Computational Overview of Finnish Hydronyms

Antti Leino
leino@cs.helsinki.fi
Helsinki Institute for Information Technology
Research Institute for the Languages of Finland

Abstract

The spatial distribution of a wide range of linguistic phenomena has traditionally been visualised in the form of maps. Distribution maps are very useful when dealing with only a few different phenomena at a atime, but they soon become rather unwieldy as the number of different distributions increases. This is related to what is known in the field of data analysis as the "curse of dimensionality": in general, a lot of traditional methods tend to become unusable when dealing simultaneously with a massive number of different variables.

There are ways to cope with the problems that arise from massive dimensionality. This article shows how some of these methods, most notably principal component analysis, can be applied to onomastic data. Starting with raw data that consists of all hydronyms that appear on Finnish basic maps, the goal is to find a few of the most important trends that lie behind the distributions of individual names. Some of the results are rather predictable in view of present knowledge about Finnish dialects and settlement history; others are less so.

1 Introduction

The National Land Survey of Finland has, for its own purposes of producing maps, a Geographic Names Register. A part of this register is the Place Name Register, which contains all names that appear on the 1:20 000 Basic Map (Leskinen 2002). The study leading to this presentation concentrates on common hydronyms, *common* in this case meaning those names that appear on at least ten or five municipalties. The number of names that fulfill this criterion is shown in table 1.

	Total	In data set	Municipalities
Lakes	25 178	1 492	≥ 10
Parts of lakes	17 469	939	≥ 10
Rivers	14 650	797	≥ 10
Rapids	3 460	84	≥ 5
Other parts of rivers	5 372	67	≥ 5

Table 1: Finnish National Land Survey Place Name Register

The purpose of this study was to distill an overview from this corpus of data. This problem resembles in some respects the field of dialectometry (eg. Goebl 1982; Nerbonne 2003; Nerbonne and Heeringa 2001), although there are differences between an onomastic study — like the present one — and one dealing with dialectal variation. When dialectometric researchers have studied broad, national-scale trends they have often concentrated on developing and using more and more sophisticated methods for computing the distances between dialects, based on the variation of several linguistic features.

The geographical distribution of linguistic features in dialectology — and by extension, dialectometry — is not discrete, but rather the distributions of different variants overlap. Toponyms, on the other hand, are a discrete set: for the purposes of this study it is reasonable to claim that the places and their names are known. This is a rather major difference between traditional dialectometry and the type of onomastic study presented here.

2 Methods

2.1 Principal Component Analysis

One of the well-known problems in the field of data analysis is what is called the "curse of dimensionality". That is, as the number of different variables increases most traditional statistical methods become first cumbersome and rather soon in practice entirely unusable. Often the best way to cope with a data set with a massive number of separate variables is to try to decrease the dimensionality. One of the tools commonly used for this purpose is Principal Component Analysis (eg. Mardia et al. 1979).

In short, the aim of Principal Component Analysis is to take the data and transform it so that one gets components that are not correlated with each other. These components are weighted combinations of the original variables, and they are presented in order of decreasing variance. Thus the first principal component accounts for the largest fraction of the total variance and the entire set of components accounts for all of it.

A geometrical interpretation is that one plots the data in a multidimensional space, where each axis of the coordinate system corresponds with one of the variables. To get the principal components, one turns the coordinate system so that one axis, which corresponds with the first principal component, points in the direction where the variance of the data is greatest; the second axis, while at right angles to the first, is then turned in the direction where the residual variance is the greatest, and so on. This interpretation is also useful in that it makes it intuitively clear that as one sets the direction of the axes, they could equally well be turned exactly around. Thus in principal component analysis the direction of the +/- sign in any of the components is arbitrary.

The ordering of the principal components means that in most cases the first few principal components give a rough overview of the data. Also, it is usually possible to reduce the noise of the data by concentrating on the first components and ignoring the last ones, as the latter contain relatively little real information.

2.2 Cluster Analysis

Cluster analysis (Tryon 1939) is a family of methods for organising data to structures that are, one hopes, meaningful. A good introduction to the topic is Kaufman and Rousseeuw (1990), but in a nutshell the goal is to divide the data to clusters, so that the difference between items in the same cluster is as small as possible, and the difference between items in different clusters as large as possible. There are several ways to do this, but in general clustering methods can be divided into hierarchical (often called also agglomerative or joining) and partitioning (also called divisive) methods. Both of these have their own strengths and weaknesses.

In hierarchical clustering first individual items are joined to each other, and the groups to each other, so that the result is a tree of cluster associations. In this tree, the different branches are the clusters, and one can choose the appropriate level of detail by deciding which branches are viewed as separate clusters. One of the serious problems with hierarchical clustering, especially with such data as analysed in the current study, is that small-scale variation, while in reality rather unimportant, can have a large effect on the results of the analysis: when one joins two elements at a time it is possible, and in practice common, that a larger group gets split into two branches which in turn get separated.

In partitioning (also called divisive) methods, on the other hand, the data is divided to a specified number of clusters. Here the the typical difficulty is that one has to know — or guess — the number of clusters in advance. Also, since these methods compare an item to the cluster as a whole, instead of simply two items to each other, they often do not allow one to use as wide a range of similarity measures as the hierarchical methods.

Finding the optimal clustering is in most cases what computer scientists call an NP-hard problem: that is, in practice impossible. Approximations are of course possible, but these often give slightly different clusterings each time the analysis is performed. However, Ben-Hur and Guyon (2003) note that the stability of cluster analysis can be increased by using principal component analysis as a first step. In the present study this was done; subsequently, cluster analysis was performed by the K-medoids partitioning method (Kaufman and Rousseeuw 1990, chapter 2).

3 Analysis of the Hydronyms

3.1 Lakes

The lake names were set as a matrix, with the municipalities as variables and the distributions of each name as observations. The goal, thus, was to transform the actual geographic regions to components that explain the distributions of lake names.

The maps in figures 1—3 show the weights of each municipality in the first three components, drawn in shades of gray on a map with main dialectal divisions shown as black lines; next to each map is a table of the 20 names most strongly associated with each end of the spectrum. The first component, which accounts for 13 % of the variation in name distributions, appears to be related to the division of Eastern and Western Finnish dialects. The second component, which with 4% of the total variation is already markedly less significant, is concentrated mainly in the Kainuu region, and the third component is strongest in Tavastland and Lapland.

The first component, in figure 1, can be considered an expected result: the East

—West division is the most fundamental one in Finnish dialects. The second component is rather less expected, and it may have something to do with the fact that the center is in the municipalities where the density of lakes is at its highest. The names most strongly associated with the darker end of the scale in figure 2 are consistent with the lakes being uniformly small; the names associated with the light end of the scale imply a wider variation in lake sizes. On the other hand, this area shows up rather prominently in the river data as well, so there may be other reasons besides the small size of lakes.

