Spatial epidemiology

Solymosi Norbert

Quantitative veterinary epidemiology

Department of Microbiology and Infectious Diseases

University of Veterinary Medicine Budapest

Spatial epidemiology is the branch of epidemiology that deals with the geographical processing and analysis of health-related data. Main areas:

- disease mapping
- spatial pattern analysis
- ecological analysis

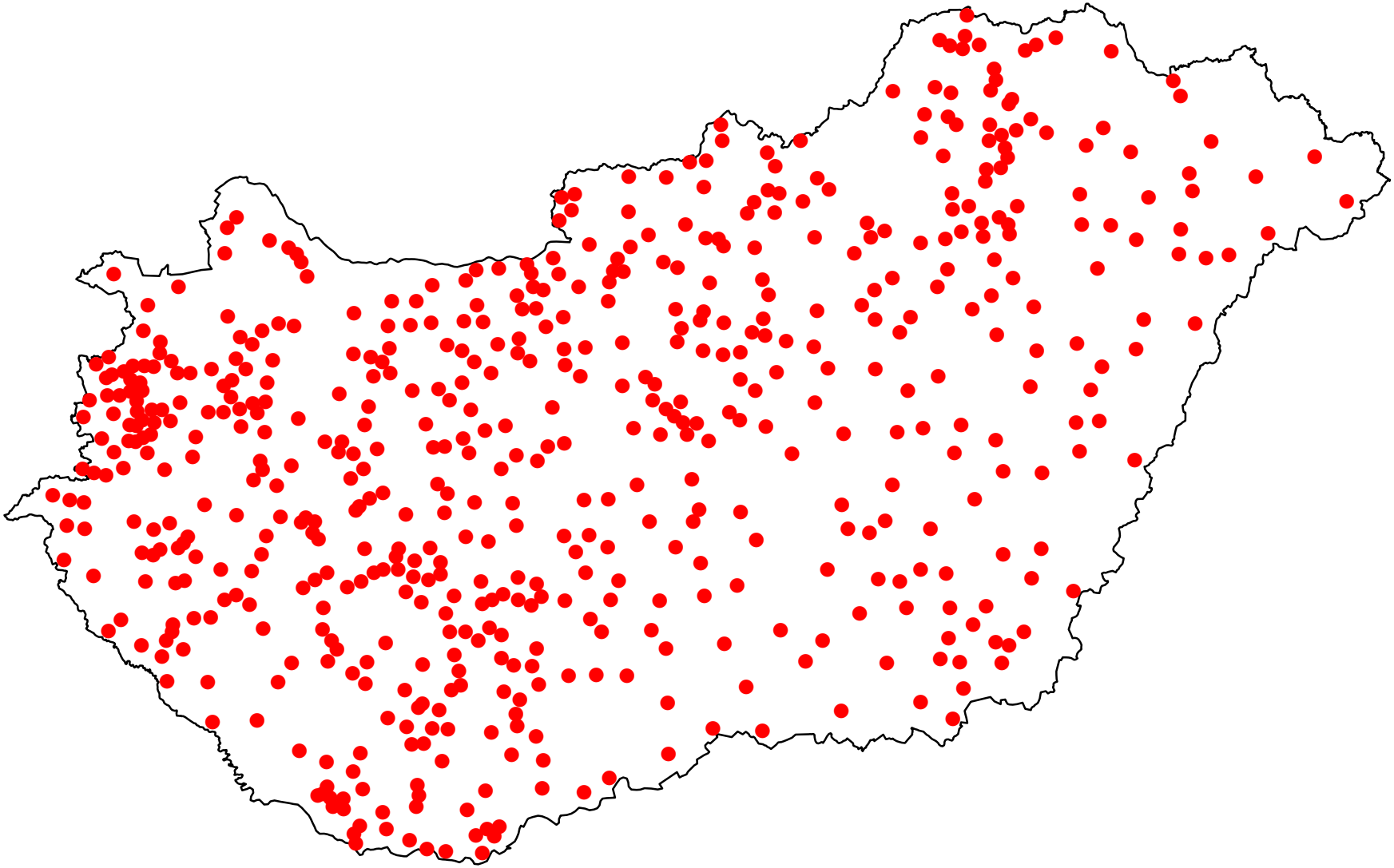
The purpose of the geographic representation of health-related data is to map the risk.

Risk mapping provides significant additional information compared to simple tabular representation.

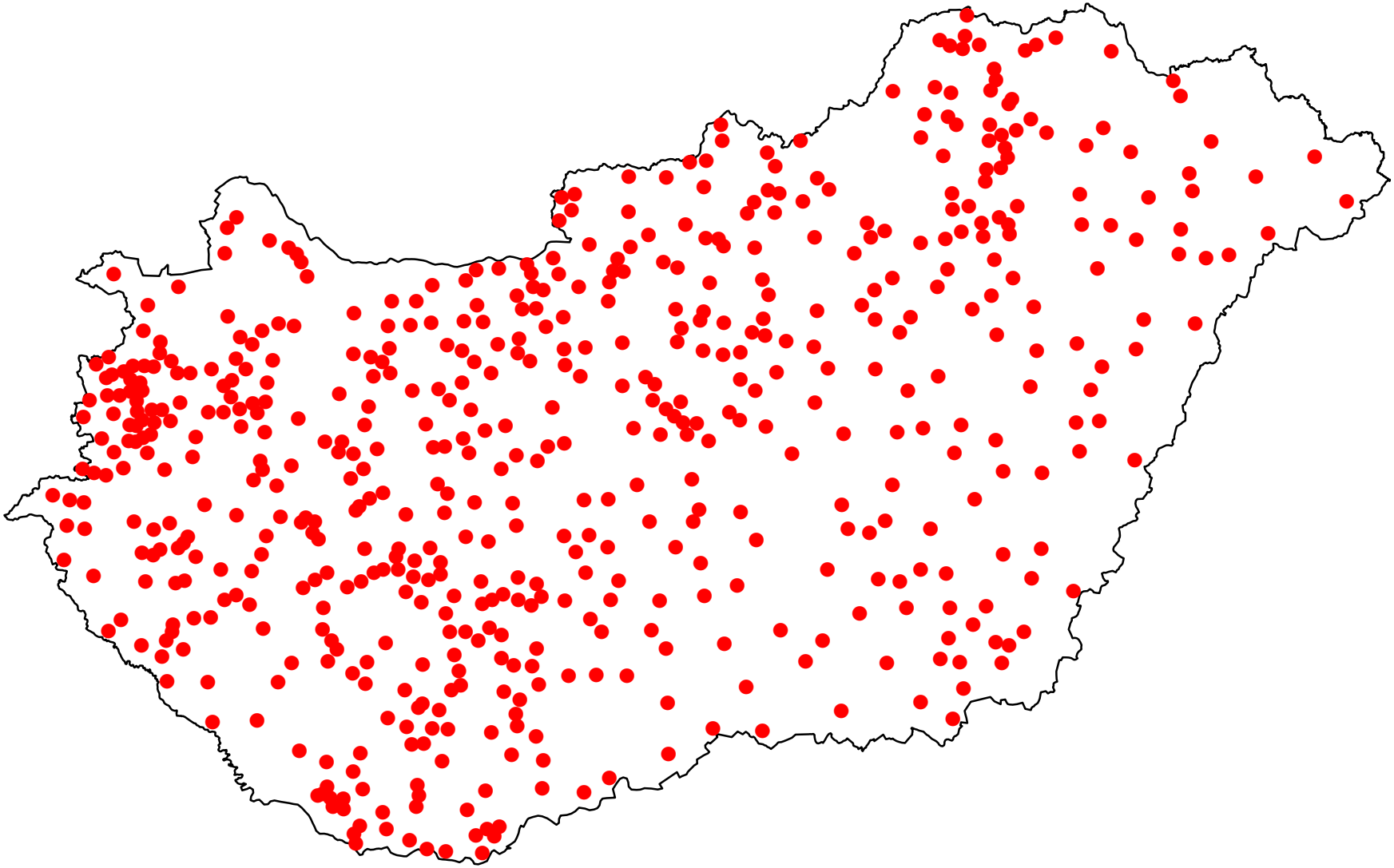
Several methods are used in epidemiology for disease mapping.

The simplest method of mapping is to mark the location of the disease with a dot on the map.

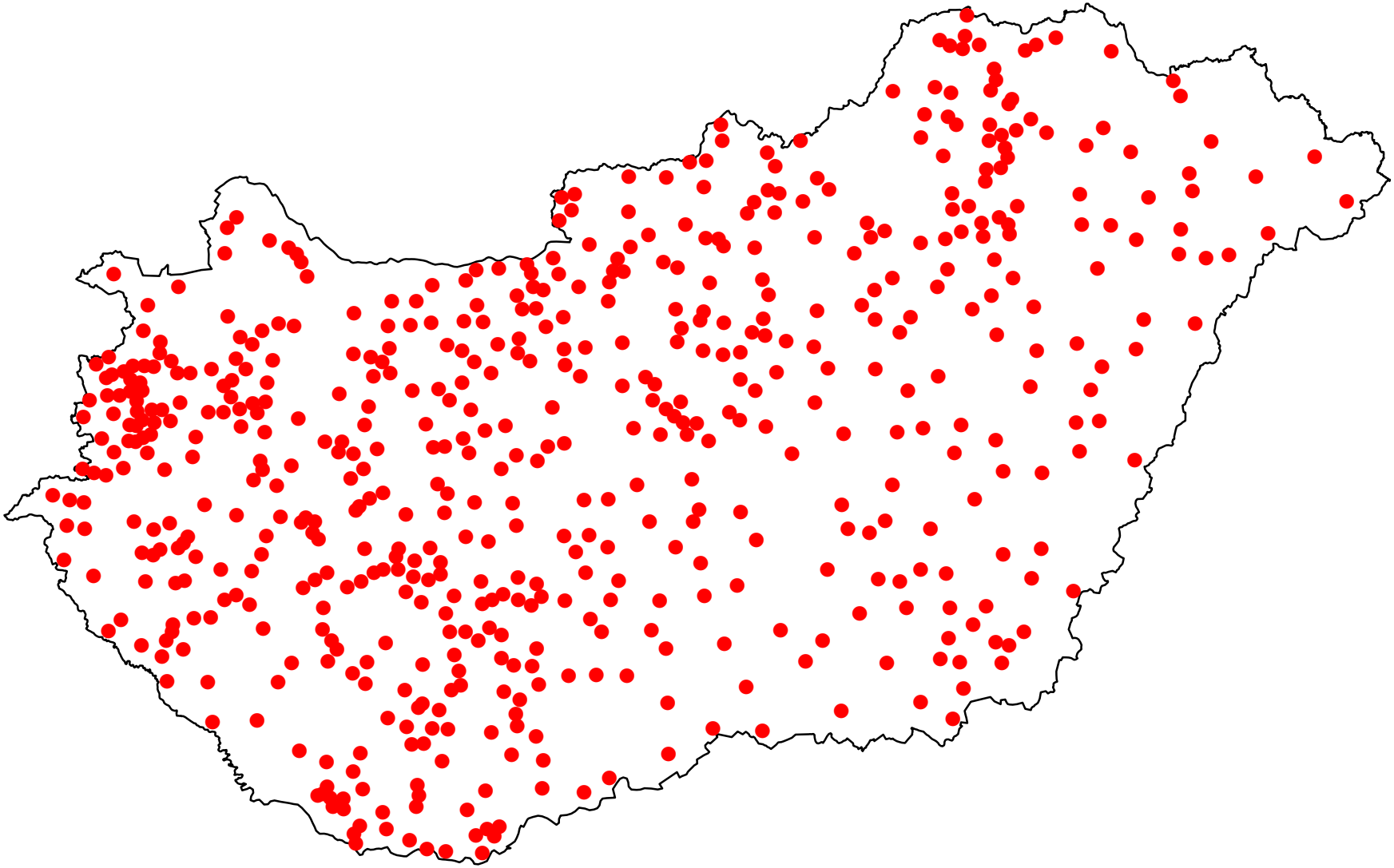
Health-related information in most of the cases has point-based geographical reference in its raw form (e.g., GPS coordinate, home address, location).



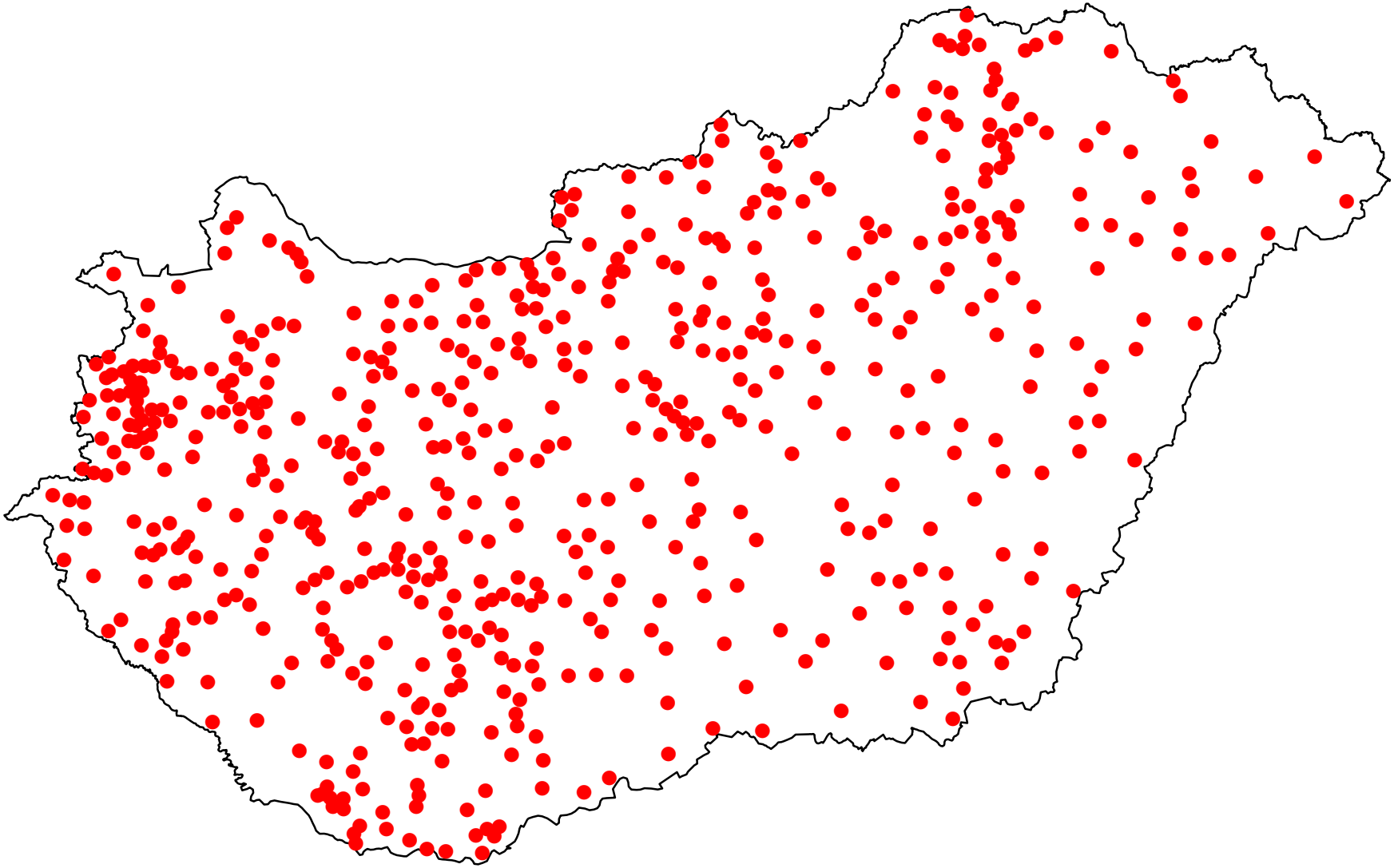
The red dots indicate the settlements from which red foxes were diagnosed with rabies during 1990.



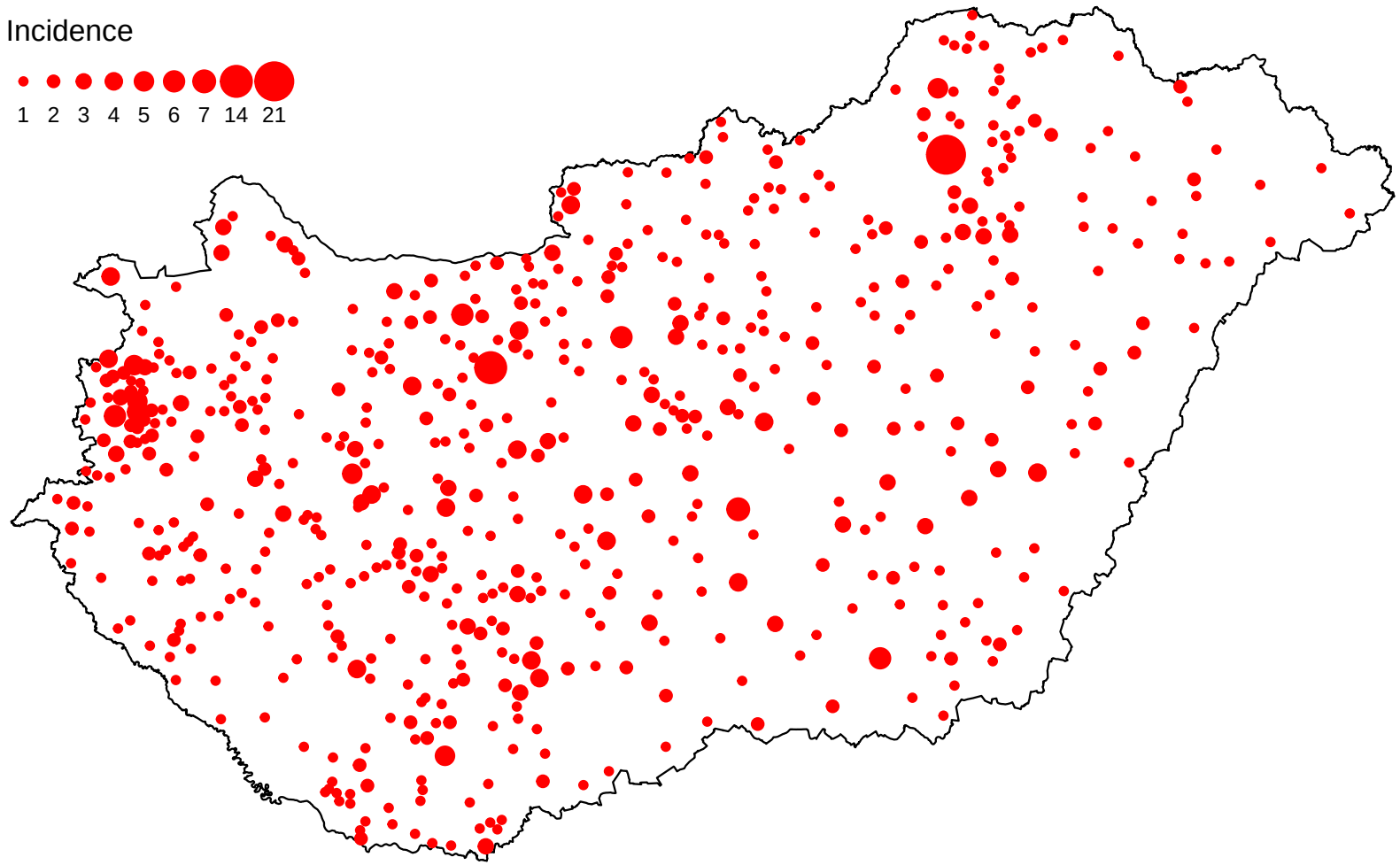
However, in most cases, limited information can be drawn from point-based mapping about the geographical pattern of risk.



