Spatial data mining as an onomastic tool

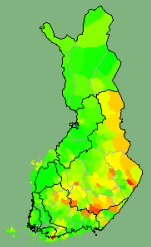
http://www.cs.helsinki.fi/u/leino/jutut/ungegn-03.pdf

Antti Leino

Helsinki Institute for Information Technology, Basic Research Unit

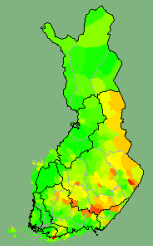
Research Institute for the Languages of Finland

10th October 2003



Introduction

- Onomastics
  - Multidisciplinary: linguistics, history, some geography
  - Let's add computer science
- Goals
  - Dependences between occurrences of different names
    - New information on how places are named
  - Homogeneous regions
    - New information on the relationships between settlement history, linguistic regions and naming
- Methods
  - Pretty straightforward application of data mining techniques to a novel data set
  - Most of this more thoroughly explained in Leino *et al.* (2003)
- Tools
  - Basic Unix/Linux tools
  - the Perl scripting language <URL:http://www.perl.org/>
  - the R statistics environment <URL:http://www.r-project.org/>, esp. with the spatial statistics packages `splancs` and `spatstat`
  - the GRASS GIS environment <URL:http://grass.itc.it/>

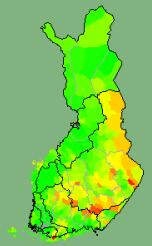


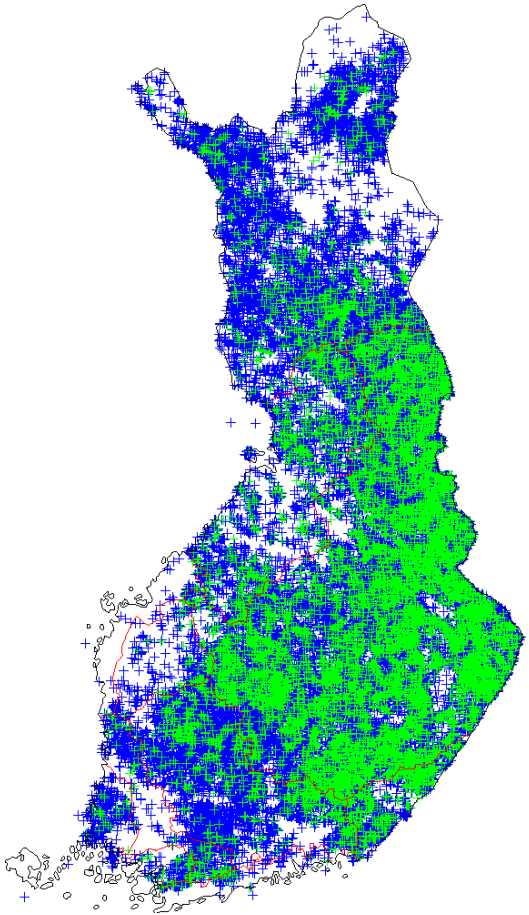
Place Name Data

- Finnish National Land Survey Place Name Register (Leskinen 2002)
  - 718 000 name instances
  - 58 000 lakes
  - 25 000 different lake names
  - 54 most common lake names: 9 008 lakes
  - 45 name endings: 55 538 lakes
- High-dimensional marked point processes
  - Spatial statistics: mostly single processes, at best low dimensionality
  - Data mining: mostly non-spatial data

