Rule Discovery and Probabilistic Modeling for Onomastic Data

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

[1] Helsinki Institute for Information Technology, Basic Research Unit
Department of Computer Science
P.O. Box 26, FIN-00014 UNIVERSITY OF HELSINKI, FINLAND
[2] Research Institute for the Languages of Finland
Sörnäisten rantatie 25, FIN-00500 HELSINKI, FINLAND
[3] Helsinki University of Technology
Laboratory of Computer and Information Science
P.O. Box 5400, FIN-02015 HUT, FINLAND
[4] University of Helsinki, Department of Finnish
P.O. Box 3, FIN-00014 UNIVERSITY OF HELSINKI, FINLAND

Abstract. The naming of natural features, such as hills, lakes, springs, meadows etc., provides a wealth of linguistic information; the study of the names and naming systems is called onomastics. We consider a data set containing all names and locations of about 58,000 lakes in Finland. Using computational techniques, we address two major onomastic themes. First, we address the existence of local dependencies or repulsion between occurrences of names. For this, we derive a simple form of spatial association rules. The results partially validate and partially contradict results obtained by traditional onomastic techniques. Second, we consider the existence of relatively homogeneous spatial regions with respect to the distributions of place names. Using mixture modeling, we conduct a global analysis of the data set. The clusterings of regions are spatially connected, and correspond quite well with the results obtained by other techniques; there are, however, interesting differences with previous hypotheses.

1 Introduction

In spatial statistics, a *point process* is a random process that produces points in the Euclidean plane. A realization of such a process, i.e., a set of points, is called a *point pattern*, or *spatial point data* [1, 2]. A *marked point process* consists of several point processes producing different types of points. The points are often also called *events*.

Marked point processes arise in many applications, such as linguistics (in the study of dialects or place names, each word, grammatical construct, name,

N. Lavrač et al. (Eds.]: PKDD 2003, LNAI 2838, pp. 291–302, 2003.

etc. corresponds to a different type of event), biodiversity studies (different types of events correspond to, e.g., different types of plants, and the locations are the places in which the plant has been observed), business applications (locations of customers etc.). There are some fundamental differences in the point data in these applications. The most relevant here is that in some of these cases the point data represents an underlying phenomenon that is contionuous (e.g. the occurrence area of a species, or the area in which a particular word is used), while in others the underlying phenomenon is itself discrete. In the current study we discuss the latter type of point processes.

The analysis of high-dimensional point processes can be quite demanding. The data is often sparse, i.e., we have only fragmentary information of the underlying phenomenon. When there are several different types of events, modeling their interaction can be complex. In many cases the observed quantities are results of several unobserved processes. The granularity and accuracy of the locations of the points can vary: sometimes the event can be localized perfectly, sometimes not.

Spatial statistics (see, e.g., the books [1, 2]) has developed several strong methods for analyzing a single point process. However, marked point processes with a high number of different types of events have received less attention.

This paper is a case study in the use of pattern discovery and mixture modeling for the analysis of a high-dimensional marked point processes.

Our application is in the area of linguistics, especially *onomastics* (the study of names), particularly place names. The naming of natural features, such as hills, lakes, springs, meadows etc., provides a wealth of information. Our example data consists of full information about place names in Finland. The names tend to be fairly old, and they provide information about the population history and linguistic conditions at the time when the names where given.

Research in onomastics has traditionally been conducted by selecting a single name, or a group or related names, drawing maps of their occurrences, and doing qualitative analysis of the patterns of occurrences. Global analyses of the spatial distributions of different names are non-existent.

Our case study concerns two major themes in onomastics. The first is dependence between occurrences of names. It has long been assumed that the name of a nearby location has an influence on the naming of a location. For example, if a lake is called “Black Lake” (usually because the water is sufficiently clear that one can see the dark bottom of the lake), then a nearby lake might be named “White Lake”. No quantitative evidence for this phenomenon is known, however. A special case of the local influence of names is *repulsion*: if a location is called B, then it makes sense to assume that other similar locations near this will *not* be called B: after all, the purpose of naming is to assign identifiers to locations. Our first goal is to study the local interactions between names.

The second theme we want to verify is the existence of relatively homogeneous spatial regions with respect to the distribution of place names. It is typically assumed that the naming conventions in nearby areas should be more or less similar, i.e., that there are clear regional trends in the style of names. The occurrence maps of individual names support this hypothesis, but virtually no global analyses exist.