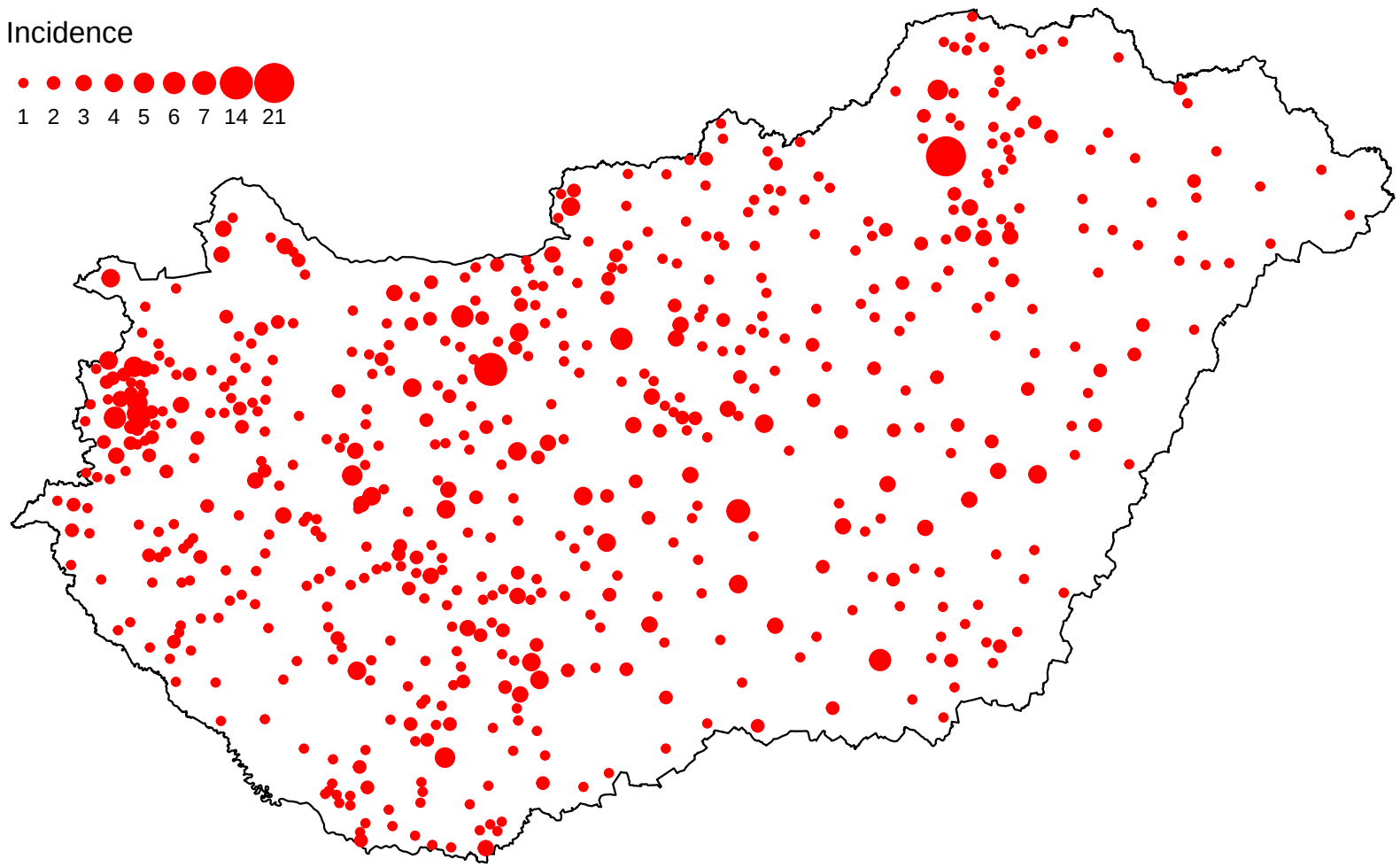
Impressions based on point maps can be significantly influenced by the resolution used in mapping, the size, shape, and colour of the points.



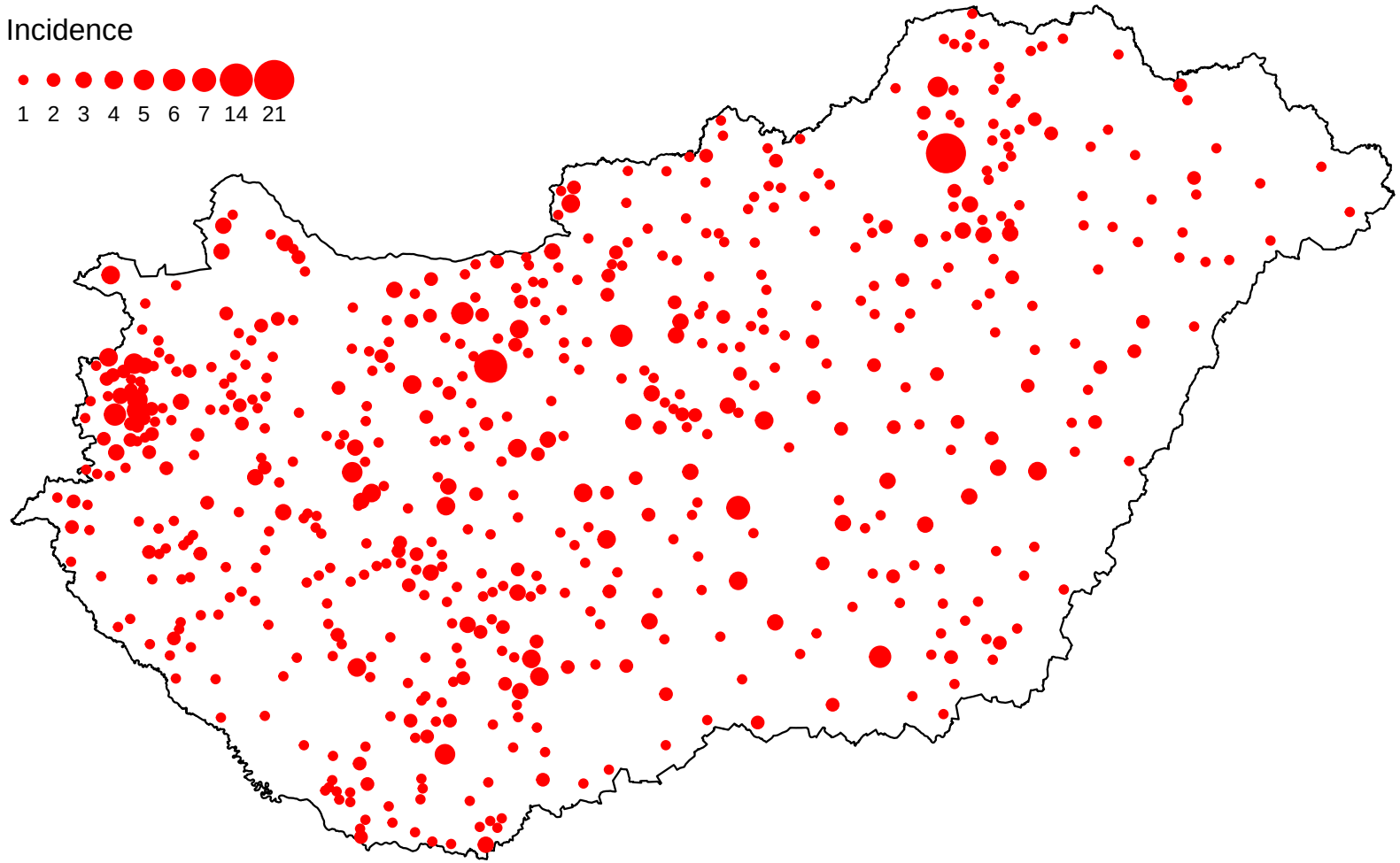
The weakness of point maps is that we cannot display the case number with them, only the positive/negative positions.



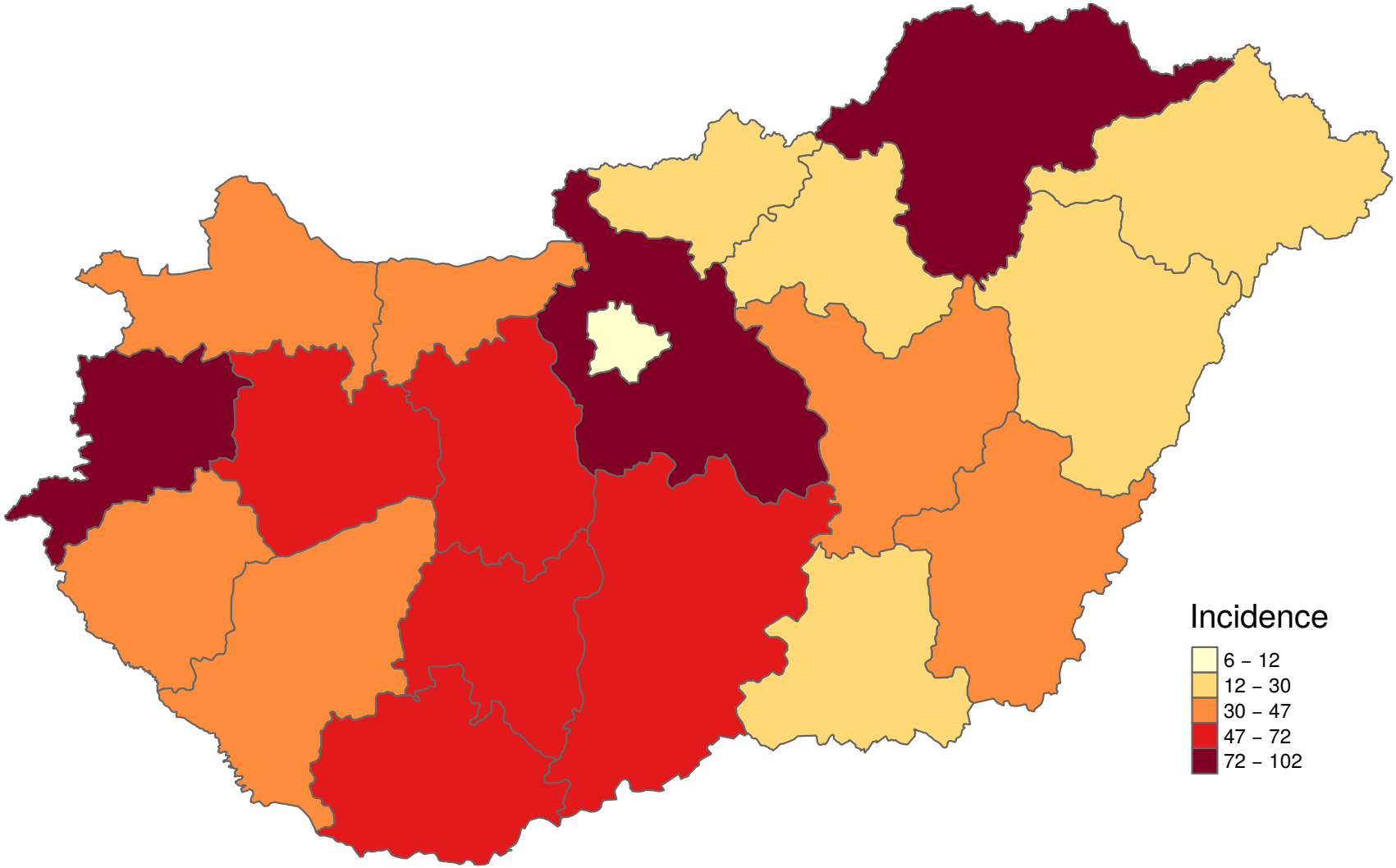
By the so-called bubble maps, it is possible to show how many cases were found in a given place or settlement in the given time period.



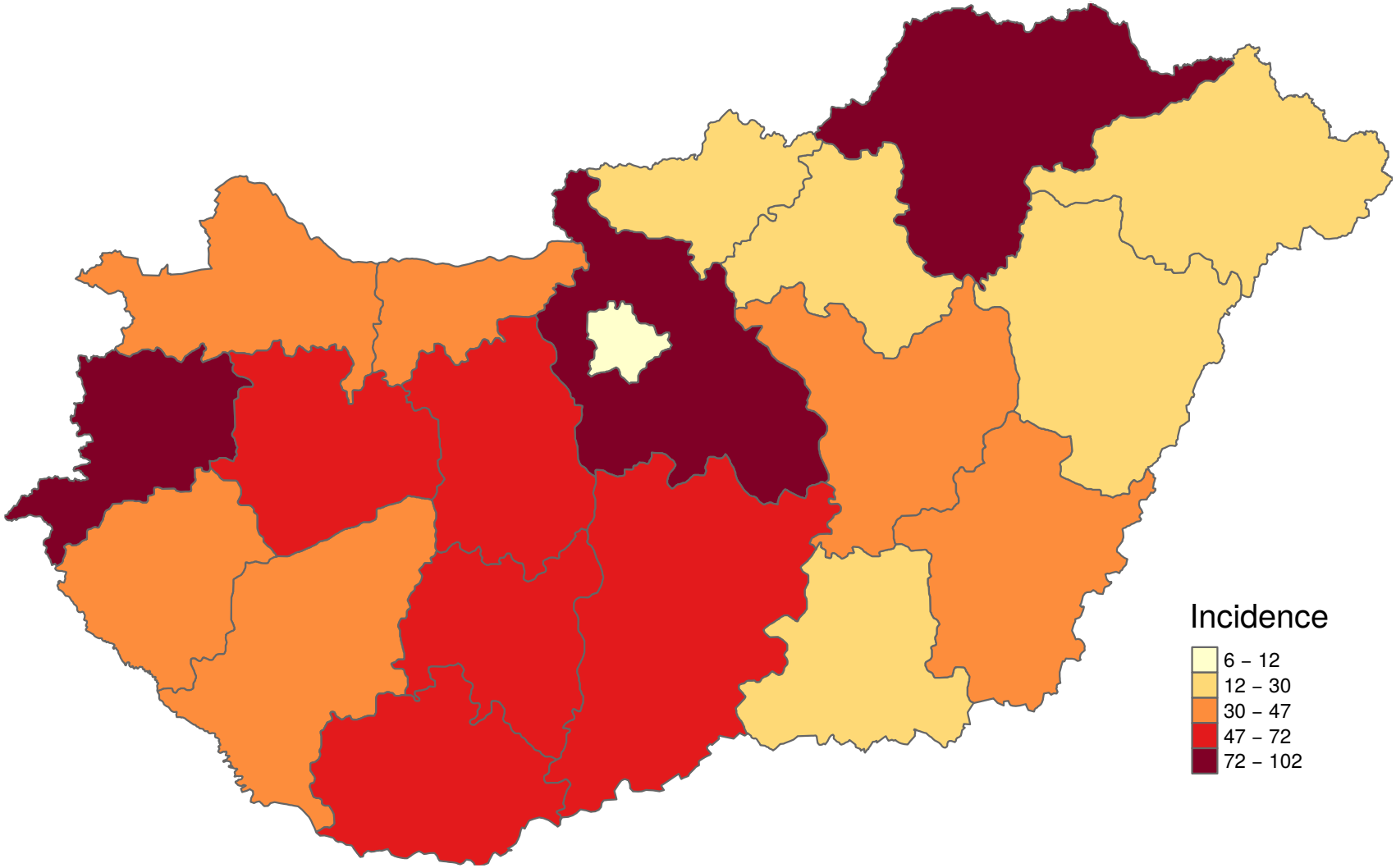
The size of the mark used on the map (in our case, a circle) shows the number of foxes diagnosed with rabies.



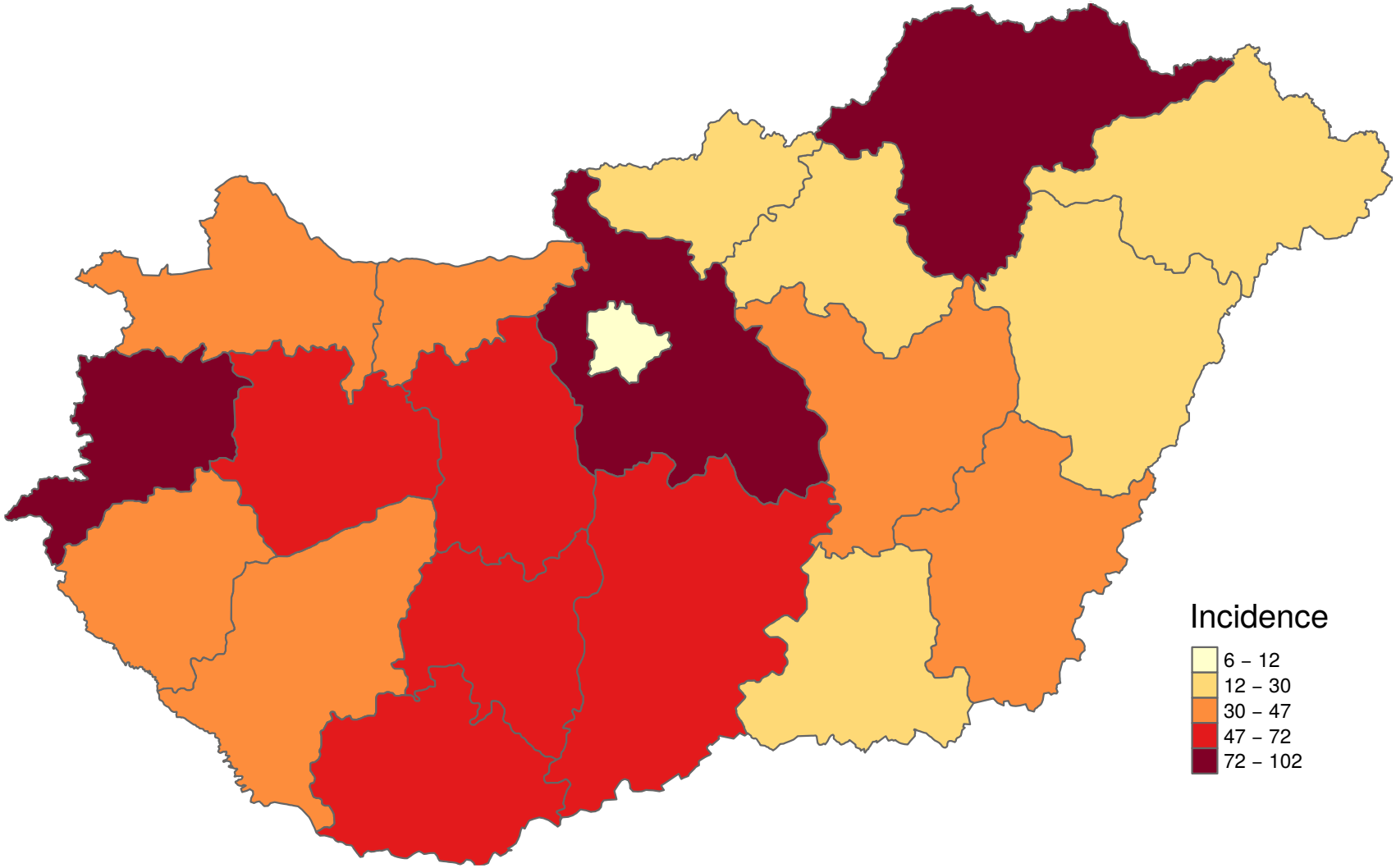
Here, in addition to the technical factors mentioned in the point maps, another distorting moment is that if there is a settlement in which there have been many cases, the positive place in its neighbourhood might be overlapped by the sign showing the number of cases.



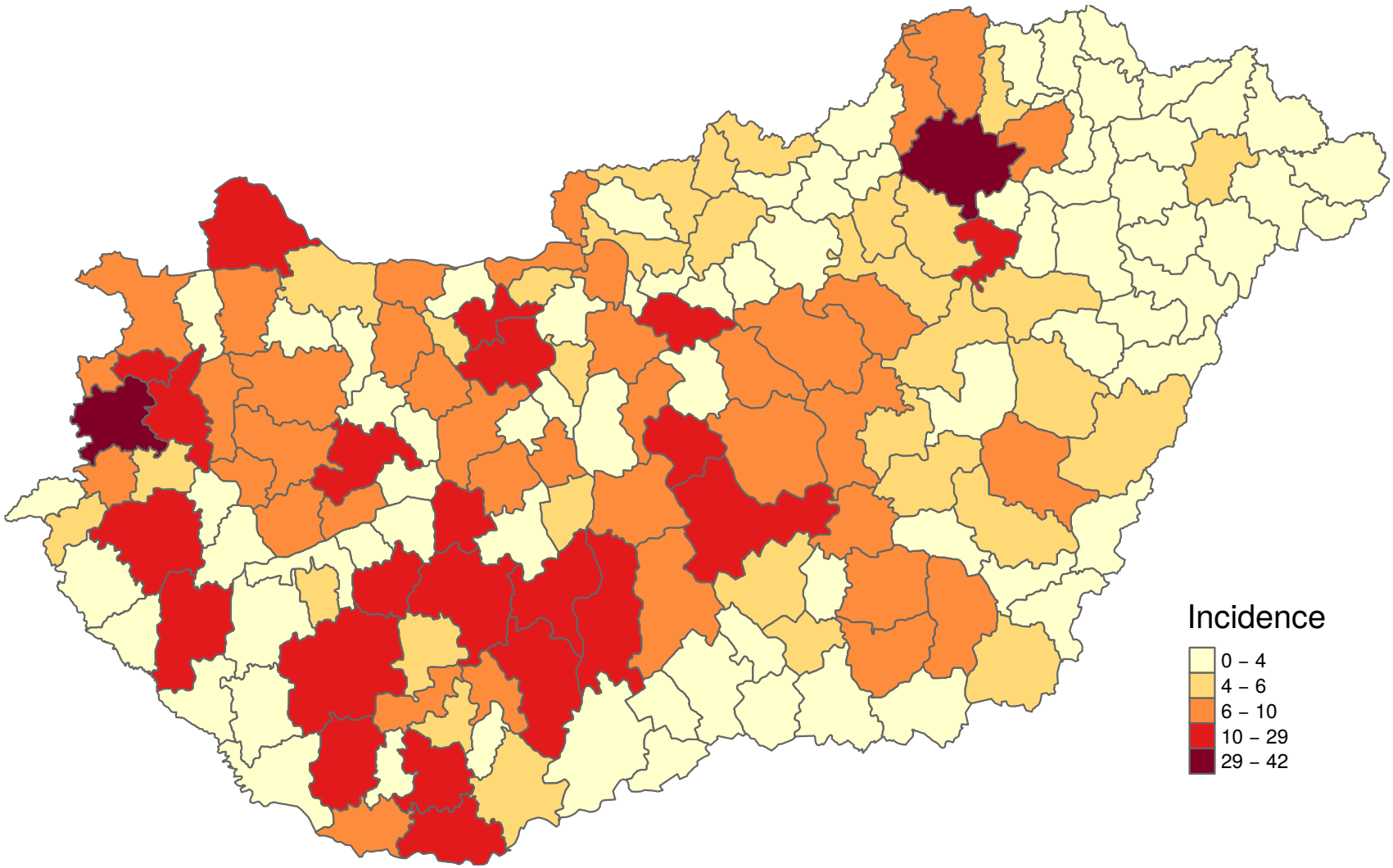
Additional possibilities in risk mapping are the so-called choropleth maps. The name is of Greek origin, khora means the region of the settlement, and plethos means the people.