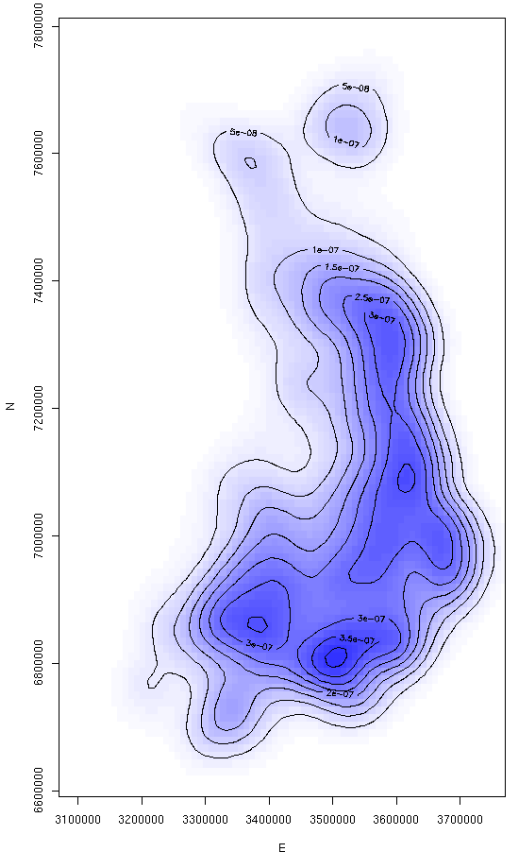
```
Pitkäjärvi;1;Suomi;410;Vakavesi;6682578;2541586;6684464;3375471;049;
Espoo - Esbo;011;Helsingin seutukunta;01;Uusimaa - Nyland;1;Uusimaa - Nyland;
1;Etelä-Suomen lääni - Södra Finlands län;204301A;19O1D4;1;
Virallinen kieli tai saame;1;Enemmistön kieli;1;Maastotietokanta;10011998;
40011998

Pitkäjärvi;6684464;3375471;049

järvi;Pitkäjärvi;6684464;3375471;049
```

Figure 1: Example of raw Place Name Register data, common names data and name endings data



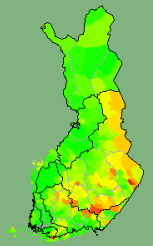


Name endings data (+)
common lake names data (+)



Kernel estimate
of the lake intensity

Figure 2: Lake names in the Place Name Register data



Association Rules

- $X \Rightarrow Y$, where $X, Y \subseteq \{A_1, \ldots, A_n\}$
  - Frequency $f(X \cup Y)$
  - Accuracy $\frac{f(X \cup Y)}{f(X)}$
- Spatial association rules
  - Various views on these (eg. Koperski and Han 1995, Estivill-Castro and Lee 2001, Huang *et al.* 2002, 2003)
  - Here: $X \Rightarrow_r Y$, where r is radius

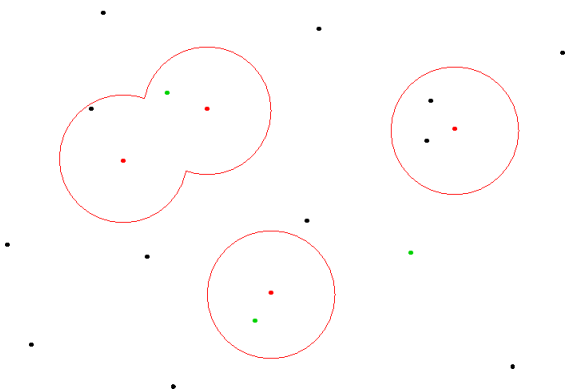
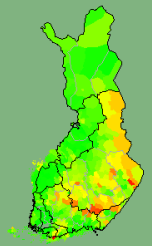


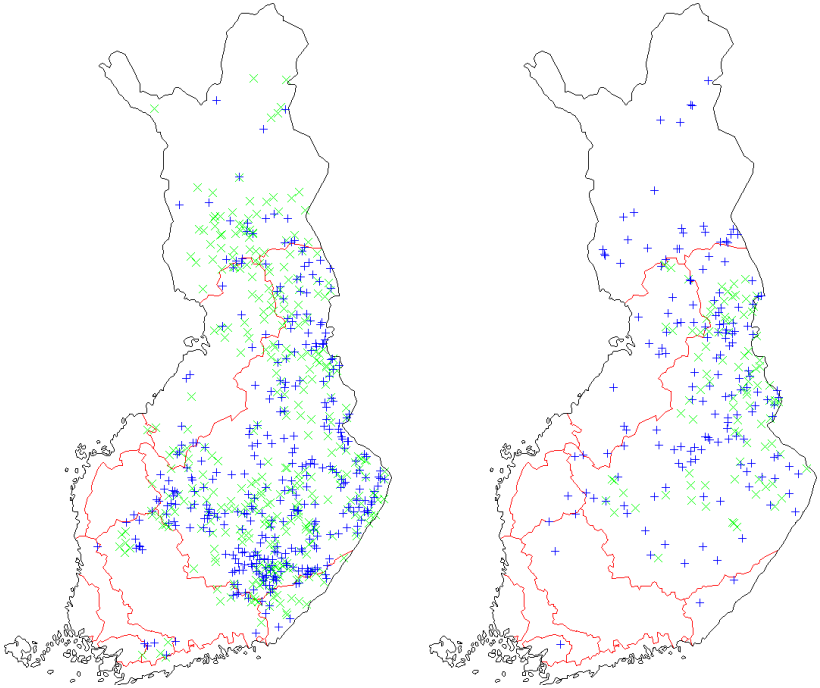
Figure 3: Spatial association rule $A \Rightarrow_r B$ as selection

