Regional variation in Finnish lake and hill names

Antti Leino

Abstract

The Finnish basic map, and the database used by the National Land Survey to produce it, contains over 300 000 different toponyms and over double that amount of named places. It is impossible to study the distributions of all these toponyms at once, and thus a large-scale onomastic analysis requires an overview of the material.

Modern computational methods can extract the most significant directions from the overall variation into a few components, and these components can subsequently be used as a basis for onomastic analysis. A brief analysis of the names of two types of natural features results in components that have interesting parallels in the history of Finnish settlement and in language contacts.

Introduction

The 1:20 000 basic map of Finland includes over 700 000 places with a Finnish name, using just over 300 000 different toponyms. These places include some 58 000 lakes and 86 000 hills. As seen in tables 1 and 2, there are some very common names, but on the other hand about half are completely unique. This massive number of names and their distributions cannot be presented in the form of traditional distribution maps. Simply discarding the least common names is not a viable option: concentrating on names that occur at least ten times – which be reasonable in that a smaller distribution does not tell very much – would still leave some 700 lake and almost 1000 hill names. In order to present an overview of the entire variation one has to do something else.

Fortunately, similar problems exist in other fields, and there are computational techniques that can be applied to linguistic variation (e.g. Leino 2004; Hyvönen et al. 2007). These techniques start with dividing the geographic area into sections – for instance, using existing administrative regions or a regular grid – and then looking at which names occur in each of these sections. A table that presents this information contains the same information as the distribution maps of all the names, but in a form that is much easier to process by computational means.

In statistic terms, the names are now considered observations and the municipalities as variables.[1] There are several statistical methods to analyse such data, of

[1]It would be possible to do the analysis the other way – with the municipalities as observations

Places per name	Lakes		Hills	
	Names	Places	Names	Places
≥ 1	25 178	58 267	46 222	86 303
≥ 2	5 154	38 243	8 999	49 080
≥ 5	1 492	29 170	2 395	32 727
≥ 10	695	24 078	970	23 657
≥ 20	331	19 230	390	16 098
≥ 50	111	12 580	83	7 030
≥ 100	45	8 168	20	2 854
≥ 200	14	3 916	3	772
≥ 500	1	522		

Municipalities per name	Lake names	Hill names
≥ 1	25 178	46 222
≥ 2	4 746	8 393
≥ 5	1 304	2 102
≥ 10	561	772
≥ 20	239	259
≥ 50	53	38
≥ 100	8	5

Table 1: The number of lake and hill names on the basic map

	Lakes		Hills	
	Name	Places	Name	Places
1.	Mustalampi	522	Palomäki	269
2.	Ahvenlampi	357	Myllymäki	257
3.	Haukilampi	346	Isomäki	246
4.	Likolampi	339	Kivimäki	167
5.	Paskolampi	293	Pitkämäki	148
6.	Kuikkalampi	265	Hautakangas	147
7.	Sammakkolampi	247	Rajamäki	145
8.	Umpilampi	243	Multamäki	140
9.	Särkilampi	239	Palovaara	127
10.	Pitkälampi	225	Palokangas	127

Table 2: Ten most common lake and hill names

which I have used two for the present article. The first of these, Principal Component Analysis, is well established, and the method was originally presented by Hotelling (1933). The second, Independent Component Analysis (Hyvärinen et al. 2001) is much more recent, but it has been successfully applied to different kinds of data. In both these methods, the main goal is to transform the municipalities into underlying coponents that best explain the bulk of the overall toponymic variation.

In some ways, these methods resemble traditional dialectology. A language is typically divided into dialects by combining the isoglosses of different features, but these features are not treated as equals. More important features are given more weight than less important ones, and this weighting depends on the judgement of the dialectologist. Principal Component Analysis does the same thing, but automatically: in determining the first component, the weights of the features are determined so that the difference between the two extremes is as great as possible, and for the subsequent components this process is repeated, with the additional requirement that the resulting component is uncorrelated with the previous ones. Thus each component explains the maximal amount of variation that is left over from the previous components. Independent Component Analysis uses a somewhat different criterion: now the goal is to find a specified number of components that are independent of each other – in essence, instead of giving a series of divisions into two the analysis gives a single multi-way division.

For this article I have applied both these methods to three different toponymic corpora. The first two are based on the Place Name Register maintained by the National Land Survey for purposes of map-making (Leskinen 2002). As the first corpus I have selected from this register all Finnish lake names that have at least five occurrences each; as the second one I have a similar selection of hill names. To complement this, I have used the Toponym Atlas currently being compiled by the Research Institute for the Languages of Finland (Ainiala 2007).

Dividing the geography

The various methods for component analysis start with a matrix that contains the names on one axis and some sort of geographical regions on the other. In my initial attempts (Leino 2004) I used municipalities as the geographical division, partly because that is traditional within Finnish linguistics and partly because this information was already present in the Place Name Register. At first glance the results were encouraging. A similar attempt at analysing lexical variation was also quite successfull (Leino et al. 2006).

However, a closer look at the results reveals that this division has its drawbacks. The problem is that the size of the Finnish municipalities varies enough to affect the analysis. When looking at individual distribution maps, the different size of

and names as variables – but the results are much the same. This appears to have something to do with the properties of linguistic variation, as a simliar phenomenon is apparent in dialect data (Leino et al. 2006), even though this interchangeability does not apply generally.

the administrative regions is not a major problem, as the human eye is very good at recognising patterns in the distributions. However, performing a computational analysis of a large number of such distributions is a different matter: the larger an area is, the more names one can expect to find. The differences in sizes introduce an additional source for noise in the data.

In some cases there are no alternatives. For instance, the lexical analysis had to be done by municipalities, as the original data had been collected that way. In the case of lexical variation this approach is not as much of a problem, though, as one should not expect the vocabulary used in a single village to be orders of magnitude smaller than that used in the entire municipality. In toponyms the case is different, although even there one may have to use an administrative division. For instance, the material in the Toponym Atlas is organised in this manner.

