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

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

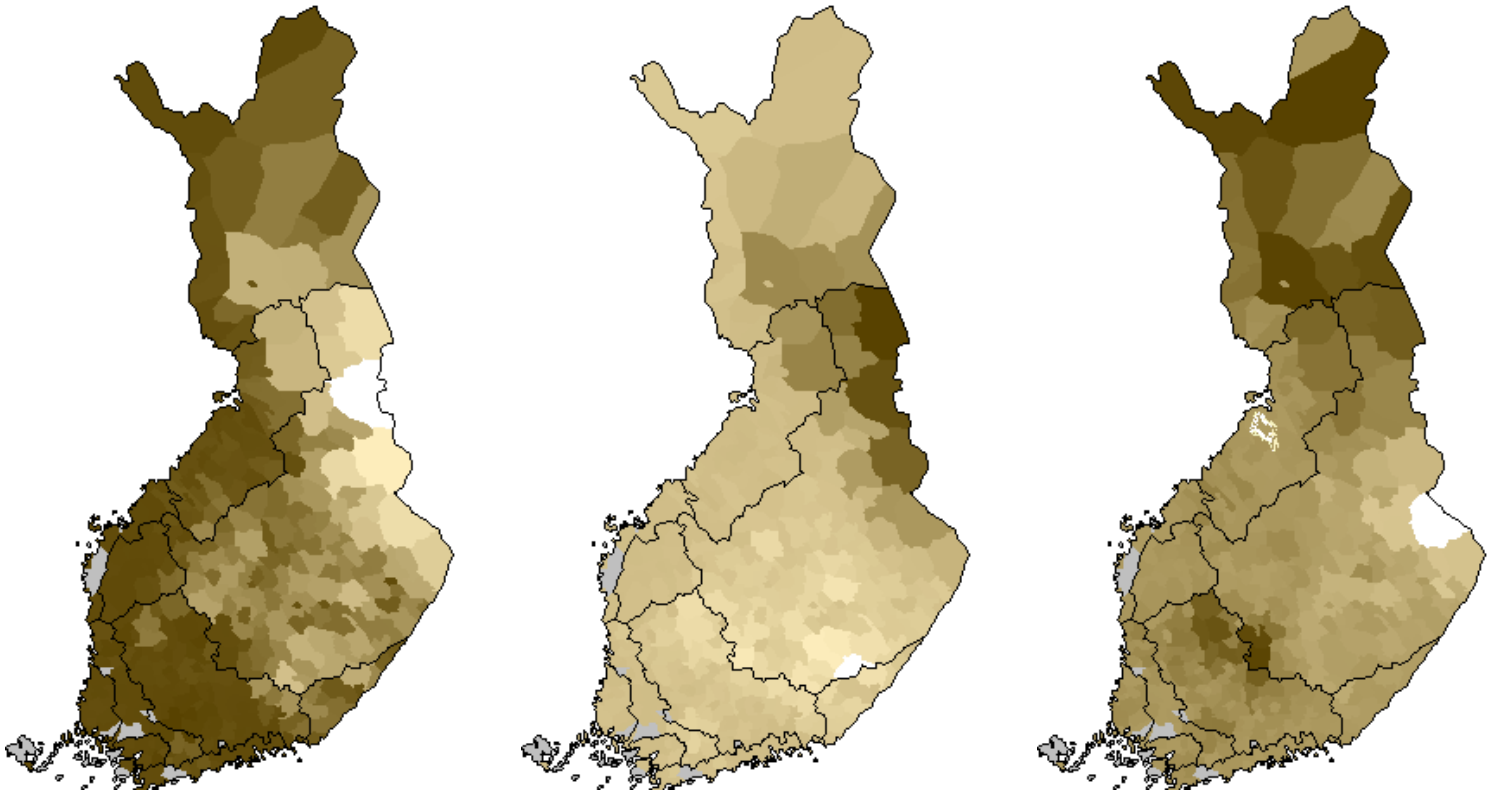
Principal Component Analysis

- 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

Cluster Analysis

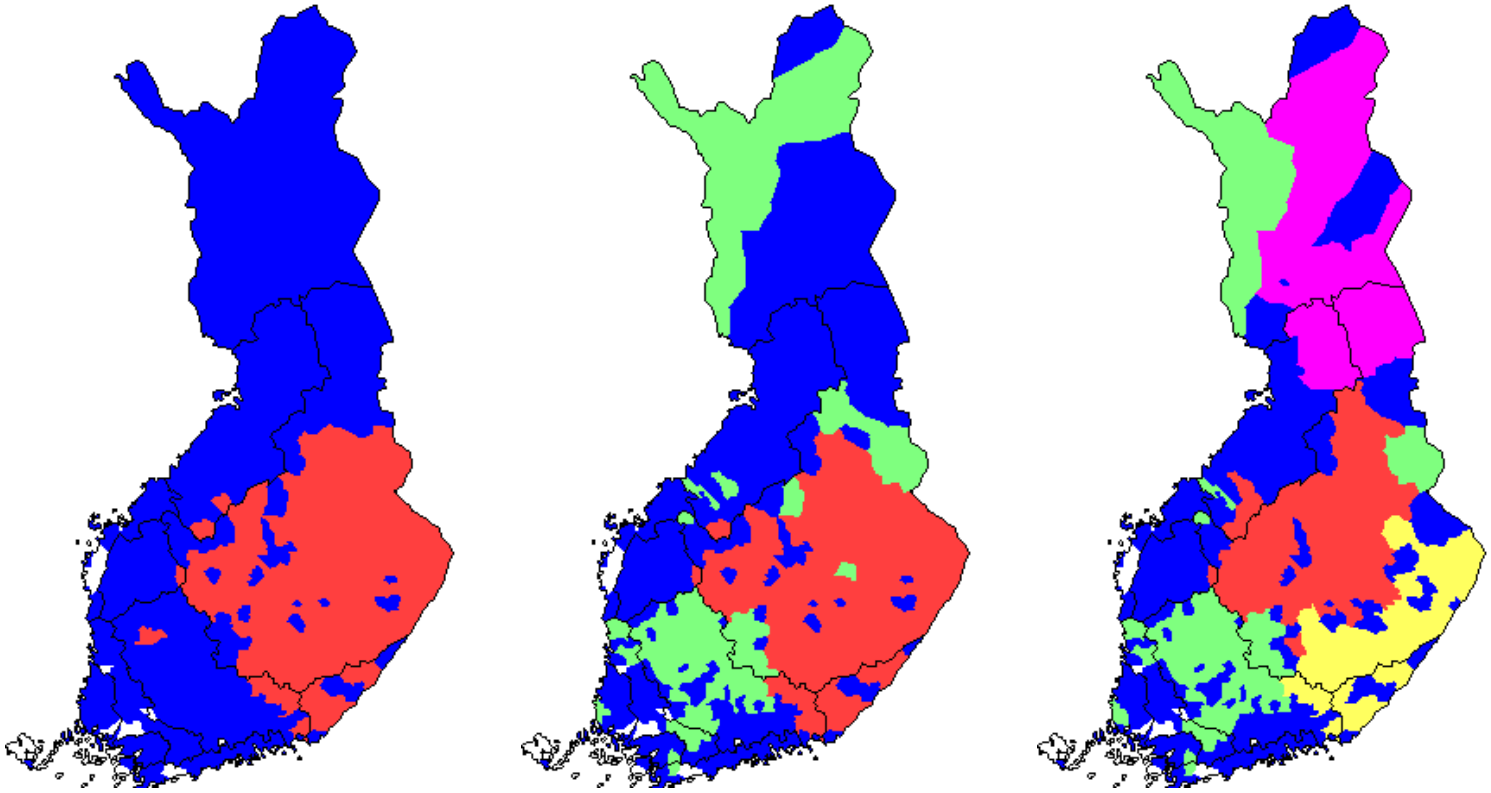
- Main goal: divide data to sections, called clusters, so that
  - items in same cluster as similar as possible
  - items in different clusters as different as possible
- Hierarchical vs. partitioning methods
- Hierarchical clustering usually not very robust
- Optimal partitioning not feasible, but approximations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

