Rule Discovery and Probabilistic Modeling for Onomastic Data

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

Antti Leino [1] [2]
Heikki Mannila [1] [3]
Ritva Liisa Pitkänen [2] [4]

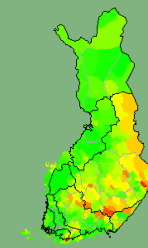
[1] Helsinki Institute for Information Technology, Basic Research Unit

[2] Research Institute for the Languages of Finland

[3] Helsinki University of Technology,
Laboratory of Computer and Information Science

[4] University of Helsinki, Department of Finnish

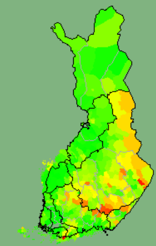
25th September 2003



Rule Discovery and Probabilistic Modeling for Onomastic Data
Leino, Mannila, Pitkänen
PKDD 2003

Introduction

- High-dimensional marked point processes
  - Spatial statistics: mostly single processes, at best low dimensionality
  - Data mining: mostly non-spatial data
- Onomastics
  - Study of names, in this case place names
  - Multidisciplinary: linguistics, history, some geography
- 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



Place Name Data

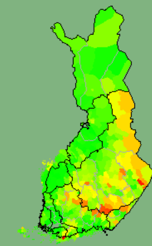
- Finnish National Land Survey Place Name Register [1]
  - 718 000 name instances
  - 58 000 lakes
  - 25 000 different lake names
  - 54 most common lake names: 9 008 lakes
  - 45 name endings: 55 538 lakes

```
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;1901D4;1;
Virallinen kieli tai saame;1;Enemmistön kieli;1;Maastotietokanta;10011998;
40011998

Pitkäjärvi;6684464;3375471;049

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

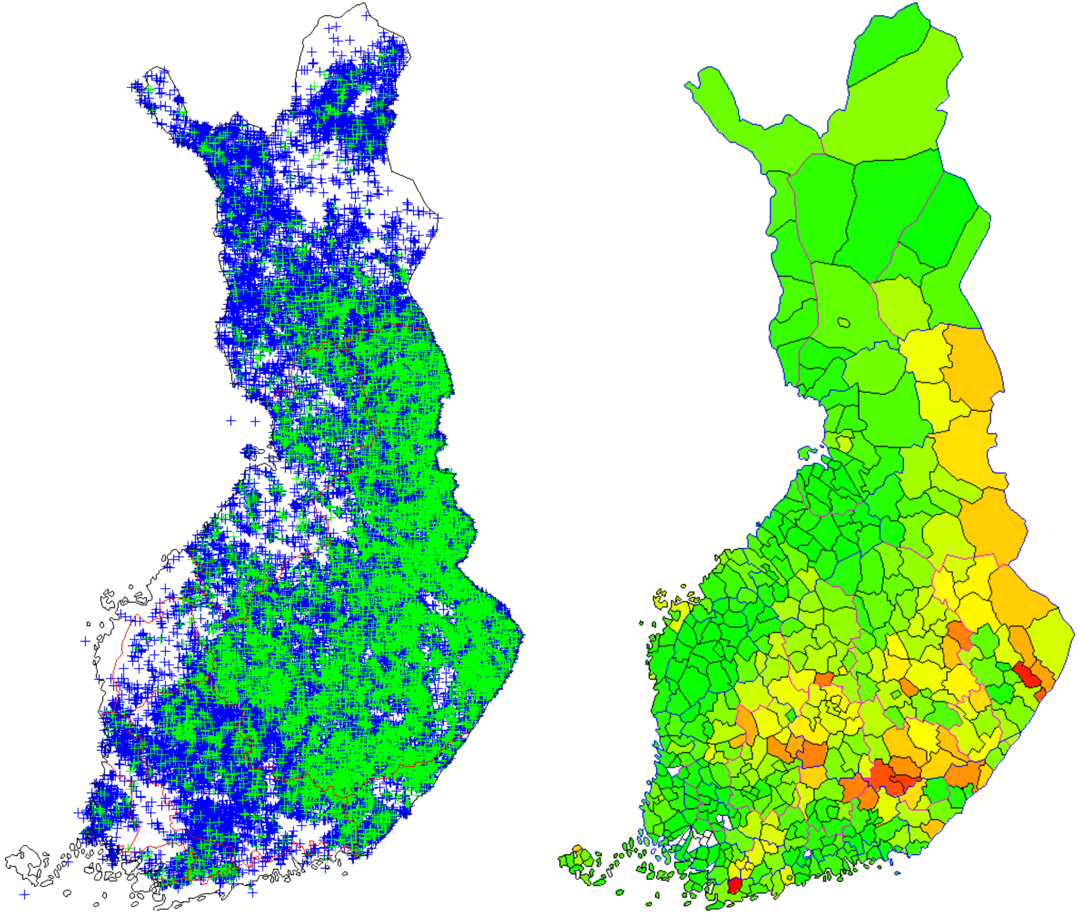
Figure 1: Example of raw Place Name Register data, common names data and name endings data



