Jan* and *Maria*), and the rise of fashion names in the twentieth century (see figure 2). The first names with fashion properties were traditional names but in an abbreviated form, already known for a long time but never adopted or accepted officially (*Nelly* from *Cornelia*). In the 1940s really new names (new for The Netherlands) became popular (*Ingrid*). Followed by numerous others. There is a tendency for the life time of fashion names to become shorter and shorter (*Jayden*).



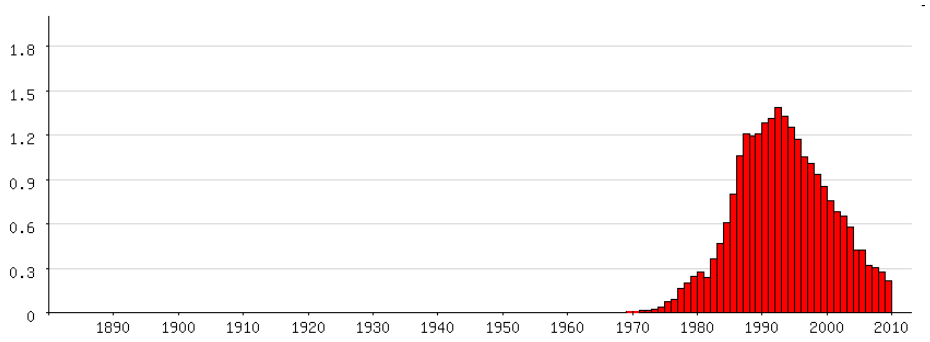

Decline of traditional naming: *Jan*



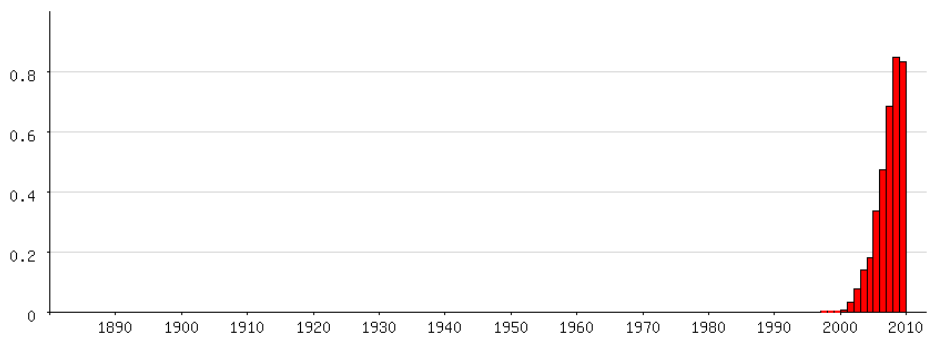

Abbreviated calling form of a traditional name: *Nelly*




New names for The Netherlands: *Ingrid*




The rise and decline of English names: *Kevin*




Short hype names: *Jayden*

**Figure 2**. Typical Dutch popularity distributions over the twentieth century, with the annual percentage names given on the vertical axis. 1% in The Netherlands roughly constitutes 1000 children per gender. Where in the past, Jan accounted for 7% of the boys (and Johannes and Maria even much more), today the most popular name hardly exceeds 1%.

### II.1.b. FAMILY NAMES

For 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 living. These data were linked to the data from the 1947 census, which is available in digital format on the aggregation level of the province (number of persons with the family name); the figures for municipality of living in 1947 are available but not digitized. The 16 million persons proved to be named with 314,000 unique surnames. The website presenting the surnames was launched in December 2009 on www.meertens.knaw.nl/nfb. It provides the number of persons with the target name in 1947 and in 2007, the geographic distribution in 1947 at the province level and in 2007 at the municipality level (468 municipalities), and documentation references for several thousands of names, see figure 3. Orthographic and onomastic relations between names are presented in a hierarchical network structure with the most frequent name on top.