In this paper we address both these themes. We first show how one can modify the basic ideas of association rule techniques to obtain local descriptions of the dependencies between the occurrences of names. The results show that indeed there are statistically significant associations between the occurrences of names. As for repulsion effects, we show that they are far less noticeable than expected.

For the second theme, we demonstrate the use of mixture modeling for the data at the granularity of municipalities, and show that the resulting clusters of municipalities are spatially extremely coherent. Thus the results verify the basic hypothesis that spatial homogeneity exists and provide new data for further onomastic research into the naming processes that cause the phenomenon.

The rest of this paper is organized as follows. The data set is described in Section 2. In Section 3 we show how the basic ideas of association rules can be generalized to the case of spatial point patterns, and give a sample of the results. Section 4 describes how mixture modeling applies to this data set, and discusses the results briefly. Section 5 is a brief conclusion.

2 The data set

Our example data set is a subset of the Finnish names occurring in the National Place Name Register, a part of the Geographic Names Register kept by the National Land Survey of Finland. The register contains all place names that appear on the 1:20 000 Basic Map and is maintained for the purposes of creating these maps. The size of the register, as well as that of our subsets, can be found in table 1, which shows the total number of Finnish names (or name instances), the number of different names, and the number of different municipalities in which these names are found.

	Name	Different	Municipalities
	instances	names	
Entire Register	717 747	303 626	447
Lakes	58 267	25 178	408
Common lake names	9 008	54	315
Name endings	55 538	45	407

Table 1. National Place Name Register data

The full data model of the register is explained in [3], but for the present study it is sufficient to note that the register includes a *language* field, a *feature type* field and the spatial information in different formats, including two co-ordinate systems and several administrative divisions. The *feature type* categorizes geographical features into such classes as *lake or pond*, or *river*, or *stretch* *of river,* or *forest*. For lakes, the location is fixed to be a selected point inside of the lake.

For our study we selected first all lake names in Finnish. This selection we pruned further along two different lines. For our primary data set we chose the names that have at least 90 instances. While our aim was to concentrate on the most common names, the limit of 90 instances is somewhat arbitrary. To supplement the primary data set we selected for clustering purposes a second data set, consisting not of complete place names but of derivational suffixes and final parts of compound names.

The two different subsets were selected mainly for onomastic reasons. Our working hypothesis was that spatial associations are in a large part related to the phenomenon of contrastive names — that is, pairs of names that refer to similar geographical features and differ only by the first part of the name in some sort of contrastive manner. To study this we needed to search for spatial associations for full names. Similarly, both intuition and onomastic consensus would say that there is a repulsion effect between two instances of the same name which is closely related to the use of place names to identify a place: a name cannot normally be used by the same group of people to denote two different places of the same type.[5] Again, this means we have to study the full names. In either case it seems appropriate to restrict ourselves to relatively common names, to make sure there are enough instances of each of them to get valid results.

With clustering the situation is somewhat different. The obvious way to start is to use full names, like we do with the association rules, and there is no reason to doubt that this approach works. However, it is also reasonable to postulate that by studying word endings — both derivative suffixes and end-parts of compound names — we can get insight into differences in naming practices. Using name endings is thus an attempt to do cluster analysis based on the distribution of various name types, not just names as such.

3 Spatial association rules

In this section we consider the first theme: finding local effects between the occurrences of different names. As an example, consider Figures 1—3 showing the occurrences of certain pairs of names. How do the occurrences of one name affect the probability of occurrence of another name? It is fairly clear that the maps alone cannot answer the question.

In spatial statistics questions such as this have been addressed by using, e.g., nearest neighbor distances or the K function and its derivatives [4, 2]. Here we describe a similar approach, but using the terminology of association rules.

Given a set of observations over 0-1 attributes $A_1, \ldots, A_n$, an *association rule* is an expression $X \Rightarrow Y$, where $X, Y \subseteq \{A_1, \ldots, A_n\}$. Given a set X of attributes, the frequency $f(X)$ of X is the fraction of observations that have a 1 in

[5] It is, however, relatively common to name e.g. a farm after a nearby lake.

all attributes of X. The frequency of the rule is defined to be $f(X \cup Y)$, and the accuracy (confidence) of the rule is $f(X \cup Y)/f(X)$.