Fortunately, in the case of the Place Name Register there is an alternative way of organising the geography. The register contains coordinates for each of the named places, and it is possible to use these. I have thus divided Finland into a grid of 40×40 km squares, and used these squares as the geographical units. Limiting myself to those names that occur at least five times each, I got two matrices: one of 1 492 lake names by 261 squares and another of 2 395 hill names by 254 squares. The different number of squares is due to the fact that there are some squares that do not contain any names of these types at all, mostly because of geography. In comparison, the Toponym Atlas gives a matrix of 239 names by 617 municipalities.

Principal components: directions of variation

The results of the analysis are shown as maps: those in Figures 1–2 include the first four principal components in lake names and those in Figures 7–8 the first four in hill names. In comparison, Figures 14–15 show the first four components in the Toponym Atlas. In each of these maps, the component is shown as a colour scale with deep blue at one end and bright yellow at the other. It is not important which of the ends of the scale is blue and which one is yellow: the analysis assigns the plus/minus sign arbitrarily, and so the only meaningful aspect of the color scale is that the ends are opposite each other.

In Principal Component Analysis, the first component incorporates as much of the original variance as possible. The amount of the total variance that is included in the first four components is shown in Table 3; in the corpora originating in the Place Name Register the first component includes a far smaller fraction of the total variation than the Toponym Atlas corpus. This is likely due to the Toponym Atlas corpus being only about one tenth of the size of the other two.

Table 3 also shows the correlation between each of the components and the number of names per region.[2] The first component in each of the corpora has a very strong correlation: in essence, it tells us only the density of names. While this

[2]Whether this correlation is positive or negative is not important, since the plus/minus sign in Principal Component Analysis is arbitrary.

	Lakes		Hills		Atlas	
Component	Variance	Correlation	Variance	Correlation	Variance	Correlation
1	6.3 %	0.945	5.4 %	-0.807	20.3 %	-0.992
2	3.2 %	0.239	2.9 %	0.173	6.3 %	-0.085
3	2.6 %	-0.175	2.1 %	-0.351	2.8 %	0.010
4	1.9 %	-0.045	2.1 %	-0.363	2.2 %	0.019

Table 3: The percentage of total variance and the correlation with the amount of names for the first components

is not very interesting in itself, it is useful in that this factor is now isolated in the first component, and the others should be easier to interpret.

In both lake and hill names the first component can also be interpreted as the effects of the physical environment, as approximated by the density of lakes and hills. In the case of hill names, the second component can be interpreted this way as well: the component in Figure 7 b shows a difference between Ostrobothnia and the rest of the country. In terms of physical geography, Ostrobothnia is a uniform, flat stretch of old sea bottom; the typical hill names in the region end in *-saari* 'island'.

In the Toponym Atlas, the next components reflect dialectal variation – in fact, components 2 and 3, shown as Figures 14 b and 15 a, are quite close to those appearing in a lexical corpus (Leino et al. 2006: 43). The opposition between Eastern and Western dialects can also be seen in the hill names, in Figure 8 a. While the first component in lake names, in Figure 1 a, is adequately explained by the density of lakes, it also matches the traditional dialect division. It seems likely that the component captures the effect of both these factors; the analysis cannot distinguish between two causes that contribute to the same effect. The rest of the principal components seem to be related to effects that are more clearly apparent in the independent components.

Independent components: centres in variation

The independent components derived from all three corpora show mostly the same regions. This is somewhat surprising, given that Principal Component Analysis found rather different-looking components from each of them. Nevertheless, not all of the independent components in the different corpora are the same – or more properly, not all of the componets can be seen in every corpus.

The differences between the corpora can for the most part be explained by environmental factors. For instance, the lake names do not give any components that are centered on the coastal regions where there are few lakes. On the other hand, near the eastern border, where the density of lakes is at its largest, they give the most detailed picture.

The components can by and large be interpreted in terms of settlement history.

There are six regions that show particularly clearly:

Tavastian hunting grounds: The maps show as black squares the iron-age fortified hills listed by Taavitsainen (1990). In Tavastland, the region just north of this belt of early settlements shows clearly in all three corpora: in lakes Figure 5 a, in hills Figure 12 a and in the Toponym Atlas in Figure 17. This was the core region of the hunting and fishing grounds for the adjoining agricultural lands, and it was settled permanently only later.

Southern Carelia: The early-settled regions in Southern Carelia show also clearly in all three corpora, although the Basic Map only covers the present-day Finland. In lakes, this region is seen in Figure 4 a, in hills in Figure 11 a and in the Toponym Atlas in Figure 19 b; in addition, the hill name corpus also shows the adjoining region in Figure 9 a.

Finland Proper: The old province of Finland Proper does not appear in the lake name corpus, most likely because there are very few lakes in the area. In the other two corpora it shows clearly, in the hill names in Figure 13 and in the Toponym Atlas in Figure 18 b. In the former map, the component covers the Uusimaa province as well, which would imply that the component can also be seen as a more general indication of linguistic contact with early Scandinavia – the coast south of this region was settled from Sweden during the early middle ages.

West coast: Slightly north of Finland Proper, the Toponym Atlas reveals the region of Lower Satakunta in Figure 18 a. The settlement names in this region are very characteristic, including numerous village names that are loans from Swedish or in some cases earlier Germanic languages. This is not reflected in the lake and hill names, however – or at least, not sufficiently to show in the analysis of these corpora. The hill names, on the other hand, show Southern Ostrobothnia as a compact region in Figure 10 b; this region has similarly clear and sharp borders with the neighbouring dialects.

Eastern border: The border against Russia – or more properly, the Carelian language – shows clearly in all corpora. The hill name and Toponym Atlas components in Figures 12 b and 17 b include the entire stretch from the Finnish province of Northern Carelia to that of Kainuu. The lake name corpus, however, gives a more detailed picture. This is not surprising, considering that this region has the greatest density of lake names per square kilometre.