The third component seems again linguistically or culturally related. The dark region in the northernmost part of Lapland in figure 3 may be an anomaly caused by the fact that the area was originally Lappish-speaking, and so the Finnish names are either new or translations of old Lappish ones. However, the dark regions slightly more south in Kainuu and the southernmost one in Tavastland are possibly related: for instance, Talvio (2002) lists 11th century coin hoards in both areas but not in the region between.

A two-way clustering based on the first three components, shown on the right-most map of figure 4, results in a division of Finland into the Eastern region, in light gray, and the Western one, in darker gray. As the number of clusters increases, first the Western cluster splits to, on one hand, Tavastland and the area around the Tornio river in Lapland, shown an intermediate shade, and on the other hand the rest; this division appears consistent with the settlement history of the Tornio river valley (cf. Vahtola 1980). Later on the rest of Lapland, shown in very dark gray, splits off from the Western cluster and the Eastern cluster splits into the old provinces, very light, and the region that was settled in the 17th century, slightly darker.

Figures 5—7 and 8 show similar maps based on the names of parts of lakes, such as bays. The three principal components are roughly similar, but the clustering is geographically somewhat less consistent. One contributing factor is likely that this data set is smaller than that of lake names, so one should not expect quite as thorough results. Another partial explanation is that the names in Lapland — the area where the clustering results are least consistent — are generally much younger than those in the south.

3.2 Rivers

The maps in figures 9—11 showing principal components of river names show also drainage basins as white lines. One can see that the first principal component appears to be correlated on whether the municipality is up- or downriver. The second component is concentrated on the basins of the Oulu and Kemi rivers, or more generally in Northern Finland; the third, like the second component in lake names, is again concentrated in Kainuu.

Figures 12 and 13 show clusterings based on river name components. Figure 12 shows a two-way clustering based on different numbers of components; it is interesting how the one based on only two first components assigns Kainuu to the same cluster as the coastal regions. With a larger set of components one cluster, shown in light gray, would seem to include the northern Bothnia and Kainuu in addition to the traditionally settled regions in the south. As noted above in the discussion about lake names, it is perhaps not altogether impossible to see in this last map a rough reflection of the areas under permanent Finnish influence in late Viking age, although it is not clear that this in fact is the reason for the results.

The three-way clustering, on the two leftmost maps in figure 13, starts to look somewhat more understandable: the old hunting regions appear as a separate cluster in light gray, the old agricultural lands in the south as another in an intermediate shade, and the coastal regions as the third one in dark gray. In the five-way clustering on the rightmost map, Lapland and the old Savolax separate as the very dark and second-lightest clusters.

All in all, the distributions of river names do not combine into quite as expected structures as was the case in lake names. One possible reason is that river names are more closely related to physical phenomena; another one would be that river names were treated differently from lakes in the old hunting cultures. Yet another one would simply suggest that the problem is in the data: the coordinates for rivers are given as a point in the mouth of the river, which may at least partially account for the first component.

The data sets of names of rapids and other parts of rivers were much smaller, so it is understandable if the analyses are less definite than with the other data. Also, the difference between these two types is not necessarily clear; from the names it is apparent that places that are — or have been — viewed as rapids by the local people are classified as other parts of rivers. On the other hand, there is no obvious reason to suspect that this ambiguity affects the analysis.

All in all, the first components derived from the rapids data, shown in figures 14—16, parallel those from the river names, except for Lapland. This is a reasonably strong argument against the hypothesis that the difference in distribution between lake and river names is caused by the coordinate encoding in the data: the problems inherent in representing a river by a point near its mouth do not apply to rapids.

The principal components in the other parts of rivers, which are shown in figures 17—19, have also some similarity to the river and rapids names. However, there is even more noise apparent in the data than in the rapids names. This is not surprising, considering that this is the smallest data set.

While the principal components show structures that support the analysis on river names, cluster analysis, as shown in Figure 20, resulted in very little interesting information. Essentially the only interesting structure can be seen in the five-way clustering on the names of rapids, where Lapland emerges as a separate cluster shown in very dark gray.

4 Conclusions

For the most part, the methods used in this study would appear to work. Analyses on the larger data sets resulted in clusters that were geographically homogeneous, even though the methods themselves did not use any geographical information before the last step of actually drawing the map. The resulting maps were close to traditional dialectal borders, which also supports the validity of the results; on the other hand, they were also sufficiently different from these that the results are interesting.

The names of lakes, and also parts of lakes, have an overall distribution that closely follows dialectal variation. This is not surprising, and neither is it surprising that names appear somewhat more conservative than the language currently spoken, so that the regions can be interpreted in terms of Finnish settlement history. The results obtained are more or less in line with what has already been known.

River names, however, are different. Are the reasons for this difference rooted in the old hunting culture, or is this because of the distribution of natural features? Some further study would seem to be warranted. Another interesting result is that in all data sets the difference between Kainuu and the rest of the country shows up within the first three principal components. There is no immediately obvious reason for this, so again further study seems indicated.

References

Ben-Hur, A. and Guyon, I. (2003). Detecting stable clusters using principal component analysis. In M. Brownstein and A. Kohodursky, editors, *Methods in Molecular Biology*, pages 159—182. Humana press.

Goebl, H. (1982). *Dialektometrie: Prinzipien und Methoden des Einsatzes der numerischen Taxonomie im Bereich der Dialektgeographie*. Wien: Österreichischen Akademie derWissenschaften.

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley-Interscience.

Leskinen, T. (2002). The geographic names register of the National Land Survey of Finland. In *Eighth United Nations Conference on the Standardization of Geographical Names.*

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate Analysis.* Academic Press.

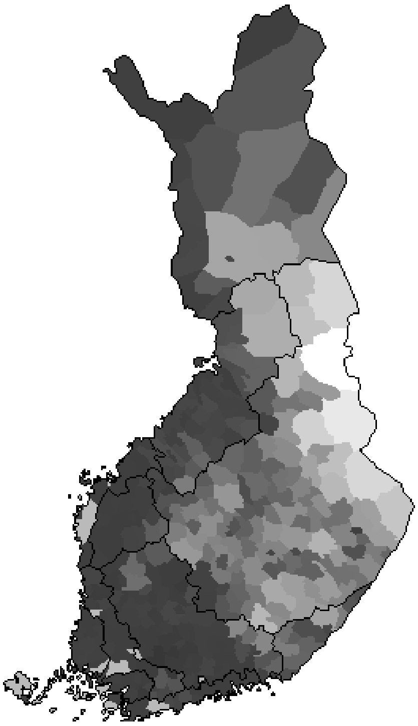
Nerbonne, J. (2003). Linguistic variation and computation. In *Proceedings of the 10th Meeting of the European Chapter of the Association for Computational Linguistics*, pages 3—10.