Introduction
Place Name Data
Association Rules
Probabilistic . . .
Conclusions and . . .
References

Page 3 of 15

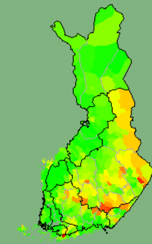
Go Back
Full Screen
Close
Quit



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

Lakes / km^2 in
Finnish municipalities

Figure 2: Lake names in the Place Name Register data



Rule Discovery and Probabilistic Modeling for Onomastic Data

Leino, Mannila, Pitkänen

PKDD 2003

Introduction
Place Name Data
Association Rules
Probabilistic . . .
Conclusions and . . .
References

Page 4 of 15

Go Back
Full Screen
Close
Quit

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 [2, 3, 4, 5]
  - Here: $X \Rightarrow_r Y$, where r is radius

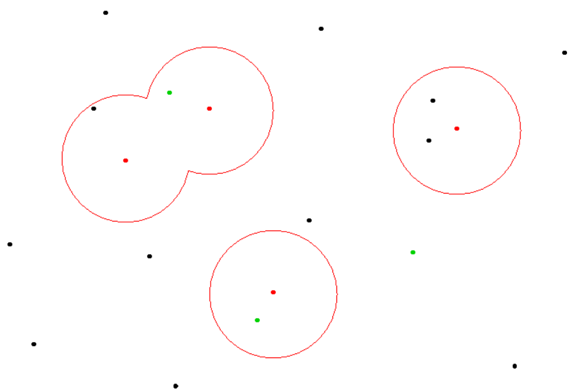
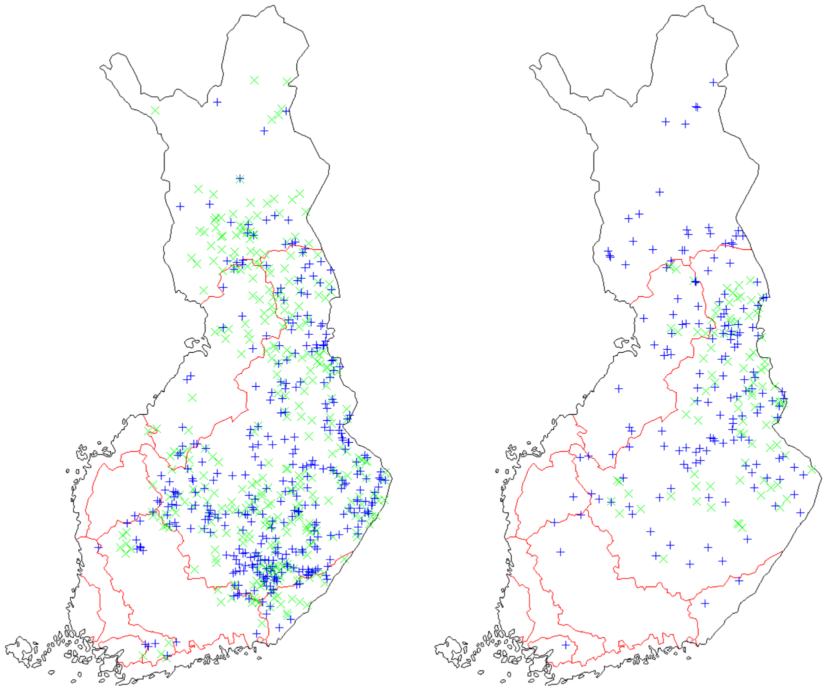


Figure 3: Spatial association rule $A \Rightarrow_r B$ as selection

- 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. The distributions look relatively similar.



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

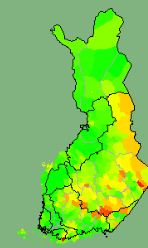
Figure 4: Distribution of two pairs of names

- Figure 5 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

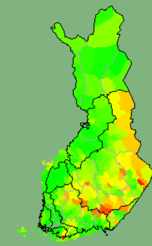
```
Ahvenlampi => Haukilampi:
+ At 1 km found 20; p(n<20) = 1.0000 (corrected 1.00)
+ At 2 km found 40; p(n<40) = 1.0000 (corrected 1.00)
+ At 3 km found 51; p(n<51) = 1.0000 (corrected 0.99)
+ At 4 km found 75; p(n<75) = 1.0000 (corrected 1.00)
+ At 5 km found 92; p(n<92) = 1.0000 (corrected 0.97)
+ 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 5: Associations in two pairs of names