Figure 3 a shows the northern part of Kainuu. This component may reflect contact with the Sámi languages; its southern edge follows the old border of the region where late medieval Russia had the right to tax the Sámi. Immediately south of it, Figure 5 b shows the central and southern parts of Kainuu. This region has traditionally had close contacts with the northern dialects of the Carelian language.

Still south of that, Figure 4 b shows the Finnish province of Northern Carelia, with its contacts with the southern dialects of Carelian.

Central Finland: The region northeast from the old Tavastian hunting grounds shows again in all the corpora. In the hill and Toponym Atlas ones it results in one component each, shown in Figures 9 b and 17 b. In the lake name corpus there are two components, shown in Figures 3 b and 6, and each of these corresponds with one of the other corpora. It would seem natural to link these with different phases of the colonisation of these regions from Savonia in the 17th century.

Conclusions

In terms of methodology, it is clear that various component analyses can be used to present an overview of onomastic variation. However, these methods have to be used with some care. It is clear that these methods do not give just one clear truth; rather, they present different views of the data, to be used as a basis for a further onomastic analysis.

The choice of the method is important, as different methods give quite different results. This does not reflect any fault in the methods, but rather it should be considered as intentional: the two analyses outlined here have each different goals, and the choice between them should be based on the objectives of the onomastic research. These two are not the only such methods, either.

Onomastically, this brief analysis shows several regions in Finnish toponyms. This regional variation reflects in part language contacts with the Swedish population along the south-western coast, the Sámi population in Lapland and the two main dialect groups of Carelian. Equally importantly, it also reflects the history of the settlement of Finland, roughly from the Viking age up to the 16th century. Furthermore, it reveals interesting regions around Tavastland and the Russian border; in these areas, further onomastic research is clearly indicated.

References

Ainiala, Terhi 2007: Ortnamns utbredningsområden beskrivna i en ortnamnsatlas. In *Nordiske navnes centralitet og regionalitet. NORNAs 35. symposium, Bornholm den 4.-7. maj* 2006, pp. 11–18. København.

Hotelling, Harold 1933: Analysis of a Complex of Statistical Variables into Principal Components. *Journal of Educational Psychology*, 24:417–441, 498–520.

Hyvärinen, Aapo – Karhunen, Juha – Oja, Erkki 2001: *Independent Component Analysis*. John Wiley & Sons.

Hyvönen, Saara – Leino, Antti – Salmenkivi, Marko 2007: Multivariate Analysis of Finnish Dialect Data. *Literary and Linguistic Computing*, 22(3):271–290.

Leino, Antti 2004: Computational Overview of Finnish Hydronyms. In Dzintra Hirsa (ed.), *Onomastica Lettica. II*, pp. 239–268. LU Latviesu valodas instituts.

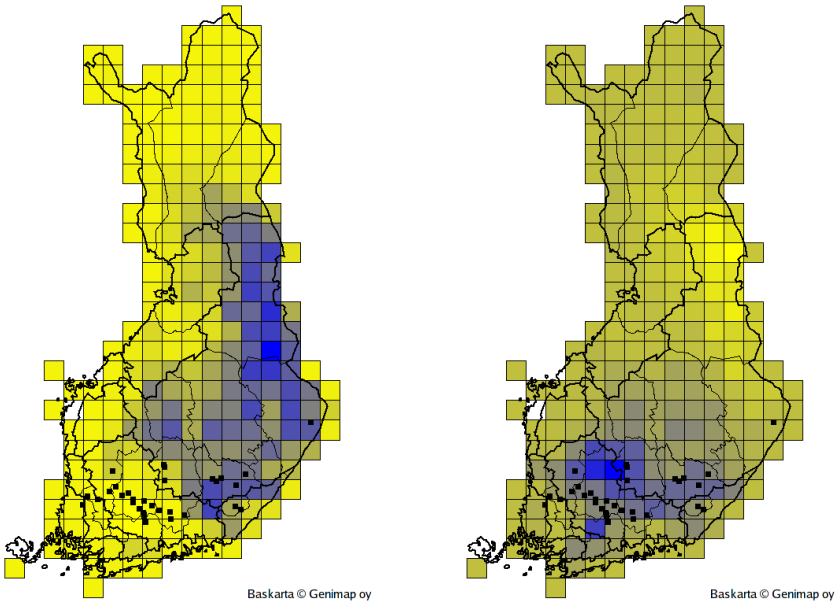
Leino, Antti – Hyvönen, Saara – Salmenkivi, Marko 2006: Mitä murteita suomessa onkaan? Murresanaston levikin kvantitatiivista analyysiä. *Virittäjä*, 110(1):26–45.

Leskinen, Teemu 2002: The Geographic Names Register of the National Land Survey of Finland. In *Eighth United Nations Conference on the Standardization of Geographical Names*.

Taavitsainen, J.-P. 1990: *Ancient Hillforts of Finland. Problems of Analysis, Chronology and Interpretation with Special Reference to the Hillfort of Kuhmoinen*. Suomen muinaismuistoyhdistyksen aikakauskirja 94. Suomen muinaismuistoyhdistys.