Lakes: Principal Components



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

Lakes: Clusters



2 clusters

based on 3 PC’s

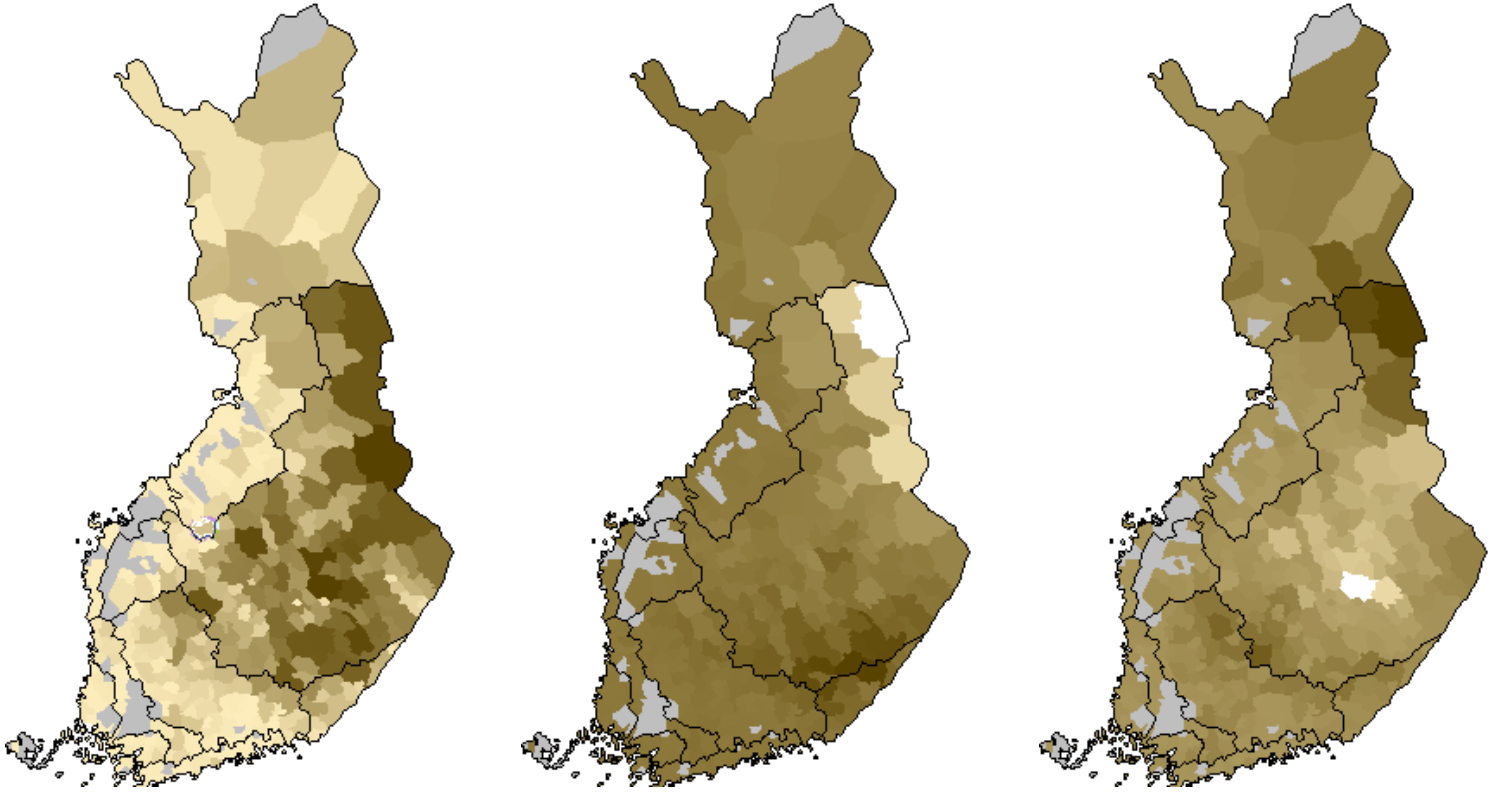
3 clusters

based on 4 PC’s

5 clusters

