Antti Leino
Helsinki Institute for Information Technology
Research Institute for the Languages of Finland

Pikes and perches go together: A data-analytical view on Finnish lake names

1. Introduction

The existence of a systematic structure – a grammar of sorts – in place names and the importance of existing patterns in naming were pointed out by some onomasticians a while ago (eg. Šrámek 1972, Kiviniemi 1977), and these ideas have been steadily gaining acceptance during the past couple of decades. One interesting part of this system of place names is the phenomenon of contrastive and variational naming. While onomasticians have in the past been aware of these there have been relatively few studies on the subject, the one best known to a Finnish audience being Kiviniemi (1971). In a similar vein, the interactions of nearby place names have been studied briefly (eg. Santakivi 1979, Eskelinen 2002), but no systematic study on these phenomena have been made. The purpose of this article is to show how new insights can be gained on these issues by applying modern computer science to the already somewhat multidisciplinary field of onomastics.

Computer science includes a field called data mining that specialises in finding ways to extract new information from large corpora of data. Such techniques would seem appropriate for finding regularities in place names, especially as the University of Helsinki has one of the leading data mining research communities worldwide. Toponym data is also interesting from the point of view of computer science, in that the mining of spatial data has not been as extensively studied as that of other types of data.

The Finnish National Land Survey has compiled a comprehensive Place Name Register (Leskinen 2002) that is an excellent starting point for onomastic data analysis. The register contains all names on the Finnish Basic Map, but the analysis for the present article was performed on a subset of the register containing 54 most common lake names, that is, those names that are used for at least 90 lakes each. The size of these data sets is shown in table 1. Although the data set consists of the occurrences of 54 names, not all are covered in this article: as seen later on, the methods used here involve sifting through data and looking for regularities, and the present article covers some of the more interesting of these.

	Different Finnish names	Name instances
Entire Register	303 626	717 747
Lakes	25 178	58 267
Common lake names	54	9 008

Table 1: Place Name Register

2. Two pairs of names

Figure 1 shows two distribution maps, each with two lake names. At first glance one would say that the semantic relationship between the two names is somewhat similar in the pairs *Ahvenlampi* 'Perch Lake' – *Haukilampi* 'Pike Lake' and *Hanhilampi* 'Goose Lake' – *Joutenlampi* 'Swan Lake'. The distributions of the names in each pair would also appear to be relatively similar, although the distributions of the two pairs differ from each other. It is, however, difficult to fully assess the geographical dependences between the names by just looking at the distribution maps. Fortunately, it is possible to adapt existing data mining methods to this situation. The key concept here is **association rule** (cf. eg. Mannila and Toivonen 2002), that is, a rule saying that the presence of one phenomenon indicates that another one is likely to appear as well.