Nerbonne, J. and Heeringa,W. (2001). Computational comparison and classification of dialects. *Dialectologia et Geolinguistica*, 9, 69—83.

Talvio, T. (2002). *Coins and Coin Finds in Finland AD 800—1200*. Number 12 in ISKOS. Finnish Antiquarian Society.

Tryon, R. C. (1939). *Cluster Analysis*. Edwards Brothers.

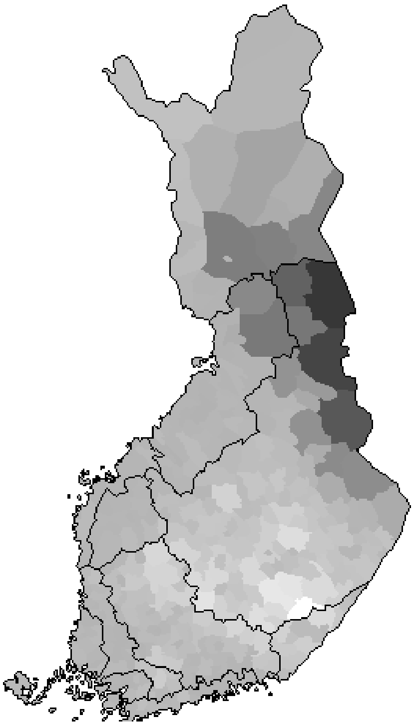
Vahtola, J. (1980). *Tornionjoki- ja Kemijokilaakson asutuksen synty: nimistötieteellinen ja historiallinen tutkimus*. Number 3 in Studia historica septentrionalia. Pohjois-Suomen historiallinen yhdistys.



	Light	Dark
1	Mustalampi	Kakarilammi
2	Ahvenlampi	Väärälammi
3	Haukilampi	Hanhilammi
4	Paskolampi	Hirvilammi
5	Sammakkolampi	Vehkalammi
6	Tervalampi	Pitkälammi
7	Särkilampi	Takalammi
8	Likolampi	Härjänsilmä
9	Heinälampi	Korpilammi
10	Pahalampi	Tervalammi
11	Koiralampi	Koukkulammi
12	Pitkälampi	Haaralammi
13	Kangaslampi	Valkealammi
14	Kortelampi	Rapalammi
15	Vehkalampi	Hautalammi
16	Umpilampi	Laihalammi
17	Saarilampi	Rimminlammi
18	Syvälampi	Kiimalammi
19	Lehmilampi	Keskinenjärvi
20	Myllylampi	Kivilammi

Geographical distribution **Top 20 names**

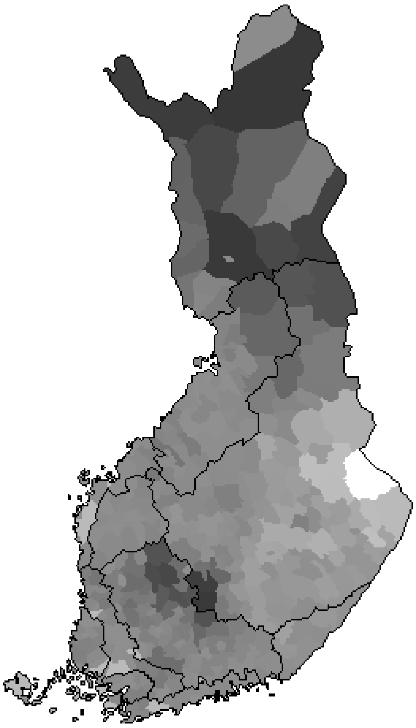
Figure 1: Lakes / Principal Component 1: 13 % of total variation



	Light	Dark
1	Kaakkolampi	Rytilampi
2	Pitkäjärvi	Hamppulampi
3	Mustalampi	Kaakkurilampi
4	Paskolampi	Raatelampi
5	Hirvijärvi	Kaivoslampi
6	Likolampi	Kokkolampi
7	Valkjärvi	Rimpilampi
8	Särkijärvi	Pikkulampi
9	Vuorilampi	Teerilampi
10	Vääräjärvi	Telkkälampi
11	Kalaton	Liejulampi
12	Haukilampi	Porolampi
13	Ahvenlampi	Salmilammit
14	Valkeajärvi	Latvalampi
15	Vehkalampi	Hanhilammit
16	Vuorijärvi	Koppelolampi
17	Saarijärvi	Takkulampi
18	Myllyjärvi	Konttilampi
19	Kärmelampi	Niittylampi
20	Pahalampi	Nuottilampi

Geographical distribution **Top 20 names**

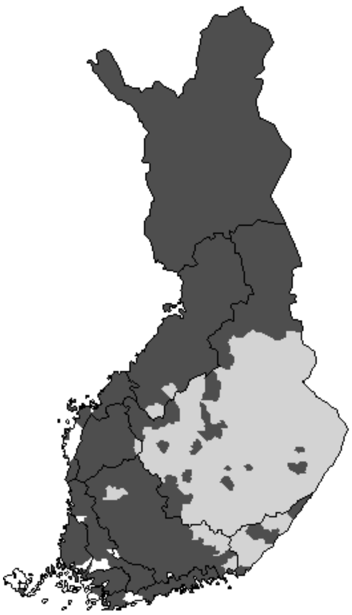
Figure 2: Lakes / Principal Component 2: 4 % of total variation



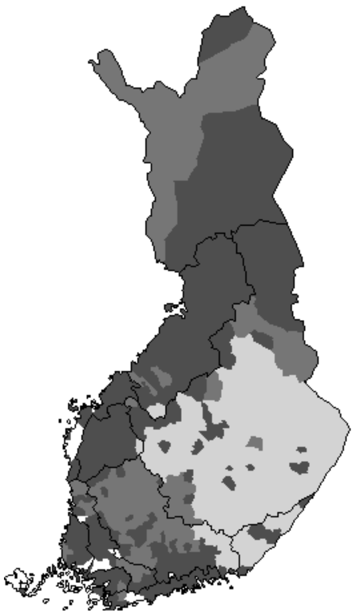
	Light	Dark
1	Likolampi	Vähäjärvi
2	Hepolampi	Särkijärvi
3	Valkealampi	Saarijärvi
4	Riihilampi	Haukijärvi
5	Aluslampi	Ahvenjärvi
6	Mustikkalampi	Syväjärvi
7	Valkeinen	Salmijärvi
8	Vehkalampi	Kalliojärvi
9	Valkeislampi	Kivijärvi
10	Väärälampi	Pitkäjärvi
11	Louhilampi	Mustajärvi
12	Sikolampi	Kaakkurilampi
13	Pieni Särkilampi	Kaitajärvi
14	Iso Valkeinen	Pirttijärvi
15	Pohjalampi	Latvajärvi
16	Kaatiolampi	Alajärvi
17	Lehmilampi	Paskolammi
18	Orilampi	Valkeajärvi
19	Pieni Heinälampi	Kotajärvi
20	Tetrilampi	Kortejärvi