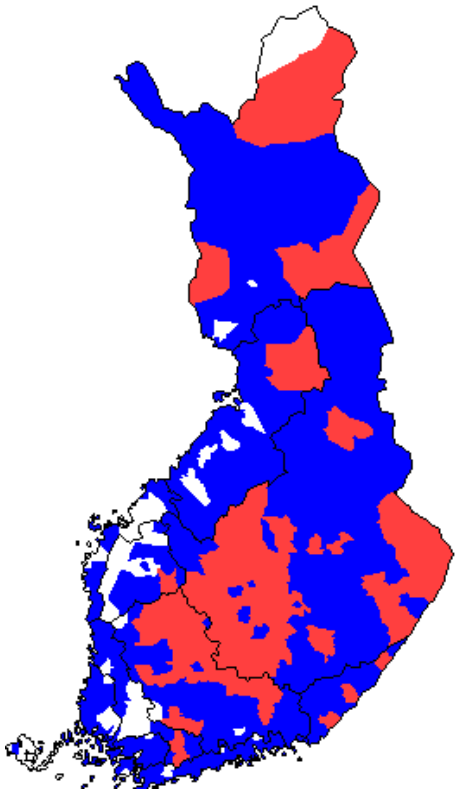
based on 6 PC’s

Parts of Lakes: Principal Components



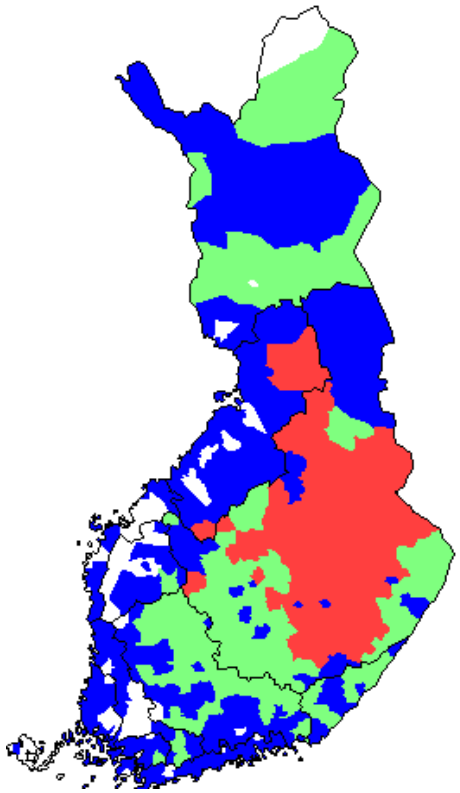
Component 1	Component 2	Component 3
15 % of variation	3 % of variation	2 % of variation