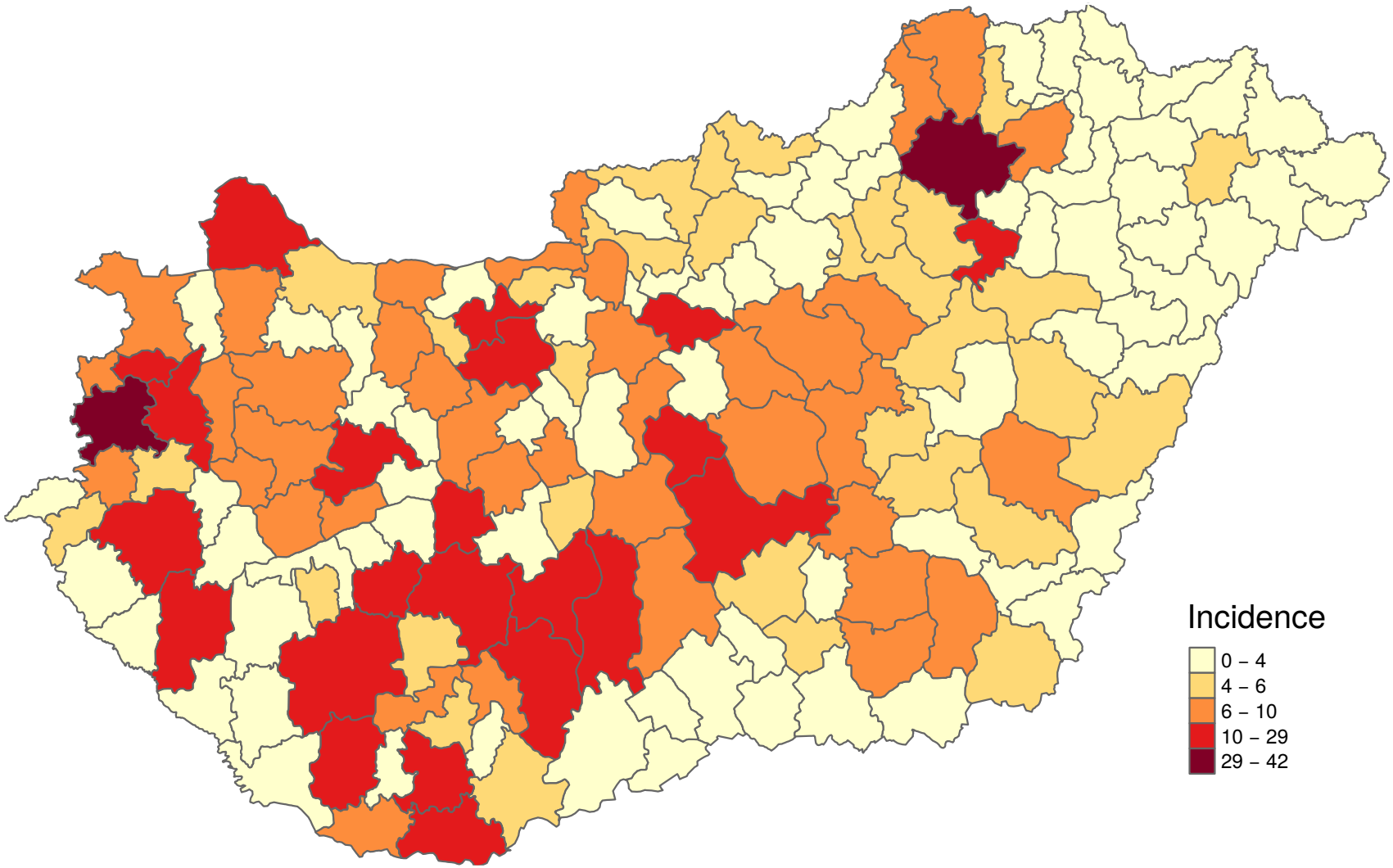
In disease mapping, choropleth maps accordingly represent the risk of occurrence aggregated by geographical units, regions.



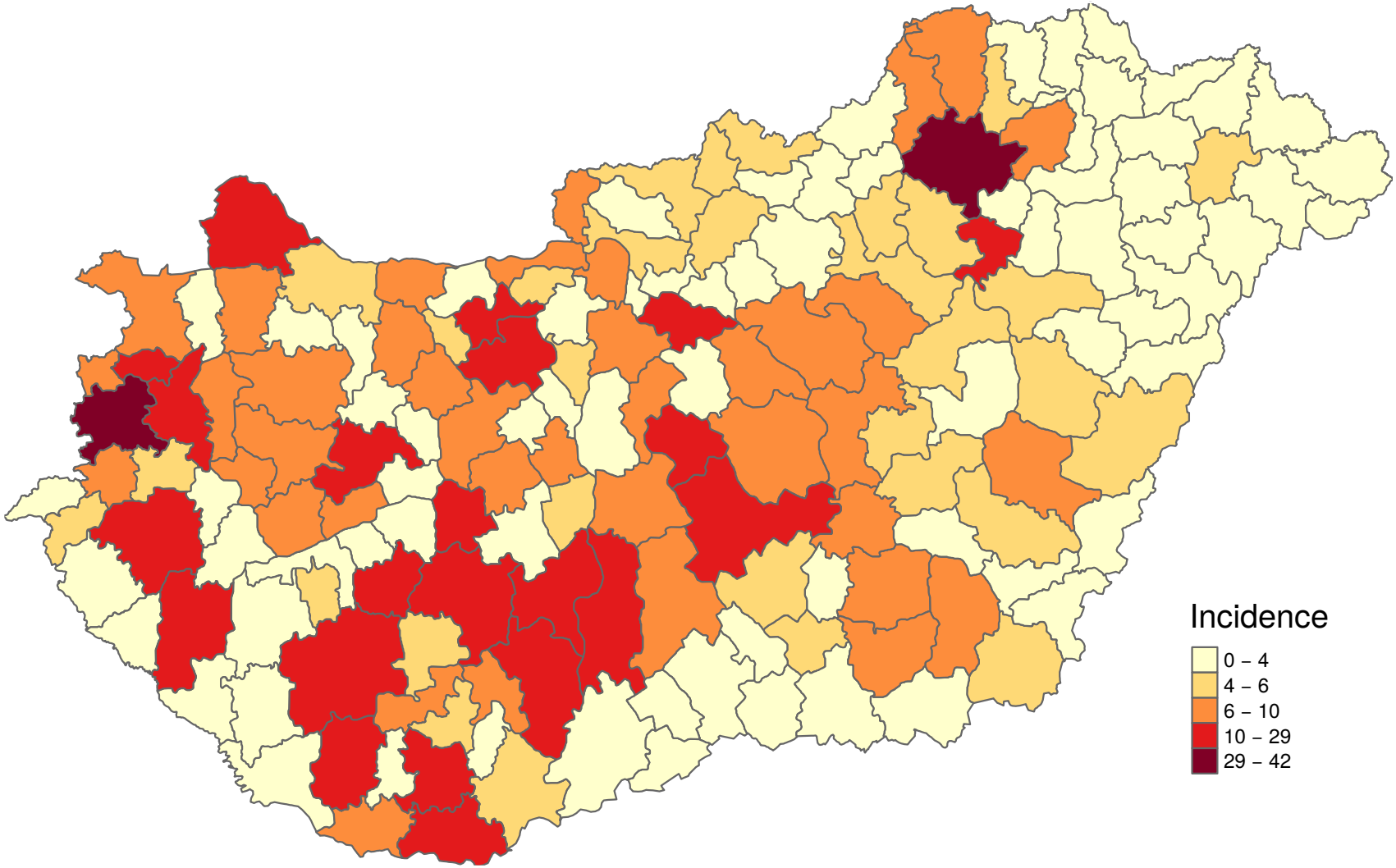
The basis for aggregation can be arbitrary, although it can significantly influence our conclusions drawn from the map. The map shows county-level aggregation of rabies cases during 1990.



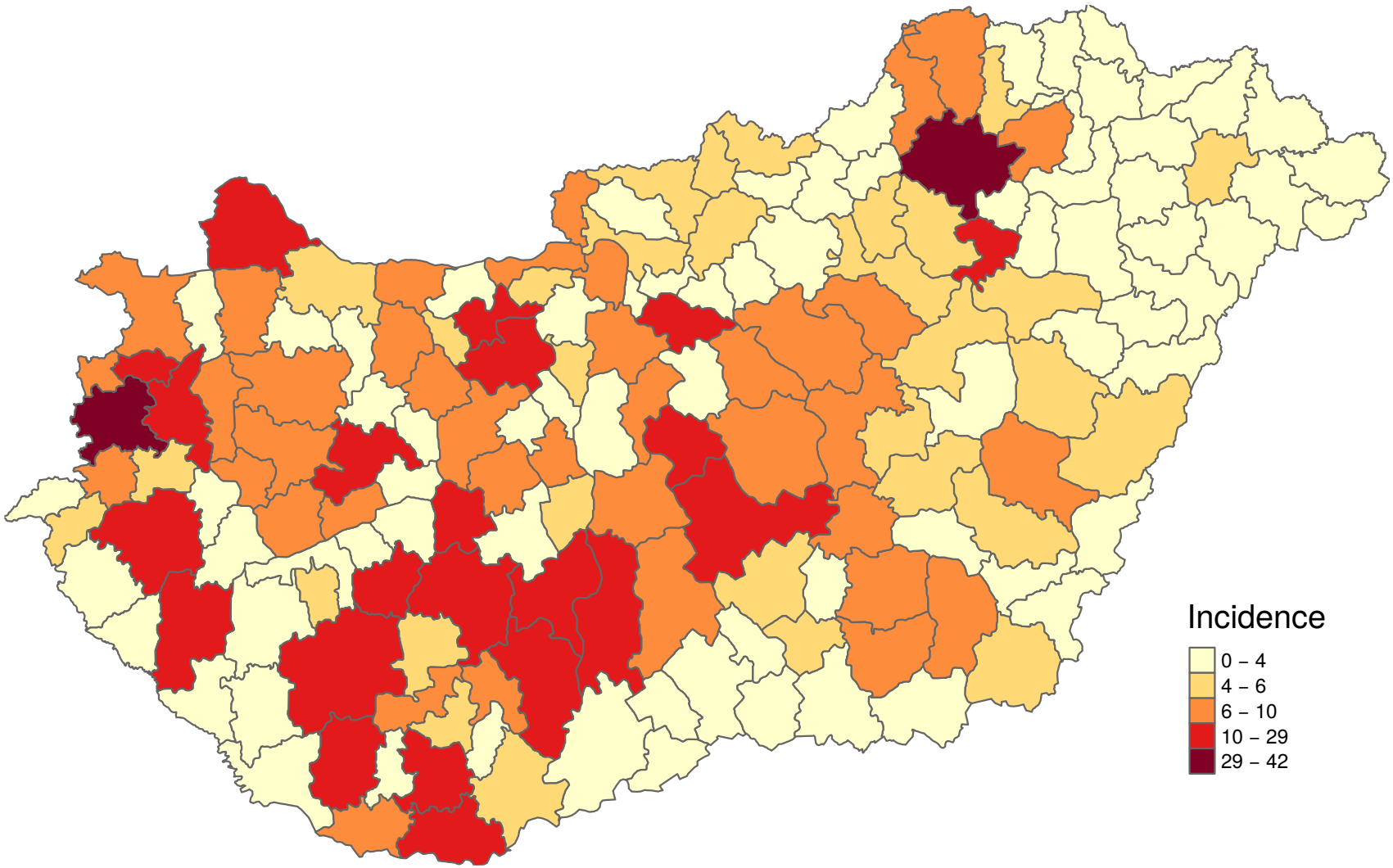
The same rabies cases aggregated on administrative subregions of Hungary.



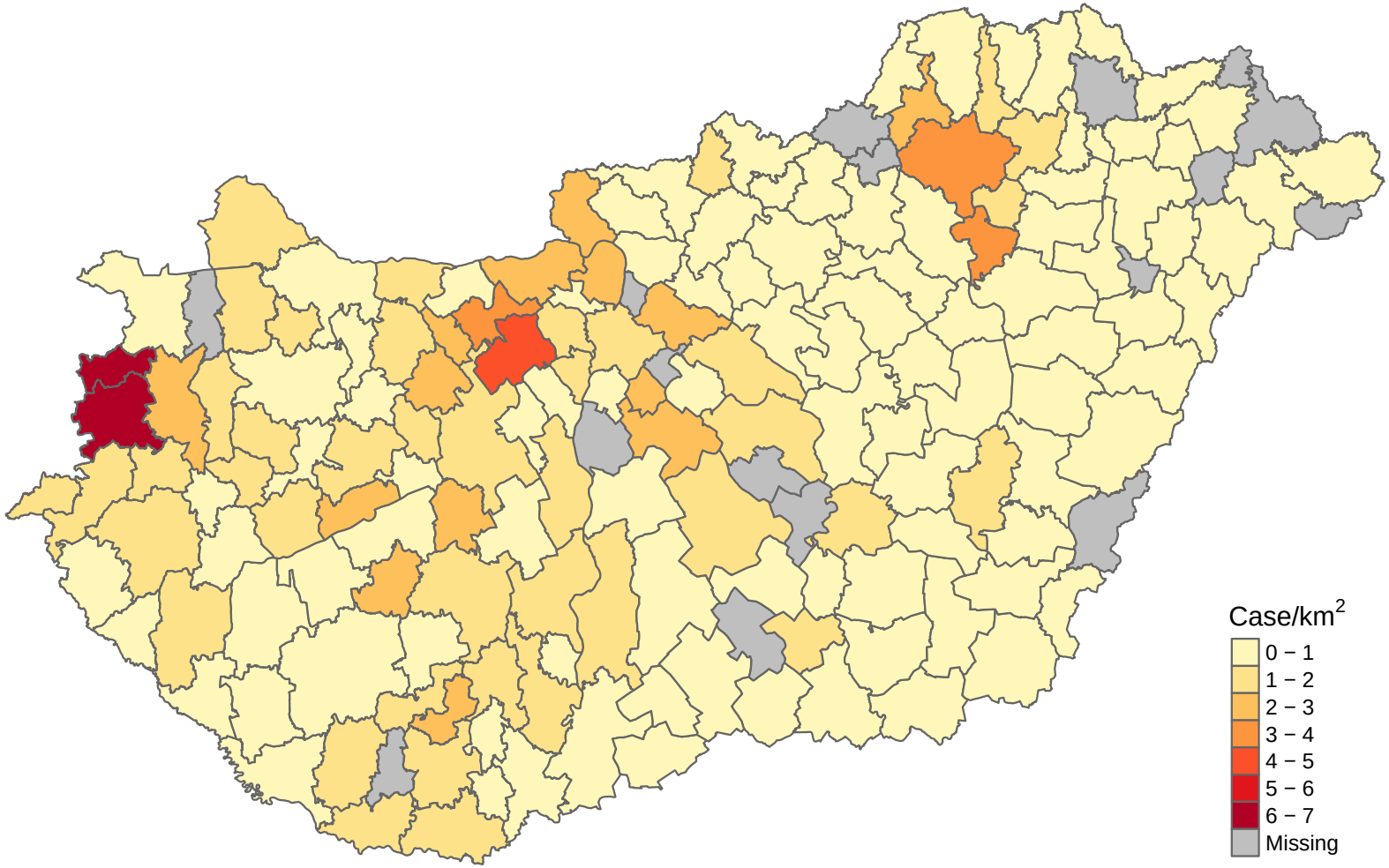
From a technical (Geographic Information System, GIS) point of view, the basic units of aggregation and representation on choropleth maps are polygons. These polygons are in close contact with their neighbours and fill the study area.



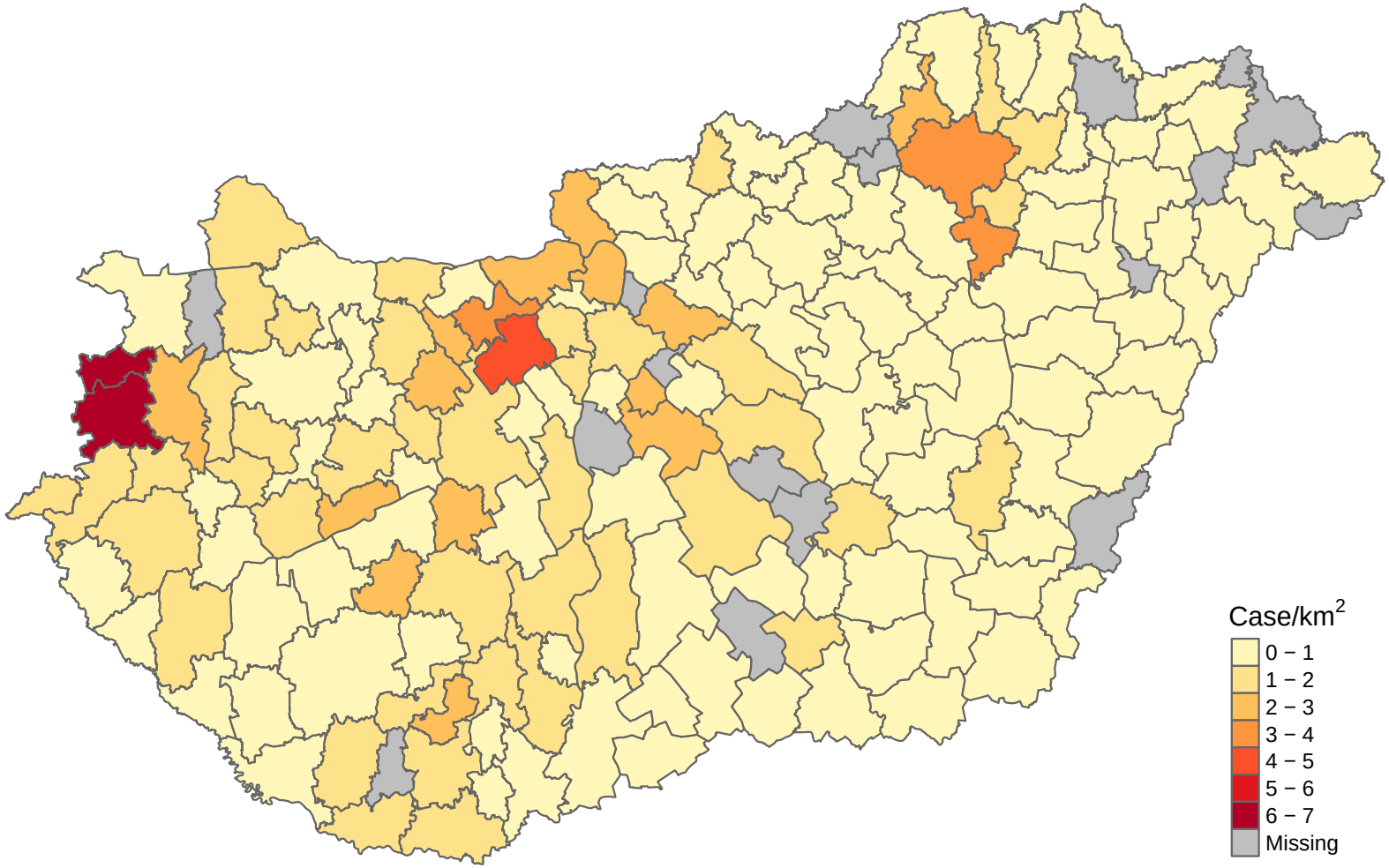
This avoids overlapping the adjacent data mentioned in the bubble map. Furthermore, this allows us to represent our discontinuous data on a continuous surface, which makes it easier to assess the risk in space.