We consider spatial association rules of the form $A \Rightarrow_r B$. The interpretation of such a rule is that given a location (x, y) in which event of type A occurs, one is likely to see at least one event of type B within distance r from (x, y). This definition is close to the ones used by [5–9]. From an onomastic point of view it seems prudent to start with restricting ourselves to associations between two names.

To test the significance of a rule $A \Rightarrow_r B$ we start with a set of places named A and another set of places named B. We want to evaluate whether the occurrence of a B is more likely in the context of a nearby A than in general. Note, however, that a B can only occur if there is a suitable natural feature present: we cannot observe a "Pike Lake" at position (x, y) unless there is a lake at (x, y). To take this into account we consider as a set of reference points all points belonging to the the same type of feature as B; call this set C_B. In our case we used all Finnish lakes as C_B.

The probability that a given place that belongs to C_B is named B is $P(B) = \frac{N(B)}{N(C_B)}$, where $N(B)$ is the total number of places named B and $N(C_B)$ is the total number of all the places of the same type. We now select the places belonging to set C_B which are within the given radius r of a place named A. We denote the size of this selection by $n(C_B)$ and the number of B places in it by $n(B)$. As null hypothesis we can now assume that the occurrences of A and B are independent. Under this hypothesis our selection can be viewed as a random sample, which can be approximated by the Poisson distribution, $X \sim \text{Poisson}(\lambda)$, where $\lambda = n(C_B)\frac{N(B)}{N(C_B)}$. To correct for multiple testing, we use the Bonferroni correction.

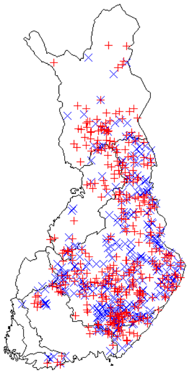
Repulsion Repulsion is essentially a special case of a spatial association rule $A \Rightarrow_r B$, where $A = B$. However, in this situation we select points based on the spatial distribution of A; it is not immediately obvious that this can be considered a random sample with regard to A. We have therefore used another method to confirm the results on repulsion.

In the general case we again start with two kinds of points, A and B, the latter of which belong to set C_B. The overall number of points B and C_B is $N(B)$ and $N(C_B)$, respectively; the probability of a given C_B point being a B point is $p = \frac{N(B)}{N(C_B)}$.

Within a given radius of the ith point with name A there are $n(C_{B_i})$ points of set C_B. We use random variable X_i to denote the number of points named B in this set. If the B points are distributed independently of each other, $X_i \sim \text{Bin}(n(C_{B_i}), p)$, so $E(X_i) = n(C_{B_i})$ and $D^2(X_i) = n(C_{B_i})p(1-p)$. Summing, we obtain a variable $S_m = \sum_{i=1}^{m} X_i$, and by assuming independence of the variables X_i, we have $E(S_m) = \sum_{i=1}^{m} E(X_i)$ and $D^2(S_m) = \sum_{i=1}^{m} D^2(X_i)$. Applying the central limit theorem we can obtain confidence estimates.

Results Applying the method presented above to the common names data set gave both expected and unexpected results. As expected, most of the pairs of names had no significant associations either way. Also to be expected was that there were pairs that had significant repulsion between the names: the spatial distributions of these names just don't overlap, for various reasons related to such things as geography or variation in dialects.

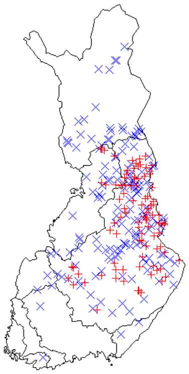
One interesting sub-category of the association rules was what can be called contrasting names. These have traditionally considered only for such pairs as *Mustalampi* "Black Lake" — *Valkealampi* "White Lake" where the contrasting element in at least one of the names refers to a notable property of the lake and there is a clear antonymic relation between the two names. Our study indicates that this kind of variation is used in the naming process more widely and with far less strict semantic constraints for the elements than onomasticians have thought. For instance, there was a group of three names, *Ahvenlampi* "Perch Lake", *Haukilampi* "Pike Lake" and *Särkilampi* "Roach Lake", all of which had significant associations with each other even over small distances. Figure 1 shows the spatial distribution for *Ahvenlampi* and *Haukilampi* on a map with main dialectal regions, along with Poisson-approximated probabilities before and after the Bonferroni correction.



```
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)

Haukilampi => Ahvenlampi:
+ 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 50; p(n<50) = 1.0000 (corrected 0.91)
+ At 4 km found 75; p(n<75) = 1.0000 (corrected 0.99)
  At 5 km found 92; p(n<92) = 1.0000 (corrected 0.88)
  At 6 km found 113; p(n<113) = 0.9999 (corrected 0.73)
  At 7 km found 131; p(n<131) = 0.9996 (corrected 0.00)
  At 8 km found 154; p(n<154) = 0.9998 (corrected 0.53)
  At 9 km found 175; p(n<175) = 0.9999 (corrected 0.64)
  At 10 km found 195; p(n<195) = 0.9999 (corrected 0.80)
```