Geographical distribution **Top 20 names**

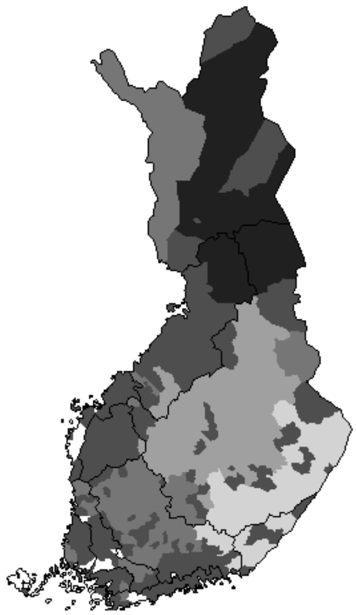
Figure 3: Lakes / Principal Component 3: 3 % of total variation



2 clusters based on 3 PC's

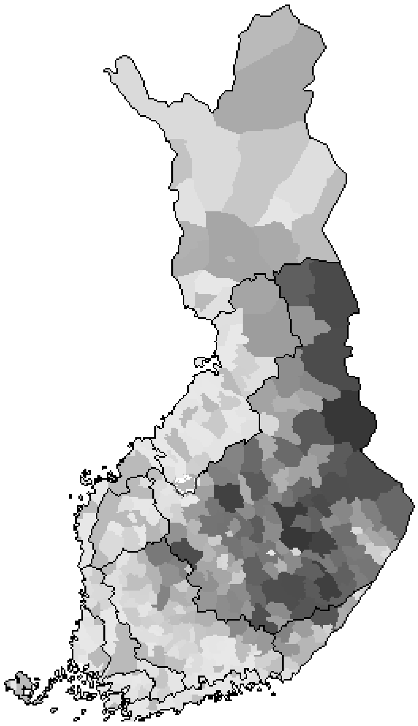


3 clusters based on 4 PC's



5 clusters based on 6 PC's

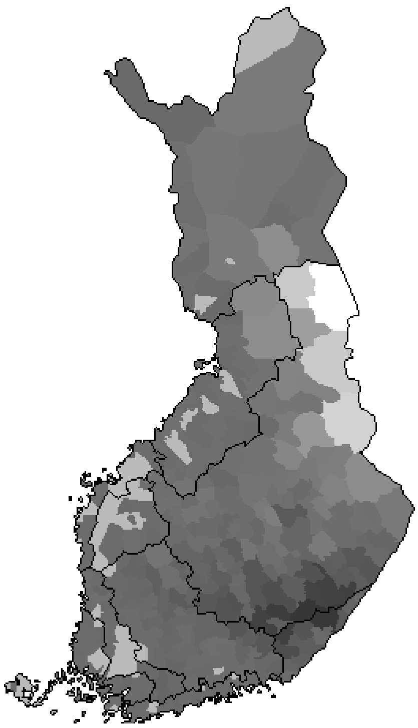
Figure 4: Lakes / Clusters



	Light	Dark
1	Isoperä	Mustalahti
2	Santaviiki	Pitkälahti
3	Joutavalahti	Likolahti
4	Hakalanlahti	Savilahti
5	Keinolahti	Levälahti
6	Salmenperä	Syvälahti
7	Pohislahti	Kylmälahti
8	Hietaperä	Saunalahti
9	Loukaslahti	Myllylahti
10	Letonlahti	Jokilahti
11	Soukanpohja	Hietalahti
12	Luusua	Kortelahti
13	Mustaperä	Kotalahti
14	Kaakkurilahti	Kivilahti
15	Maijanlahti	Riihilahti
16	Ojalanlahti	Talvilahti
17	Ruonanlahti	Suolahti
18	Vaarinlahti	Tervalahti
19	Korvensalmi	Laajalahti
20	Lepistönlahti	Haukilahti

Geographical distribution **Top 20 names**

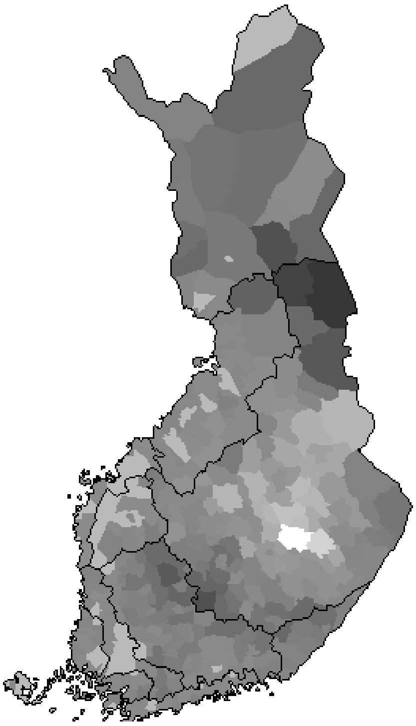
Figure 5: Parts of Lakes / Principal Component 1: 15 % of total variation



	Light	Dark
1	Itälahti	Leppälahti
2	Kokkolahti	Uitonsalmi
3	Niskalahti	Sammallahti
4	Kuikkalahti	Niittulahti
5	Rytilahti	Vuorilahti
6	Kumpulahti	Leviälahti
7	Juurikkalahti	Pitkänpohjanlahti
8	Pitkäperä	Majalahti
9	Autiolahti	Taipaleenlahti
10	Lammaslahti	Soukanlahti
11	Etelälahti	Haapalahti
12	Talvilahti	Kärmelahti
13	Palolahti	Myllylahti
14	Hoikkalahti	Mutalahti
15	Luodelahti	Hirvilahti
16	Teerilahti	Lehmälahti
17	Ahvenlahti	Katiskalahti
18	Pikkulahti	Lähdelahti
19	Vaaralahti	Ruokosalmi
20	Isolahti	Koilahti

