Clustering of Species Ranges

Hannes Heikinheimo

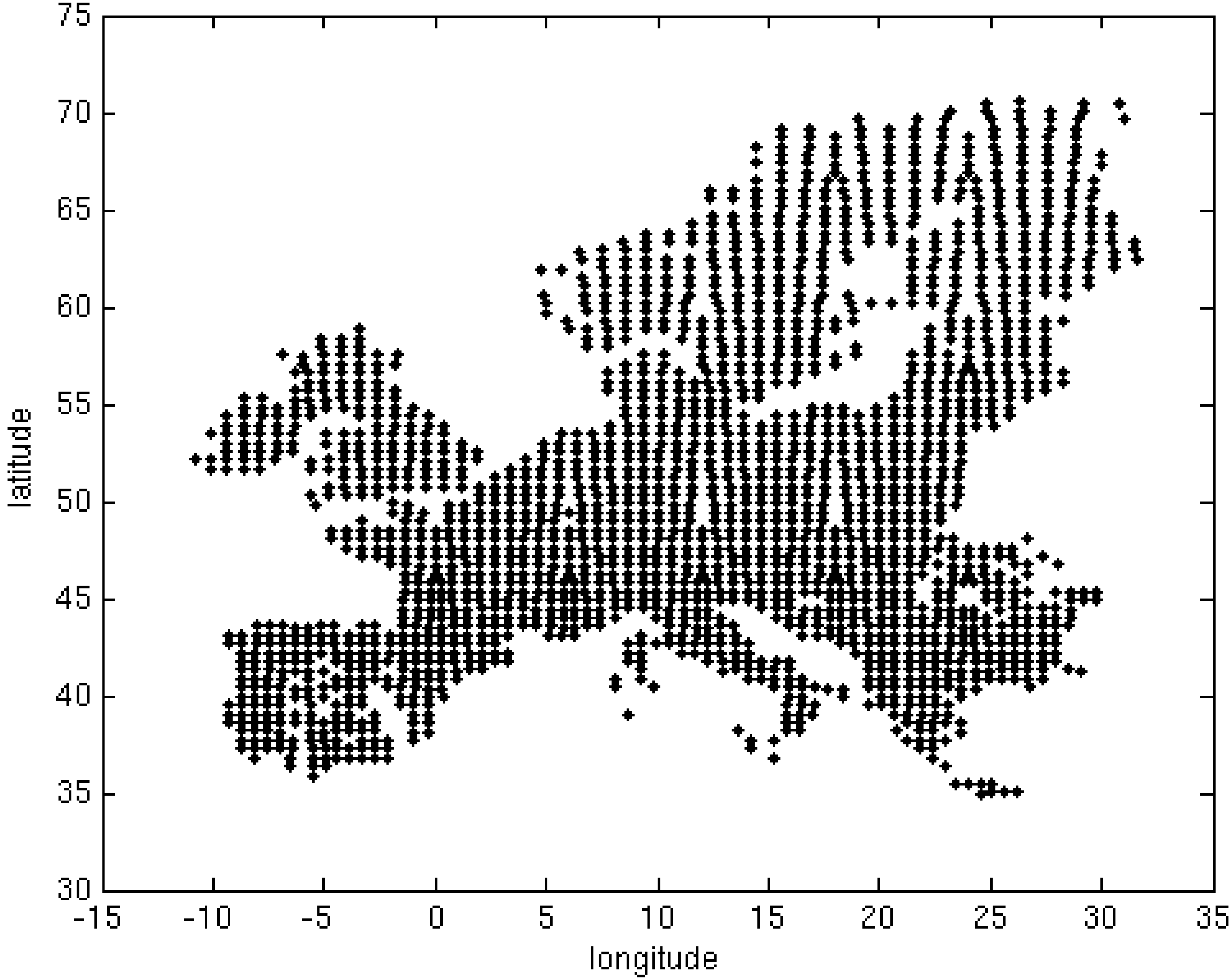
22.3.2006

Overview of this presentation

- Description of the data.
- What is the domain specific interest?
- Comparing species ranges.
- Assesing cluster validity.