- Various interesting questions on the characteristics of contrastive / variational names
- Other interesting pairs of names as well
  - *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



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 6 can.

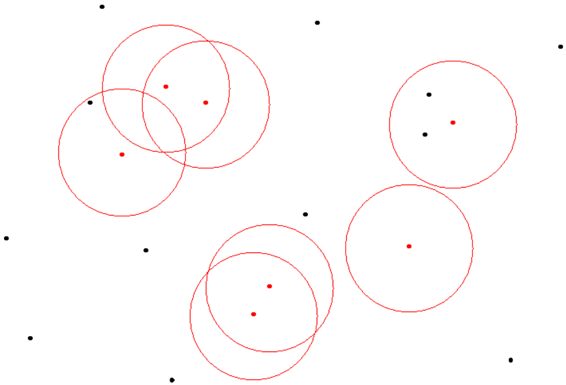
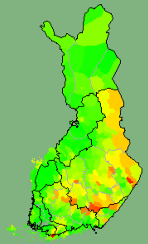


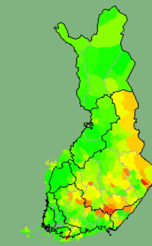
Figure 6: Spatial association rule $A \Rightarrow_r A$ as a series of selections



- Repulsion appears to be rare; this is surprising.
- There are even cases like *Umpilampi* 'Closed Lake' where there is significant attraction (cf. Figure 7). Evidently each of these names is actively used by a very small group of people, likely just a single farm.