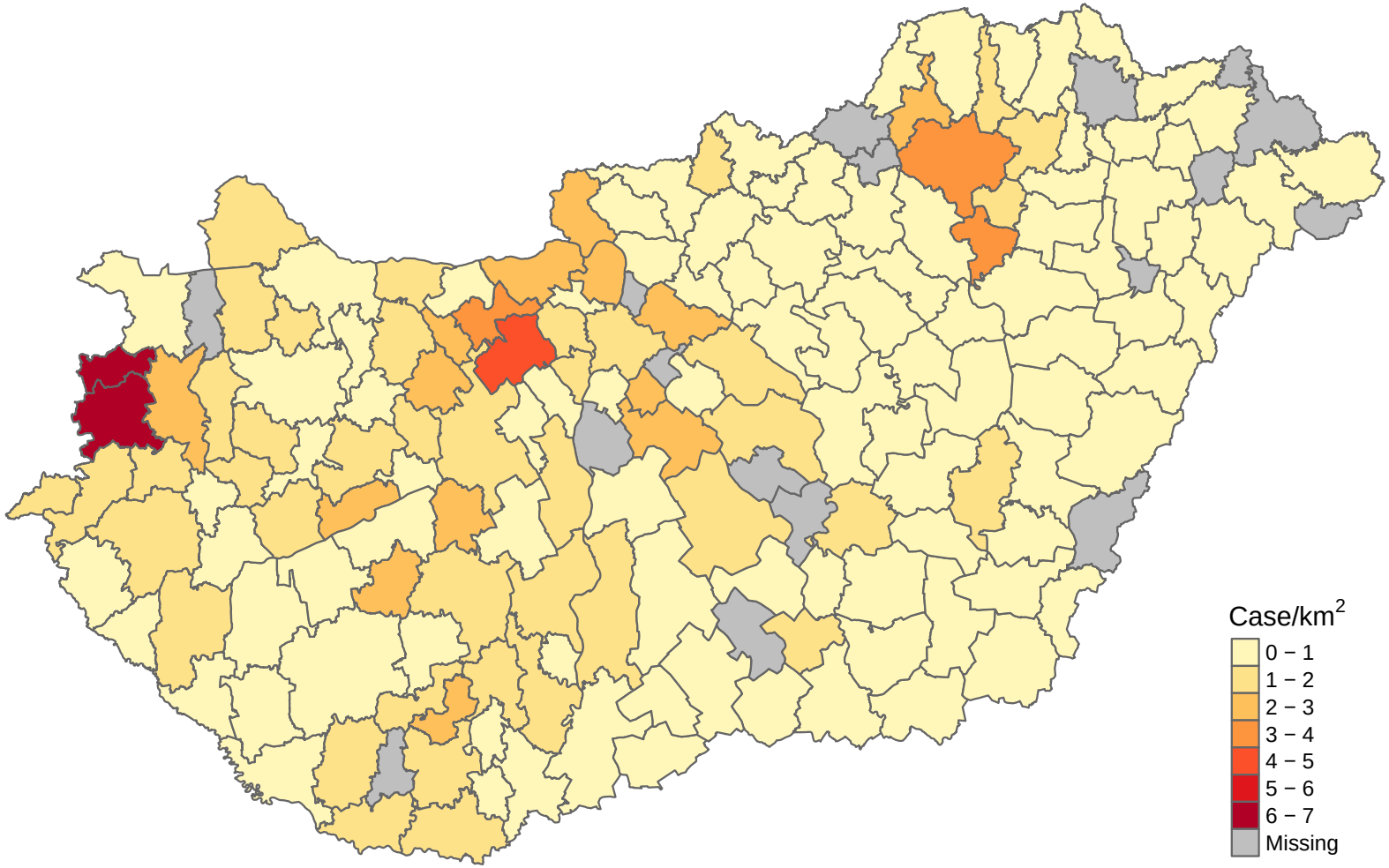
Based on the map, subregions with larger area may appear having a higher risk. Before we accept this, let us think that if the rabies risk of foxes were homogeneous in the country, we would have more cases in larger subregions.



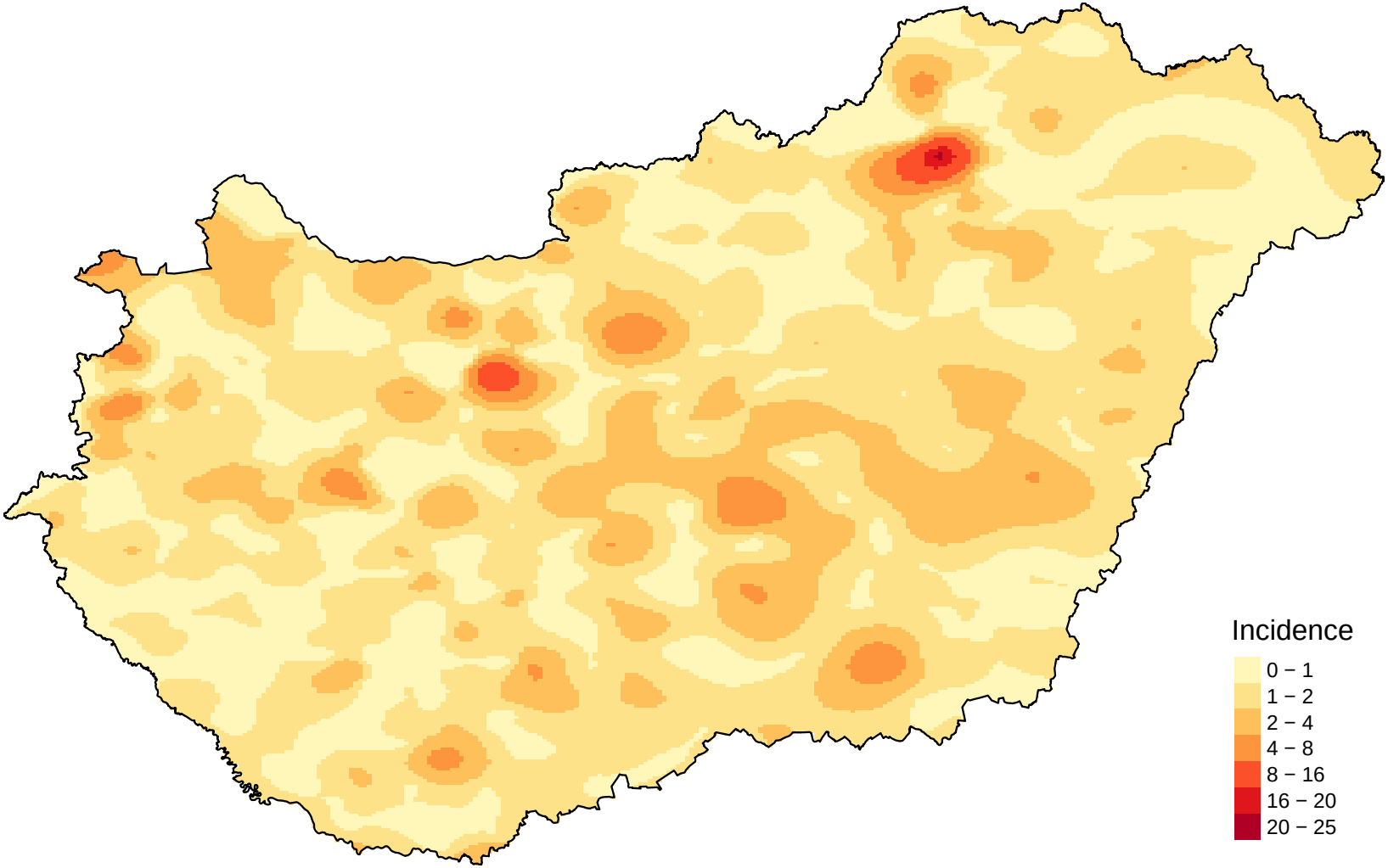
Because reliable population data for foxes are not available, we use the size of the subregions instead of the population to normalize the number of cases.



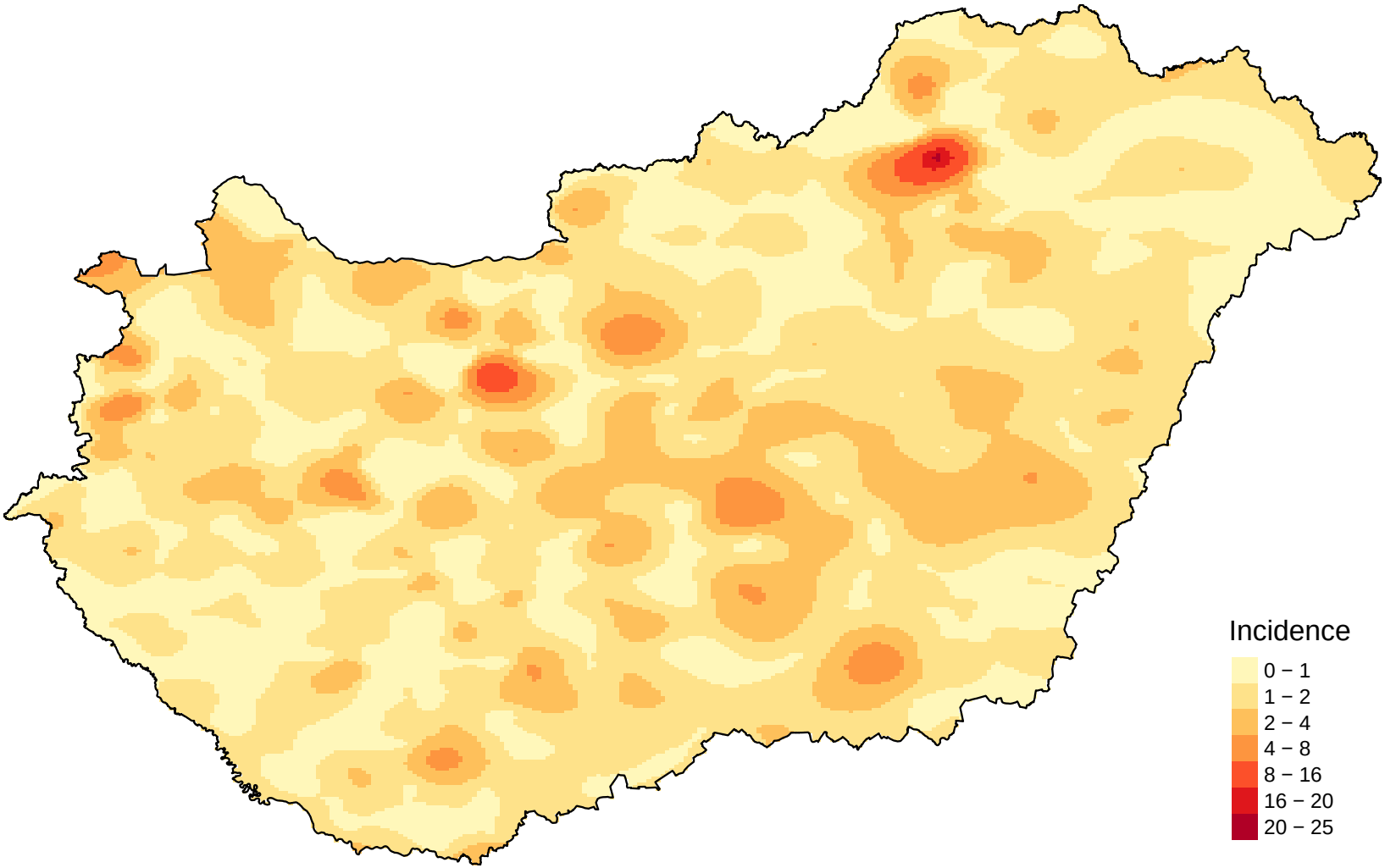
Based on this, it seems that there was a very high incidence in the Szombathely subregion in 1990, and some subregions with moderated risk in Transdanubia and Miskolc environment.



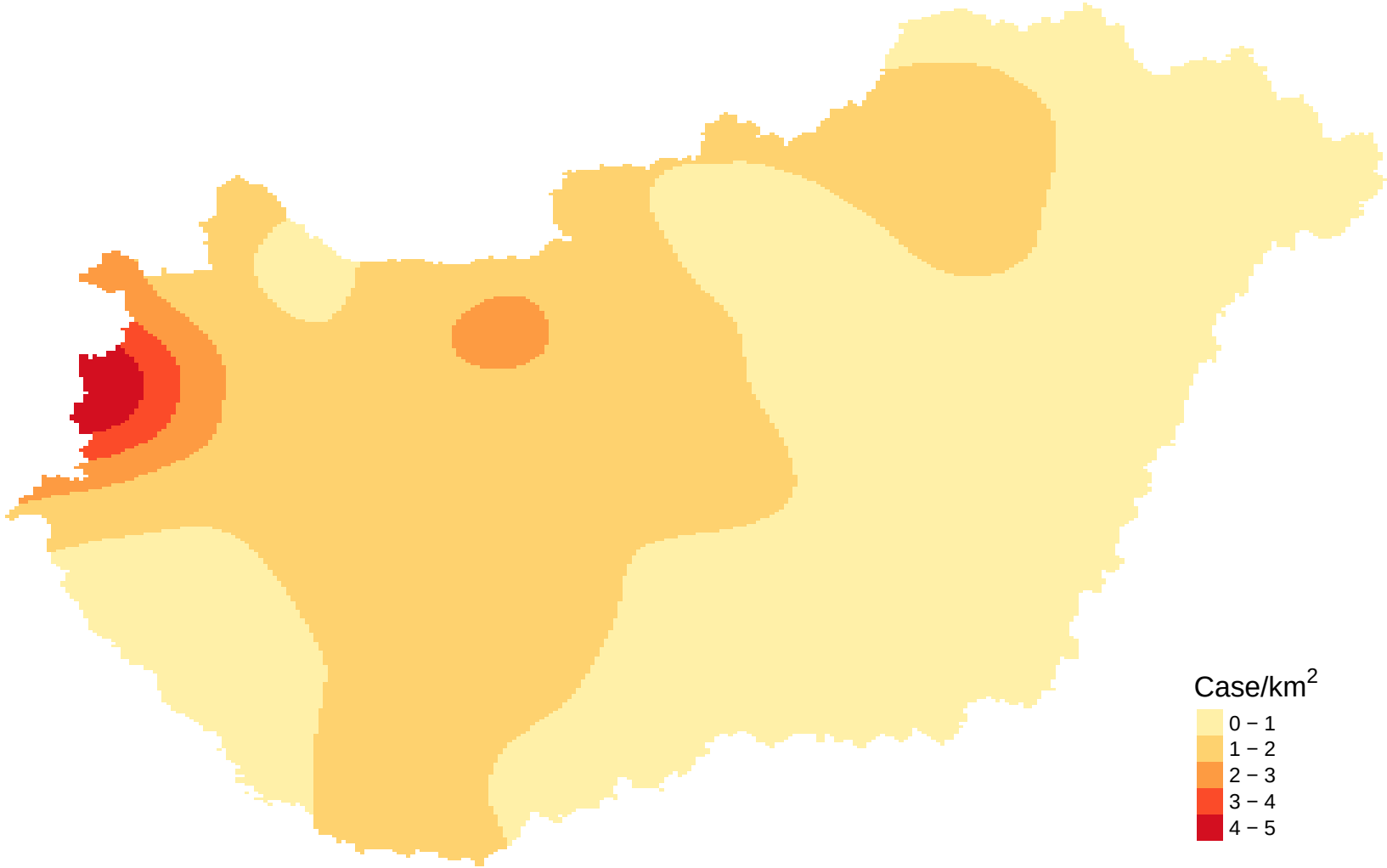
The evaluability of choropleth maps can be significantly influenced by the colour scale used painting territorial units, and the choice of value ranges for categories.



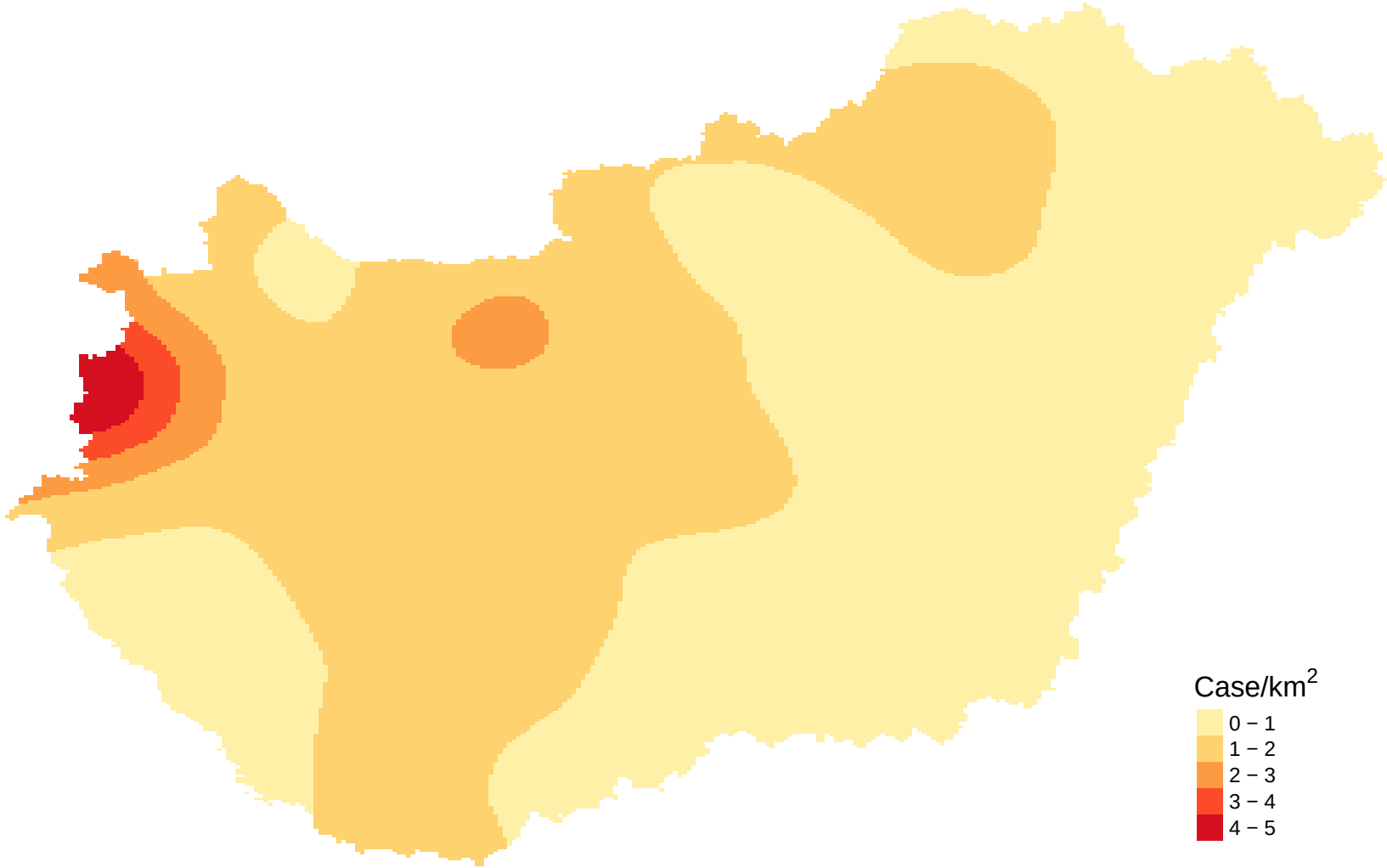
Further spatial risk visualization opportunity is the so-called isopleth mapping. When creating isopleth maps, the point-like source data are interpolated in the two-dimensional space by some smoothing method.