Fig. 1. Spatial distribution of *Haukilampi* (x) and *Ahvenlampi* (+)

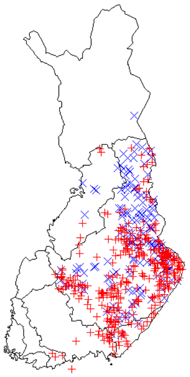
There were, however, other pairs that would at first glance appear to be similarly contrasting, but whose associations are somewhat weaker and start to show at significantly longer distances. In fact, the question arises whether there is a connection in the naming process or whether the names just have a similar distribution. One such case is the pair of *Joutenlampi* "Swan Lake" and *Hanhilampi* "Goose Lake", as shown in Figure 2. The reasons for the difference between this pair and that of *Ahvenlampi* — *Haukilampi* are not very obvious, and further onomastic study of these phenomena is needed.



```
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)

Joutenlampi => Hanhilampi:
  At 1 km found 0; p(n<0) = 0.0000 (corrected 0.00)
  At 2 km found 3; p(n<3) = 0.8542 (corrected 0.00)
  At 3 km found 3; p(n<3) = 0.4347 (corrected 0.00)
  At 4 km found 5; p(n<5) = 0.4496 (corrected 0.00)
  At 5 km found 9; p(n<9) = 0.6805 (corrected 0.00)
  At 6 km found 20; p(n<20) = 0.9968 (corrected 0.00)
  At 7 km found 25; p(n<25) = 0.9981 (corrected 0.00)
  At 8 km found 33; p(n<33) = 0.9998 (corrected 0.49)
  At 9 km found 35; p(n<35) = 0.9990 (corrected 0.00)
  At 10 km found 40; p(n<40) = 0.9992 (corrected 0.00)
```

Fig. 2. Spatial distribution of *Hanhilampi* (x) and *Joutenlampi* (+)

Then there are pairs of names that have a significant association but are not contrasting, like *Lehmilampi* "Cow Lake" and *Likolampi* "Retting Lake",[6] as shown in Figure 3. In some cases another reason for the association can be seen; here, for instance, both names have similar agricultural origins. Although one can make such guesses about the reasons for the association, the phenomenon itself is a new discovery, and again further study would be strongly indicated.



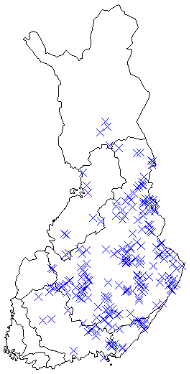
```
Lehmilampi => Likolampi:
  At 1 km found 8; p(n<8) = 0.9998 (corrected 0.48)
+ At 2 km found 21; p(n<21) = 1.0000 (corrected 1.00)
+ At 3 km found 34; p(n<34) = 1.0000 (corrected 1.00)
+ At 4 km found 45; p(n<45) = 1.0000 (corrected 1.00)
+ At 5 km found 56; p(n<56) = 1.0000 (corrected 0.99)
+ At 6 km found 69; p(n<69) = 1.0000 (corrected 0.99)
+ At 7 km found 87; p(n<87) = 1.0000 (corrected 1.00)
+ At 8 km found 104; p(n<104) = 1.0000 (corrected 1.00)
+ At 9 km found 125; p(n<125) = 1.0000 (corrected 1.00)
+ At 10 km found 143; p(n<143) = 1.0000 (corrected 1.00)

Likolampi => Lehmilampi:
  At 1 km found 8; p(n<8) = 0.9999 (corrected 0.72)
+ At 2 km found 19; p(n<19) = 1.0000 (corrected 1.00)
+ At 3 km found 30; p(n<30) = 1.0000 (corrected 1.00)
+ At 4 km found 37; p(n<37) = 1.0000 (corrected 0.98)
  At 5 km found 44; p(n<44) = 1.0000 (corrected 0.88)
  At 6 km found 48; p(n<48) = 0.9991 (corrected 0.00)
  At 7 km found 61; p(n<61) = 0.9999 (corrected 0.71)
  At 8 km found 72; p(n<72) = 1.0000 (corrected 0.93)
+ At 9 km found 85; p(n<85) = 1.0000 (corrected 1.00)
+ At 10 km found 93; p(n<93) = 1.0000 (corrected 1.00)
```