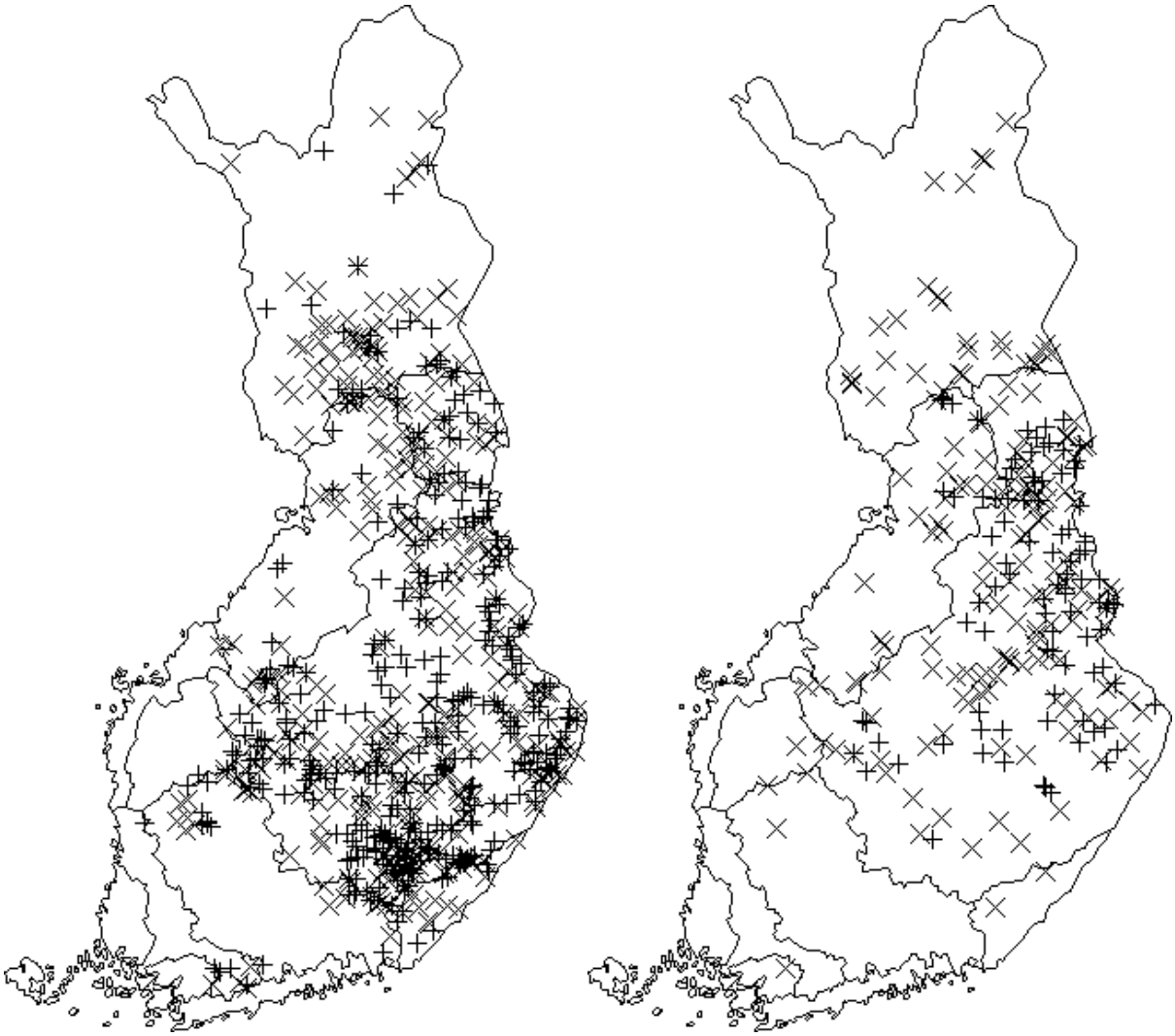

Figure 1: Distribution maps of the names *Ahvenlampi* (x) – *Haukilampi* (+), left, and *Hanhilampi* (x) – *Joutenlampi* (+), right.

The methods used here are more thoroughly explained elsewhere (Leino et al. 2003), but the association rule between two place names can be summarised by figure 2. One starts with two names (say, *Ahvenlampi* and *Haukilampi*), shown in the figure as x and +, respectively. It is then possible to select all lakes (shown in the figure as either x, +, or ·) within a given radius of any *Ahvenlampi* and count the frequency of the name *Haukilampi* first in the overall population of all lakes and second in the vicinity of lakes called *Ahvenlampi*. If the names appear independently of each other, the selection is a random sample with regard to the occurrence of the name *Haukilampi* and the frequencies should be about the same, save for random variation. More precisely, the number of instances of *Haukilampi* in the selection should follow the Poisson distribution, $X \sim \text{Poisson}(\lambda)$, where the parameter λ is the number of all lakes in the selection multiplied by the frequency of *Ahvenlampi* in the overall population. This makes it possible to give probabilistic estimates of the strength of the association between the names. There are also ways to refine this method, for example by applying the Bonferroni correction to get an estimate of how likely it is to get a significant-looking result if one simply conducts 54×54 of these tests.