The data

- Grided occurance data of present day European land mammals.
- Forms a 2670 (cells) times 194 (species) 0-1 matrix.
  - Mean number of species per cell is 31.3 with standard deviation of +- 16.6.
  - Mean number of cells per species is 430.9 with standard deviation of +-520.7.


latitude
longitude
75
70
65
60
55
50
45
40
35
30
-15
-10
-5
0
5
10
15
20
25
30
35

What is the domain specific interest?

- Ecologists are interested in what kind of properties species (meta)communities have and how they are related to their environment.
- For paleontogists species communities of the present day give an interesting reference point for fossil studies.

Related questions?

1. Are the species ranges some how clustered?
2. Are the clusters geographically or/and enviromentally distict?
3. What kind of species dynamics are present in the clusters?

- In this presentation we will concentrate on the first question.

Binary distances

- Binary distances are distances defined between two equal length binary vectors.
- For binary vectors **x** and **y** we define the contingency matrix as

$$M(\mathbf{x}, \mathbf{y}) = \begin{bmatrix} m_{00} & m_{01} \\ m_{10} & m_{11} \end{bmatrix},$$

- where for i, j in {0, 1} the quantity m_{ij} is the number of components k of **x** and **y** such that $x_k = i$ and $y_k = j$.

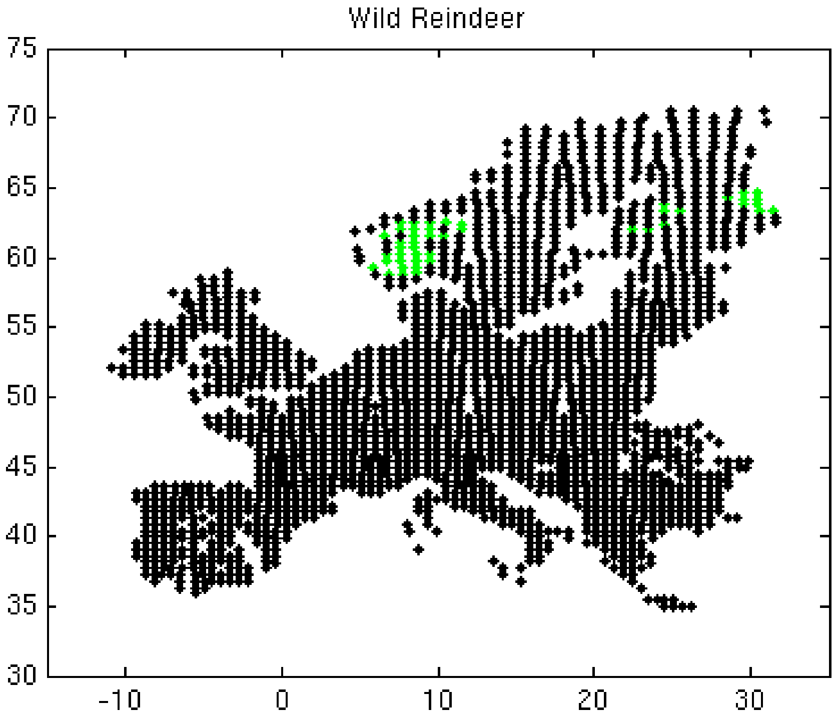
Binary distances

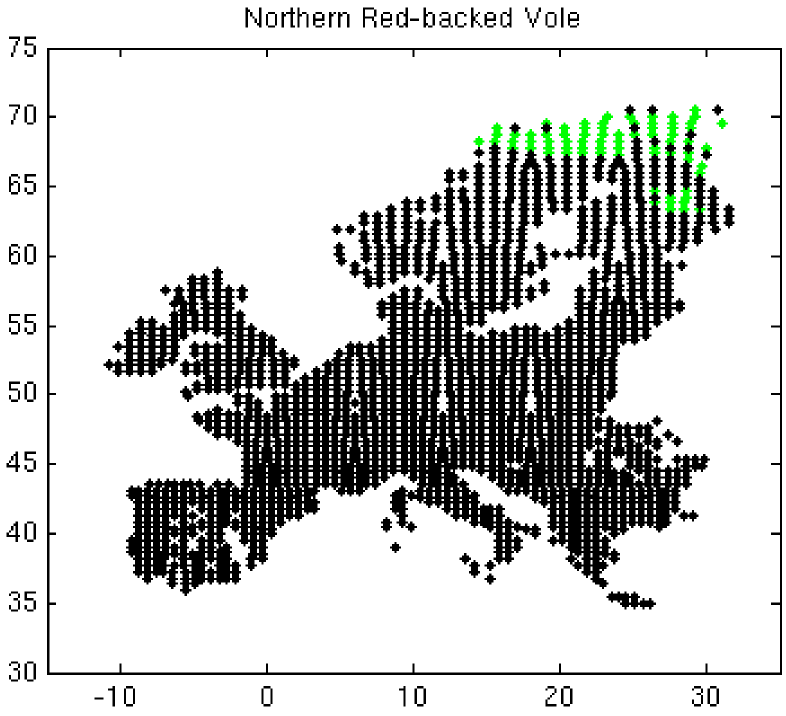
- Hamming: $d_h(\mathbf{x},\mathbf{y}) = \frac{m_{01}+m_{10}}{m_{00}+m_{01}+m_{10}+m_{11}}$
- Jaccard: $d_j(\mathbf{x},\mathbf{y}) = 1 - \frac{m_{11}}{m_{01}+m_{10}+m_{11}}$
- 2. Kulczynski:

$$d_k(\mathbf{x},\mathbf{y}) = 1 - \frac{1}{2}(\frac{m_{11}}{m_{01}+m_{11}} + \frac{m_{11}}{m_{10}+m_{11}})$$

- Correlation: $d_{corr}(\mathbf{x},\mathbf{y}) = 1 - corr(\mathbf{x},\mathbf{y})$

