Computational Overview of Finnish Hydronyms

http://www.cs.helsinki.fi/u/leino/jutut/riga-04/

Antti Leino

leino@cs.helsinki.fi

Helsinki Institute for Information Technology
Basic Research Unit

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 presentation 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; others are less so.

Introduction

- Finnish National Land Survey Place Name Register

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

- How to compile a simple, easy-to-read overview?
- Traditional distribution maps won't work: too many names

The National Land Survey of Finland has, for its own purposes of producing maps, a Geographicn 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* meaning those names that appear on at least ten or five municipalties. The number of names that fulfill this criterion is shown on the slide.

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. Where dialectometric researchers has 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.

Principal Component Analysis

- Curse of dimensionality — how to reduce the number of variables
- PCA: transform the data to get underlying components
  - not correlated
  - ordered by decreasing variation
- So principal component #1 is the most significant one, &c.
- Can be used to reduce noise: make further analysis on the first few components

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 the variation.

In practice this means that often the first few principal components can 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.

Cluster Analysis

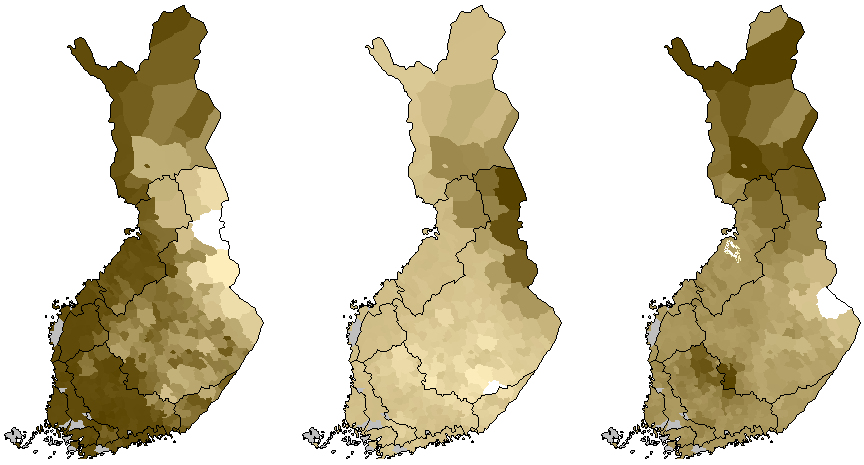
- Main goal: divide data to sections, called clusters, so that
  - items in same cluster as similar as possible
  - items in different clusters as different as possible
- Hierarchical vs. partitioning methods
- Hierarchical clustering usually not very robust
- Optimal partitioning not feasible, but approximations possible
- Here: partitioning based on a few principal components.

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 items in the same cluster are similar to each other and items in different clusters are different from each other. Clustering methods are commonly divided in two. In hierarchical (often called also agglomerative or joining) 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 partitioning (also called divisive) methods, on the other hand, the data is divided to a specified number of clusters.

One problem with hierarchical clustering, especially with data like 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. With partitioning methods, on the other hand, the typical difficulty is that one has to know — or guess — the number of clusters in advance.

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. On the other hand, 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).

Lakes: Principal Components



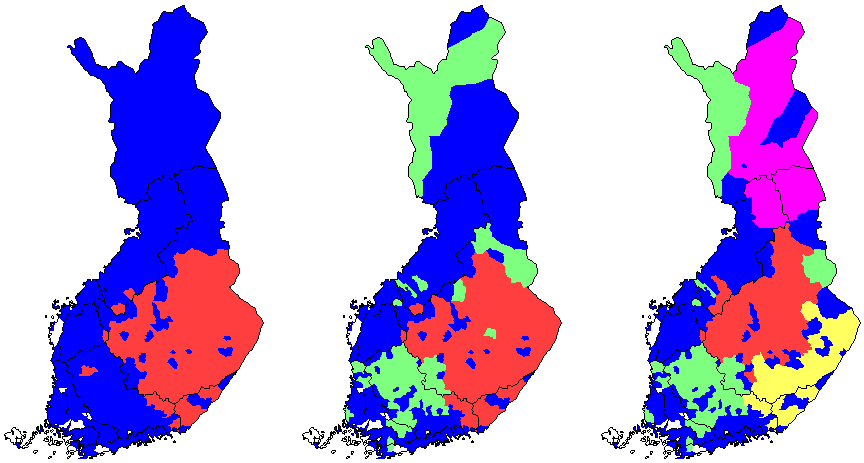
Component 1	Component 2	Component 3
13 % of variation	4 % of variation	3 % of variation

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 show the weights of each municipality in the first three components, drawn in shades of brown on a map with main dialectal divisions shown as black lines. 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 can be viewed as expected, as the East—West division is the most fundamental one in Finnish dialects. The second component is rather less so, 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 third component seems again linguistically related, as there were historical contacts between the two regions where the component peaks.