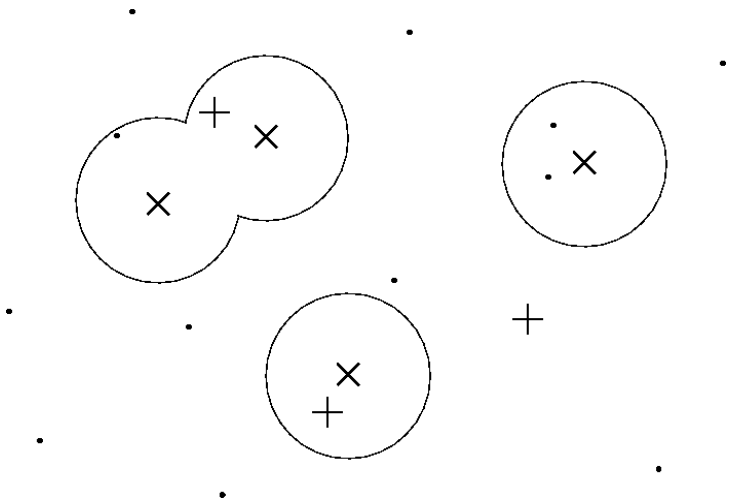

Figure 2: Selecting names for finding association rules

Table 2 shows that neither pair of names appears independently of each other in all radii. However, the table also shows that the association rules *Ahvenlampi* ⇒ *Haukilampi* and *Hanhilampi* ⇒ *Joutenlampi* are statistically significant using very different radii: while the first is significant at small distances the second gains significance only at larger distances. The peak of significance in the first case is at distances of 10 km and less; in the second case it is at somewhere around 8–20 km.

Granted, the probability calculations assume that the frequency of each name is invariant throughout the country, and this is obviously not true. A further study should take into account the variations in the overall distribution of each name. However, this simplification mostly means that one would expect a peak in the probability between names with similar overall distributions somewhere in the range of some tens of kilometers; a peak at radii in the kilometer range, in turn, would be more consistent with a naming pattern at work.

```
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)
+ At 15 km found 263; p(n<263) = 0.9991 (corrected 0.00)
+ At 20 km found 301; p(n<301) = 0.9970 (corrected 0.00)
+ At 25 km found 326; p(n<326) = 0.9954 (corrected 0.00)
+ At 30 km found 335; p(n<335) = 0.9831 (corrected 0.00)
+ At 35 km found 340; p(n<340) = 0.9588 (corrected 0.00)
  At 40 km found 344; p(n<344) = 0.9308 (corrected 0.00)
  At 45 km found 345; p(n<345) = 0.8857 (corrected 0.00)

Hanhilampi => Joutenlampi:
  At 1 km found 0; p(n>0) = 0.2091 (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)
+ At 15 km found 60; p(n<60) = 1.0000 (corrected 0.99)
+ At 20 km found 73; p(n<73) = 1.0000 (corrected 0.86)
+ At 25 km found 83; p(n<83) = 0.9997 (corrected 0.14)
+ At 30 km found 90; p(n<90) = 0.9991 (corrected 0.00)
+ At 35 km found 93; p(n<93) = 0.9965 (corrected 0.00)
+ At 40 km found 93; p(n<93) = 0.9838 (corrected 0.00)
+ At 45 km found 93; p(n<93) = 0.9609 (corrected 0.00)
  At 50 km found 93; p(n<93) = 0.9348 (corrected 0.00)
```

Table 2: The strength of association rules *Ahvenlampi* ⇒ *Haukilampi* and *Hanhilampi* ⇒ *Joutenlampi* at different radii

It is also true that the two pairs are not exactly similar. Pikes and perches are among the most common predatory fish in Finland, and one can expect to find them in any reasonably-sized lake. Geese and swans do not live next to each other like this. Nevertheless, the pairs are similar in that in each the names differ from each other by one semantic feature, and one would expect them to appear close to each other if the naming process involved this kind of pattern. The exact overall frequency of each name is not in itself relevant to the method used; rather, the fundmental question is whether the frequency peaks significantly above normal. The existence of this peak, and the distance at which it occurs, indicates that something onomastically interesting may be going on.

Thus it seems likely that the pair *Ahvenlampi–Haukilampi* is a result of a naming process where the variation of a common theme plays an important role – the names appear very close to each other so often than it is difficult to come to a different conclusion. The pair *Hanhilampi–Joutenlampi* is different: the association rule holds only at longer distances. This in turn means that it can be adequately explained by the fact, seen already on the distribution map, that the overall distributions of these two names coincide.

3. Other observations

In addition to the two pairs of names seen above, several other interesting association rules were also found in the data. A complete analysis of these would require a longer treatise than a brief article, and thus the following is merely a short selection.

- *Mustalampi* 'Black Lake' ⇒ *Valkealampi* 'White Lake' and vice versa.
  This rule was expected, as it is perhaps the most typical example of contrastive naming in the Finnish lakes. It does not in itself tell anything new, but the fact that highly significant association rules were found both ways is an indication that the method works.