Fig. 3. Spatial distribution of *Lehmilampi* (x) and *Likolampi* (+)

The repulsion between different instances of the same name does not seem to be a very common phenomenon. Onomastically, this is rather surprising. It is

[6] The name refers to a step in the processing of flax into linen.

true that our data set contains such names as *Pahalampi* "Evil Lake"[7] (shown in Figure 4) or *Palolampi* "Burnt Lake",[8] where there are no instances within 2 km of each other. However, the area covered by such selections is rather small, and most of these findings cannot be considered significant. The repulsion effects are for the most part insignificant even without the Bonferroni correction. One possible explanation for the scarcity of significant repulsion is that the body of Finnish lake names is relatively large and the distance a name needs to retain its usefulness as an identifier quite small: the name of a typical small lake is only used within a single village. The latter of these two factors may be sufficient to keep the repulsion small enough to disappear into the random variation caused by the former.



```
Pahalampi => Pahalampi:
  At 1 km found 0; p(n>0) = 0.7161 (corrected 0.00)
  At 2 km found 0; p(n>0) = 0.9899 (corrected 0.00)
  At 3 km found 8; p(n>8) = 0.6203 (corrected 0.00)
  At 4 km found 16; p(n<16) = 0.5101 (corrected 0.00)
  At 5 km found 41; p(n<41) = 0.9997 (corrected 0.17)
+ At 6 km found 63; p(n<63) = 1.0000 (corrected 1.00)
+ At 7 km found 76; p(n<76) = 1.0000 (corrected 1.00)
+ At 8 km found 87; p(n<87) = 1.0000 (corrected 1.00)
+ At 9 km found 103; p(n<103) = 1.0000 (corrected 1.00)
+ At 10 km found 119; p(n<119) = 1.0000 (corrected 1.00)
```

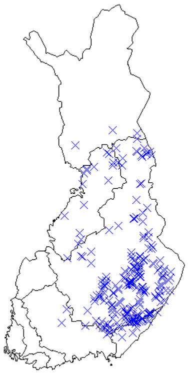
Fig. 4. Spatial distribution of *Pahalampi*

With all this in mind, it is still somewhat surprising to find that there are cases like *Umpilampi* "Closed Lake"[9] (shown in Figure 5) where there is a visible association even at distances of 1 km or less. Again, one can guess for the reasons why this is possible — these are mostly small ponds, and in many cases the need to refer to one of them exists only within one farmer family — but nevertheless this would appear to contradict the onomastic consensus that the basic unit for name use in rural areas is one village.

[7] Some of these — possibly even a large amount — are euphemisms for a vulgar name that the locals considered too offensive to tell outsiders they perceived as being of a higher social standing, such as visiting onomasticians or geographers.

[8] These names are related to the agricultural method of burn-beating, practiced in some places in Finland until the early 20th century.

[9] That is, a small lake overgrown with weeds.



```
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)
```

Fig. 5. Spatial distribution of *Umpilampi*

4 Probabilistic modeling

We now turn to the second onomastic theme, the existence or nonexistence of homogeneous regions with respect to place names. We tested this hypothesis by considering the municipalities as observations, and using mixture modeling and the EM algorithm to obtain a clustering of the municipalities.

In more detail, we took the 315 municipalities, and created 54 variables, one for each of the names in the common names data set. This gives us 54-dimensional data set, where each column indicates the number of occurrences of the name in the municipality. We then took the 407 municipalities and 45 name endings, and conducted a similar test on that set.

We use mixture modeling to this data set [10, 11]. A (finite) mixture model assigns a probability $P(\mathbf{x}|\Theta)$ to an observation $\mathbf{x}$ as weighted sum $\sum_j P(\mathbf{x}|\theta_j)$ of component distributions $P(\mathbf{x}|\theta_j)$ for $j = 1, \ldots, K$, where the weights (or mixing proportions) π_j satisfy $\pi_j \geq 0$ and $\sum \pi_j = 1$.