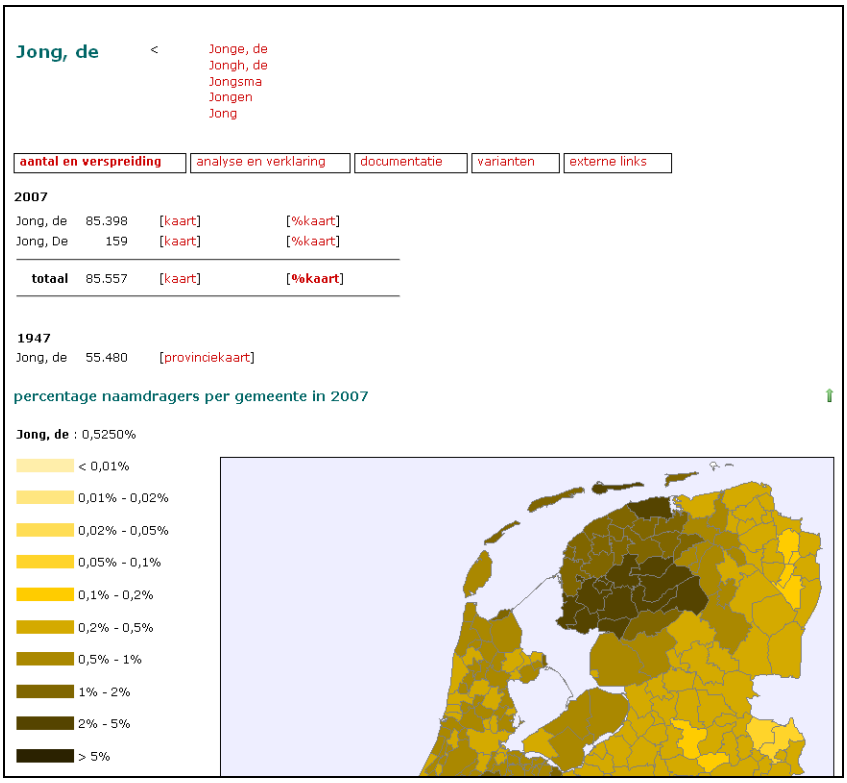



**Figure 3**. Partial screenshot of the website of Dutch surnames, showing information on the name de Jong, and a map of the percentage per municipality in the Northern part of the country.

Both websites reached 15 million page views in the first month after launching, and a stableone million page views per month afterwards, showing a very high public interest.

## *II.2. Historic Civil Registration*

From the early nineties of the 20th century, in The Netherlands hundreds of volunteers have been working on digitization of all names and dates from the historical registers of birth, marriage and decease. The civil registration system started in 1811 and was based on Napoleonic law. Registers are public, with a delay, today are available: registers of birth until 1909, marriage registers until 1934 and registers of decease until 1959. All digitized data are publicly accessible through www.wiewaswie.nl. At present, about half of the job is done,

over 16 million registers having been digitized, containing information on about 70 million (not unique) persons.

Automatic reconstruction of families from these data is now in progress in the LINKS project (*Link*ing *s*ystem for historical family reconstruction). The project started in 2009 and is based at the International Institute of Social History in cooperation with Utrecht University, the Meertens Institute and the Leiden Institute of Advanced Computing. For more information about the LINKS project, see http://www.iisg.nl/hsn/news/links-project.php

Ideally, the goal of LINKS is to identify uniquely all individuals mentioned in the certificates, and, just like modern CR, to tag them with a persistent ID and the IDs of their parents. By using the same techniques, in theory it will be possible to link our historical 'population registration' with the current one. However, for reasons of privacy protection we do not foresee such a match in the near feature. In the LINKS project we systematically explore techniques for automatic record linkage under the condition of incomplete and partly inaccurate data. Geographic, onomastic and phonetic-linguistic knowledge in combination with fuzzy learning techniques will be applied in the reconstruction process.