Parts of Lakes: Clusters



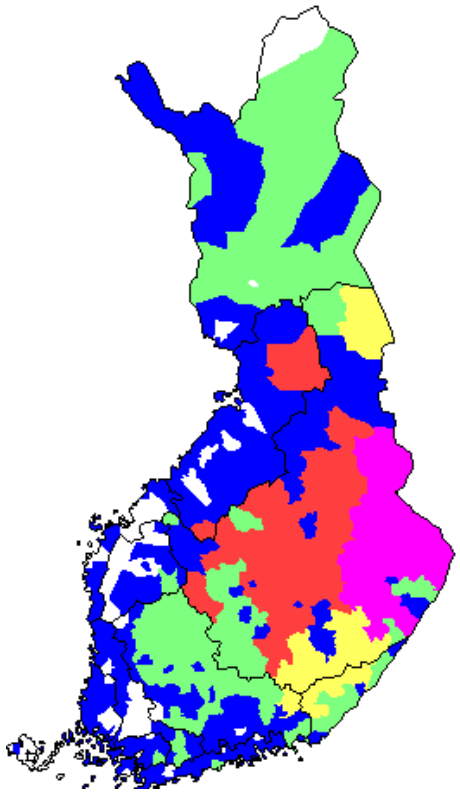
2 clusters

based on 4 PC's



3 clusters

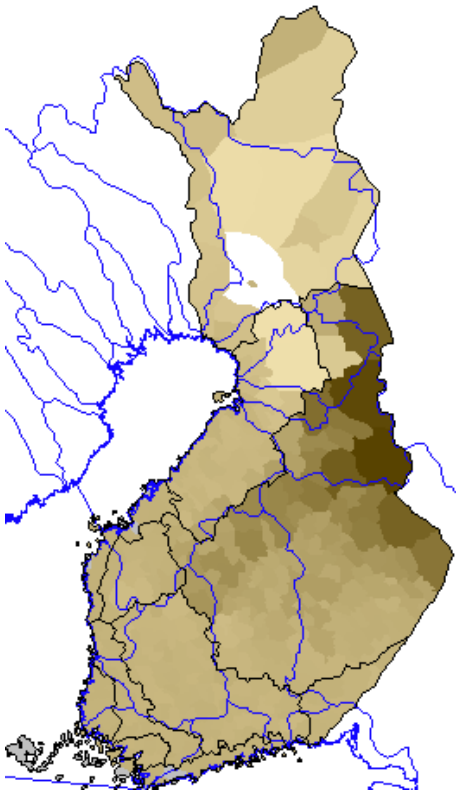
based on 4 PC's



5 clusters

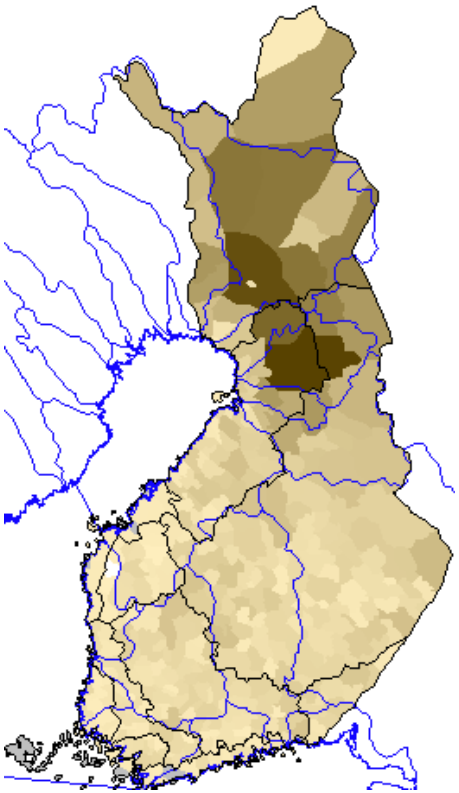
based on 6 PC's

Rivers: Principal Components



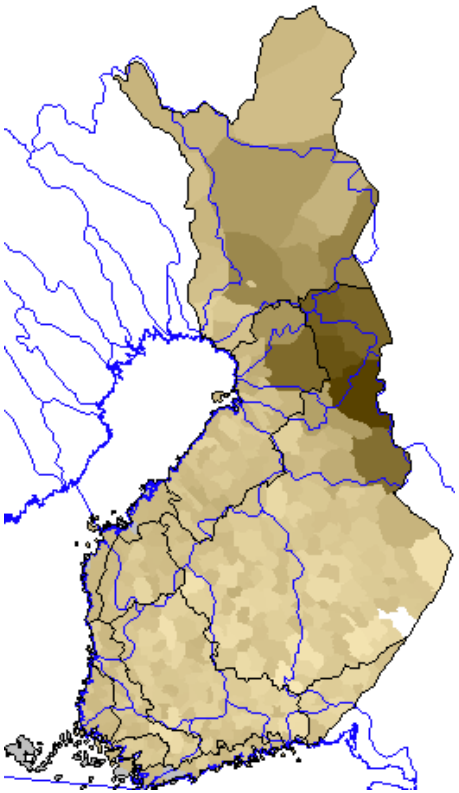