Lakes: Clusters



2 clusters
based on 3 PC's

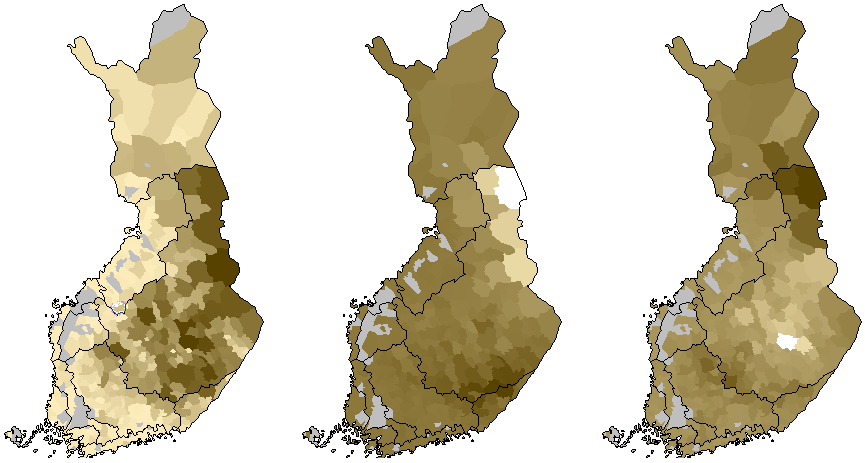
3 clusters
based on 4 PC's

5 clusters
based on 6 PC's

A two-way clustering based on the first three components results in a division of Finland into the Eastern region, in red, and the Western one, in blue. 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 in green, and on the other hand the rest. Later on Lapland, shown in purple, splits off from the Western cluster and the Eastern cluster splits into the old provinces, in yellow, and the region that was settled in the 17th century, in red.

The next two slides 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.

Parts of Lakes: Principal Components

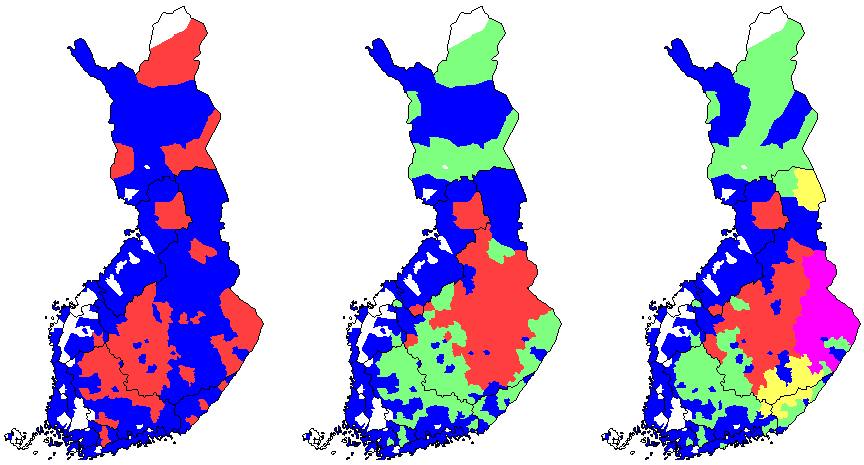


Component 1
15 % of variation

Component 2
3 % of variation

Component 3
2 % of variation

Parts of Lakes: Clusters

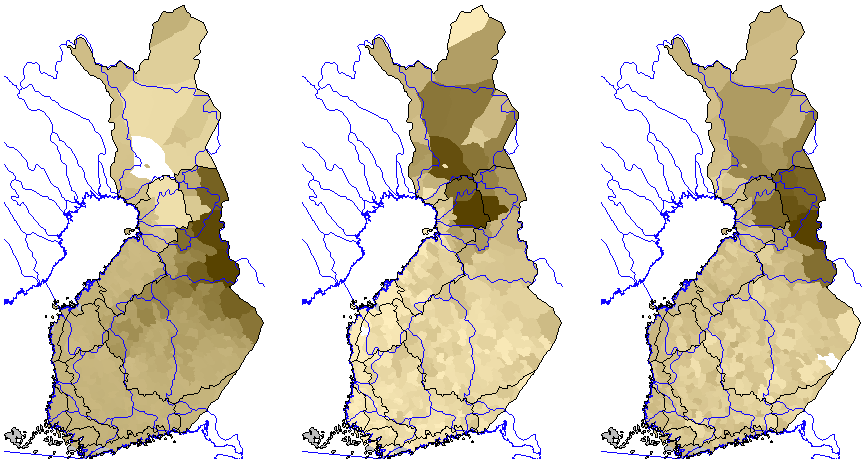


2 clusters
based on 4 PC's

3 clusters
based on 4 PC's

5 clusters
based on 6 PC's

Rivers: Principal Components



Component 1	Component 2	Component 3
10 % of variation	5 % of variation	3 % of variation