For each single component of the model for an observation $\mathbf{x} = (x_1, \ldots, x_d)$ we assume independence between variables and use the multinomial Bernoulli distribution

$$P(\mathbf{x}|\theta) = \prod_{i=1}^{d} \theta_i^{x_i}$$

with the constraint $\sum_{i=1}^{d} \theta_i = 1$. A finite mixture of multivariate Bernoulli probability distributions is thus specified by the equation

$$P(\mathbf{x}|\Theta) = \sum_{j=1}^{K} \pi_j P(\mathbf{x}|\theta_j) = \sum_{j=1}^{K} \pi_j \prod_{i=1}^{d} \theta_{ji}^{x_i}$$

with the parameterization $\theta = \{\pi_1, \ldots, \pi_K, (\theta_{ji})\}$ containing $K(d+1)$ parameters for data with d dimensions.

Given a data set R with d binary variables and the number K of mixture components, the parameter values of the mixture model can be estimated using the Expectation Maximization (EM) algorithm [12–14]. The EM algorithm has two steps which are applied alternately in an iterative fashion. Each step is guaranteed to increase the likelihood of the observed data, and the algorithm converges to a local maximum of the likelihood function [12, 15]. The method gives for each component and each observation a probability of the observation stemming from that component.

We applied mixture modeling to the data described above; for each municipality $\mathbf{x}$ and component j we can compute the probability of the observation $\mathbf{x}$ stemming from component j by

$$P(\mathbf{x}|j) = \frac{P(\mathbf{x}|\theta_j)}{\sum_i P(\mathbf{x}|\theta_i)}.$$

For most municipalities there is clearly one component j which gives the municipality the highest probability. Example results are shown in Figures 6 and 7. The different clusters are shown in shades of grey; white municipalities have no lakes in the data set.[10]

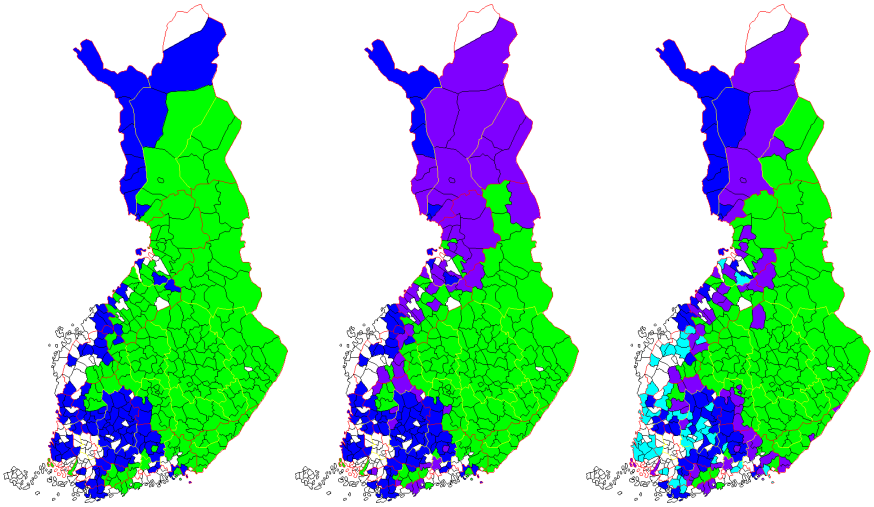


Fig. 6. Clustering based on the most common lake names

Several features are of interest. First of all, the clusters of municipalities obtained in this way are spatially very well connected. Note that the method in itself has no information about the locations of the municipalities, and hence the spatial connectedness of the clusters is interesting. Second, as the number

[10] This is mostly because the common names data set contains only 15% of the lakes, but also because Finland is a bilingual country, and there are some municipalities that are uniformly Swedish.