- If no association (ie. A and B independent of each other), selection in Figure 3 is a random sample



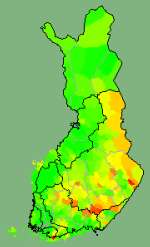
Results

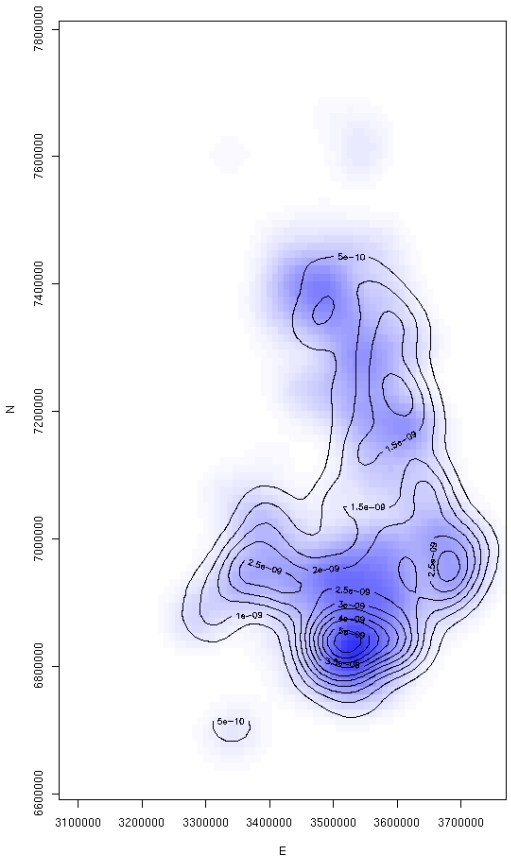
- Figure 4 shows the distribution of two pairs of names and figure 5 their kernel estimates. The distributions look relatively similar.



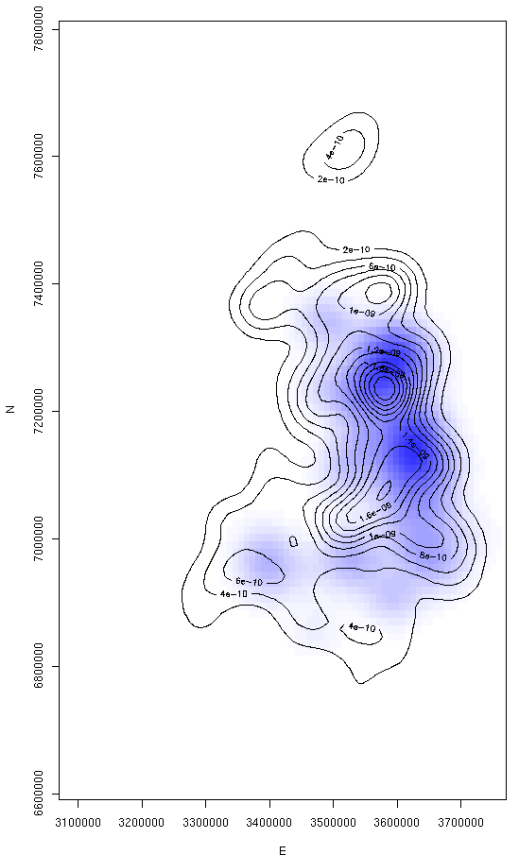
Ahvenlampi 'Perch Lake' (x) *Joutenlampi*'Swan Lake' (x)
Haukilampi 'Pike Lake' (+) *Hanhilampi* 'Duck Lake' (+)