Geographical distribution **Top 20 names**

Figure 6: Parts of Lakes / Principal Component 2: 3 % of total variation

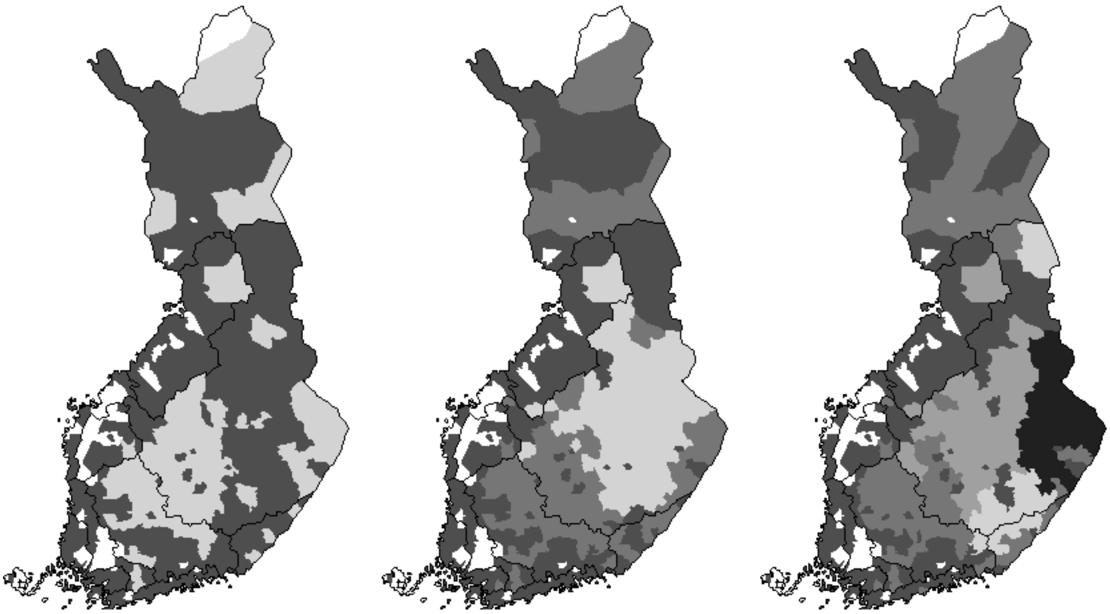


Geographical distribution

	Light	Dark
1	Pohjoislahti	Leveälahti
2	Tulilahti	Kotalahti
3	Itälahti	Syvälahti
4	Laajalahti	Myllylahti
5	Hiekkalahti	Isolahti
6	Etelälahti	Savilahti
7	Laajanlahti	Mustalahti
8	Hiekkakaarre	Pikkulahti
9	Ruokolahti	Isoselkä
10	Sammakkolahti	Kirkkolahti
11	Jokilahti	Takalahti
12	Luodelahti	Kylmälahti
13	Levälahti	Isosalmi
14	Viitalahti	Hietalahti
15	Tulisalmi	Haapalahti
16	Kangaslahti	Hangaslahti
17	Kylmäkaarre	Niittulahti
18	Jynkänlahti	Kutulahti
19	Kannaslahti	Santalahti
20	Murtolahti	Likalahti

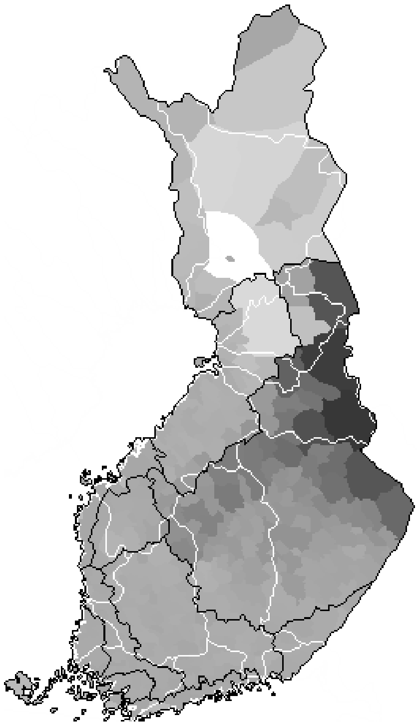
Top 20 names

Figure 7: Parts of Lakes / Principal Component 3: 2 % of total variation



2 clusters based on 4 PC's | **3 clusters based on 4 PC's** | **5 clusters based on 6 PC's**

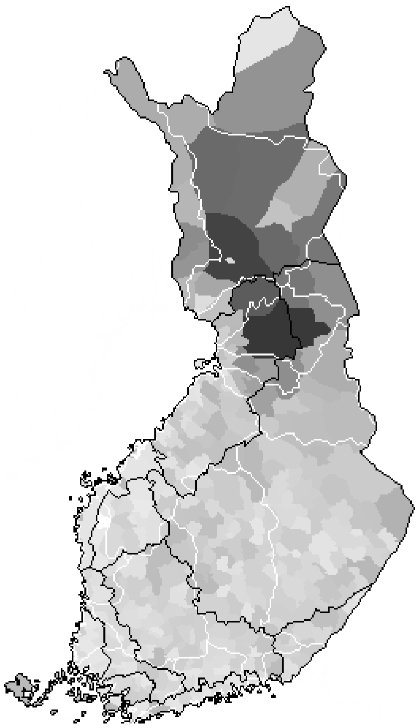
Figure 8: Parts of Lakes / Clusters



	Light	Dark
1	Myllypuro	Myllyoja
2	Kylmäpuro	Pahaoja
3	Kivipuro	Mustaoja
4	Mustapuro	Särkioja
5	Tervapuro	Kivioja
6	Vehkapuro	Hanhioja
7	Haarapuro	Välioja
8	Kortepuro	Karhuoja
9	Välipuro	Saukko.oja
10	Kalliopuro	Peuraoja
11	Koirapuro	Korteoja
12	Heinäpuro	Ruosteoja
13	Ruunapuro	Palo.oja
14	Hepopuro	Sammaloja
15	Välijoki	Pikkuoja
16	Pajupuro	Rytioja
17	Myllyjoki	Hirvioja
18	Karhupuro	Hirvasoja
19	Haukipuro	Säynäjäoja
20	Palopuro	Lammasoja

Geographical distribution **Top 20 names**

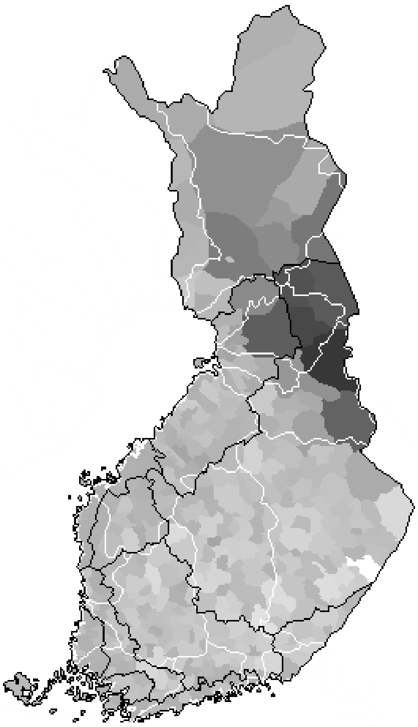
Figure 9: Rivers / Principal Component 1: 10 % of total variation



	Light	Dark
1	Sarviluoma	Myllyoja
2	Kiviluoma	Välijoki
3	Korpiluoma	Kivioja
4	Koivuluoma	Pahaoja
5	Saaranoja	Välioja
6	Varsanoja	Myllypuro
7	Vehkaluoma	Mustaoja
8	Päkinoja	Kylmäoja
9	Heiniluoma	Syväoja
10	Kivisoja	Myllyjoki
11	Varsoja	Korteoja
12	Kissanoja	Särkioja
13	Teyripuro	Saukko.oja
14	Isonnevanoja	Kivipuro
15	Koskutjoki	Ahvenoja
16	Pitkänsillanoja	Kaivosoja
17	Kuusjoki	Hanhioja
18	Saarenoja	Rytioja
19	Kohisevanjoki	Alajoki
20	Kylänjoki	Kylmäpuro