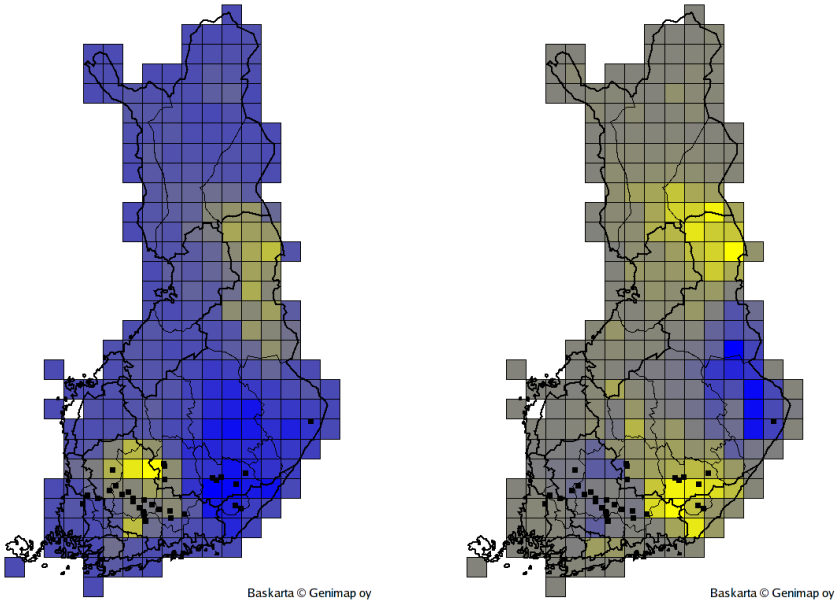
Maps

Lake names



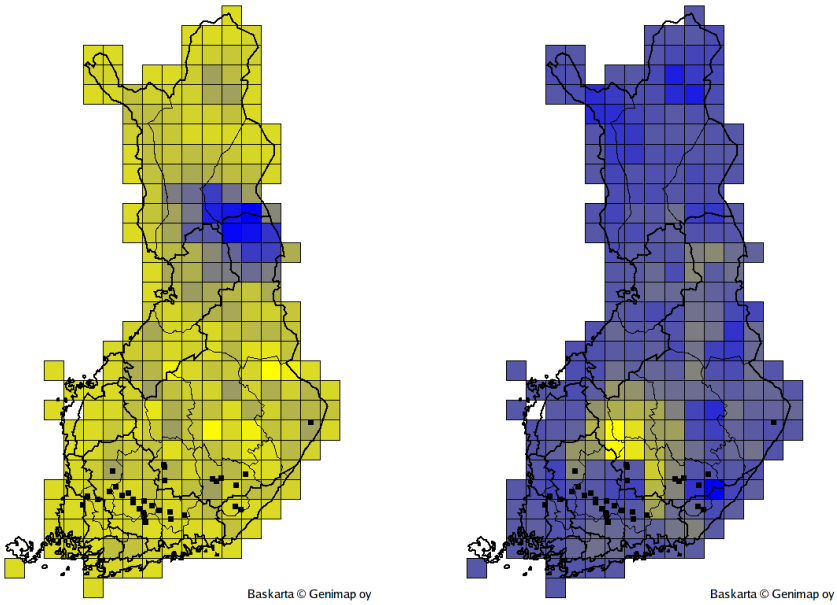
a. Component 1 b. Component 2

Figure 1: Principal components 1–2 in lake names



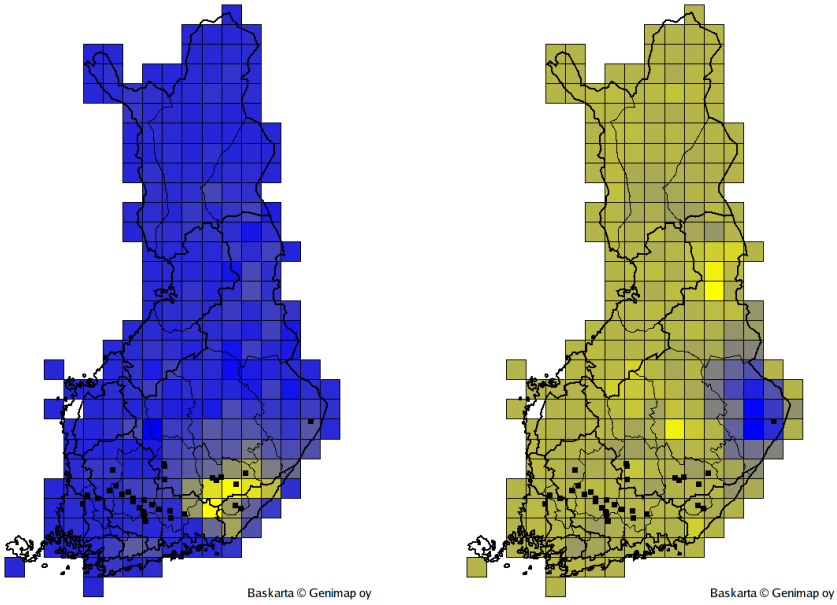
a. Component 3 b. Component 4

Figure 2: Principal components 3–4 in lake names



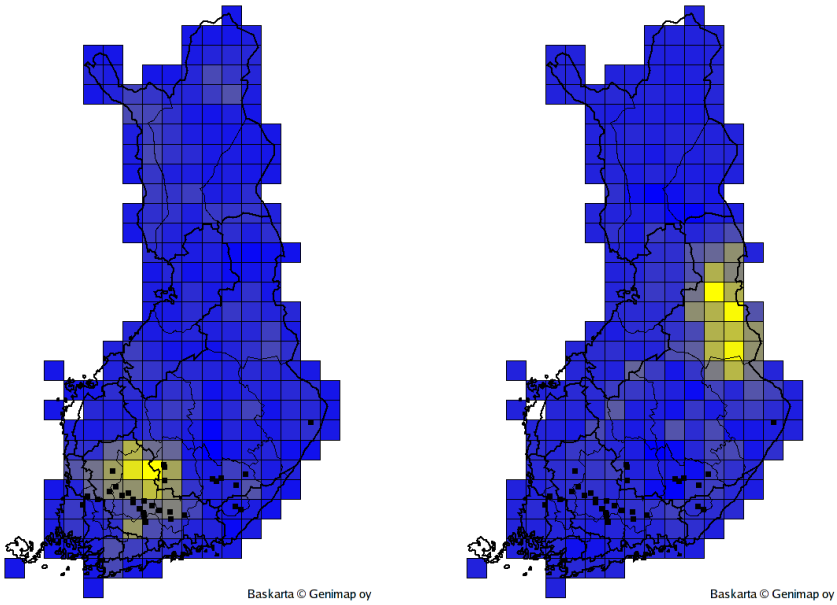
a. Component 1 b. Component 2