Figure 4: Distribution of two pairs of names



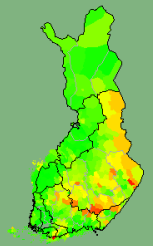


Ahvenlampi 'Perch Lake' (■)
Haukilampi 'Pike Lake' (—)



Joutenlampi'Swan Lake' (■)
Hanhilampi 'Duck Lake' (—)

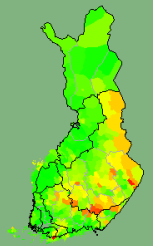
Figure 5: Kernel estimates of the names in figure 4



- Figure 6 shows the Poisson-approximated probabilities.
  - *Ahvenlampi* $\Rightarrow_r$ *Haukilampi*: a strong association at small radii
  - *Hanhilampi* $\Rightarrow_r$ *Joutenlampi*: much weaker and at longer radii

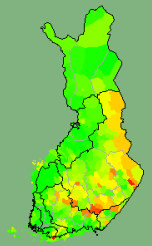
```
Ahvenlampi => Haukilampi:
+ At 1 km found 20; p(n<20) = 1.0000 (corrected 1.00)
+ At 2 km found 40; p(n<40) = 1.0000 (corrected 1.00)
+ At 3 km found 51; p(n<51) = 1.0000 (corrected 0.99)
+ At 4 km found 75; p(n<75) = 1.0000 (corrected 1.00)
+ At 5 km found 92; p(n<92) = 1.0000 (corrected 0.97)
+ At 6 km found 116; p(n<116) = 1.0000 (corrected 0.98)
+ At 7 km found 137; p(n<137) = 1.0000 (corrected 0.95)
+ At 8 km found 170; p(n<170) = 1.0000 (corrected 1.00)
+ At 9 km found 181; p(n<181) = 1.0000 (corrected 0.96)
+ At 10 km found 204; p(n<204) = 1.0000 (corrected 0.98)

Hanhilampi => Joutenlampi:
  At 1 km found 0; p(n<0) = 0.0000 (corrected 0.00)
  At 2 km found 3; p(n<3) = 0.9259 (corrected 0.00)
  At 3 km found 3; p(n<3) = 0.6418 (corrected 0.00)
  At 4 km found 5; p(n<5) = 0.6983 (corrected 0.00)
  At 5 km found 9; p(n<9) = 0.8927 (corrected 0.00)
  At 6 km found 18; p(n<18) = 0.9990 (corrected 0.00)
  At 7 km found 21; p(n<21) = 0.9985 (corrected 0.00)
+ At 8 km found 31; p(n<31) = 1.0000 (corrected 0.98)
  At 9 km found 33; p(n<33) = 1.0000 (corrected 0.91)
  At 10 km found 37; p(n<37) = 1.0000 (corrected 0.91)
```

Figure 6: Associations in two pairs of names



- Other interesting pairs of names as well
  - *Mustalampi* 'Black Lake' $\Rightarrow_r$ *Valkealampi* 'White Lake': expected, but an indication that the method works
  - *Lehmilampi* 'Cow Lake' $\Rightarrow_r$ *Likolampi* 'Retting Lake': association results from cultural connection
  - *Likolampi* 'Retting Lake' $\Rightarrow_r$ *Pitkälampi* 'Long Lake': association but no obvious reason
- Various interesting questions on the characteristics of contrastive / variational names
  - Quite a few cases where the names contrast with regard to one of the semantic features of the modifier
  - Difference between the terms **kontrastnamn** and **variationsnamn**?
  - Possibly a partial answer to Pamp (1991):

    > Det finns skäl att förmoda att analogien vid bildning av naturnamn verkar också när namnen har kommit till på saklig grund. Problemet är bara att den här fungerar så diskret att den oftast är mycket svår att påvisa.



Repulsion

- A special case of association rules, $A \Rightarrow_r A$
- Not obvious that a sample like in Figure 3 could be considered random. However, the sum of samples in Figure 7 can.