### *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 eventually 78.000 persons (born 1812-1922) on a manual basis. The research persons were sampled from the birth certificates. In addition to standard personal data, also religious affiliation, occupation, household composition, literacy, social network, and migration history are collected from the civil certificates and population registers. By providing this representative dataset the HSN not only supports research with micro-data into social developments in the 19th and 20th centuries, but also a) provides a control group or groups which researchers can compare with their own research population, b) develops the expertise which individual researchers usually cannot acquire in the limited time at their disposal and c) offers the possibility for researchers to use the existing HSN dataset as a basis for their own research projects (Mandemakers 2001)

Collecting new data is realized by taking the database as a starting point for further research, both through increasing the number of individuals included (oversampling) and by deepening by means of recording supplementary variables for a specific group of research subjects. Scholars thus kill two birds with one stone. Not only can they use the data already recorded, the software and expertise developed by the HSN are available as well. This expertise can also be considered an important byproduct of the data entering of the past ten years. For using its software and already recorded data, the HSN sets the precondition that new data must be added to the data set, so that they will eventually become available to other researchers too. Over eighteen projects have been realized now collecting additional information, especially in the fields of migration, both inbound and outbound (East Indies). By way of oversampling another 20,000 research persons are added to the HNS-database.

More information can be found on www.iisg.nl/hsn. The HSN supported a wide range of investigations with hundreds of publications in the areas of historical demography, history of the family, and historical sociology.

## III. Data mining, tools and examples

### *III.1. Geographic distribution*

Ideally, families can be reconstructed with fair accuracy from 1811 onwards. Before that year, one has to rely on parish registers and other sources, and reconstruction – though not impossible – becomes difficult. Since more than 70% of the population already had a family name in 1811, for many families the founder of the family (who started a hereditary surname) lived much earlier. He may have had many descendants in 1811 with unknown mutual relations, but usually with a common surname (including spelling variations). These family branches are then still genetically related. Although the size (today and in 1811) and geographic distribution of the family can give an indication whether identified branches have a common founder, this is not guaranteed. A surname may have been come into existence independently in different places. In that case there is no consanguinity among families with the same surname. This especially will be the case for patronymics, occupational names and provenance names.

The current geographic distribution 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 for search by *regular expression*, properties of all kinds of *sets* of surnames can be shown as well, see the example in figure 4. 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.

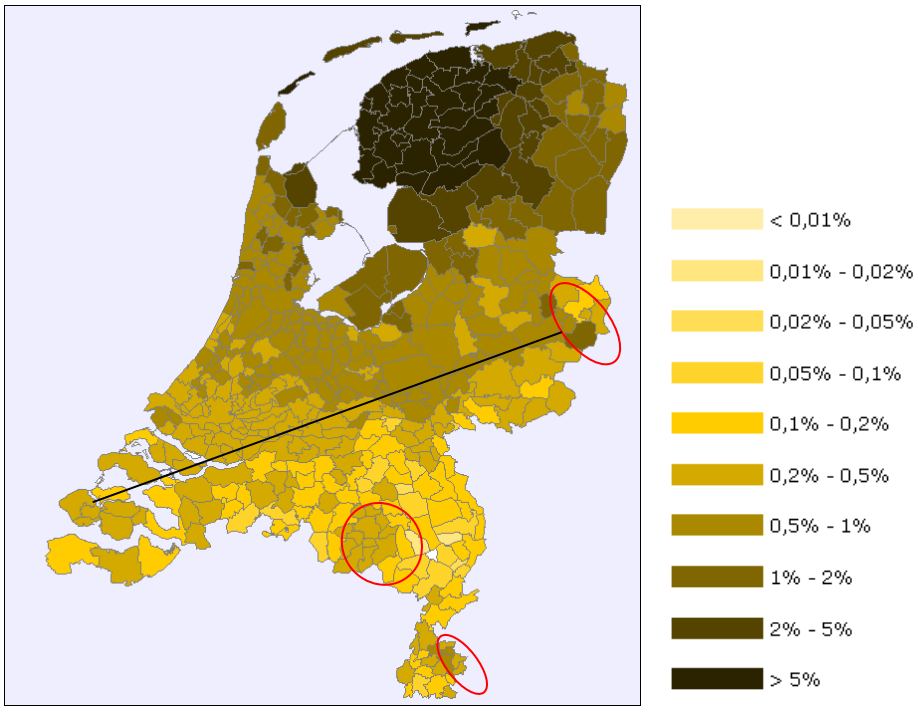



**Figure 4**. Geographic distribution of all surnames that fulfil the regular expression ‘stra$’, implying names ending in –stra, in percentage per municipality. This is a typical Frisian name ending, expressing ‘coming from’. The map shows the high presence in the province of Friesland, the circular shape of the decrease of the presence of the name in the North, a relative sharp boundary with the Catholic south of the country (below black line) - with exceptions in areas of industrial development (in the coal mines of Limburg in the south, around Eindhoven (Philips), and the textile factories in the eastern part; red ovals).