Figure 3: Independent components 1–2 in lake names



a. Component 3 b. Component 4

Figure 4: Independent components 3–4 in lake names



a. Component 5 b. Component 6

Figure 5: Independent components 5–6 in lake names

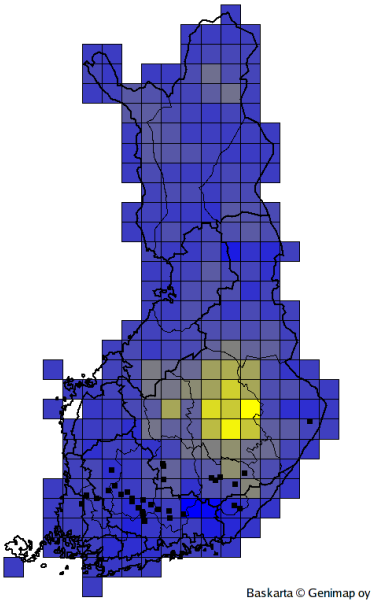
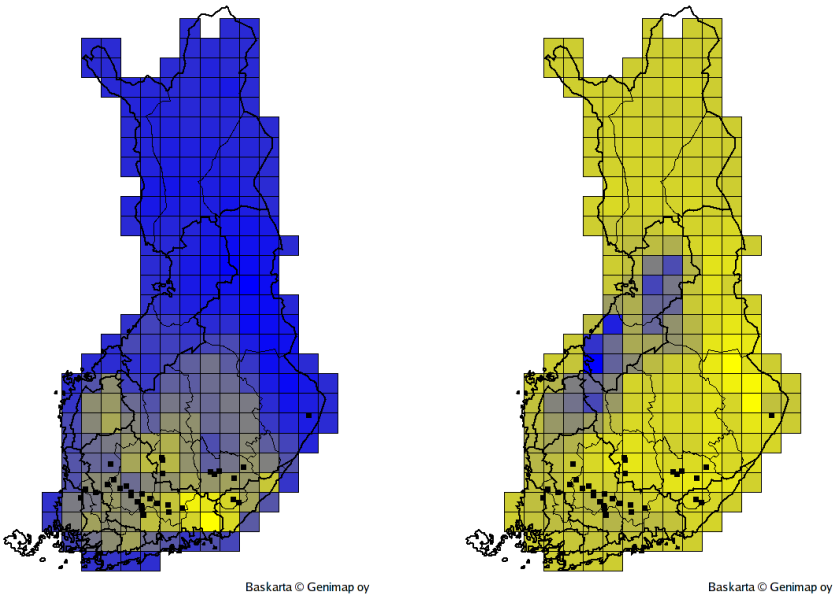


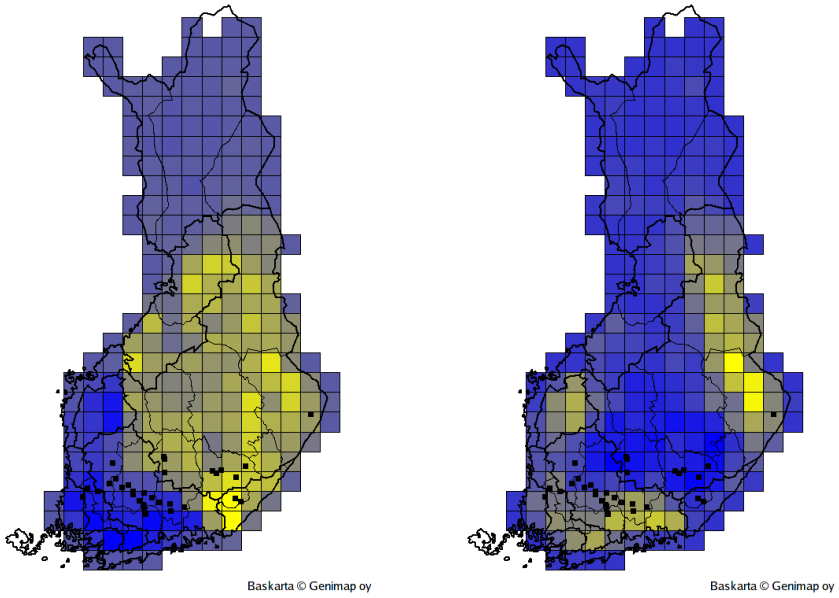
Figure 6: Independent component 7 in lake names