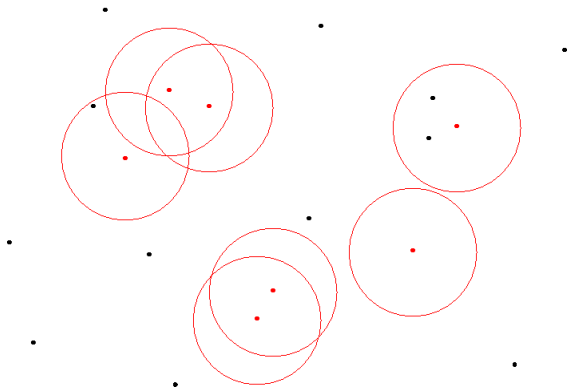
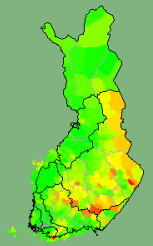


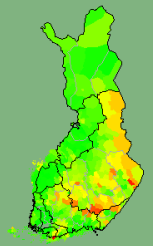
Figure 7: Spatial association rule $A \Rightarrow_r A$ as a series of selections



- Repulsion appears to be rare; this is surprising.
- There are even cases like *Umpilampi* 'Overgrown Lake' where there is significant attraction (cf. Figure 8). Evidently each of these names is actively used by a very small group of people, likely just a single farm.

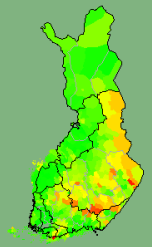
```
Umpilampi => Umpilampi:
  At 1 km found 9; p(n<9) = 0.9999 (corrected 0.66)
+ At 2 km found 32; p(n<32) = 1.0000 (corrected 1.00)
+ At 3 km found 66; p(n<66) = 1.0000 (corrected 1.00)
+ At 4 km found 82; p(n<82) = 1.0000 (corrected 1.00)
+ At 5 km found 103; p(n<103) = 1.0000 (corrected 1.00)
+ At 6 km found 126; p(n<126) = 1.0000 (corrected 1.00)
+ At 7 km found 136; p(n<136) = 1.0000 (corrected 1.00)
+ At 8 km found 154; p(n<154) = 1.0000 (corrected 1.00)
+ At 9 km found 164; p(n<164) = 1.0000 (corrected 1.00)
+ At 10 km found 171; p(n<171) = 1.0000 (corrected 1.00)
```

Figure 8: Conspicuous absence of repulsion between instances of *Umpilampi*



Probabilistic Modeling

- View the data as a matrix, with municipalities as rows and names (or name endings) as columns; each cell has the frequency of these names in the municipality.
- Apply the EM clustering algorithm (Dempster *et al.* 1977, Redner and Walker 1984, McLachlan 1996):
  - Assign random component weights
  - E-step: For each data point, compute the probability that the data resulted from the model
  - M-step: Compute the component weights according to the results of the E-step
  - Iterate the E and M steps as necessary
- Observations
  - Clusters spatially well connected.
  - As the number of clusters increases, new divisions appear — but the old boundaries mostly stay in place.
  - Clusters correspond with onomastic and historical information.
  - The old Western Finnish habitation shows fairly well
  - Also the boundary between the Eastern and Western dialect groups; names reflect an older demographic state than current dialects
  - Interesting parallels to dialectometric maps (Wiik 1999)