Reverse analyses, in which we seek surnames with a common geographic distribution can be done on the available (modern) data but have not been studied yet.

### *III. 2. Migration*

Once a full reconstruction of the Dutch population from 1811 onwards would become available, migration studies can be performed easily by tracing the places of birth of subsequent generations. This could be done for a family but also for the inhabitants of some village or town.

We performed such an analysis on the basis of our first name corpus from modern civil registration. We did not use the first names themselves, but the information on place and year of birth of a person, and the ID of the parents for the intergenerational links. Starting with all present inhabitants of some municipality with an age between 30 and 50 years, we mapped their places of birth, and the places of birth of their parents, and of their (great) grandparents. Also, it is possible to start with all persons born between 1880 en 1900 in some municipality and map the places of birth of their children, and grandchildren. An interactive online application (see www.meertens.knaw.nl/migmap) has all these possibilities, see an example in figure 5.

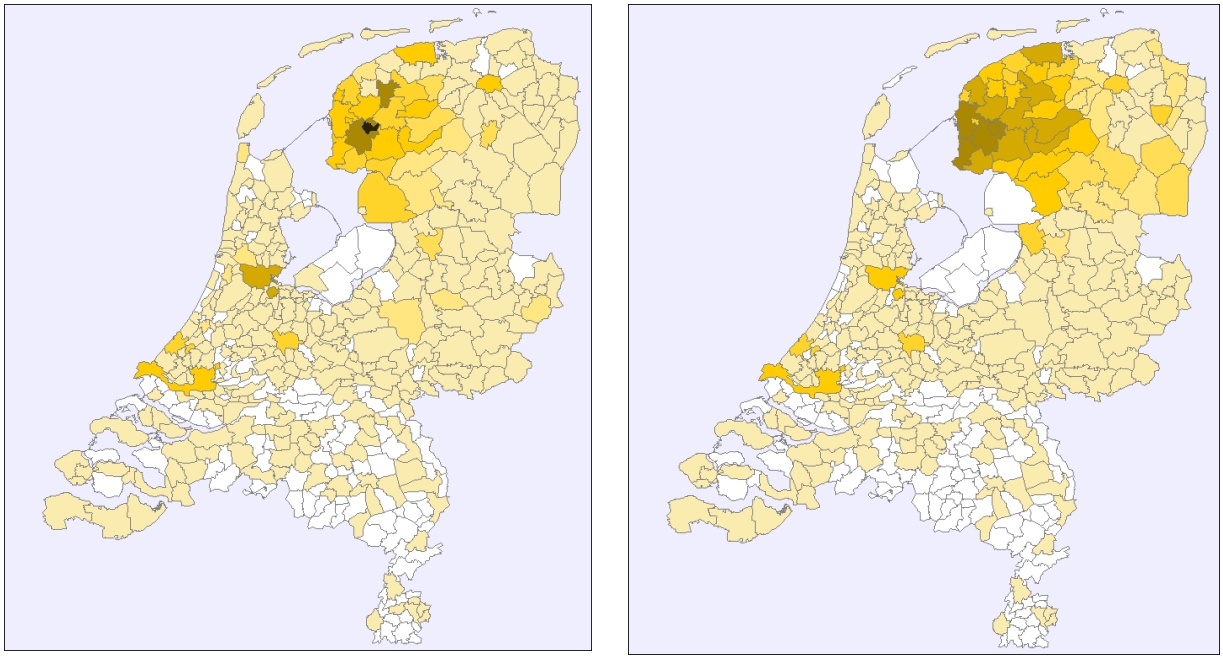


**Figure 5**. Places of birth (percentage) of male inhabitants of the town of Sneek (dark spot), between 30-50 years of age, in the left-hand panel; and those of their great-grandfathers in the right-hand panel.