Component 1

10 % of variation



Component 2

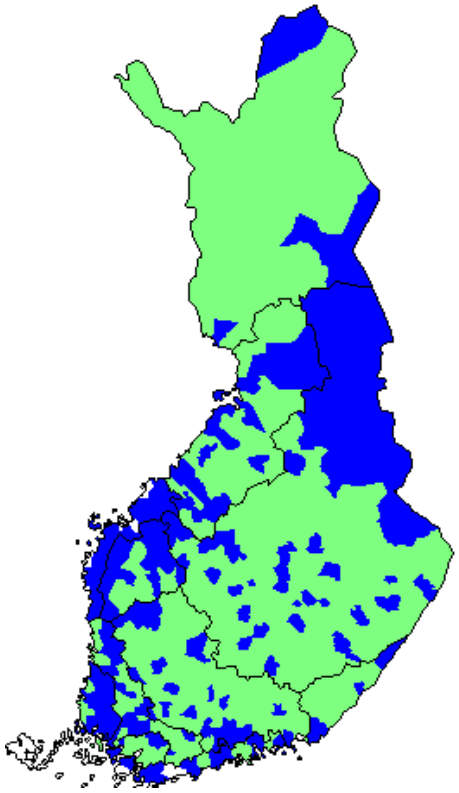
5 % of variation



Component 3

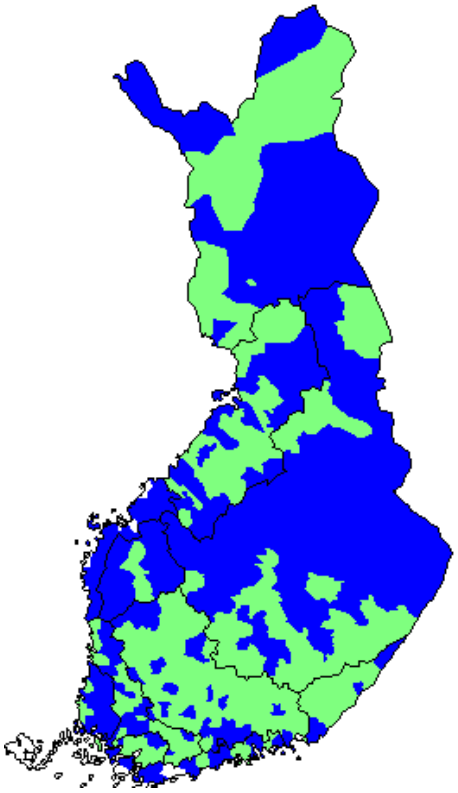
3 % of variation

Rivers: 2 Clusters



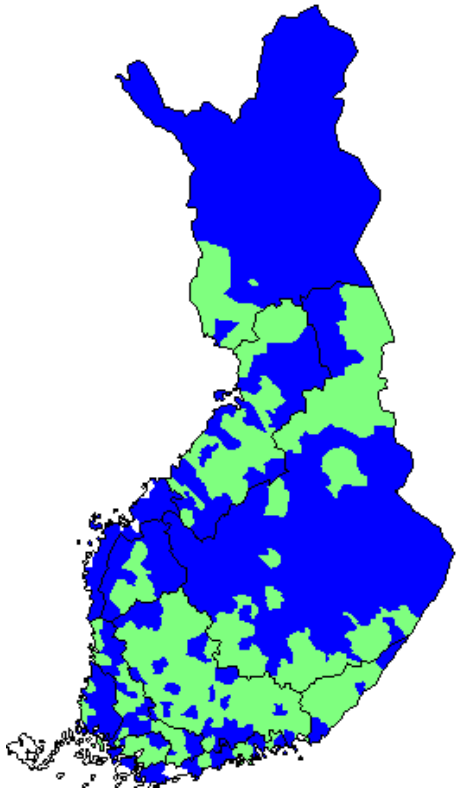
2 clusters

based on 3 PC's



2 clusters