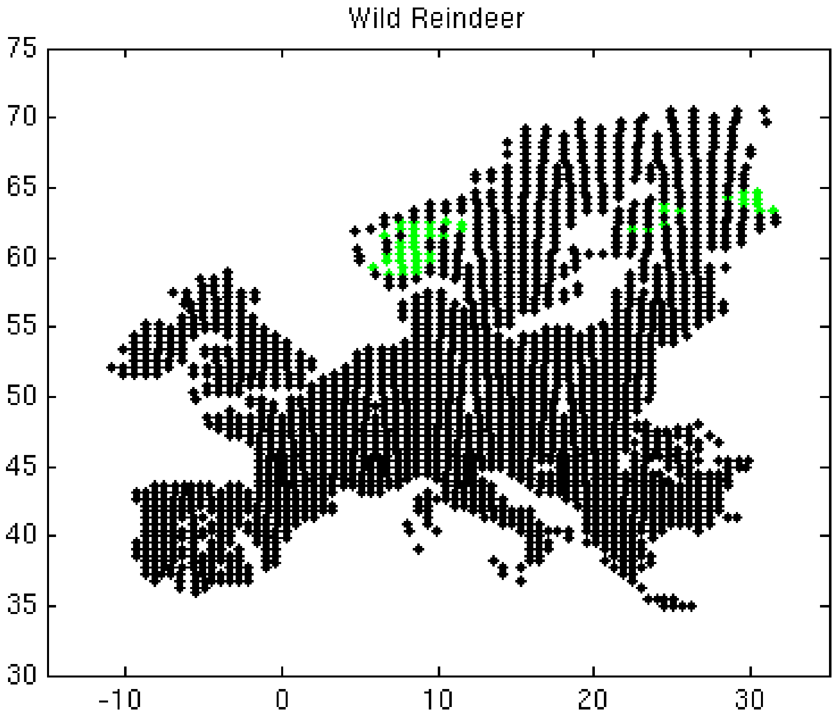
Example

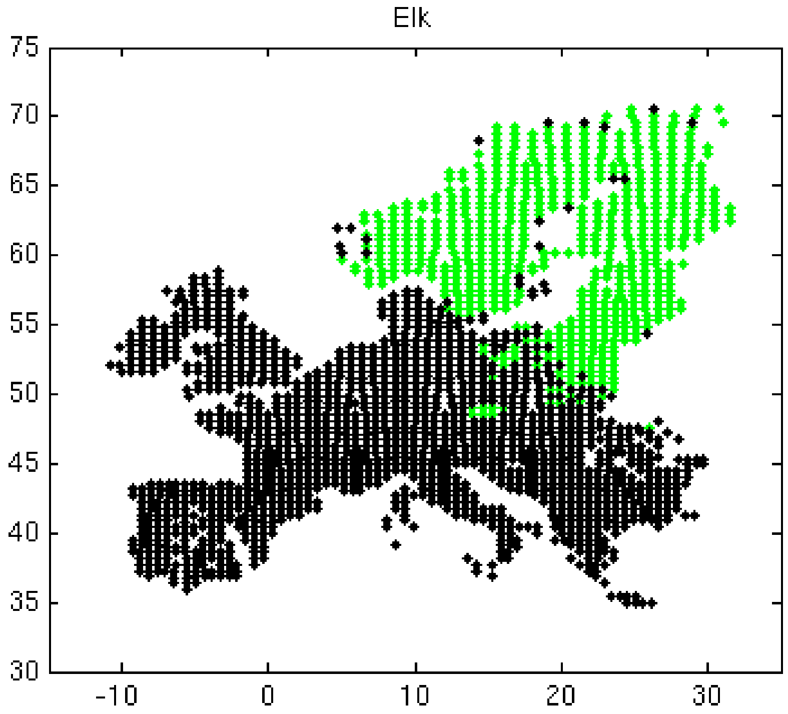




Hamming: 0.05, Jaccard: 0.99, 2. Kulczynksi: 0.98, Corr.dist./2: 0.5

Example





Hamming: 0.28, Jaccard: 0.93, 2. Kulczynksi: 0.48, Corr.dist./2: 0.39

Clustering:

```
[idx] = kmeans(data,9,’Distance’,d);
```

Cluster validity

- **Question**: How do we know that the clustering result we obtained is some how relevant?
- **Answer**: We can compare the result to results obtained for randomized data and see if the original clustering result is significantly better.

Cluster validity

- Invent a randomized procedure to generate credible random data sets.