### *III.3. Regional surnames, co-variation in place*

If persons do not migrate much, the surname of a family may stay in the area of origin for many centuries. Whether this still is the case for certain surnames and certain areas can be investigated by analysing the current distribution of surnames (Bloothooft 2011). We define a regional surname as a name for which 50% of all bearers live within a radius of 30 km from a centre. Names that fulfil this criterium are likely monogenetic, and bearers may still live in the area where their ancestor once adopted or got the surname. We computed the percentage of persons per municipality with such a name and present this in figure 6 in a map of the Netherlands. It shows that in rural areas in the south and the east of the country, percentages

of regional names can be as high as 40%. This is also the case for former fisherman villages at the North Sea and villages around the former Zuiderzee (in the middle of the country) with closed communities. In the north of the country, surnames were not adopted at a large scale until the enforcement by Napoleonic law in 1811. Less original choices of surnames were made, which are therefore not typically regional (<10% regional names). The newly reclaimed polders in the middle of the country have very low percentages as well, with the exception of the island of Urk. It still has to be investigated whether a high percentage of regional name bearers goes together with preservation of linguistic and cultural characteristics.

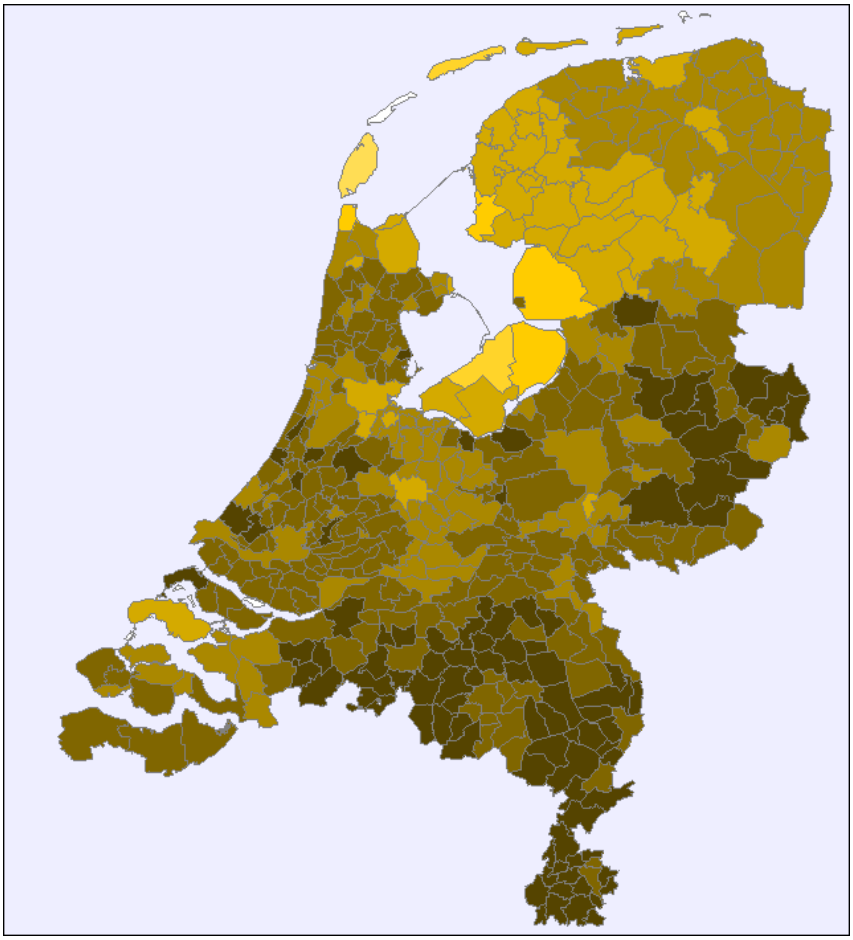


**Figure 6**. Density of regional surnames in The Netherlands. The five shades indicate 1-2% (light), 2-5%, 5-10%, 10-20%, 20-50% (dark) of inhabitants in a municipality with a regional name.

### *III.4. Co-variation of first names in families*

An important property of the data in civil registration (and reconstructed life courses) is that they give fundamental parameters of the life of individuals, such as names, and dates and places of birth, marriage and decease, but also the family relations among individuals. The individual data can be aggregated in time and/or space, to study population properties over time or among regions. But on the basis of known family relations, we can perform such studies within families and across generations. Doing this, we stay more closely to the social strata of the population.