Hill names



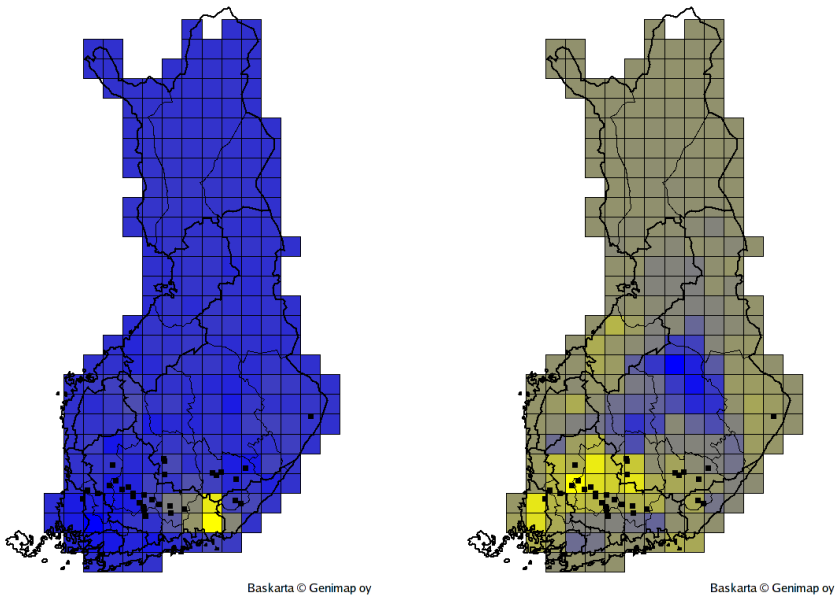
a. Component 1 b. Component 2

Figure 7: Principal components 1–2 in hill names



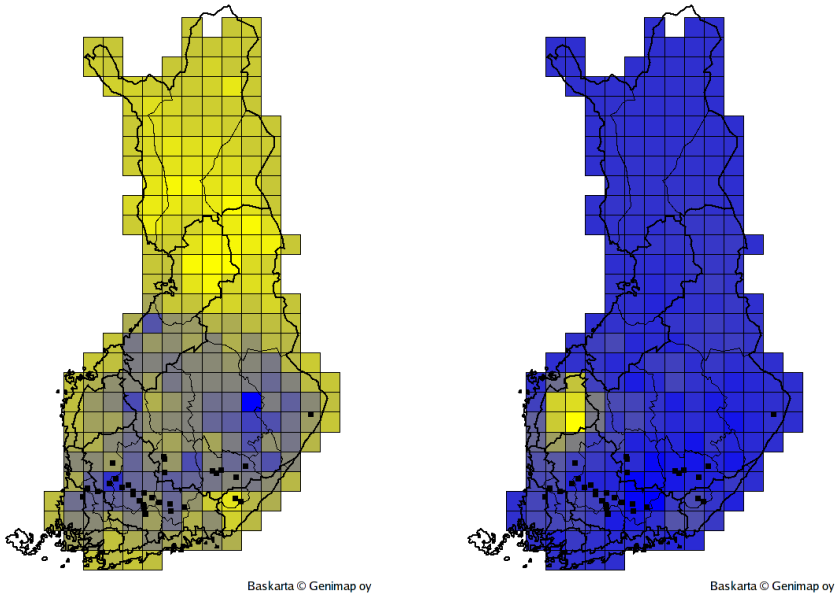
a. Component 3 b. Component 4

Figure 8: Principal components 3–4 in hill names



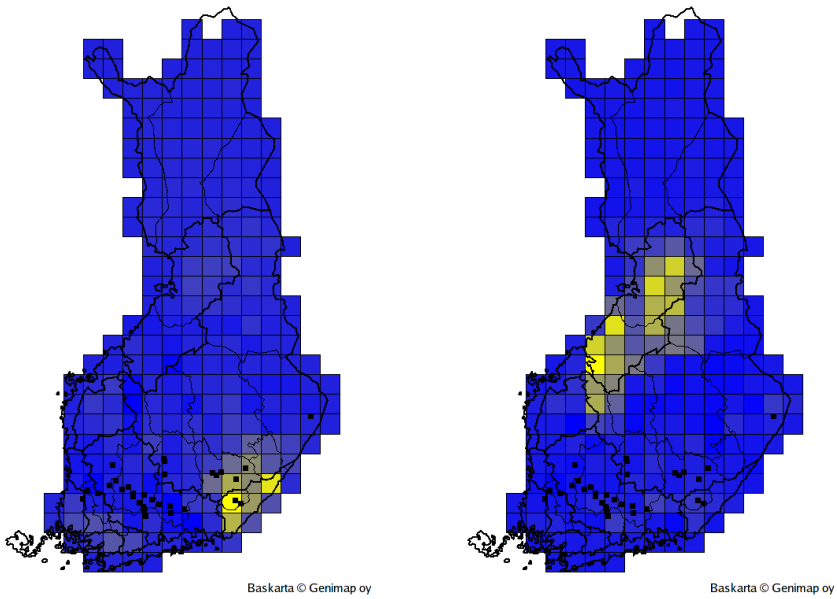
a. Component 1 b. Component 2

Figure 9: Independent components 1–2 in hill names



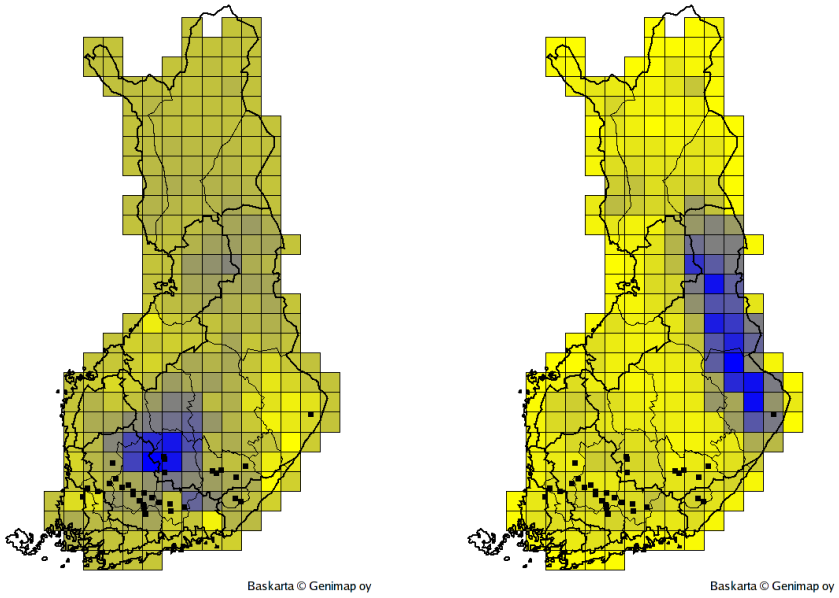