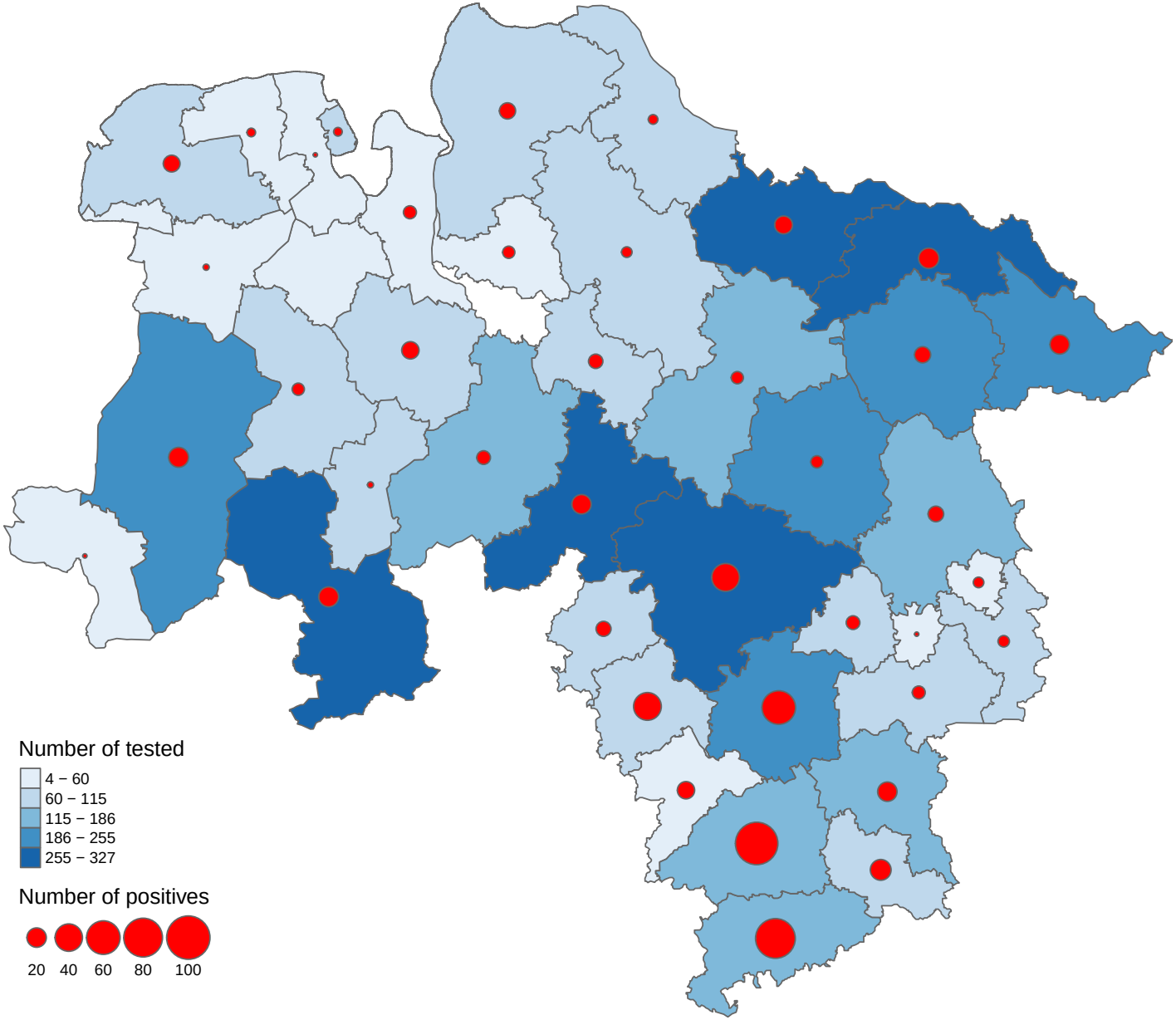
Isopleth map created from settlement-level data of rabies cases diagnosed during 1990.



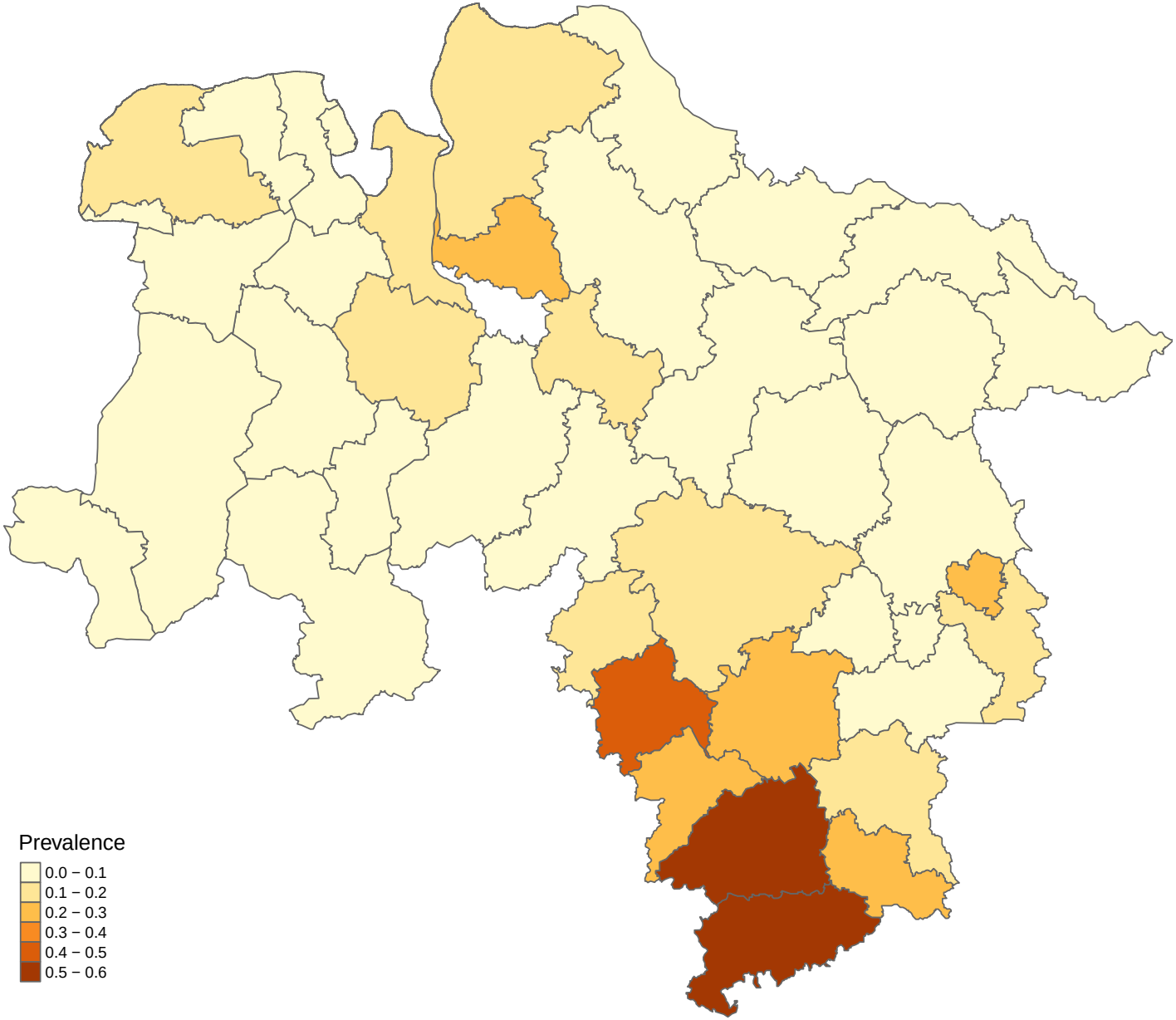
Isopleth map based on the rabies cases detected during 1990 normalized by the subregion area size of origin.



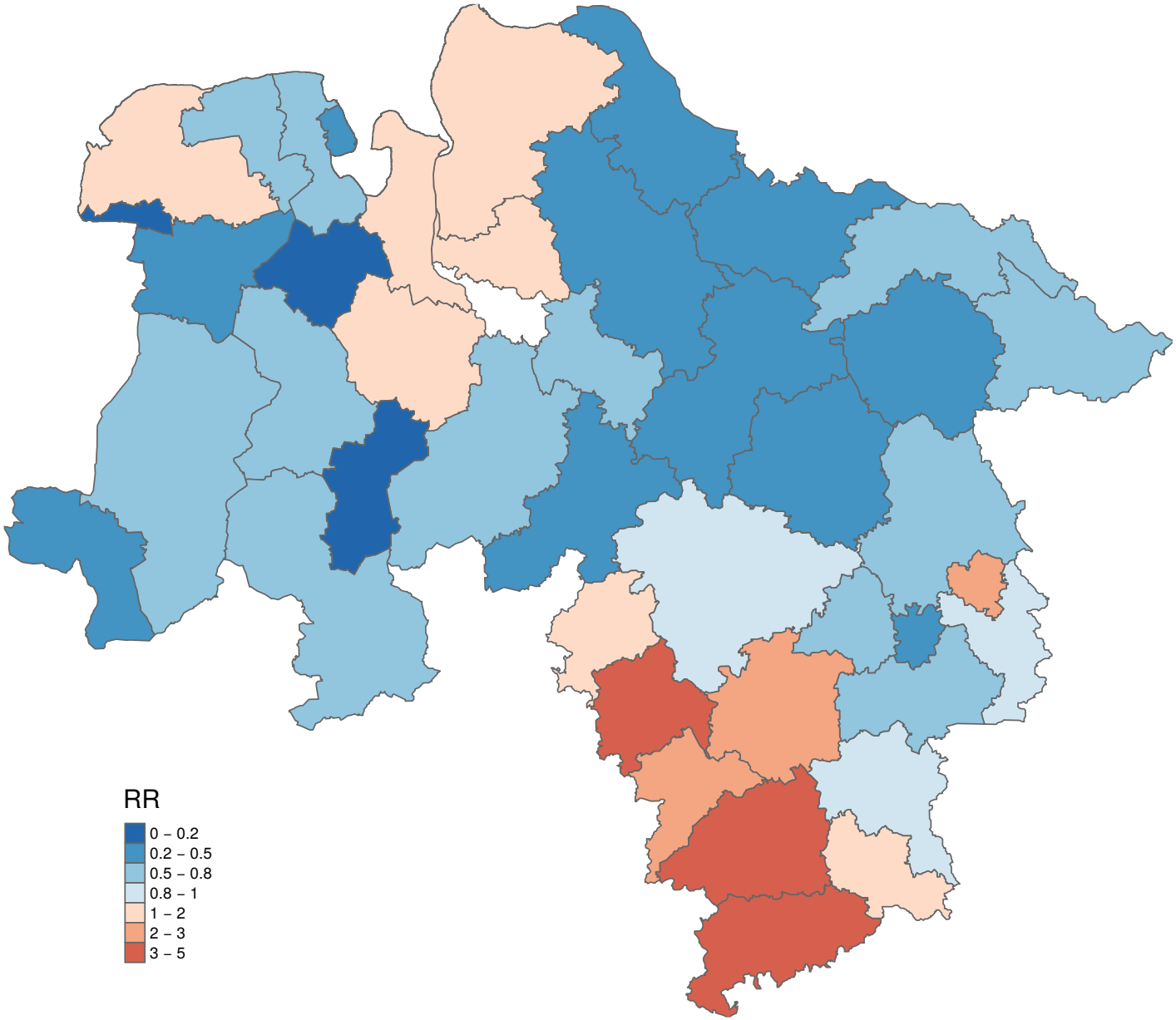
The smoothing method used to create isopleth maps and its parameterization can greatly influence the risk map, and thus our epidemiological inferences.



The previous disease mapping procedures may be combined or supplemented with other information. The number of tested and E. multilocularis positive red foxes in Lower Saxony (Berke, 2001).



E. multilocularis prevalence per district, estimating prevalence simply as the proportion of positive cases.

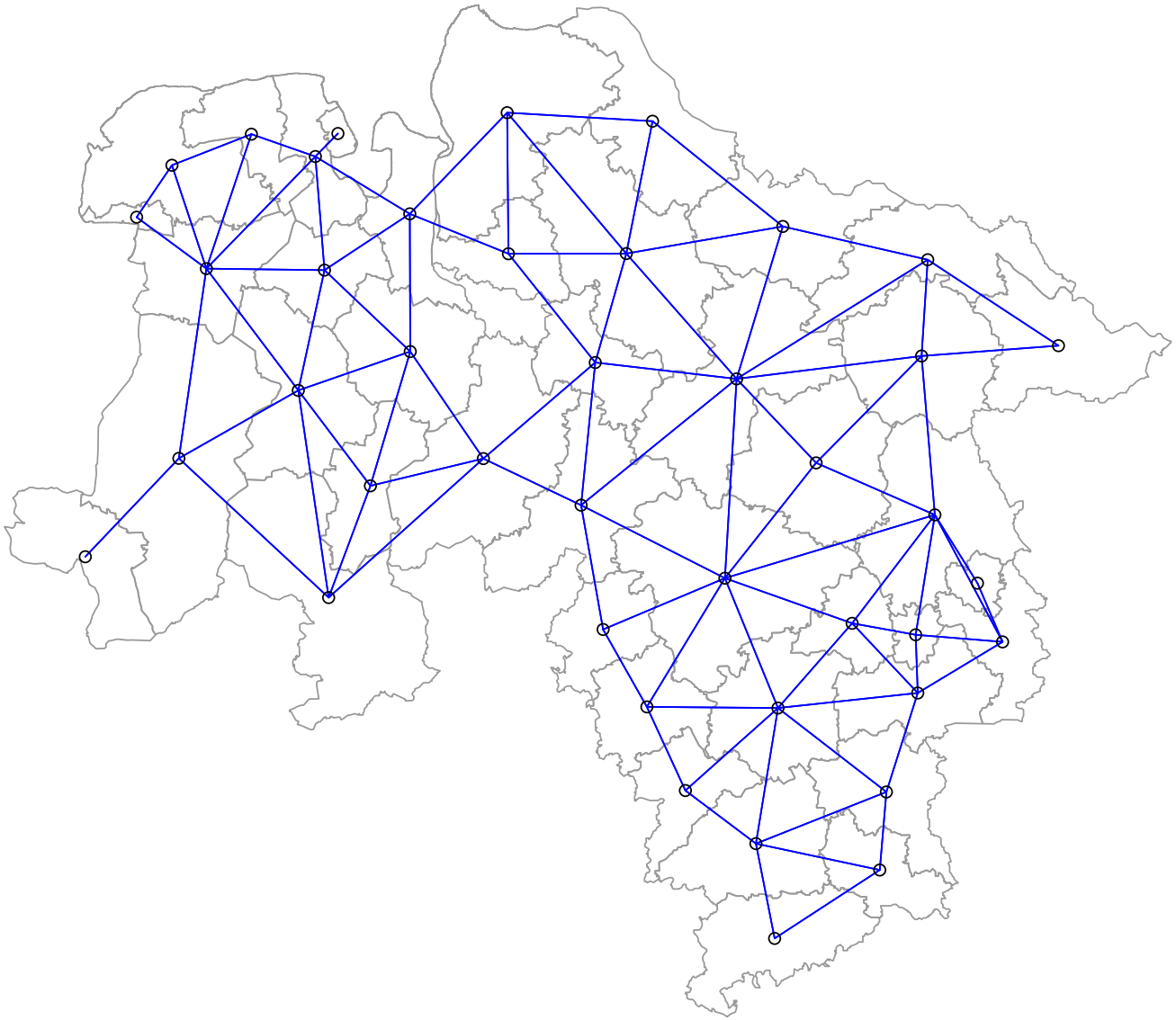


Relative risk map of E. multilocularis infection in foxes. The RR expresses the risk of each district in proportion to the average risk of the entire area.

The risk mapping approaches so far, except the two-dimensional smoothing, are not real spatial models because they do not involve the geographical distance/proximity of events from each other. So far, maps have been used for visualisation only, but their structure did not affect the estimates.

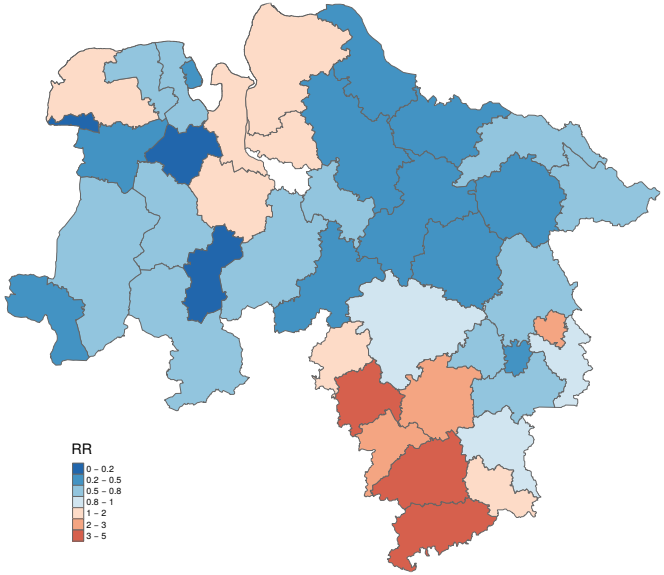
However, it is easy to recognise that the risks of E. multilocularis infection observed, at least in the direct neighbourhood, may not be completely independent of each other. Just keep in mind that administrative boundaries, in most cases, do not constitute a physical barrier or match with natural boundaries that would block the movement of foxes.

Thus, the territory of a fox shot in one district may overlap another district, or it may have moved from one district to another for some exceptional reason. Alternatively, the origin of infection may not be in the district in which the case was registered. According to these, our risk estimates may be biased (ecological bias).

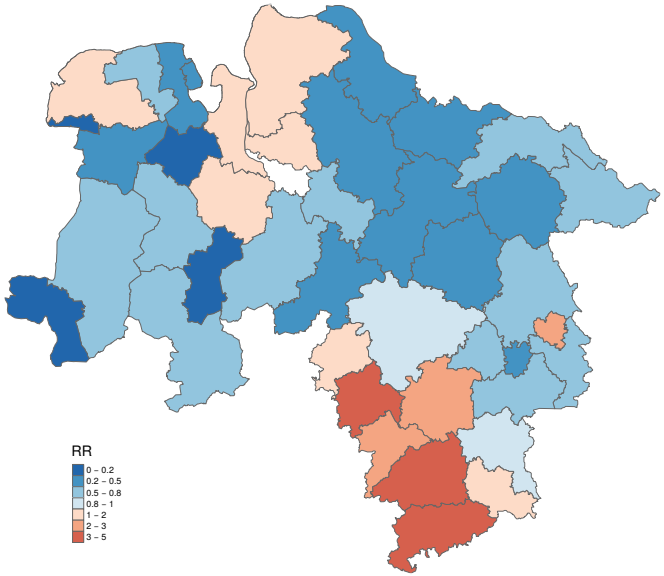


Neighbourhood map. Blue lines between the centres of the polygons indicate that the polygons in the neighbourhood structure are adjacent.

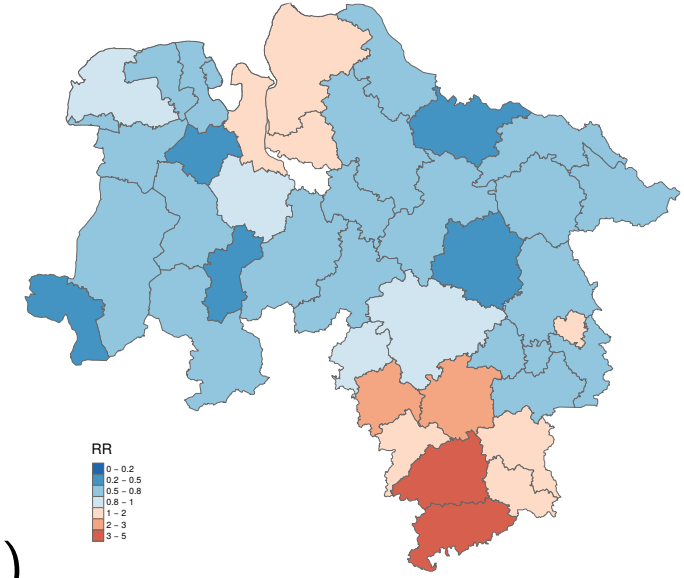
A)



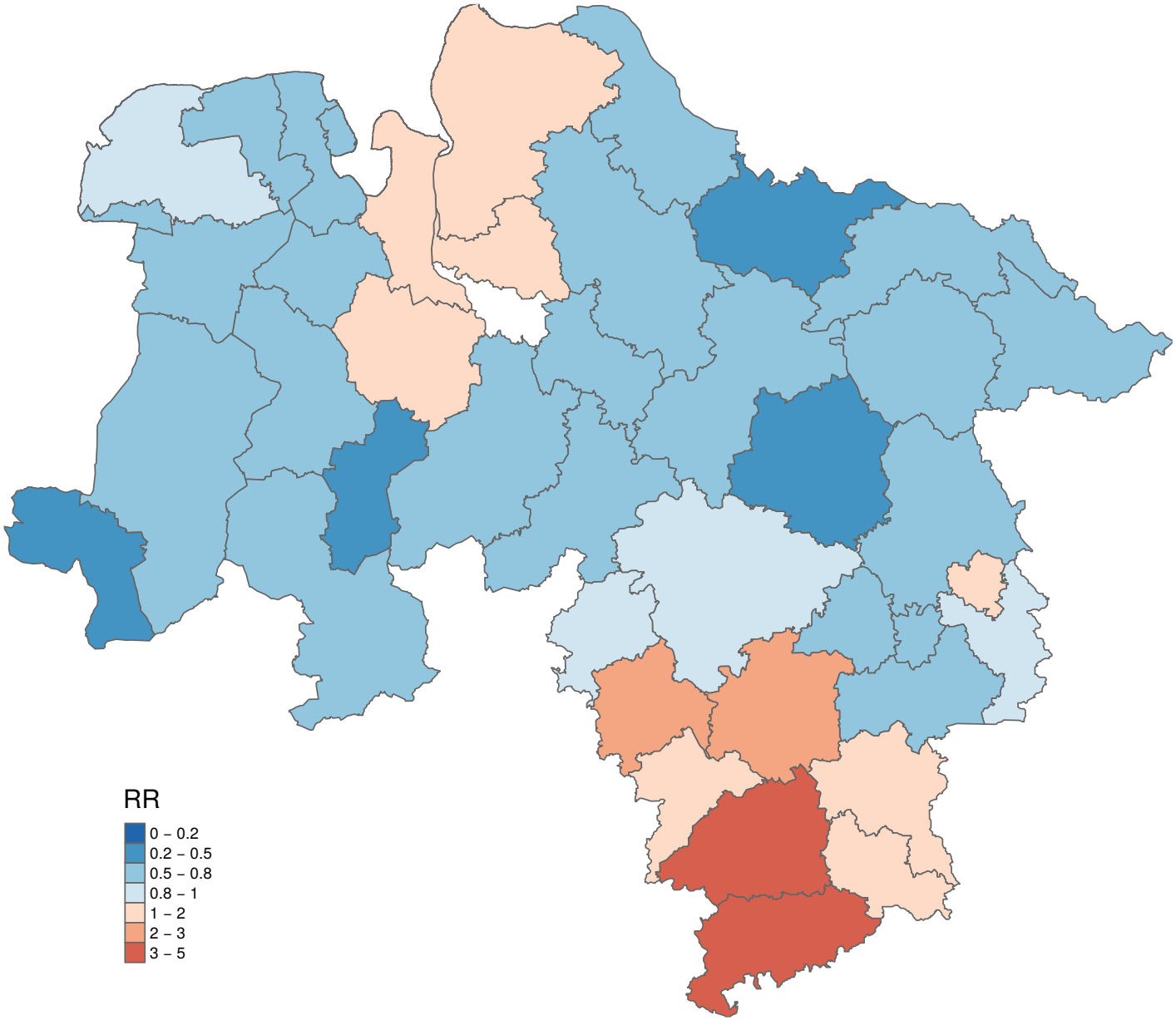
B)