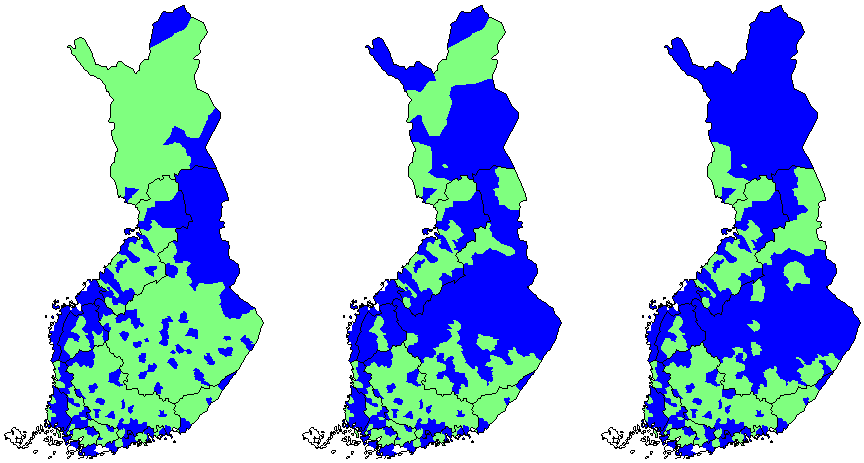
The maps showing principal components of river names show also drainage basins. One can see that the first principal component appears to be correlated on whether the municipality is up- or downriver. The second component is concentrated on the basins of the Oulu and Kemi rivers, or more generally in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

The next two slides show clusterings based on river name components. The first slide 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 green, would seem to include the northern Bothnia and Kainuu in addition to the traditionally settled regions in the south.

The three-way clustering is starting to look somewhat more understandable: the old hunting regions appear as a separate cluster in red, the old agricultural lands in the south as another in green, and the coastal regions as the third one in blue. In the five-way clustering Lapland and the old Savolax separate as the purple and yellow 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.

Rivers: 2 Clusters

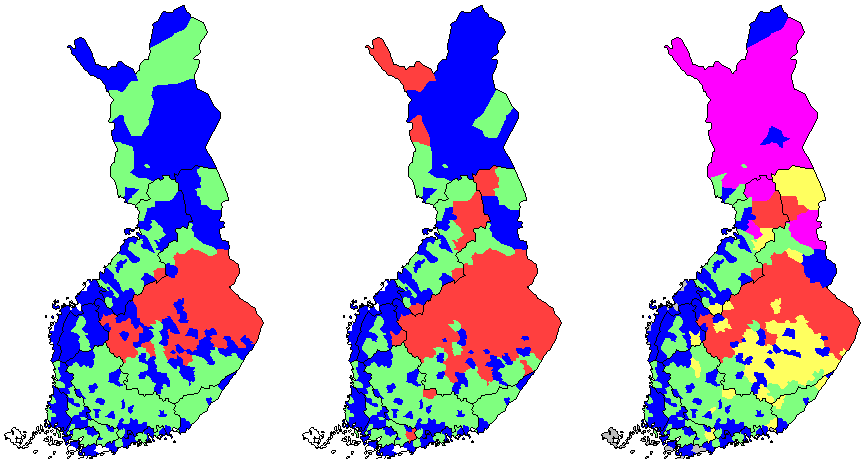


2 clusters based on 3 PC's

2 clusters based on 4 PC's

2 clusters based on 7 PC's

Rivers: More Clusters

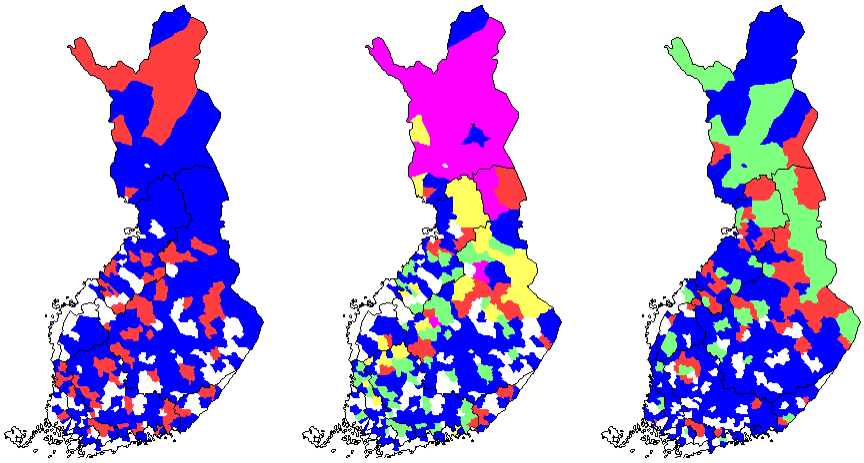


3 clusters based on 4 PC's

3 clusters based on 7 PC's

5 clusters based on 7 PC's

Parts of Rivers



Rapids:	Rapids:	Other:
2 clusters	5 clusters	3 clusters
based on 7 PC's	based on 7 PC's	based on 3 PC's

The data sets of names of rapids and other parts of rivers were so small that the analyses 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.

Conclusion

- The method appears to work with large amounts of data
- With smaller data sets (such as the parts of rivers) results are not good
  - Is this a problem in the method, or is it just that there is no overall structure?
- In lake names the primary components (and clusters) follow dialectal regions
- River names are different
  - Traces of old hunting culture ?
  - Distribution of natural features ?

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. 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? The reason may also be simply in the inaccuracies of the data, but at this point some further study would seem to be warranted.

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 der Wissenschaften.

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.

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

List of Slides