- *Lehmilampi* 'Cow Lake' ⇒ *Likolampi* 'Retting Lake' and vice versa.
  The association rule was relatively strong but is clearly not a result of variational naming. A common agricultural background would seem a likely reason for the association.

- *Likolampi* 'Retting Lake' ⇒ *Pitkälampi* 'Long Lake', *Likolampi* 'Retting Lake' ⇒ *Valkealampi* 'White Lake' and vice versa.
  These are again relatively strong association rules, but it is difficult to see an obvious reason.

- *Umpilampi* 'Overgrown Lake' ⇒ *Umpilampi*.
  The rule is quite strong even at short distances. This seems to be contrary to the common intuition that two similar places with the same name cannot exist near each other, and some kind of explanation has to be found. At this point it seems likely that these are mostly very small lakes, and the need to refer to them exists only within a single farmer family. Still, this explanation too is surprising, as onomasticians have traditionally considered a village to be the basic region for the use of small place names.

4. Conclusions

The pair *Ahvenlampi–Haukilampi* and other similar pairs found in the data suggest that the phenomenon of variational and contrastive naming is more common than previously thought and that it needs further study. As a rough working hypothesis one could suggest that there exists a relatively widespread naming pattern where a place can be named by taking the name of a nearby place of similar type and changing it so that the two names are in contrast with each other with regard to one of the semantic features of the modifier. If this hypothesis proves close to what is happening here, it would seem reasonable to use the term **contrastive names** to apply to these cases as well as to those where the modifiers are full antonyms. The term **variational names** would then be restricted to groups of names that vary the same theme in other ways – there are, for instance, pairs where the names are synonymous or where the variation is based on phonological properties of the names.

The relatively widespread existence of these contrastive pairs and the strong clustering of the instances of a single name – as in the case of *Umpilampi* – could also provide a starting point in answering the question posed by eg. Bengt Pamp (1991): how to prove the hypothesis that analogy plays an important role in the naming process even in cases where the names are also motivated by the physical features of the place.

All in all, further research is clearly indicated. This preliminary study was based on a few dozen most common lake names; a further one could take the entire corpus of Finnish lake names, and furthermore search for associations between more than just two names. Also, while the issues of contrastive naming appear immediately interesting, there were also associations that could be explained by reasons related to cultural or settlement history or by geographical features. The result of a more thorough analysis of the data should give us new insights on how the naming processes work.

References

Eskelinen, Riikka 2002: Paikannimien ja määriteosien yleisyys ja toistuminen Tervon kunnassa. Practicum paper, University of Helsinki Department of Finnish.

Kiviniemi, Eero 1971: Vastakohta- ja variointinimistä. Virittäjä 123–134.

Kiviniemi, Eero 1977: Väärät vedet. Suomalaisen Kirjallisuuden Seuran Toimituksia 337. Vaasa.

Leino, Antti, Heikki Mannila, and Ritva Liisa Pitkänen 2003: Rule Discovery and Probabilistic Modeling for Onomastic Data. Knowledge Discovery in Databases: PKDD 2003, ed. by Nada Lavrač, Dragan Gamberger, Ljupčo Todorovski, and Hendrik Blockeel, 291–302. Lecture Notes in Artificial Intelligence 2838. Springer.

Leskinen, Teemu 2002: The geographic names register of the National Land Survey of Finland. Paper presented at the Eighth United Nations Conference on the Standardization of Geographical Names.

Mannila, Heikki, and Hannu Toivonen 2002: Knowledge discovery in databases: The search for frequent patterns. Course material for *Special Course on Data Mining*. University of Helsinki Department of Computer Science.

Pamp, Bengt 1991: Onomastisk analogi. In Gordon Albøge, Eva Villarsen Meldgaard, and Lis Weise (ed.), Analogi i navngivning, 157–172. Norna-rapporter 45.

Santakivi, Pekka 1979: Paikannimien toistuminen Hauholla, Lammilla ja Tuuloksessa. M.A. thesis, University of Helsinki Department of Finnish.

Šrámek, Rudolf 1972: Zum Begriff »Modell» und »System» in der Toponomastik. Onoma 1972/73, 55–75.

Antti Leino
HIIT/BRU
Dept. of Computer Science
P.O. Box 26
FIN-00014 University of Helsinki
Finland
antti.leino@cs.helsinki.fi