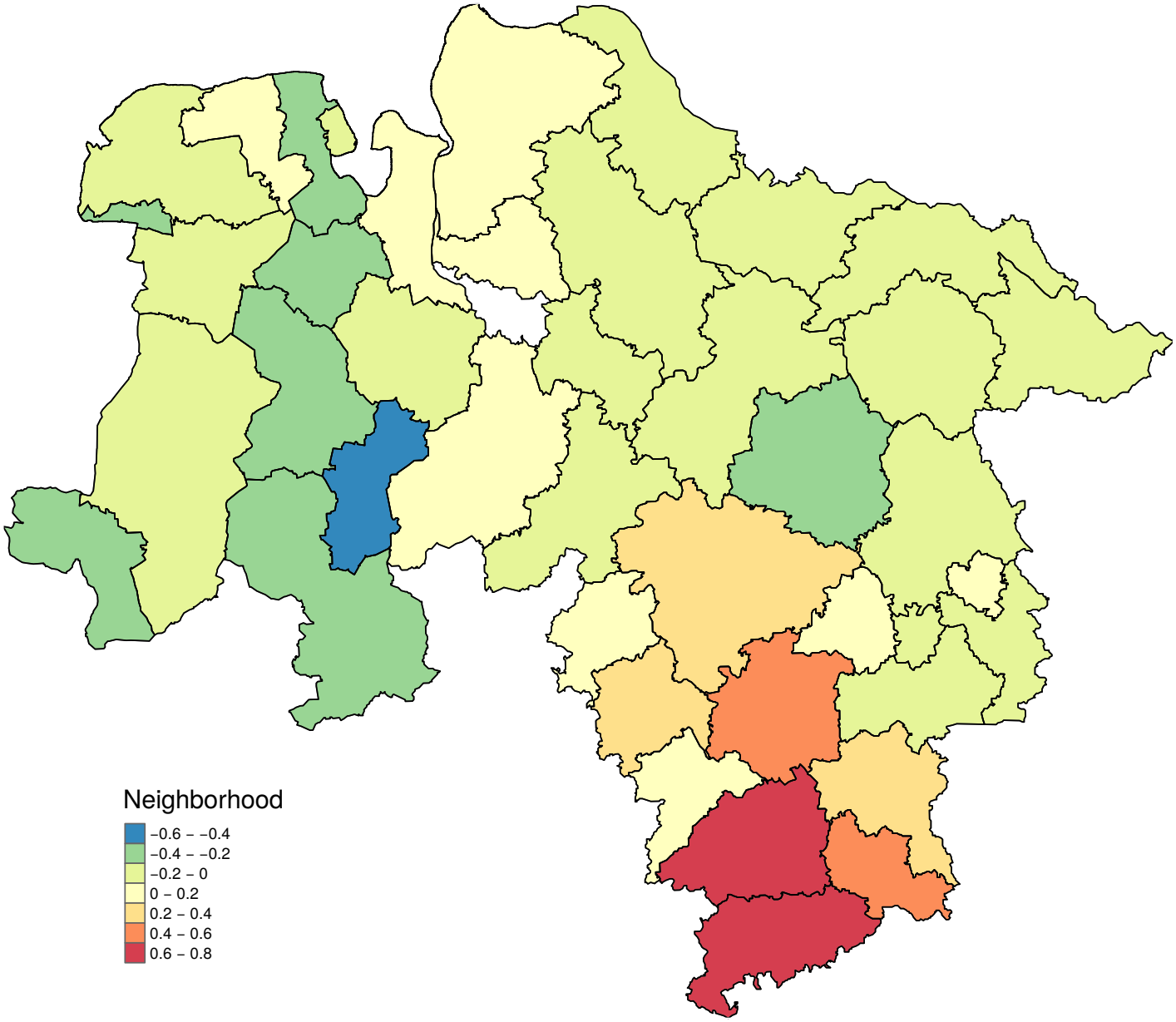
C)



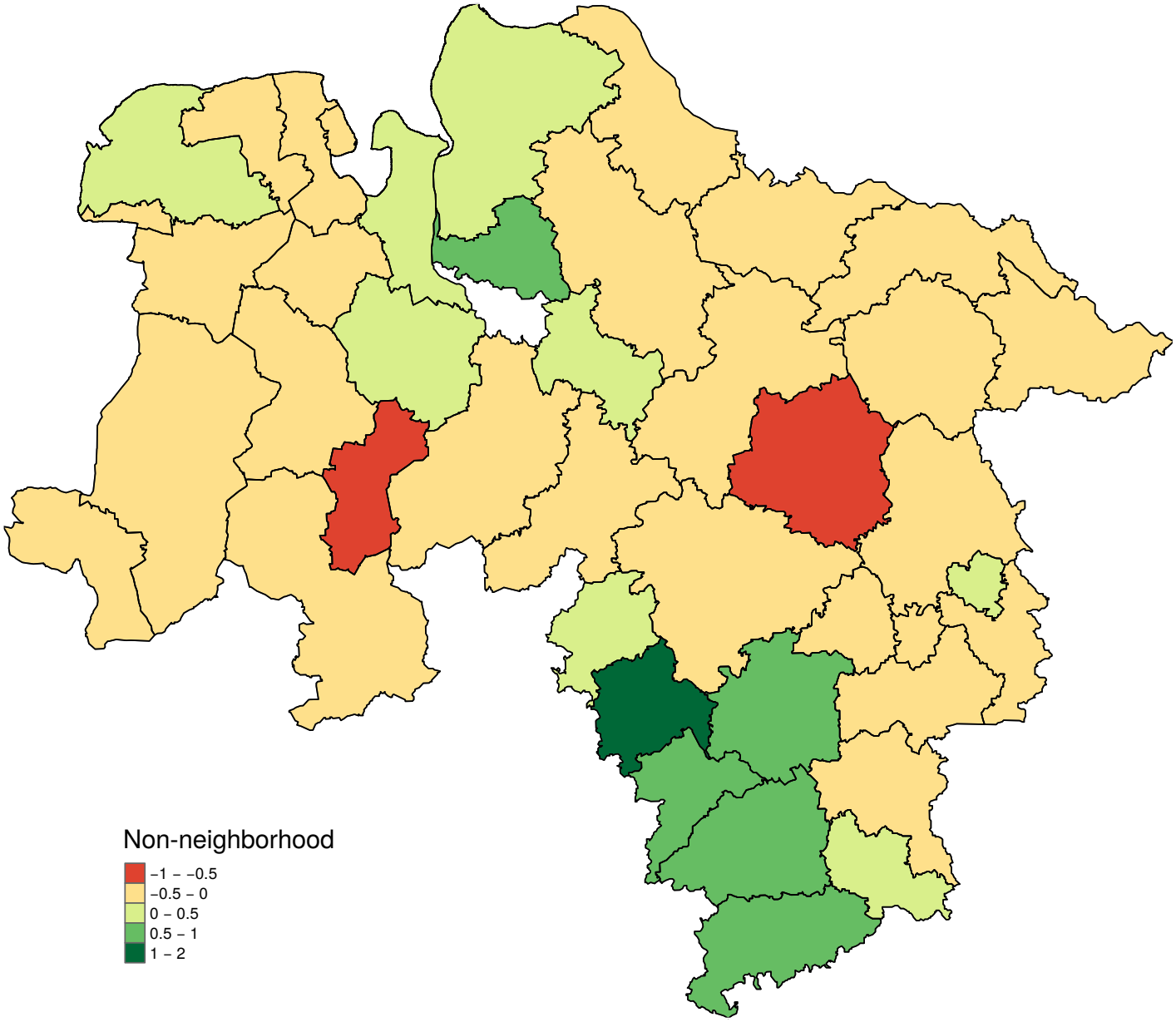
- A) Poisson-gamma model
- B) log-normal model
- C) log-normal model, with conditional autoregresive (CAR) prior



Using more complex models, the risk of district can be broken down into parts attributable to the neighborhood and non-neighbourhood effects.
Such as e.g. the Besag-York-Mollié model (Besag et al., 1991)



Using more complex models, the risk of district can be broken down into parts attributable to the **neighborhood** and non-neighbourhood effects.



Using more complex models, the risk of district can be broken down into parts attributable to the neighborhood and **non-neighbourhood** effects.

From the disease mapping examples, we could see that the map appearance can be strongly influenced by several factors (e.g. point size, colour scale, smoothing procedure, and its parameterization).

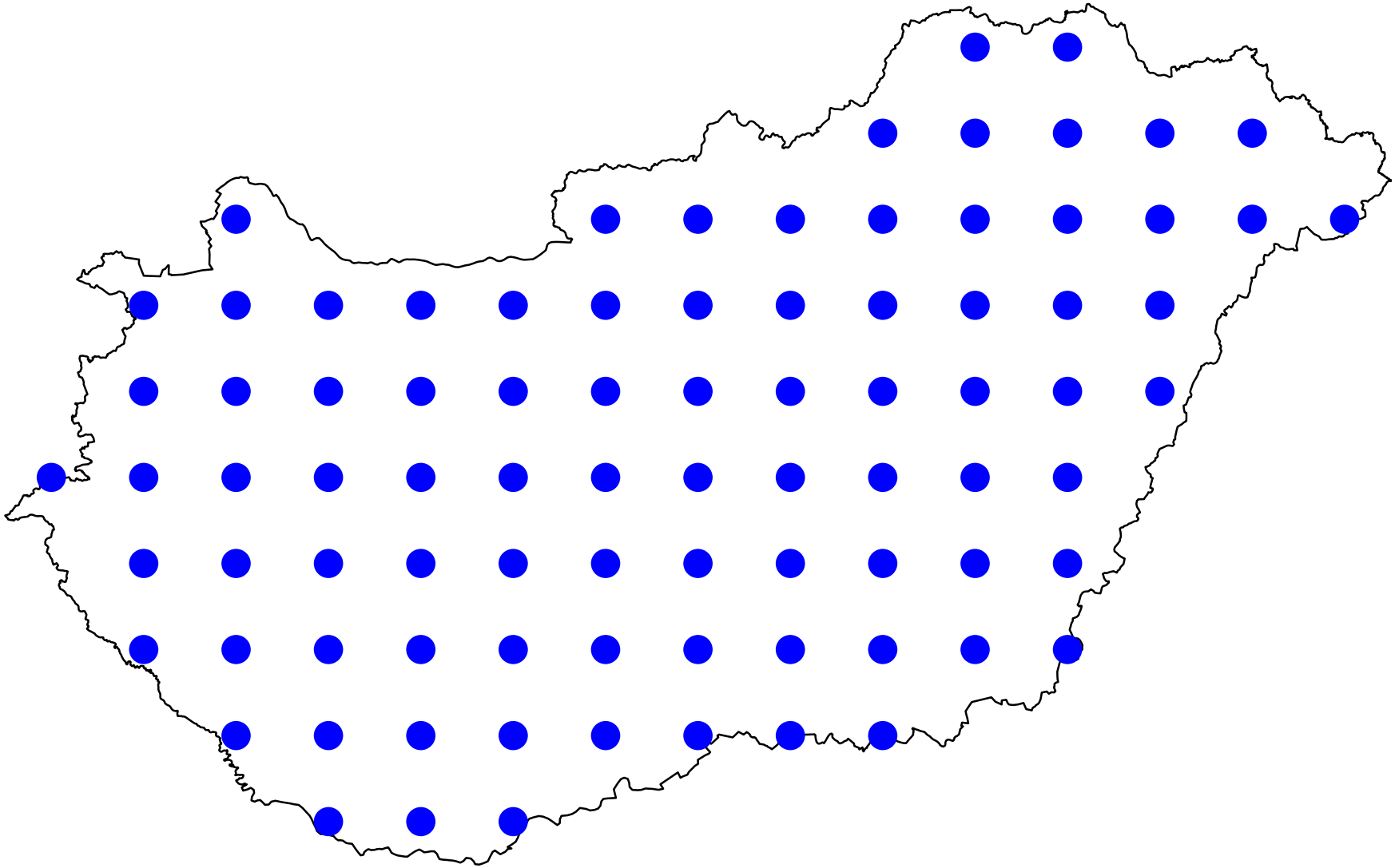
Therefore, a purely visual assessment of the geographical pattern of risk may lead to subjective conclusions.

The same risk map may be interpreted differently by different people.

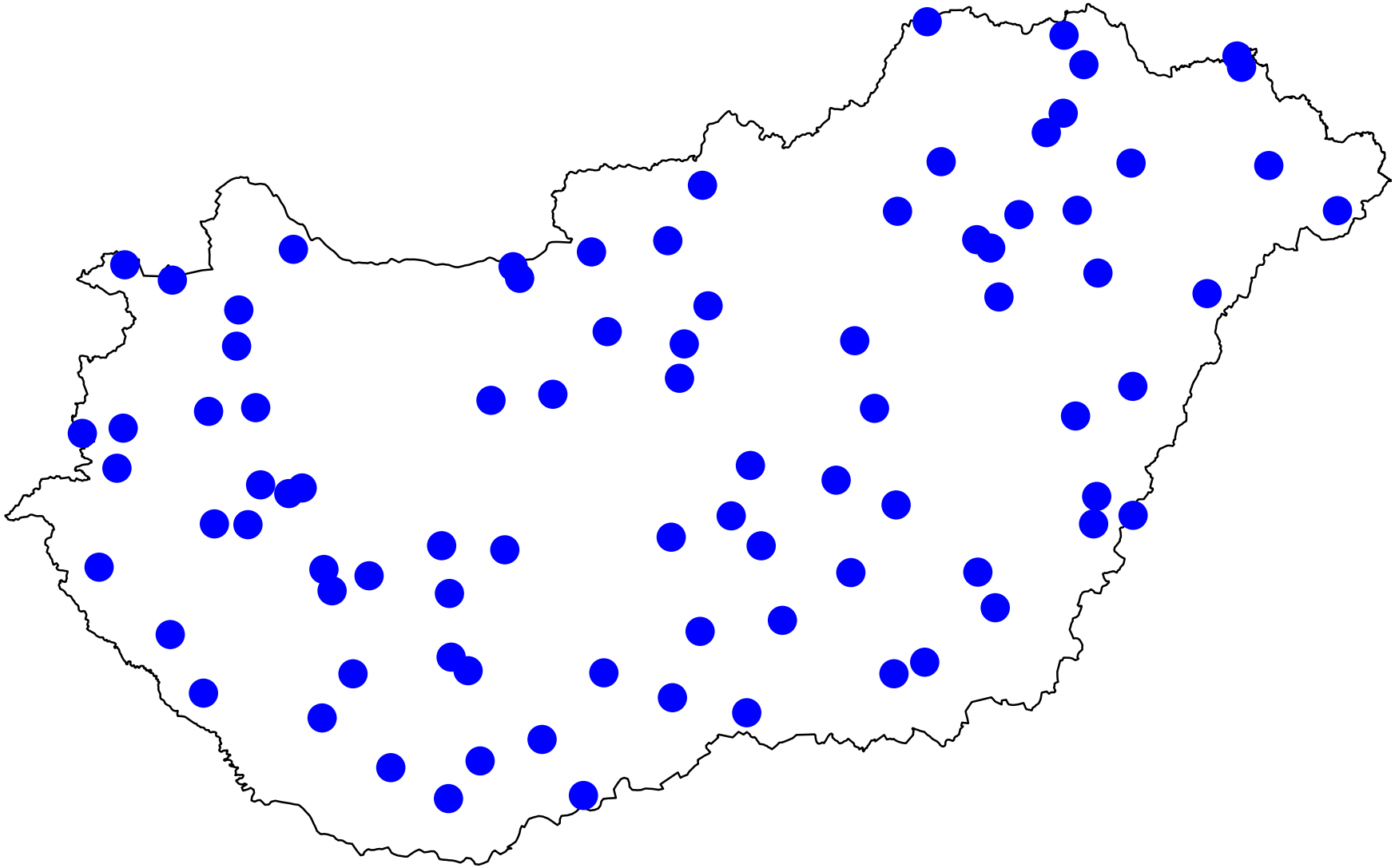
Therefore, several methods have been developed that assess spatial risk more objectively.

The spatial epidemiology area for describing the risk quantitatively in space called spatial pattern analysis.

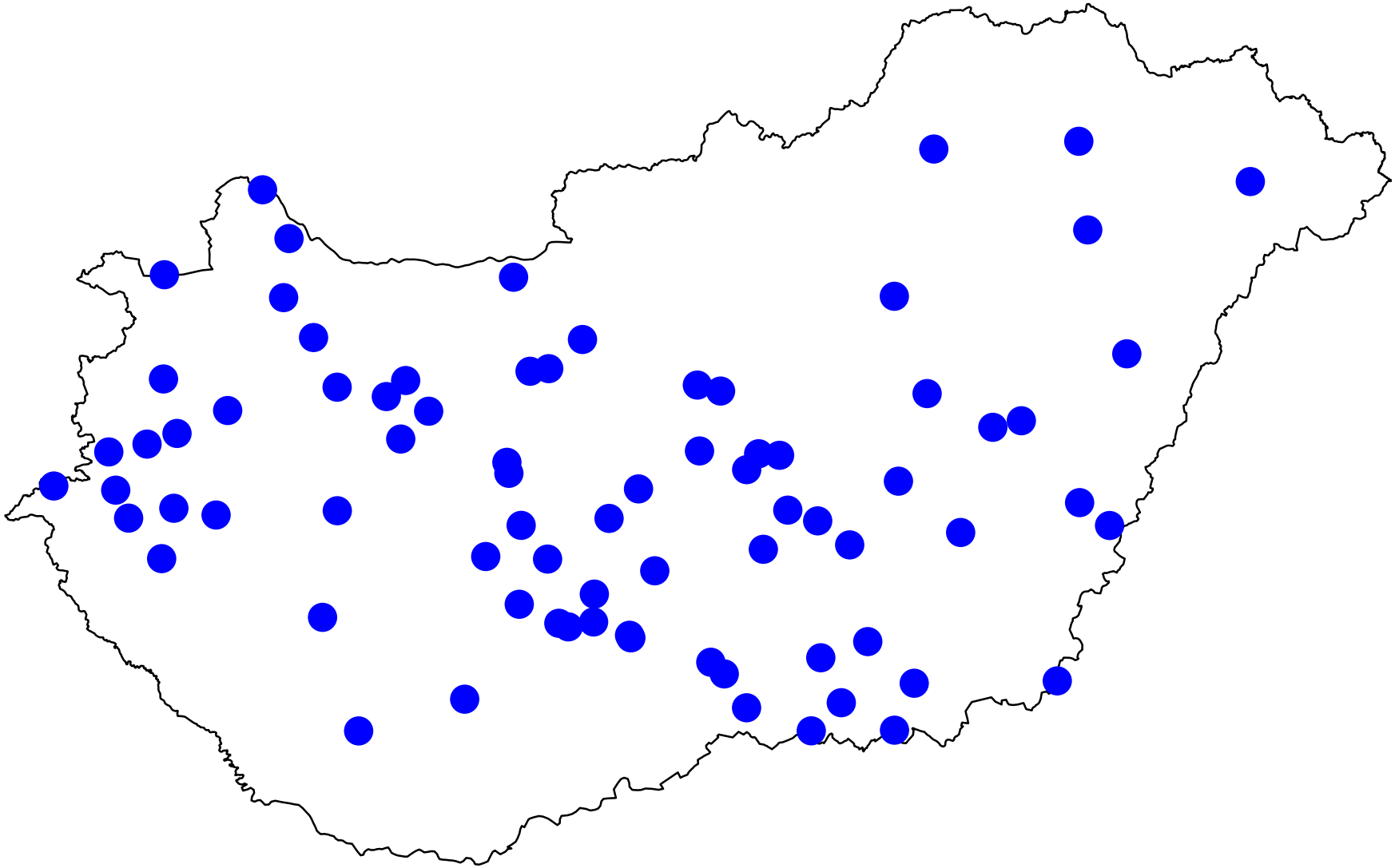
In addition to the most commonly used spatial cluster analyzes in epidemiology, barrier detection is presented below.