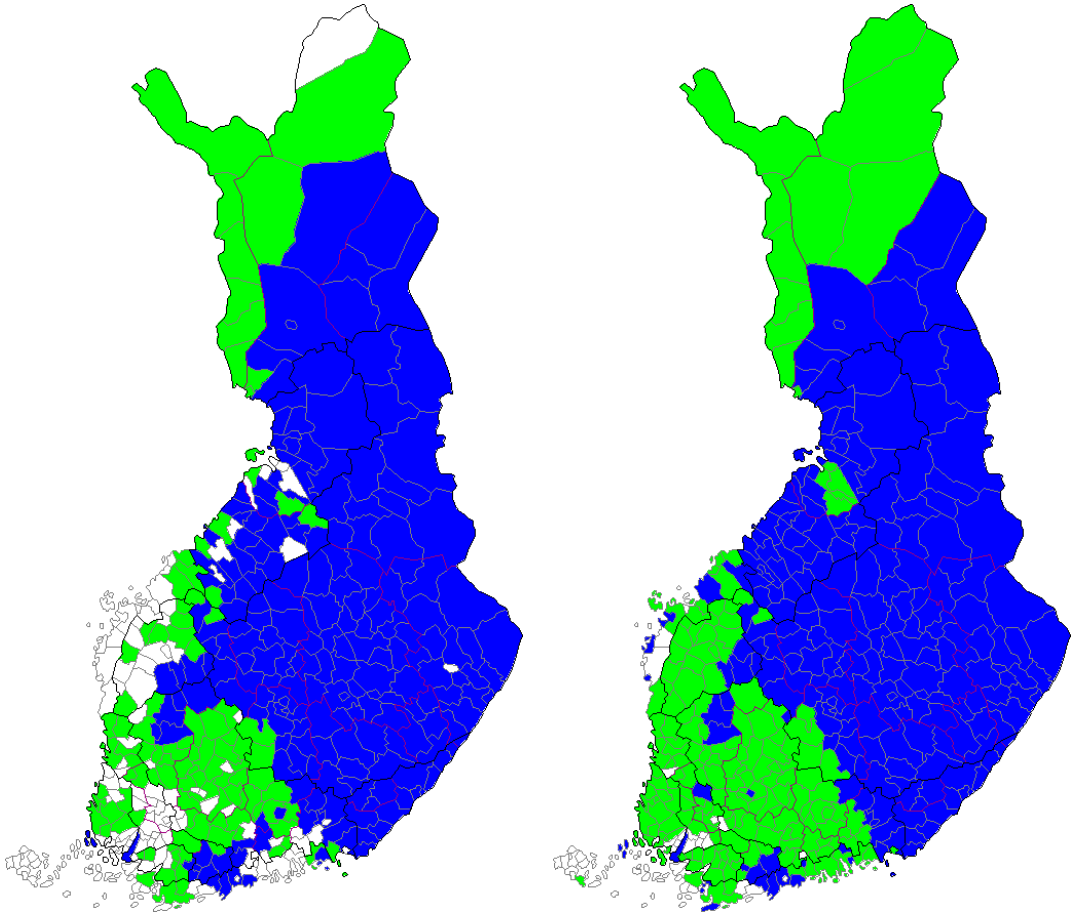
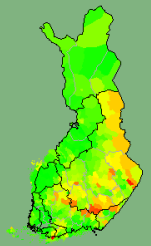


Figure 9: 2-way clustering on common names (left) and name endings (right)



Spatial data mining as an onomastic tool

Antti Leino
UNGEGN Norden Division

10.10.2003

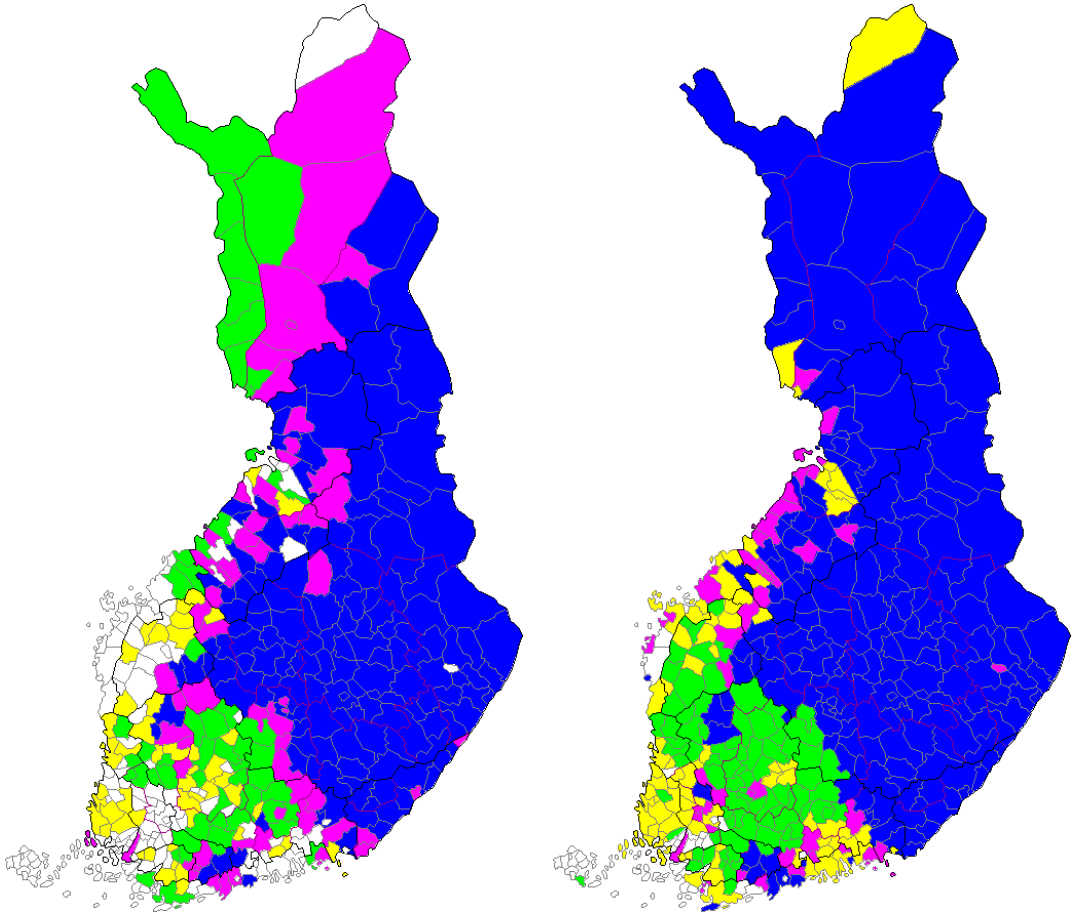
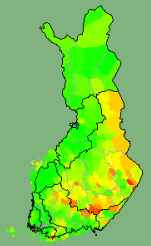
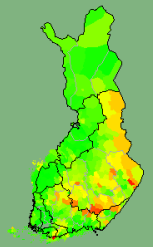