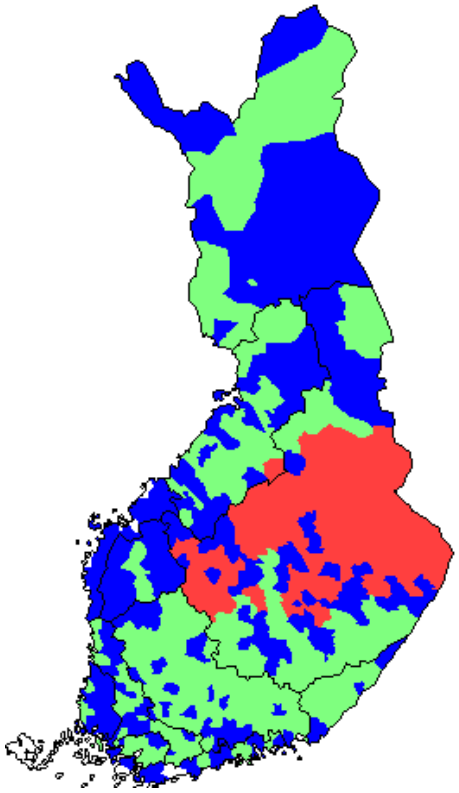
based on 4 PC's



2 clusters

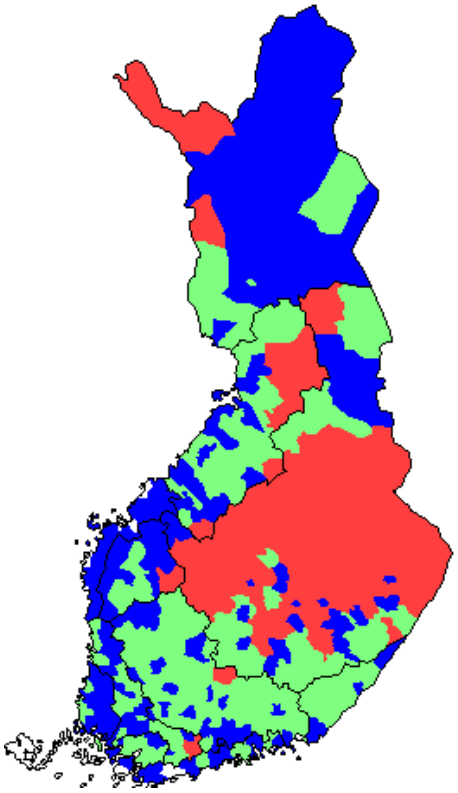
based on 7 PC's

Rivers: More Clusters



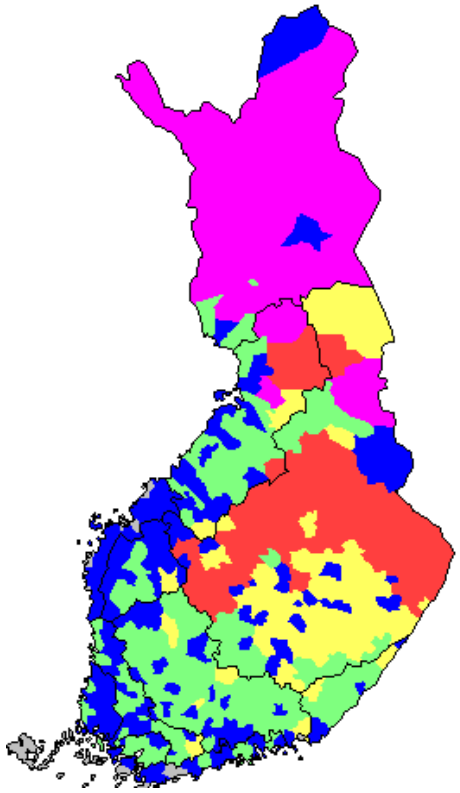
3 clusters

based on 4 PC's



3 clusters

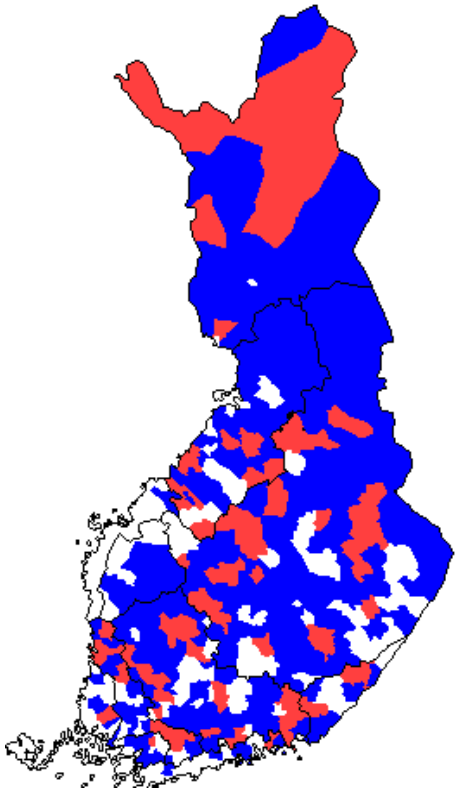
based on 7 PC's