In geostatistics, three types are usually distinguished for the distribution of points in two-dimensional space: **regular**, random, and aggregated.



In geostatistics, three types are usually distinguished for the distribution of points in two-dimensional space: regular, **random**, and aggregated.



In geostatistics, three types are usually distinguished for the distribution of points in two-dimensional space: regular, random, and **aggregated**.

The spatial cluster analysis aims to examine the geographical aggregation, clustering of cases and risks. According to the definition of CDC (1990), a cluster is an apparent or real aggregation of health-related events in space and/or time.

Several methods of spatial cluster analysis can be found in the veterinary epidemiological literature (Ward and Carpenter, 2000a,b).

Methods are usually grouped according to whether they are point-based or area-based. They are distinguished also as global, local or focused cluster analysis methods.

With global cluster analysis methods, we can obtain a test statistic value for the spatial aggregation of cases. This value indicates whether the occurred cases (or risk) in the study region were aggregated in space or not.

Local clustering methods make it possible to identify the position of aggregation in space.

Using focused clustering methods, we obtain aggregation information for a predefined geographic location.

The two most frequently used methods of global cluster analysis are the autocorrelation based Moran I and Geary c.

Autocorrelation methods analyze the neighbourhood relation between the observation units and the similarity in the variable of interest (e.g. risk).

Spatial pattern	Geary c	Moran I
Aggregated	$0 \leq c < 1$	$I > 0$
Regular	$1 < c < 3$	$I < 0$
Random	$c = 1$	$I = 0$

What can we say about the spatial clustering of *E. multilocularis* prevalence?

Spatial pattern	Geary c	Moran I
Aggregated	$0 \leq c < 1$	$I > 0$
Regular	$1 < c < 3$	$I < 0$
Random	$c = 1$	$I = 0$

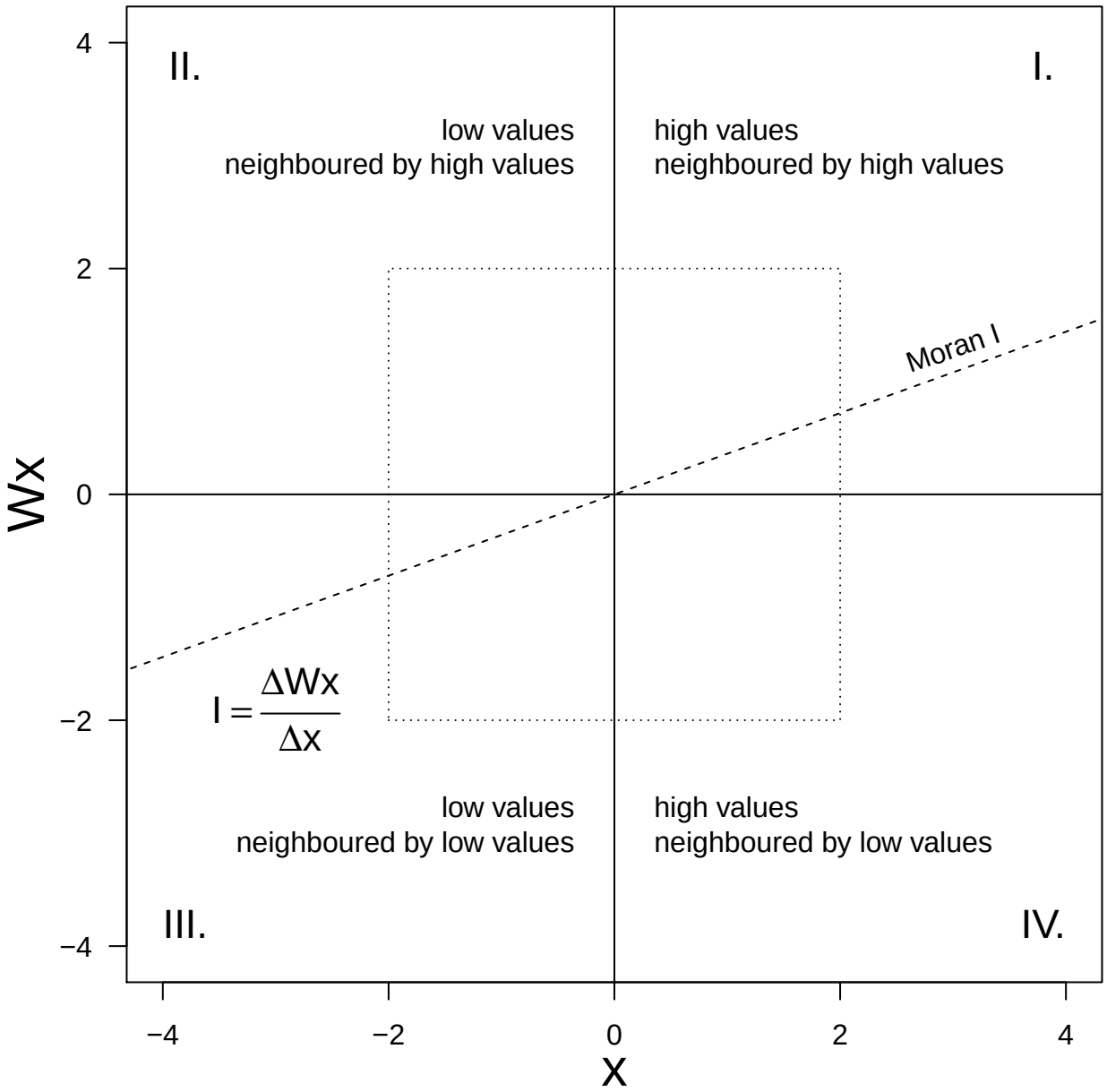
```
Moran I statistic standard deviate = 5.5582, p-value = 1.363e-08
alternative hypothesis: greater
sample estimates:
Moran I statistic       Expectation          Variance
       0.54047145       -0.02439024        0.01032791

Geary C statistic standard deviate = 4.5358, p-value = 2.87e-06
alternative hypothesis: Expectation greater than statistic
sample estimates:
Geary C statistic       Expectation          Variance
       0.43313436        1.00000000        0.01561929
```

While the global Moran I gives a single test statistic value and a p-value for all analysed districts, the local Moran I method determines those for each area unit separately (Anselin, 1992).

```
                Ii         E.Ii    Var.Ii         Z.Ii    Pr(z > 0)
23041 0.254596615 -0.02439024 0.2057601   0.61503973 2.692642e-01
23042 0.086809574 -0.02439024 0.1094258   0.33615855 3.683757e-01
23043 0.003763955 -0.02439024 0.2057601   0.06206726 4.752546e-01
23044 6.562042766 -0.02439024 0.4305401 10.03791459 5.192317e-24
23045 0.021121822 -0.02439024 0.2057601   0.10033350 4.600398e-01
23046 4.111202504 -0.02439024 0.1608041 10.31309944 3.074288e-25
.
.
.
```

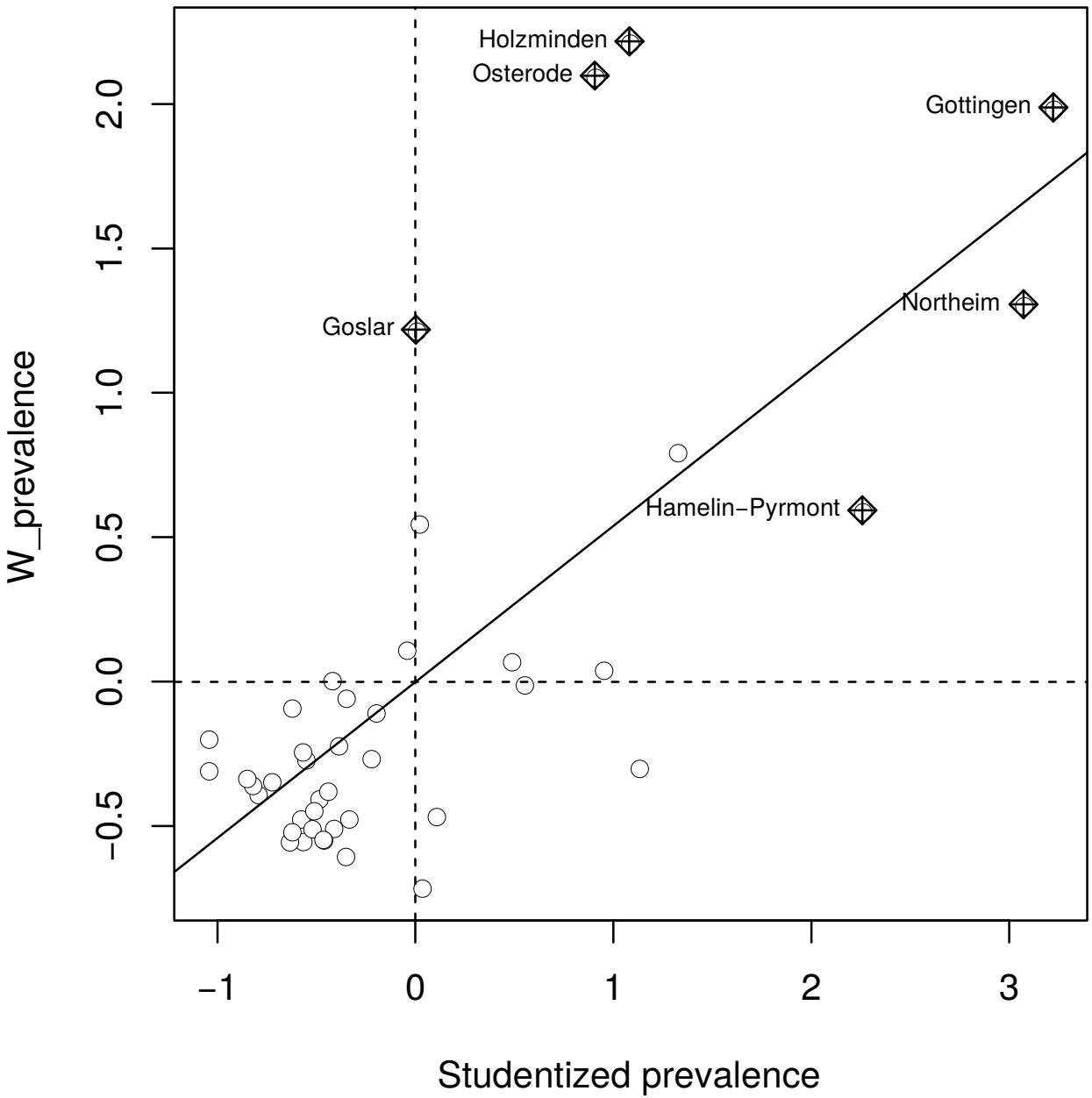
- `Ii` is the local Miran I test statistic for each district
- `Pr(z > 0)` is the p-value



Moran scatterplot.

The x-axis represents the standardized prevalence observed in units and the y-axis the standardized prevalence observed in their neighbours.

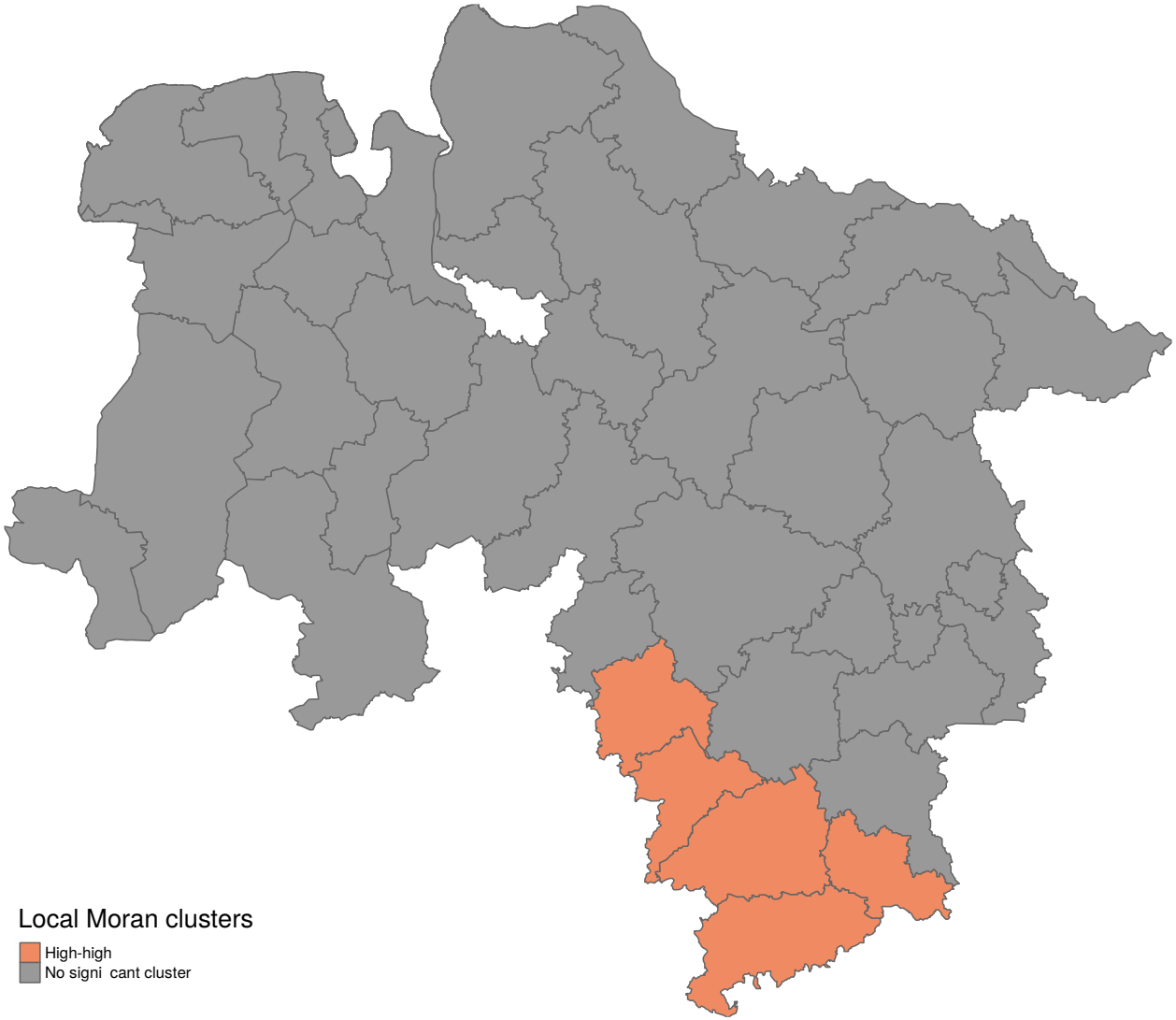
The regression coefficient of the $Wx \sim x$ linear model is the same as the global Moran I.



Moran scatterplot.

The squares in quadrant I indicate the districts in which high prevalence was observed (and in their environment).

Cluster elements with values greater than two on at least one axis are significant ($p < 0.05$).



Göttingen,
Hamelin-Pyrmont,
Holzminden,
Northeim,
Osterode districts are high-prevalence districts neighboured by high-prevalence districts.

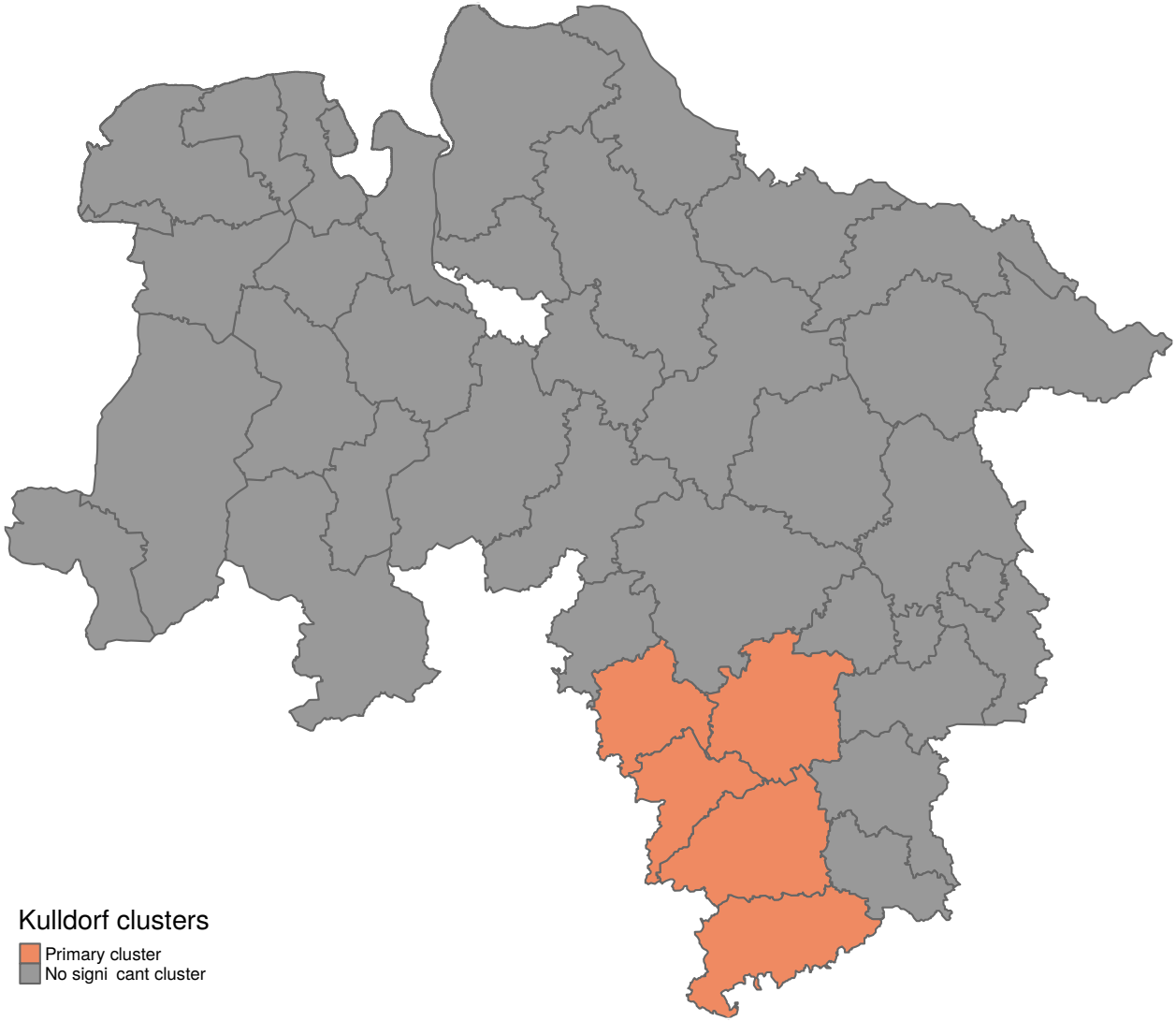
Numerous so-called spatial scan statistics are used to localize geographic and/or temporal clusters. Perhaps the most commonly used scan statistic in spatial epidemiology is the Kulldorff's one (Kulldorff and Nagarwalla, 1995).

We place a grid on the study area, from each grid point as a centre point, we form circles.