We explored this in a study of modern first names. The assumption was that parents do not choose the names of their children at random, but (perhaps unconsciously) on the basis of

what is fashion or expected in their social environment. This would imply that the names of children in the same family convey a little bit of this fashion. Traditional parents may name their children with old Dutch names like *Willem* and *Dirk*, and this combination of names will appear much more frequently than can be expected on the basis of individual probabilities of the names. By analysing 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, about fifteen clusters or name groups gave a fair description of the 1,409 most frequent names (naming 75% of all children).

The geographic distribution of each name group has significant features across the country, as shown in figure 7 for traditional Dutch names, which mainly follow the Dutch bible belt (figure 8), while short English names are preferred in the areas of Catholic dominance – which earlier chose traditional Latinized names.

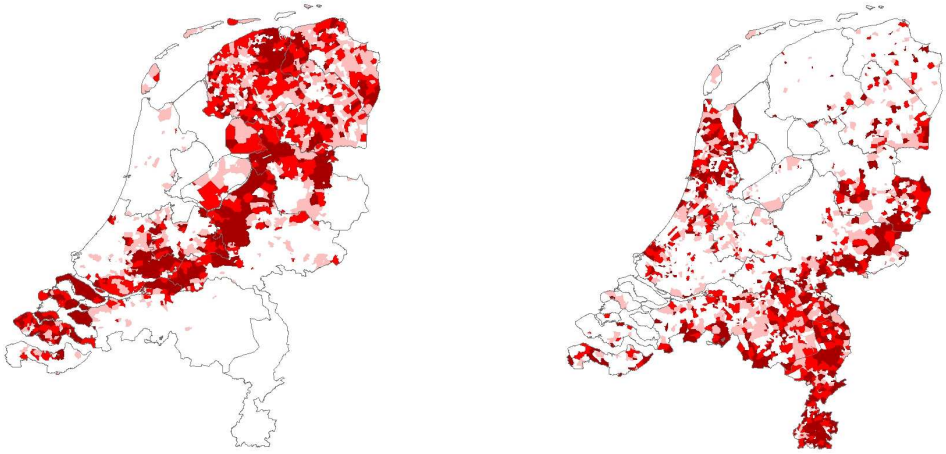


**Figure 7**. Geographic distribution of Dutch traditional first names (left) and short English names (right).

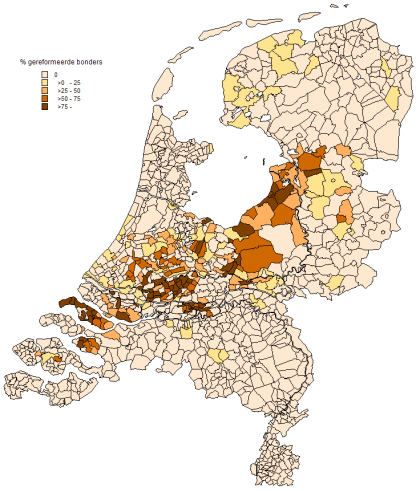



**Figure 8**. Geographic distribution of traditional Protestants, the Dutch bible belt (thanks to Hans Knippenberg).

The advantage of such an approach is that the study of names can concentrate on the properties of these 15 name groups, rather than on 1,409 individual names. It is also possible to develop a comprehensive map of The Netherlands that shows per postal code area the

name group that positively most deviates from the national average, and can be considered typical for that area (Bloothooft 2010). This map of first names in The Netherlands is shown in figure 9.