- The hard thing is to decide what properties in the data should we keep constant and what should we permutate.

- My solution was to keep the data column sums constant and to preserve the spatial autocorrelation of the species ranges in the random data.

Algorithm for generating one random species range

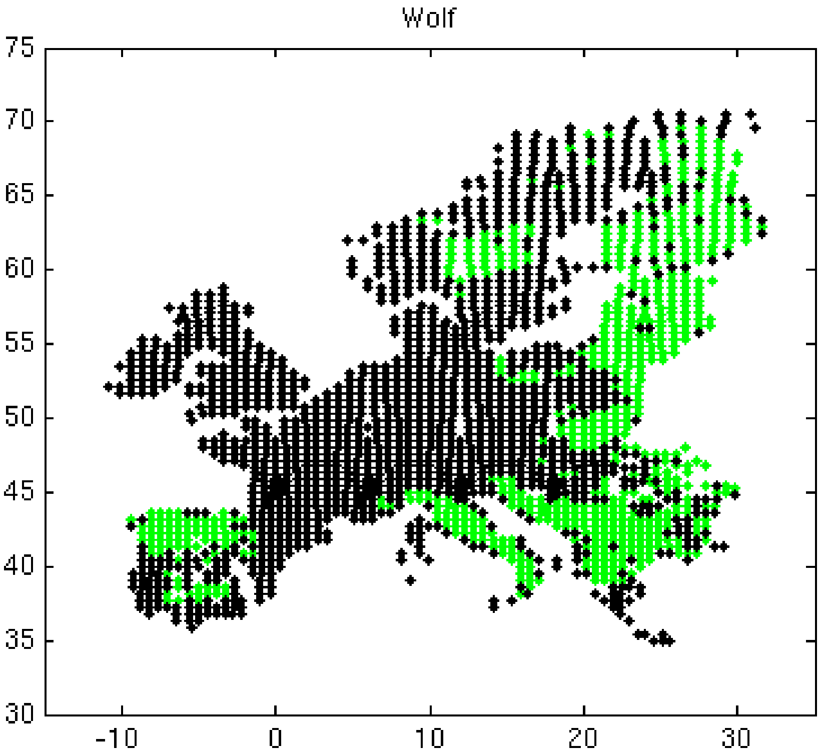
1. Define a spatial four-neighbour relations between the data cells.
2. Form a graph so that the vertices equal to the cells were the species is present and connect such neighbouring cells with edges.
3. Compute the number **c** of connected components in this graph and **Sk** the amount of edges (cells) the k:th component consists of.