Geographical distribution **Top 20 names**

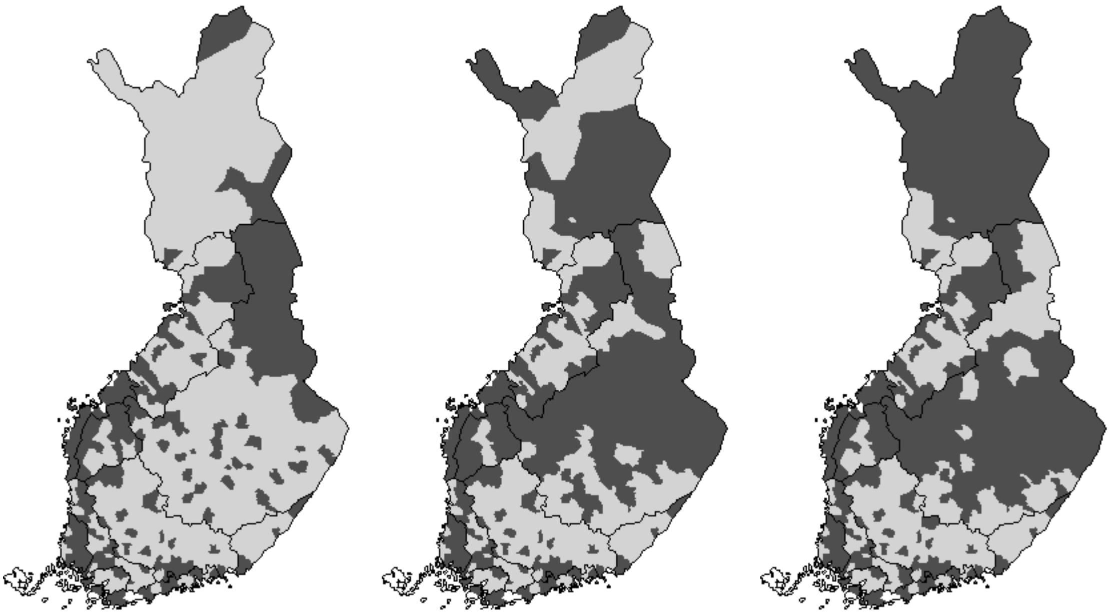
Figure 10: Rivers / Principal Component 2: 5 % of total variation



	Light	Dark
1	Myllyoja	Saarijoki
2	Myllypuro	Rytioja
3	Myllyjoki	Syrjäpuro
4	Kylmäoja	Ahvenoja
5	Vehkaoja	Latvajoki
6	Rajapuro	Kotipuro
7	Välijoki	Säynäjäjoki
8	Korvenoja	Mätäspuro
9	Kolunoja	Korteoja
10	Haapajoki	Syrjäoja
11	Sahinjoki	Säynäjäoja
12	Heinäjoki	Heteoja
13	Kivioja	Saukko.oja
14	Tervapuro	Konttipuro
15	Kolisevanoja	Lehto.oja
16	Alhonoja	Käärmepuro
17	Syväoja	Kotijoki
18	Rajajoki	Raatepuro
19	Vehkapuro	Ahvenpuro
20	Kukkopuro	Salmijoki

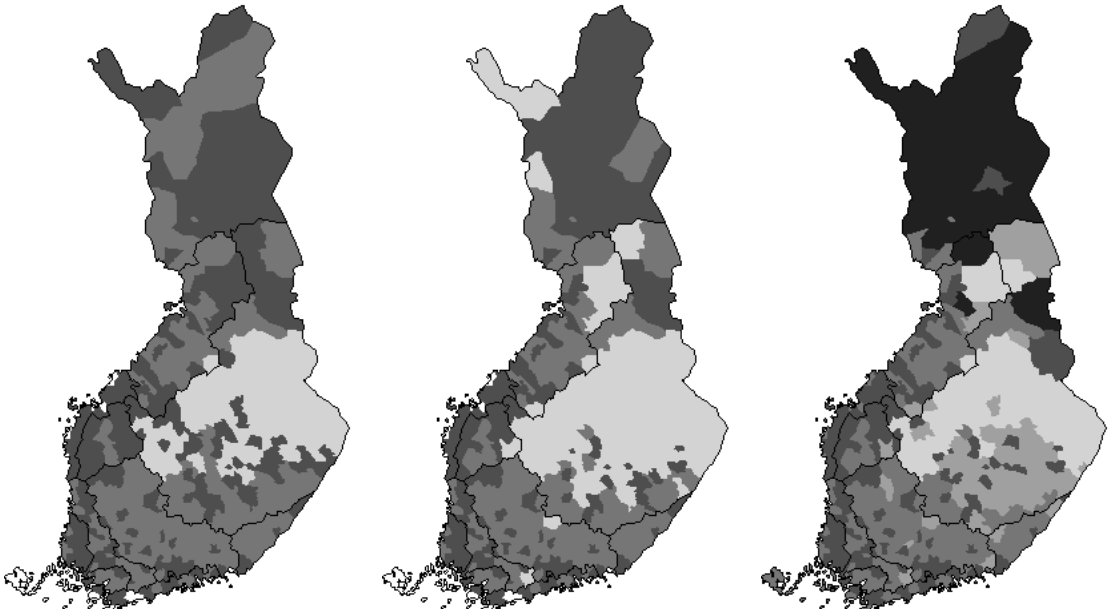
Geographical distribution **Top 20 names**

Figure 11: Rivers / Principal Component 3: 3 % of total variation



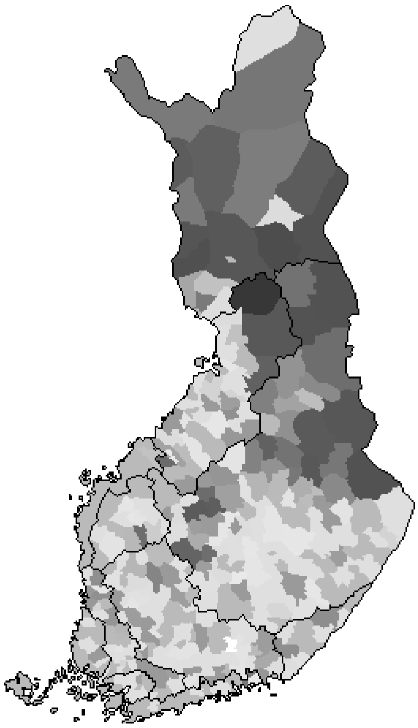
2 clusters based on 3 PC's **2 clusters based on 4 PC's** **2 clusters based on 7 PC's**