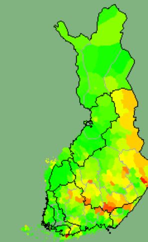
```
Umpilampi => Umpilampi:
  At 1 km found 9; p(n<9) = 0.9999 (corrected 0.66)
+ At 2 km found 32; p(n<32) = 1.0000 (corrected 1.00)
+ At 3 km found 66; p(n<66) = 1.0000 (corrected 1.00)
+ At 4 km found 82; p(n<82) = 1.0000 (corrected 1.00)
+ At 5 km found 103; p(n<103) = 1.0000 (corrected 1.00)
+ At 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 7: 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 [6, 7, 8]:
  - 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 previous 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



Leino, Mannila, Pitkänen

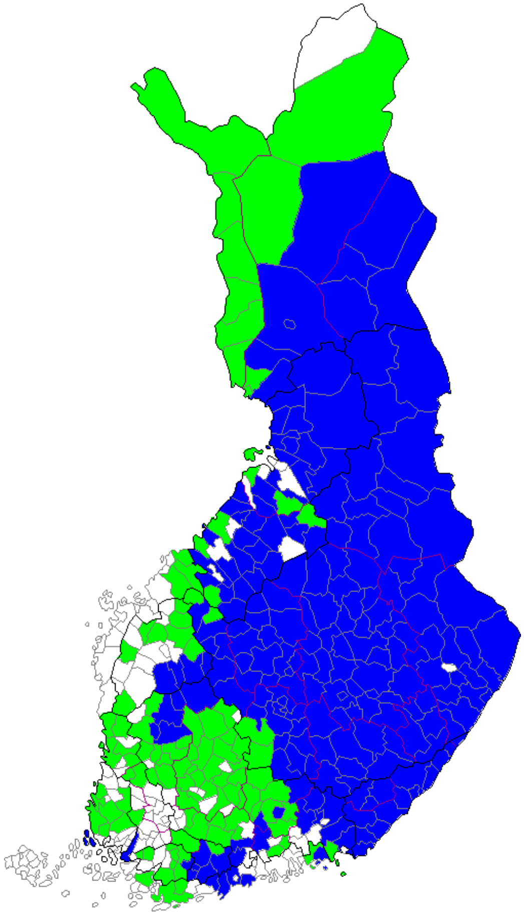


Figure 8: 2-way clustering on common names (left) and name endings (right)

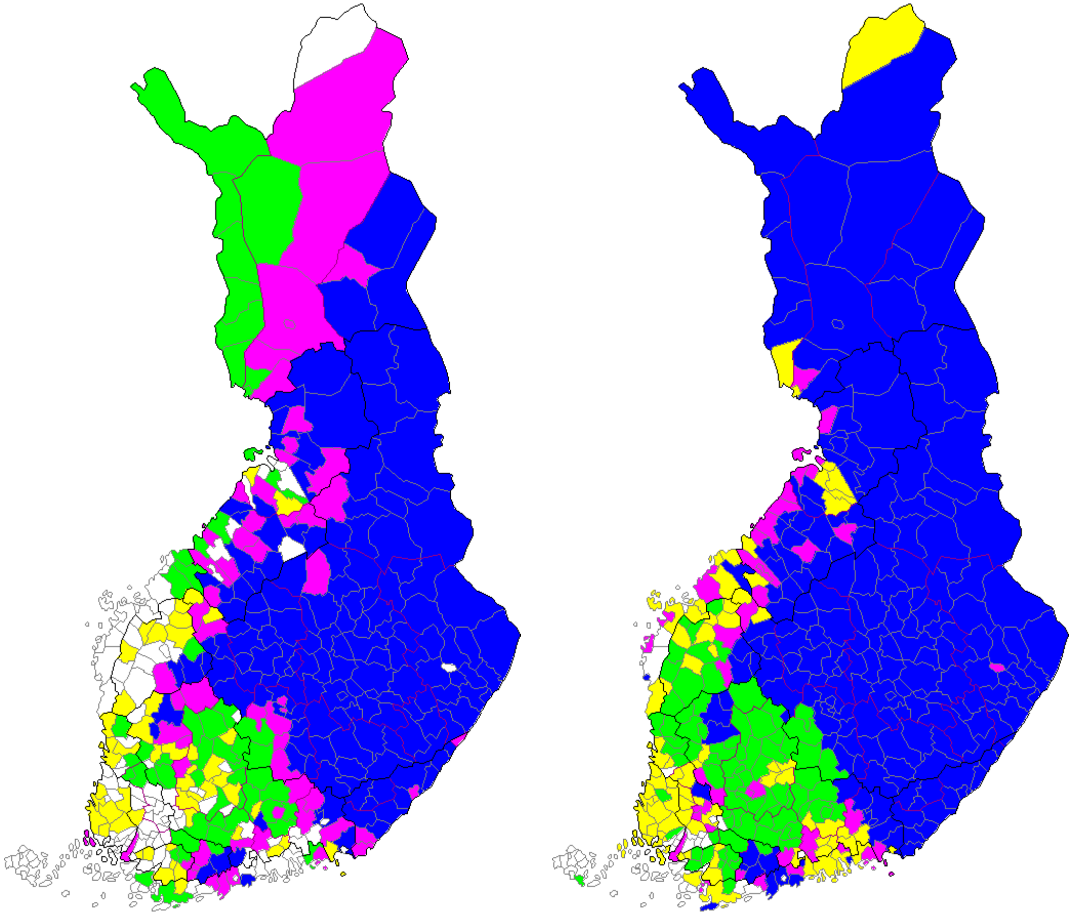
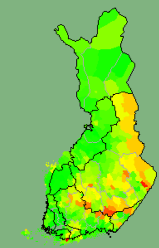


Figure 9: 4-way clustering on common names (left) and name endings (right)

Conclusions and Further Research

- Basic KDD 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: $\{A_1, \dots A_i\} \Rightarrow_r B$
  - 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 $\{A_1, \dots, A_i\}$ are often found near each other'



References

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

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

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

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

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

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

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

[8] McLachlan, G.J.: The EM Algorithm and Extensions. Wiley & Sons (1996)