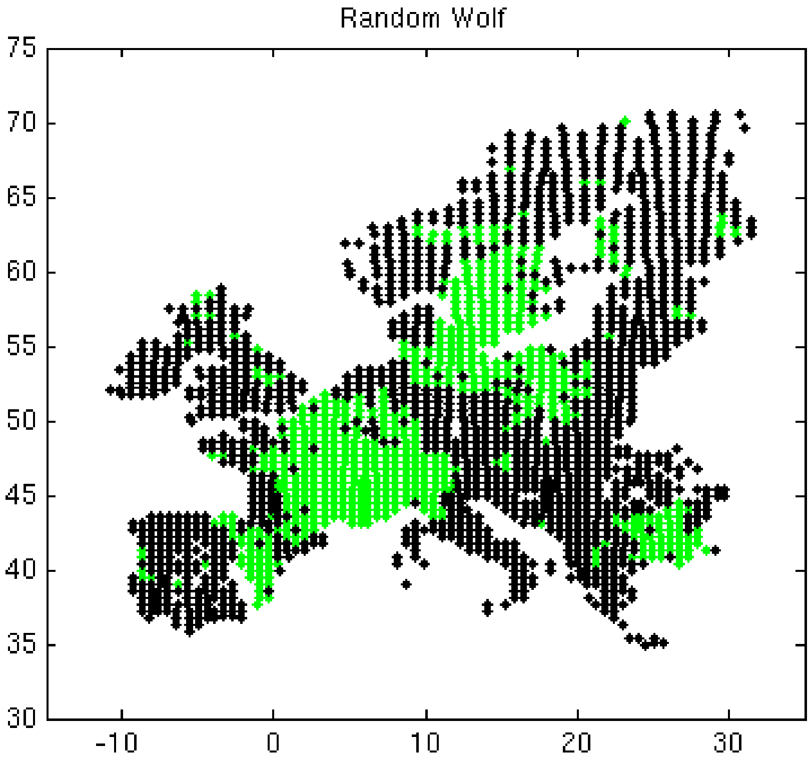
Algorithm for generating one random species range

Form a new graph for the random species range as follows:

```
4. For each k = [1,…,$\mathbf{c}$]
    5. At random turn an absence cell into a presence cell
    6. Until $\mathbf{S}_k$ absence cells are turned into presence cells
        7. turn accumulatively neighbouring absence cells into presence cells.
```

Example


Wolf
75
70
65
60
55
50
45
40
35
30
-10
0
10
20
30


Random Wolf
75
70
65
60
55
50
45
40
35
30
-10
0
10
20
30

Summary and conclusion

- Binary distance measures:
  - make sure your measuring something else than just frequency differences.
  - Some good experience from correlation distance.
- Cluster validation:
  - one possibility is randomization tests.
  - think what you want to permutated and what you want to keep constant.