a. Component 3 b. Component 4

Figure 10: Independent components 3–4 in hill names



a. Component 5 b. Component 6

Figure 11: Independent components 5–6 in hill names



a. Component 7 b. Component 8

Figure 12: Independent components 7–8 in hill names

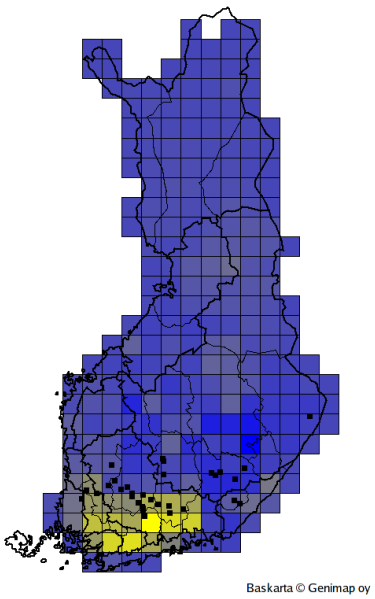
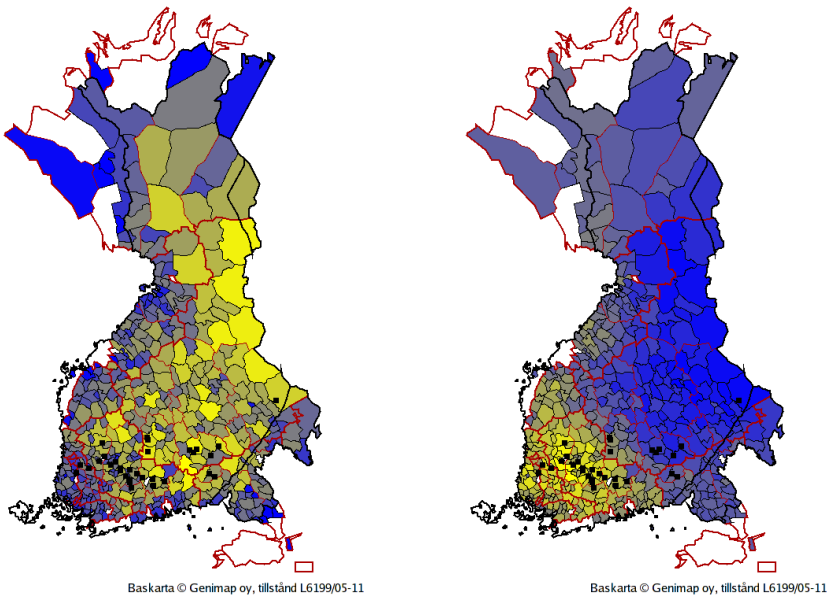


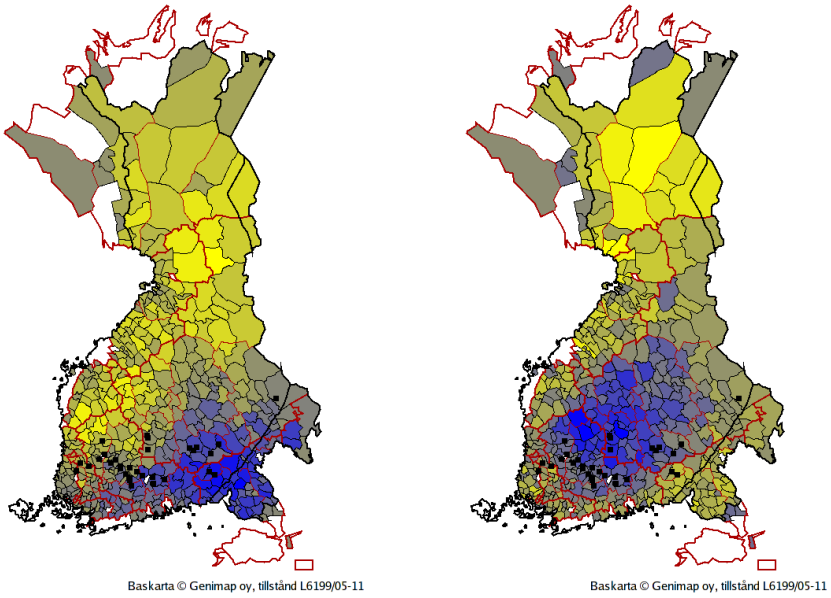
Figure 13: Independent component 9 in hill names