5 clusters

based on 7 PC's

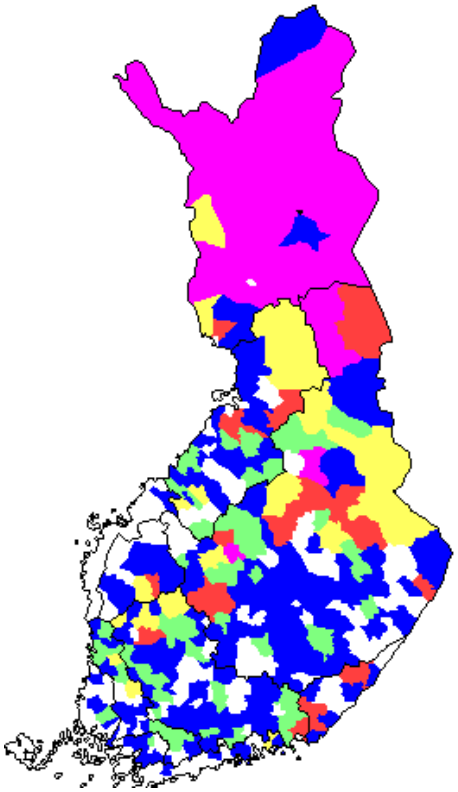
Parts of Rivers



Rapids:

2 clusters

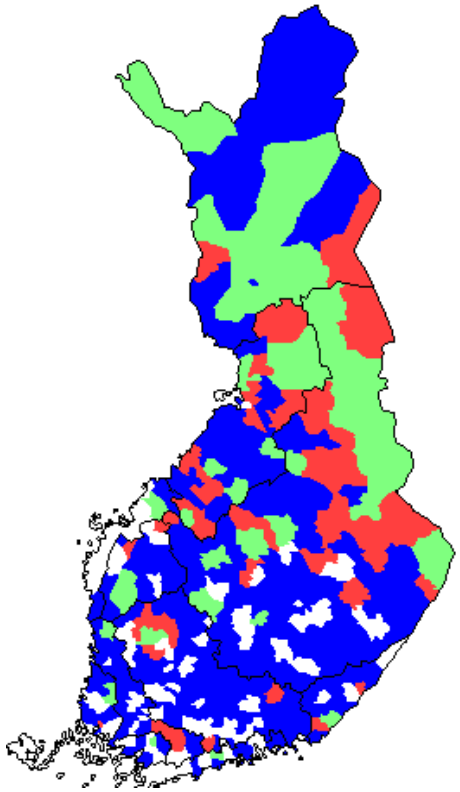
based on 7 PC’s



Rapids:

5 clusters

based on 7 PC’s



Other:

3 clusters

based on 3 PC’s

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 ?