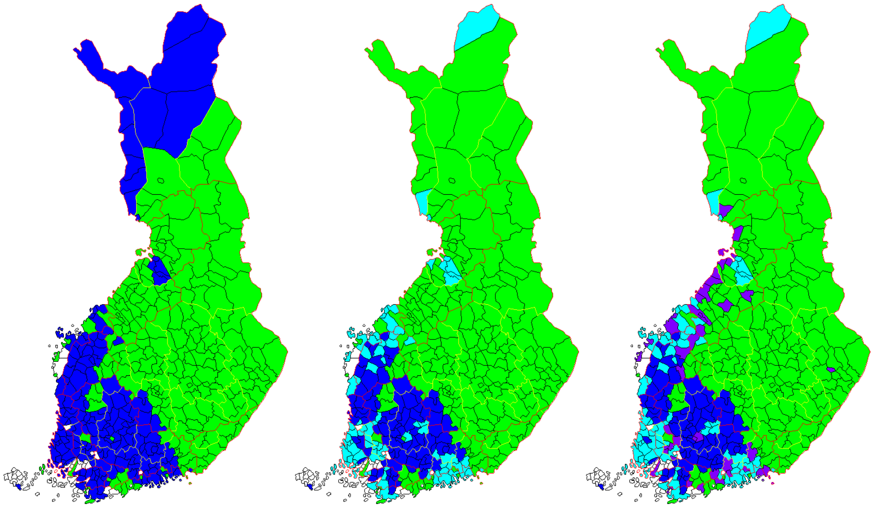


Fig. 7. Clustering based on the name ends

of clusters increases, the existing cluster boundaries tend not to change very much, but rather existing clusters split. Third, the clusters obtained correspond fairly well with the previous onomastic information about the distribution of names.

Specifically, in roughly the southernmost third of the map the boundary seen in the two-cluster maps corresponds rather well with the division between the eastern and western dialectal groups of Finnish. There is a small but noticeable deviation in Tavastland, and this is in line with our knowledge of the history of the settlement of Finland. Likewise, the western cluster continues north along the coast, and this too is in line with what we know from history. However, the middle third looks rather interesting: large regions that were designated and used as hunting grounds for the dialectally western Tavastland communities as late as the 16th century are not associated with the parent province but instead with the eastern regions, from where they were to a large extent populated in the 17th century. This would appear to imply that there is far less old influence in the names of that region than has been commonly believed, and this in turn opens up a variety of interesting onomastic questions.

5 Conclusions

We have described a case study in the area of high-dimensional spatial point processes. We showed how one can use the basic principles of rule discovery and mixture modeling to analyze an onomastic data set about place names. The discovered rules of association and repulsion between names show fascinating local effects between the occurrences. The global analysis of name distribution by using mixture modeling demonstrated that homogeneous onomastic regions do exist. The methods lead to novel onomastic results. While the computational techniques we used are fairly standard, their application was not trivial. The global and local analysis of names has been shown to be very useful, and the study is continuing in several directions.

The existing techniques can be used to answer many onomastic questions. While computational methods of this type have not been applied to onomastic data, the reactions of various researchers in that field have been promising. However, there are also computational open problems. Finding more complex local interactions between names is a particularly interesting one. If A and B occur close to each other, then C is likely to occur close, too. While straightforward generalizations of association rules of the type $AB \Rightarrow_r C$ are possible, it might be more useful to investigate rules of the form $\Gamma \Rightarrow_r C$, where Γ is a derived predicate of position, e.g., of the type "there are names of type α in the neighborhood".

A deeper issue is separating the different layers in the process leading to a particular name occurring in a particular location. In order for a lake at location (x, y) to be called "Black Pond", there has to be a lake at that location, the people who named it must use words "black" and "pond" in their dialect, their naming conventions must allow for the combined name to occur, etc. Thus the data actually is a produced by several interacting phenomena, and finding the influence of each is not easy.

References

1. Ripley, B.D.: Spatial Statistics. John Wiley & Sons (1981)
2. Bailey, T.C., Gatrell, A.C.: Interactive Spatial Data Analysis. Longman Scientific & Technical (1995)
3. 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)
4. Ripley, B.D.: The second-order analysis of stationary point processes. Journal of Applied Probability **13** (1976) 255–266
5. 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)
6. Koperski, K.: A Progressive Refinement Approach to Spatial Data Mining. PhD thesis, Simon Fraser University (1999)
7. 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)
8. 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)
9. 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)
10. McLachlan, G., Peel, D.: Finite Mixture Models. Wiley Series in Probability and Statistics. John Wiley & Sons (2000)

11. Everitt, B., Hand, D.: Finite Mixture Distributions. Monographs on Applied Probability and Statistics. Chapman and Hall (1981)
12. 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
13. Redner, R., Walker, H.: Mixture densities, maximum likelihood and the EM algorithm. SIAM Review **26** (1984) 195–234
14. McLachlan, G.J.: The EM Algorithm and Extensions. Wiley & Sons (1996)
15. Wu, C.J.: On the convergence properties of the EM algorithm. The Annals of Statistics **11** (1983) 95–103