Figure 12: Rivers / 2 Clusters



3 clusters based on 4 PC's | **3 clusters based on 7 PC's** | **5 clusters based on 7 PC's**

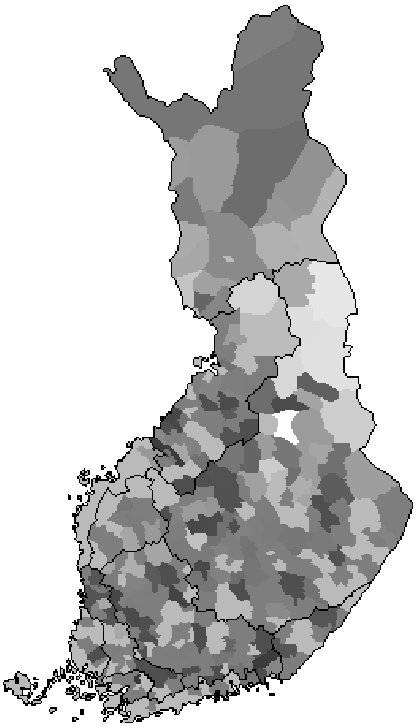
Figure 13: Rivers / 3—5 Clusters



	Light	Dark
1	Kissakoski	Myllykoski
2	Vuorikoski	Pitkäkoski
3	Härkäkoski	Saarikoski
4	Louhenkoski	Kalliokoski
5	Sahinkoski	Vääräkoski
6	Leppäkoski	Niskakoski
7	Heinäkoski	Suukoski
8	Louhikoski	Alakoski
9	Kuuskoski	Korpikoski
10	Lapinkoski	Koivukoski
11	Katajakoski	Jyrkkäkoski
12	Vääräkkä	Kattilakoski
13	Kellokoski	Haarakoski
14	Sahankoski	Siikakoski
15	Ruuhikoski	Palokoski
16	Kärppäkoski	Petäjäkoski
17	Ruukinkoski	Taivalkoski
18	Kivikoski	Tuomikoski
19	Porraskoski	Tammikoski
20	Keskikoski	Siltakoski

Geographical distribution | **Top 20 names**

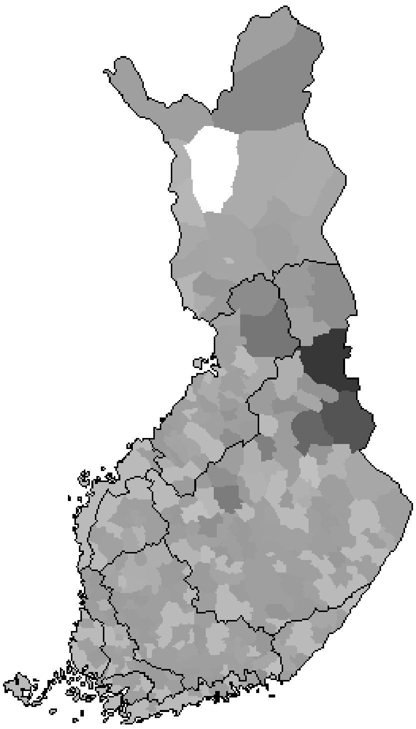
Figure 14: Rapids / Principal Component 1: 16 % of total variation



	Light	Dark
1	Pitkäkoski	Myllykoski
2	Saarikoski	Kissakoski
3	Niskakoski	Sahankoski
4	Taivalkoski	Tervakoski
5	Alakoski	Härkäkoski
6	Suukoski	Lapinkoski
7	Jyrkkäkoski	Tamppikoski
8	Vääräkoski	Kuuskoski
9	Korpikoski	Sahinkoski
10	Kurjenkoski	Lammaskoski
11	Patokoski	Hirvikoski
12	Kalliokoski	Haapakoski
13	Tammikoski	Keskikoski
14	Haarakoski	Ahokoski
15	Yläkoski	Vuorikoski
16	Koivukoski	Vääräkkä
17	Peurakoski	Porraskoski
18	Kattilakoski	Välikoski
19	Kotakoski	Ruukinkoski
20	Vääränkoski	Leppäkoski

Geographical distribution **Top 20 names**

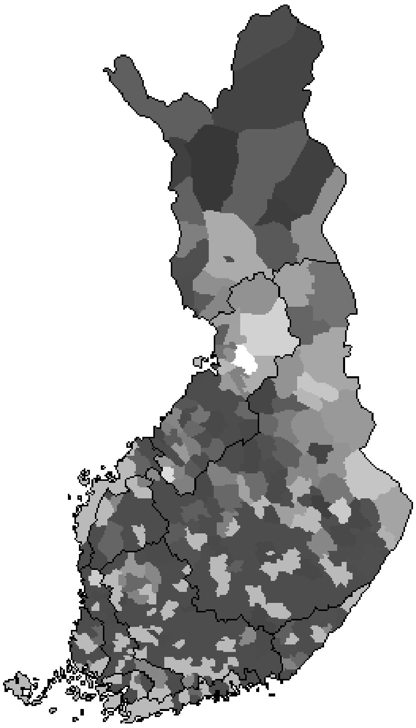
Figure 15: Rapids / Principal Component 2: 6 % of total variation



	Light	Dark
1	Kalliokoski	Niskakoski
2	Pitkäkoski	Korpikoski
3	Murtokoski	Alakoski
4	Kattilakoski	Taivalkoski
5	Mustakoski	Tammikoski
6	Vääräkkä	Koivukoski
7	Jyrkkäkoski	Leppikoski
8	Pahtakoski	Kaivoskoski
9	Tuomikoski	Palokoski
10	Ruuhikoski	Siikakoski
11	Kivikoski	Nahkakoski
12	Korkeakoski	Konttikoski
13	Porraskoski	Pajukoski
14	Ahokoski	Jäniskoski
15	Suukoski	Kokkokoski
16	Sahakoski	Saunakoski
17	Pirttikoski	Siltakoski
18	Pahakoski	Rajakoski
19	Köngäs	Vääränkoski
20	Hirvikoski	Karhukoski

Geographical distribution **Top 20 names**

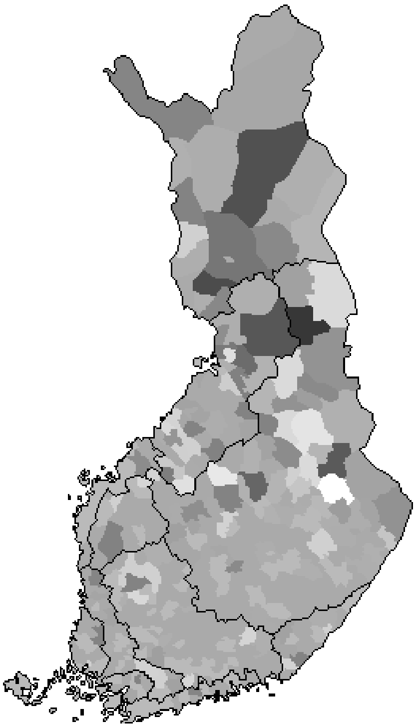
Figure 16: Rapids / Principal Component 3: 5 % of total variation