Figure 10: 4-way clustering on common names (left) and name endings (right)



Conclusions and Further Research

- Basic data mining methods can be applied to spatial point data
- Impact on onomastics
  - Certain types of contrastive names are more widespread than previously thought; theories about naming processes have to be re-evaluated
  - Repulsion appears far less noticeable than expected. This, too, has to be explained somehow.
  - Clustering seems a possible starting point for composing an onomastic overview. This can be combined with other data, such as that on dialectal variation.
- Association involving more than two names: $\{A_1, \ldots A_i\} \Rightarrow_r B$
  - How to extend known algorithms to spatial data, ie. data with no clear observations?
  - $\Gamma \Rightarrow_r B$, where $\Gamma \equiv$ 'There are names of type α nearby'
  - Combination of simple association rules and clustering: 'Names $\{A_1, \ldots, A_i\}$ are often found near each other'



References

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, **39**(B), 1–38.

Estivill-Castro, V. and Lee, I. (2001). Data mining techniques for autonomous exploration of large volumes of geo-referenced crime data. In *6th International Conference on Geocomputation*.

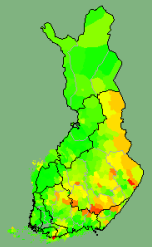
Huang, Y., Shekhar, S., and Xiong, H. (2002). Discovering co-location patterns from spatial datasets: A general approach. Submitted to IEEE Transactions on Knowledge and Data Engineering (TKDE), under second round review.

Huang, Y., Xiong, H., Shekhar, S., and Pei, J. (2003). Mining confident co-location rules without a support threshold. To appear in Proceedings of the 18th ACM Symposium on Applied Computing (ACM SAC).

Koperski, K. and Han, J. (1995). Discovery of spatial association rules in geographic information databases. In *Proceedings of the 4th International Symposium on Large Spatial Databases*.

Leino, A., Mannila, H., and Pitkänen, R. L. (2003). Rule discovery and probabilistic modeling for onomastic data. In N. Lavrač, D. Gamberger, L. Todorovski, and H. Blockeel, editors, *Knowledge Discovery in Databases: PKDD 2003*, number 2838 in Lecture Notes in Artificial Intelligence, pages 291–302. Springer.

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*.



McLachlan, G. J. (1996). *The EM Algorithm and Extensions*. Wiley & Sons.

Pamp, B. (1991). Onomastisk analogi. In G. Albøge, e. Villarsen Meldgaard, and L. Weise, editors, *Tiende nordiske navneforskerkongres, Brandbjerg 20.—24. maj 1989*, number 45 in Norna-rapporter, pages 157–174.

Redner, R. and Walker, H. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, **26**(2), 195–234.

Wiik, K. (1999). Suomen dialektometriikkaa. Turku.