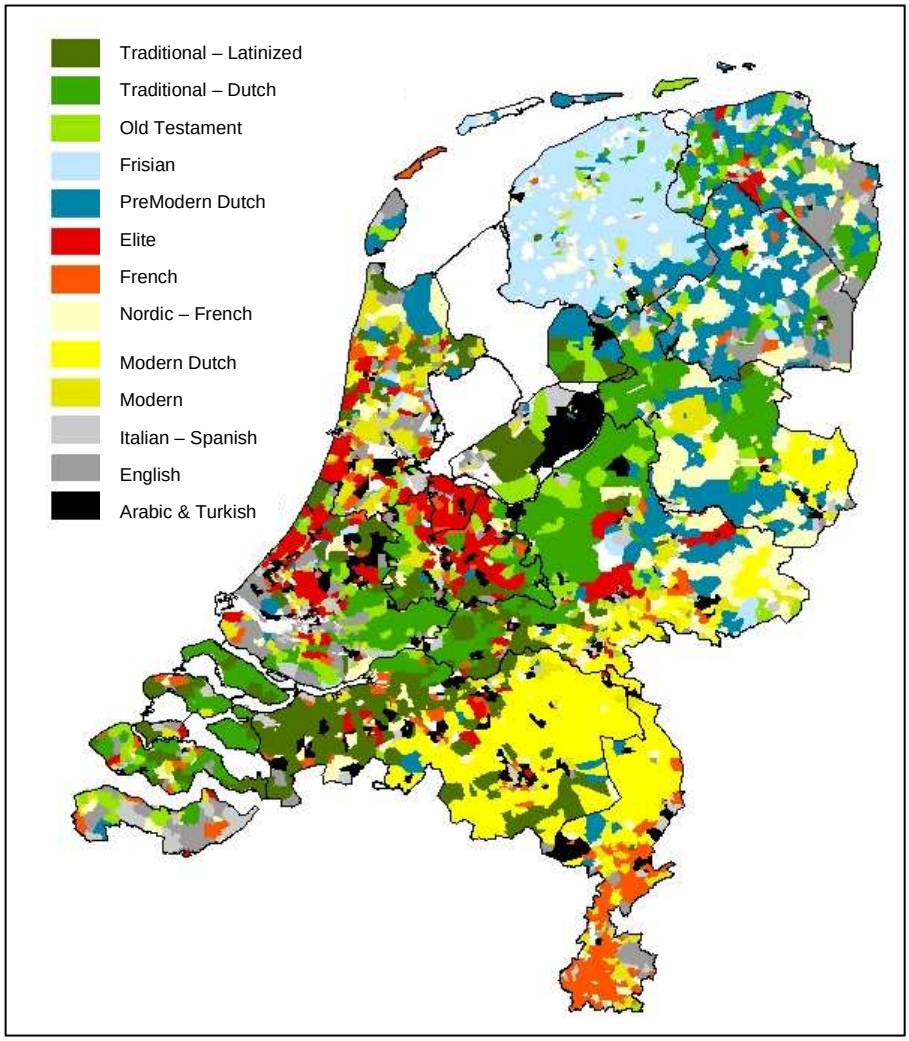



**Figure 9**. Map of first names in The Netherlands, showing the typical name group per postal code area.

## Conclusion

A number of examples are shown which demonstrate the possibilities and the power of analysis of very large and structured name databases, derived from modern Civil Registration. The same could be done on the basis of a reconstructed historical civil registration, which would provide insights in long term developments. Extension of this type of data to a European scale is something to dream of.

## References

Bloothooft, Gerrit; Groot, Loek. (2008). Name clustering on the basis of parental preferences, *Names* 56:3, 111-163

Bloothooft, Gerrit. (2010). Voornamen in kaart: een weerspiegeling van maatschappelijke veelvormigheid. In: Anton Schuurman, Onno Boonstra (ed.), *Tijd en Ruimte; Nieuwe toepassingen van GIS in de alfawetenschappen*. Utrecht: Matrijs.

Bloothooft, Gerrit; Onland, David. (2011). Socioeconomic determinants of first names, *Names* 59:1, 25-41.
Bloothooft, Gerrit. (2011). Linguistics and geography, the surname case. In: Wim Zonneveld, Hugo Quené, Willemijn Heeren (ed.), *Festschrift for Bert Schouten*. Utrecht: UiL-OTS (to appear).
Mandemakers, Kees. (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.* Minneapolis: Minnesota Population Center, 149-177.

Gerrit Bloothooft
Utrecht institute of Linguistics, Utrecht University, Utrecht
Meertens Institute KNAW, Amsterdam
International Institute for Social History KNAW, Amsterdam
Netherlands
g.bloothooft@uu.nl

Kees Mandemakers
International Institute for Social History KNAW, Amsterdam
Netherlands
kma@iisg.nl