	Light	Dark
1	Myllykoski	Peurasuvanto
2	Pitkäkoski	Kutuniva
3	Saarikoski	Saarisuvanto
4	Vääräkoski	Sahi
5	Kalliokoski	Syväsalmi
6	Korpikoski	Joenpolvi
7	Pitkäsuvanto	Polvikoski
8	Alakoski	Pajukoski
9	Louhikoski	Mukkakoski
10	Pahkakoski	Lampare
11	Kattilakoski	Alasuvanto
12	Niskakoski	Jokilampi
13	Jyrkkäkoski	Myllysuvanto
14	Hanhikoski	Haarakoski
15	Honkakoski	Hietakoski
16	Leppikoski	Mustalahti
17	Pikkukoski	Korkeakoski
18	Pirttikoski	Kuivakoski
19	Hautakoski	Savilahti
20	Rosvohotu	Haapakoski

Geographical distribution **Top 20 names**

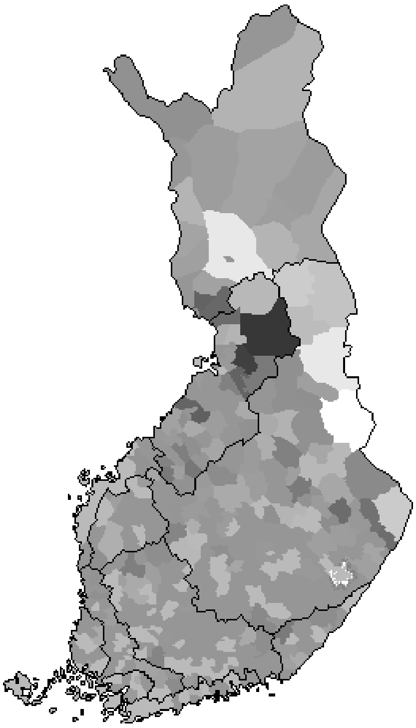
Figure 17: Other Parts of Rivers / Principal Component 1: 13 % of total variation



	Light	Dark
1	Myllykoski	Pitkäkoski
2	Korkeakoski	Saarikoski
3	Lampare	Jyrkkäkoski
4	Hanhisuvanto	Mustasuvanto
5	Ylisuvanto	Kalliokoski
6	Jokipolvi	Palokoski
7	Hautakoski	Pitkäsuvanto
8	Sahi	Louhikoski
9	Pyörre	Petäjäkoski
10	Syväsalmi	Rännikoski
11	Taivalkoski	Leppikoski
12	Koskelankoski	Suukoski
13	Tamppikoski	Patokoski
14	Koivukoski	Karhukoski
15	Kattilakoski	Pikkukoski
16	Sahakoski	Kuivakoski
17	Polvikoski	Pahkakoski
18	Alakoski	Mukkakoski
19	Honkakoski	Vääräkoski
20	Saarisuvanto	Niskakoski

Geographical distribution **Top 20 names**

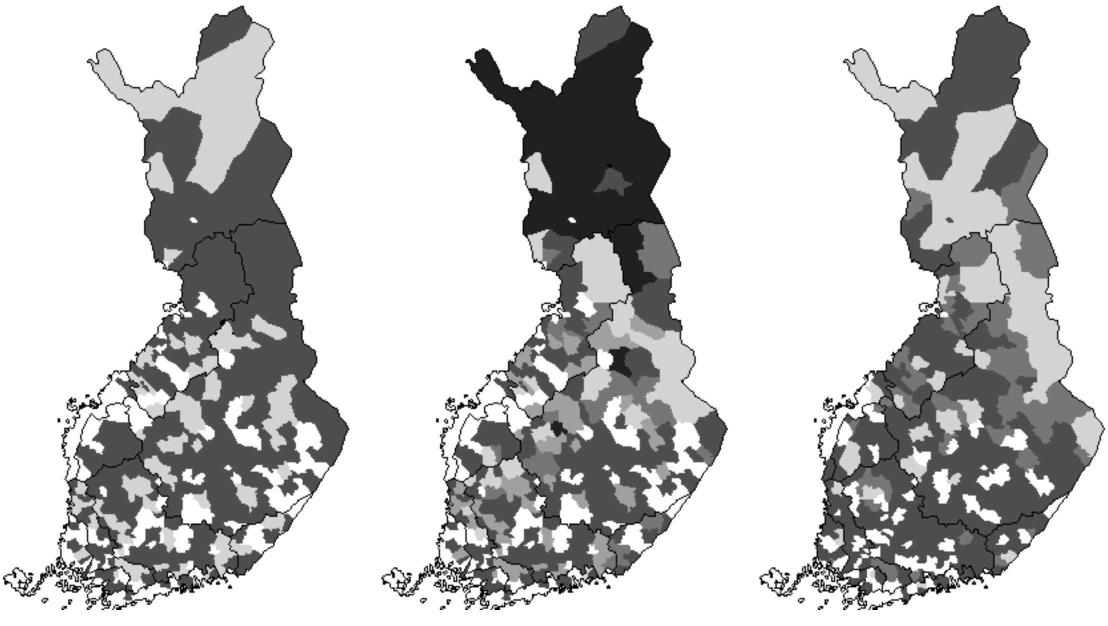
Figure 18: Other Parts of Rivers / Principal Component 2: 7 % of total variation



	Light	Dark
1	Pitkäsuvanto	Kalliokoski
2	Pitkäkoski	Vääräkoski
3	Kattilakoski	Palokoski
4	Myllykoski	Alakoski
5	Suvanto	Karhukoski
6	Mustalahti	Koivukoski
7	Myllysuvanto	Korpikoski
8	Niva	Pahkakoski
9	Rosvohotu	Pikkukoski
10	Patokoski	Haapakoski
11	Lohikoski	Leppikoski
12	Saarisuvanto	Koirakoski
13	Pyörre	Aittokoski
14	Lampare	Hautakoski
15	Siltakoski	Jyrkkäkoski
16	Kuivakoski	Suukoski
17	Niskakoski	Koskelankoski
18	Alasuvanto	Savilahti
19	Mukkakoski	Saarikoski
20	Mustakoski	Honkakoski

Geographical distribution **Top 20 names**

Figure 19: Other Parts of Rivers / Principal Component 3: 6 % of total variation



Rapids:
2 clusters
based on 7 PC's

Rapids:
5 clusters
based on 7 PC's

Other:
3 clusters
based on 3 PC's

Figure 20: Parts of Rivers / Clusters