Place Name Atlas



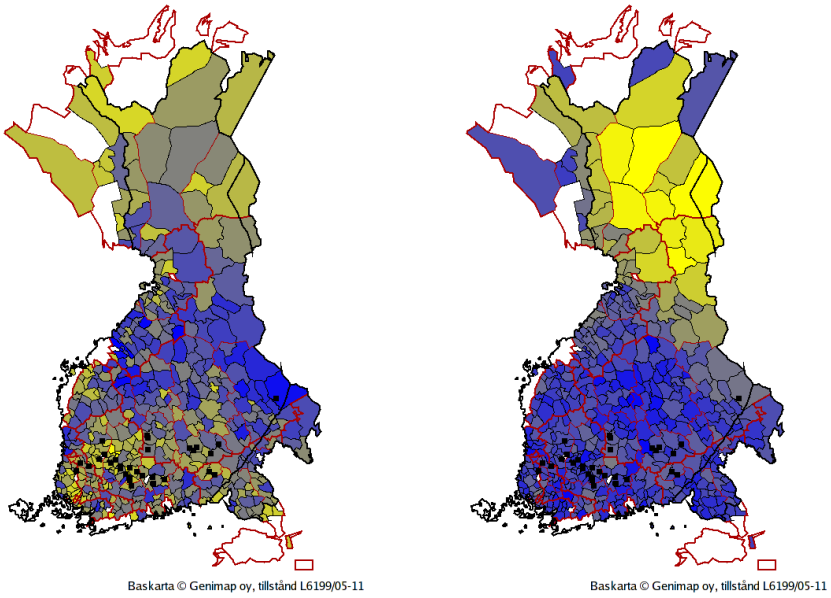
a. Component 1 b. Component 2

Figure 14: Principal components 1–2 in the Place Name Atlas



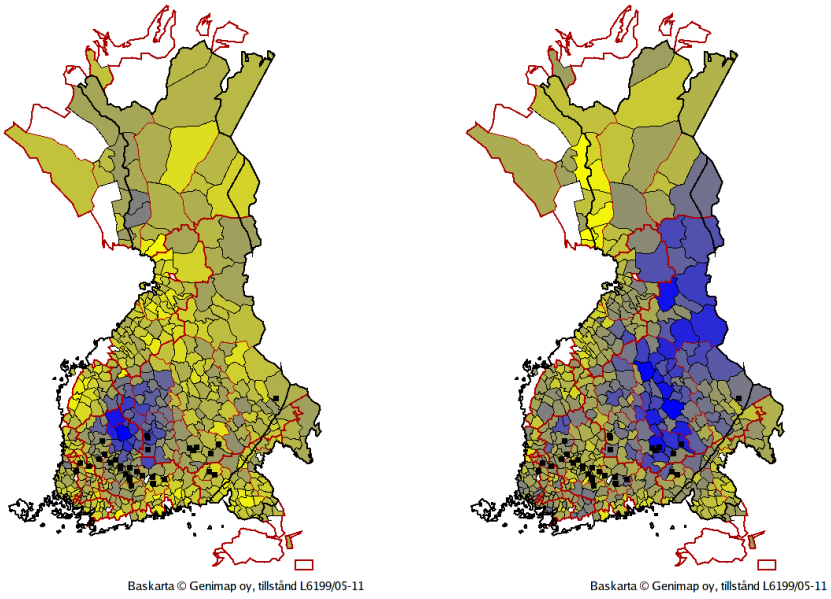
a. Component 3 b. Component 4

Figure 15: Principal components 3–4 in the Place Name Atlas



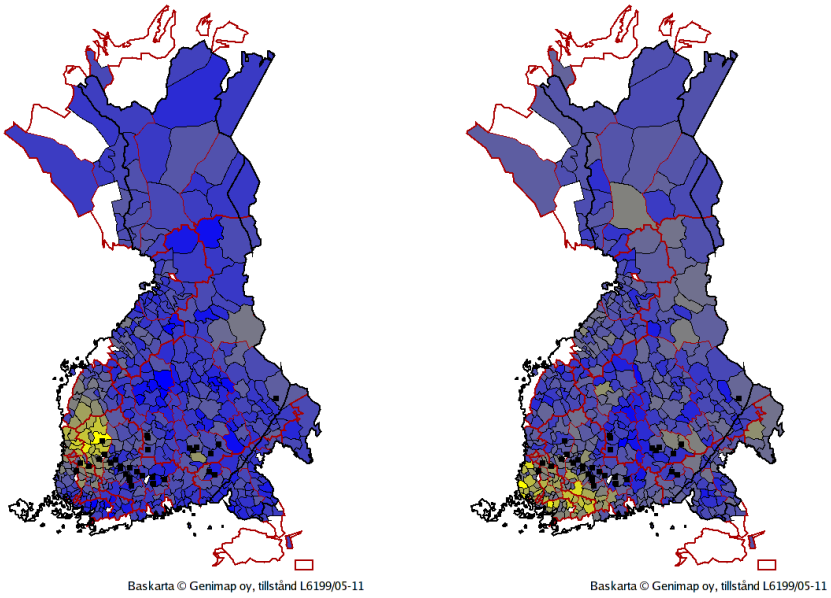
a. Component 1 b. Component 2

Figure 16: Independent components 1–2 in the Place Name Atlas



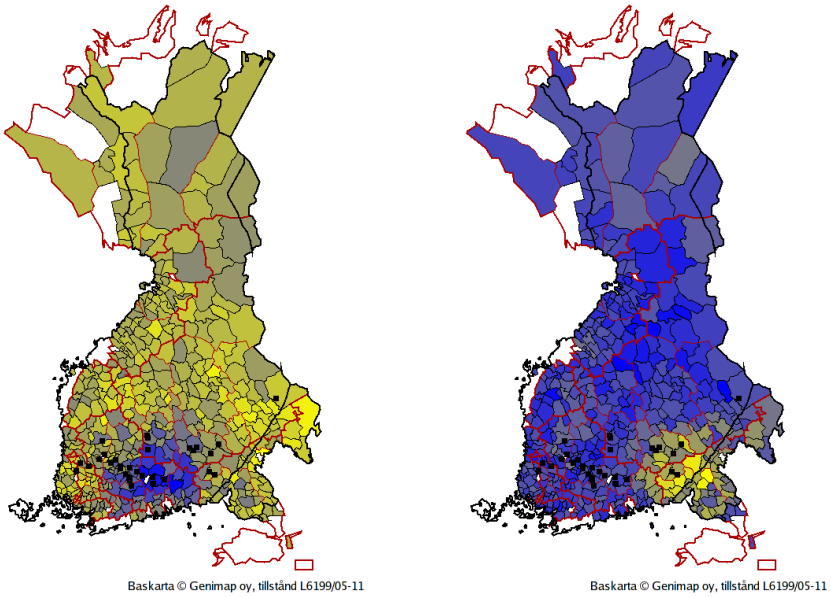
a. Component 3 b. Component 4

Figure 17: Independent components 3–4 in the Place Name Atlas



a. Component 5 b. Component 6

Figure 18: Independent components 5–6 in the Place Name Atlas



a. Component 7 b. Component 8

Figure 19: Independent components 7–8 in the Place Name Atlas

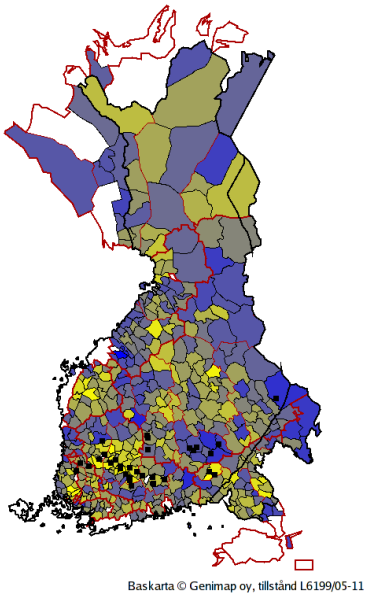


Figure 20: Independent component 9 in the Place Name Atlas