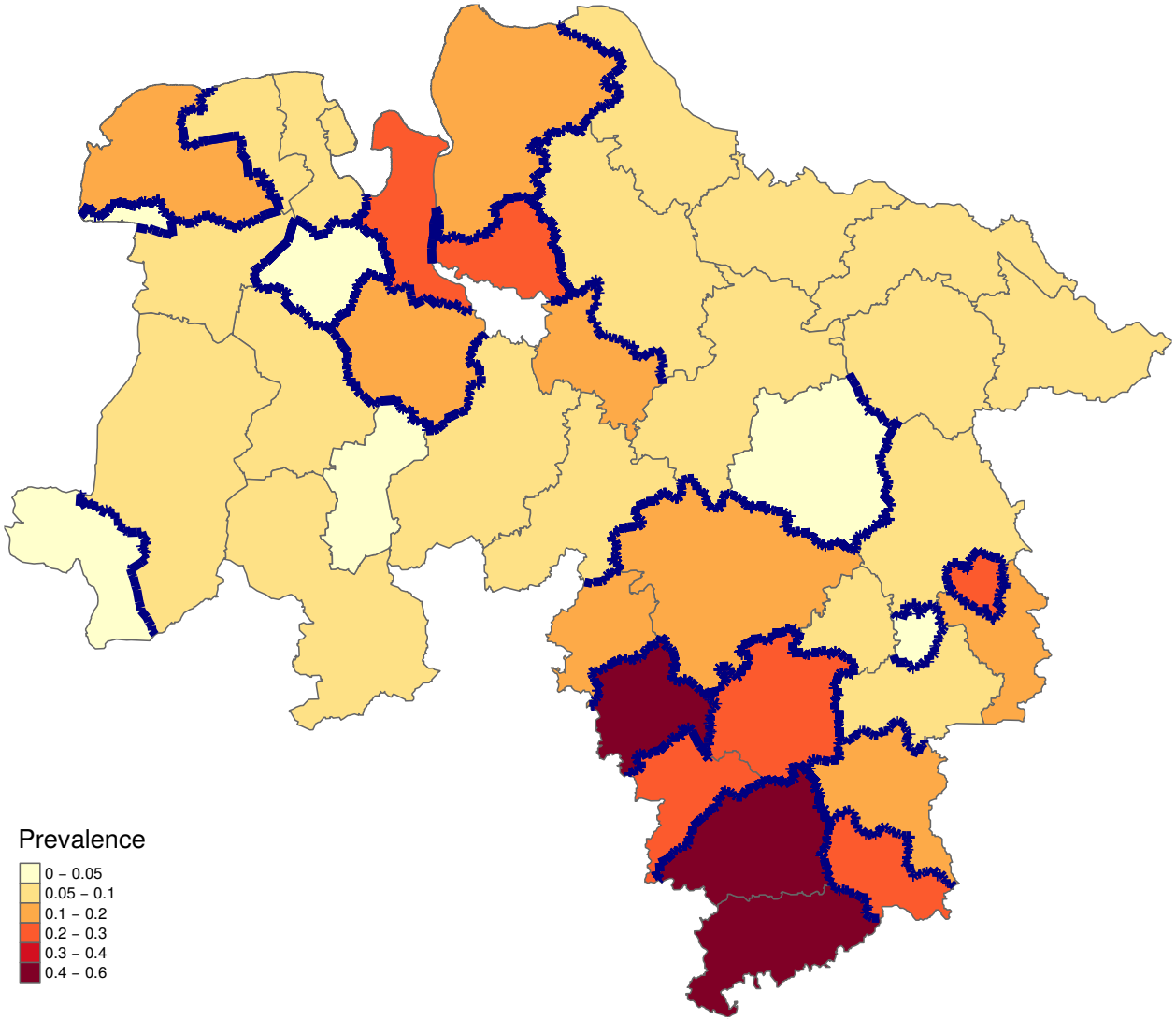
The radius of the circles is continuously increased until the upper limit (what usually is the half of the whole study area).

For each generated circle, we sum the number of cases inside and outside the circle and calculate the number of expected cases inside and outside the circle.

The observed case numbers are compared with the expected case numbers based on theoretical distributions, and the circle in which the number of cases is the least likely compared to the out-of-circle cases is identified as the primary cluster.



By the Kulldorff method, the following districts belong to a high-risk cluster: Göttingen, Hamelin-Pyrmont, Holzminden, Lüneburg, Northeim.



While cluster analyzes seek to answer the question of whether spatial aggregation of risk can be identified, barrier detection approaches draw boundaries based on differences between neighbours (Solymosi et al., 2005).

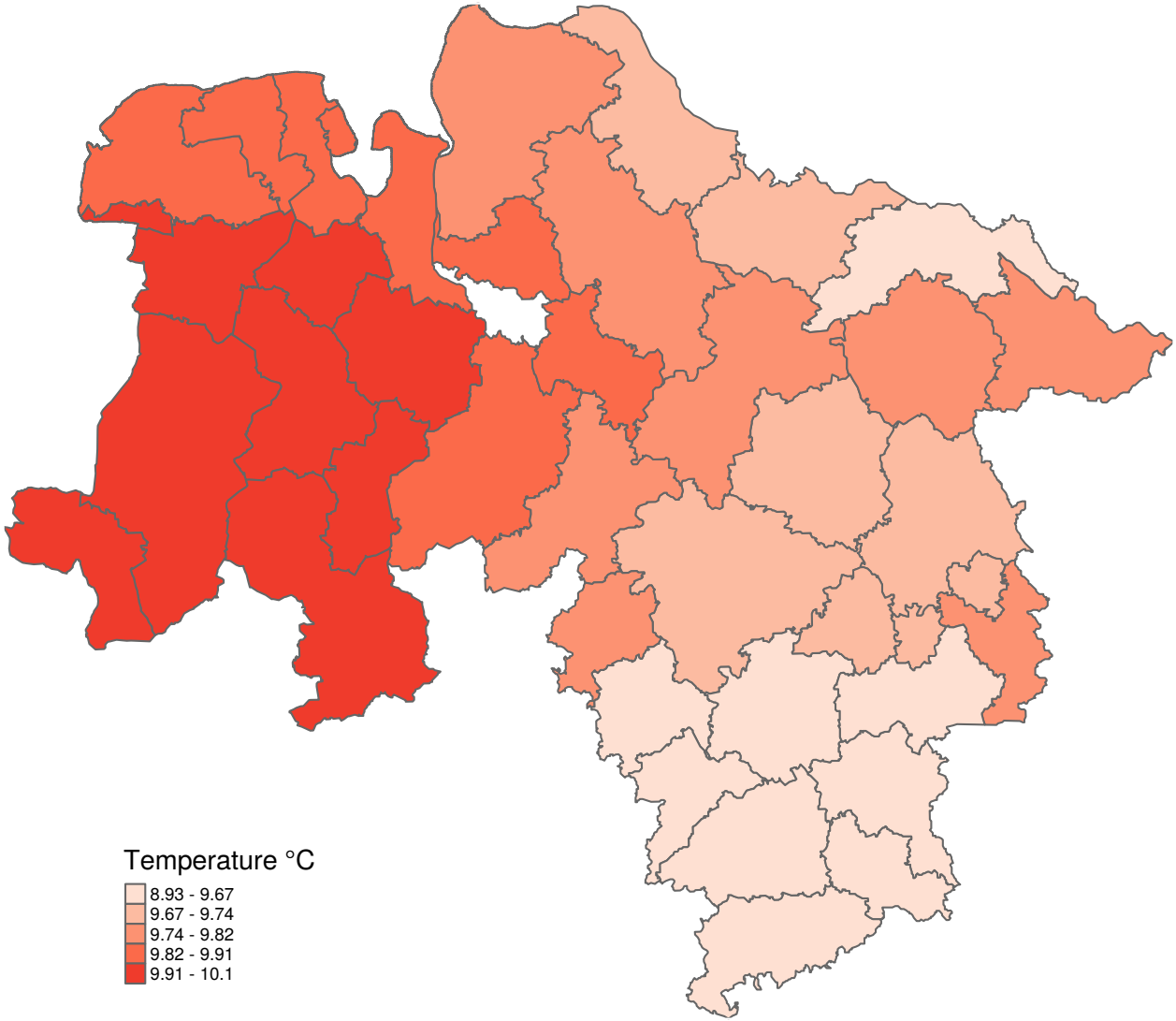
Barriers (blue) created by the method of greatest differences (Monmonier, 1973).

Based on disease mapping and pattern analysis, it seems clear that the risk of infection is much higher in the south-western districts of Lower Saxony than in most of the state. From both the animal and human health viewpoint, it would be important to know where this additional risk may come from.

There is a stage in the development of E. multilocularis when the egg is directly exposed to external environmental factors. Thus, among many factors influencing the risk of infection, external environmental conditions may play a role too.

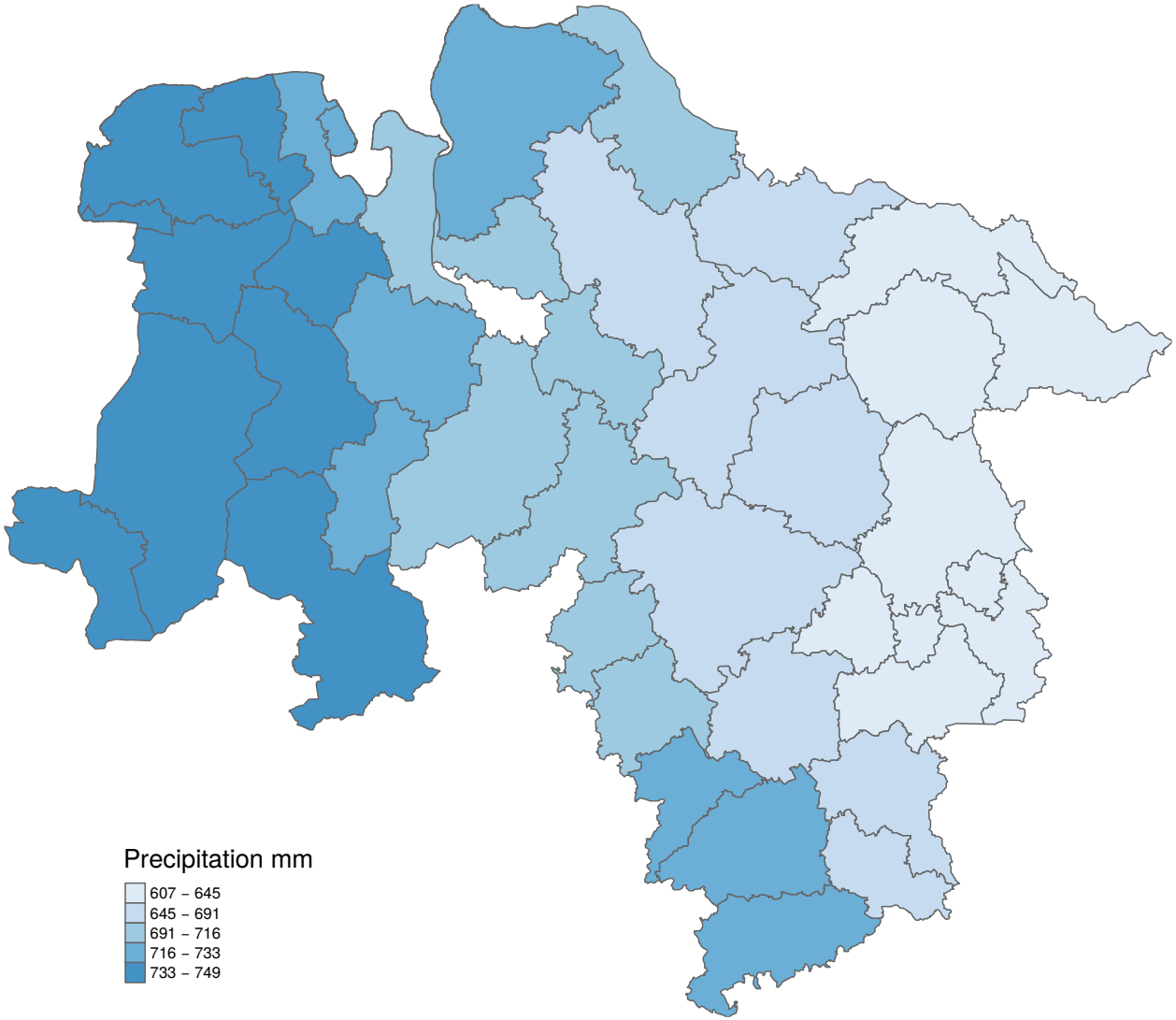
Based on literature data, the development of the parasite can be influenced by the ambient temperature, the amount of precipitation, and the soil moisture.

To analyze the dependence of the prevalence per district in Lower Saxony on these environmental variables, they must be obtained.



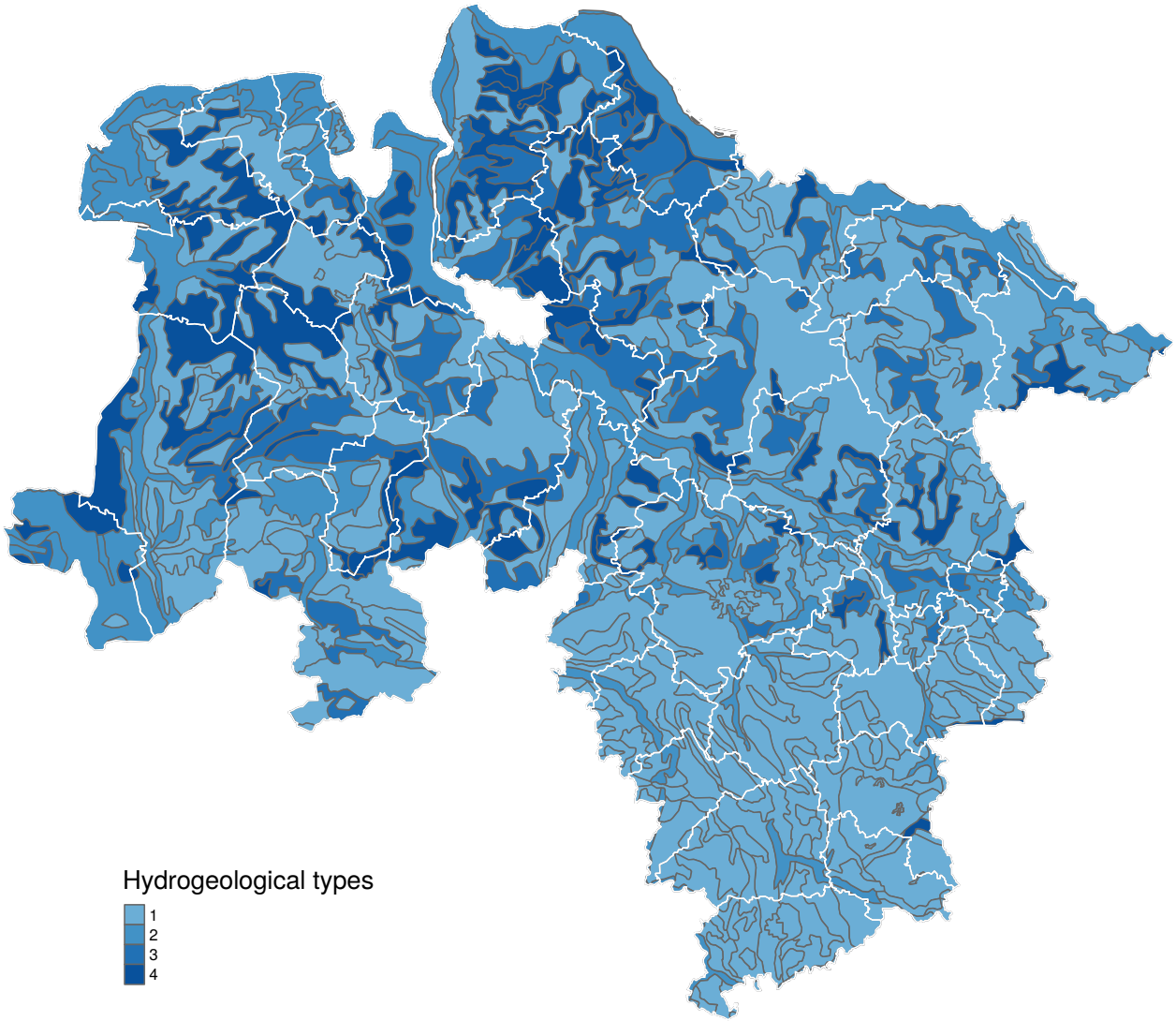
The mean of annual average temperatures over the period 1991-1997 based on the ERA-Interim reanalysis database.

Most of the districts with the lowest mean temperature are identical to the districts with the highest prevalence of E. multilocularis.



The mean of annual precipitation sums over the period 1991-1997 based on the ERA-Interim reanalysis database.

According to the map, the districts with the highest prevalence are not the districts that receive the highest amount of precipitation. But if we also compare it with the temperature map, we can see that in the districts that receive the most precipitation, the average temperature is also in the highest range.



The spatial pattern of hydrogeological types from the ESDAC database in Lower Saxony:

1. permeable subsoil, groundwater is deep: rarely moist
2. lowland soils affected by groundwater, seasonally or continuously wet or artificially drained
3. soil with a watertight layer within 80 cm is seasonally or continuously moist
4. upland or mountain soils

So far, we have formulated possible relationships between the risk of infection and some environmental factors on a subjective, visual basis.

However, this is only an introduction to the mathematical-statistical description of associations. In this approach, the dependence of a dependent variable (e.g., prevalence) on independent variables (e.g., temperature, precipitation, soil moisture) is formulated using some mathematical model.

On the one hand, these models allow us to quantify the size of the effect of independent variables on the dependent variable. On the other hand, the numerical values of these effects allow us to give predictions.

What does prediction mean in spatial epidemiology? On the one hand, it means that our model predicts what values of dependent variables are estimated for the places (e.g., districts) from which we have information about the dependent variable (e.g., prevalence).

On the other hand, since we believe that there are rules in the physical world and these rules can be formulated mathematically, we also believe that these rules are not only valid in one region but can be extended to other areas.

Our goal is to describe a model from which the prediction shows the smallest deviation from the observed values.

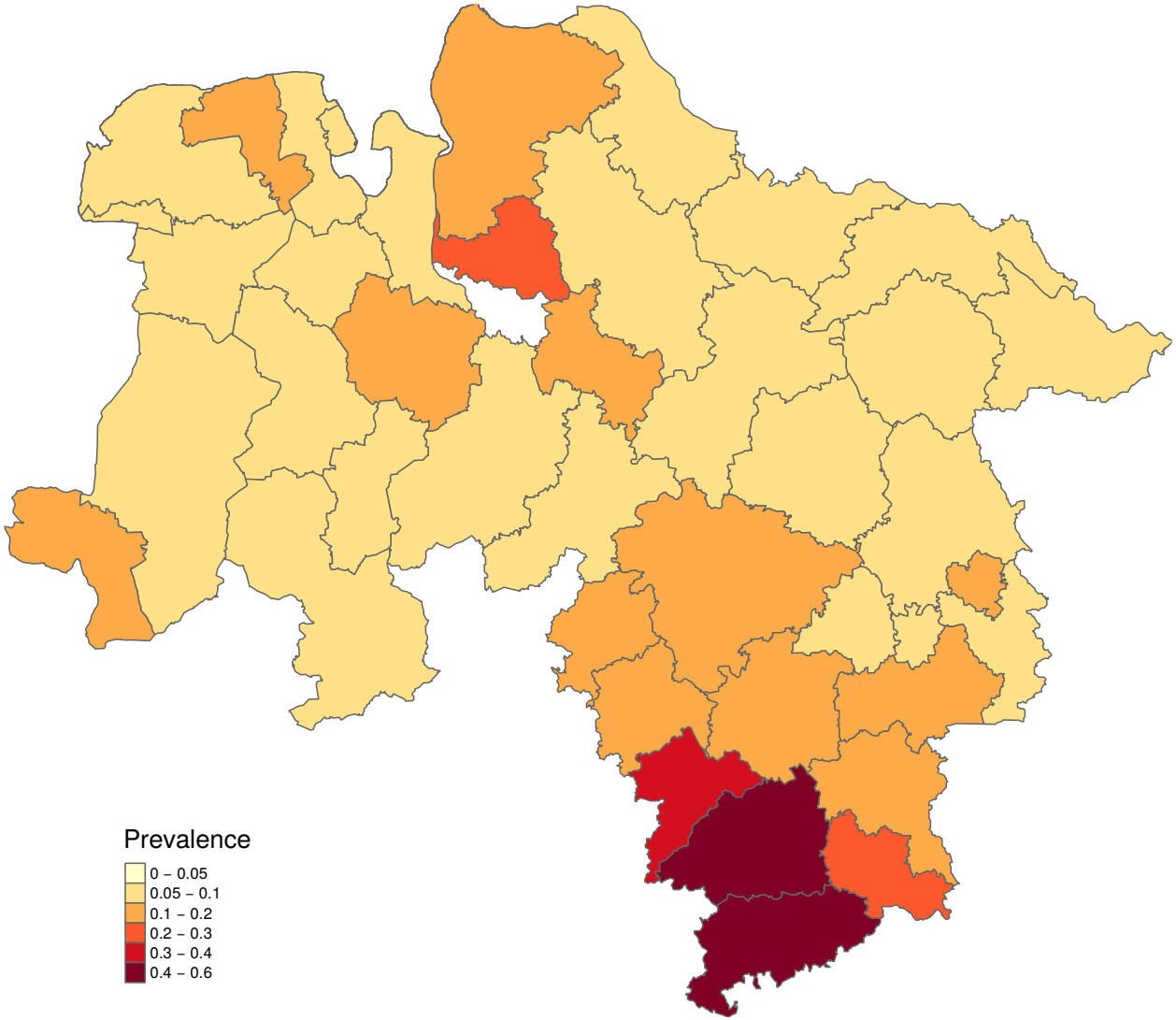
To do this, we create models from our three independent variables (temperature, precipitation, soil moisture) for all possible combinations and fit them.

Based on each fitted model, we predict the expected prevalence in each district.

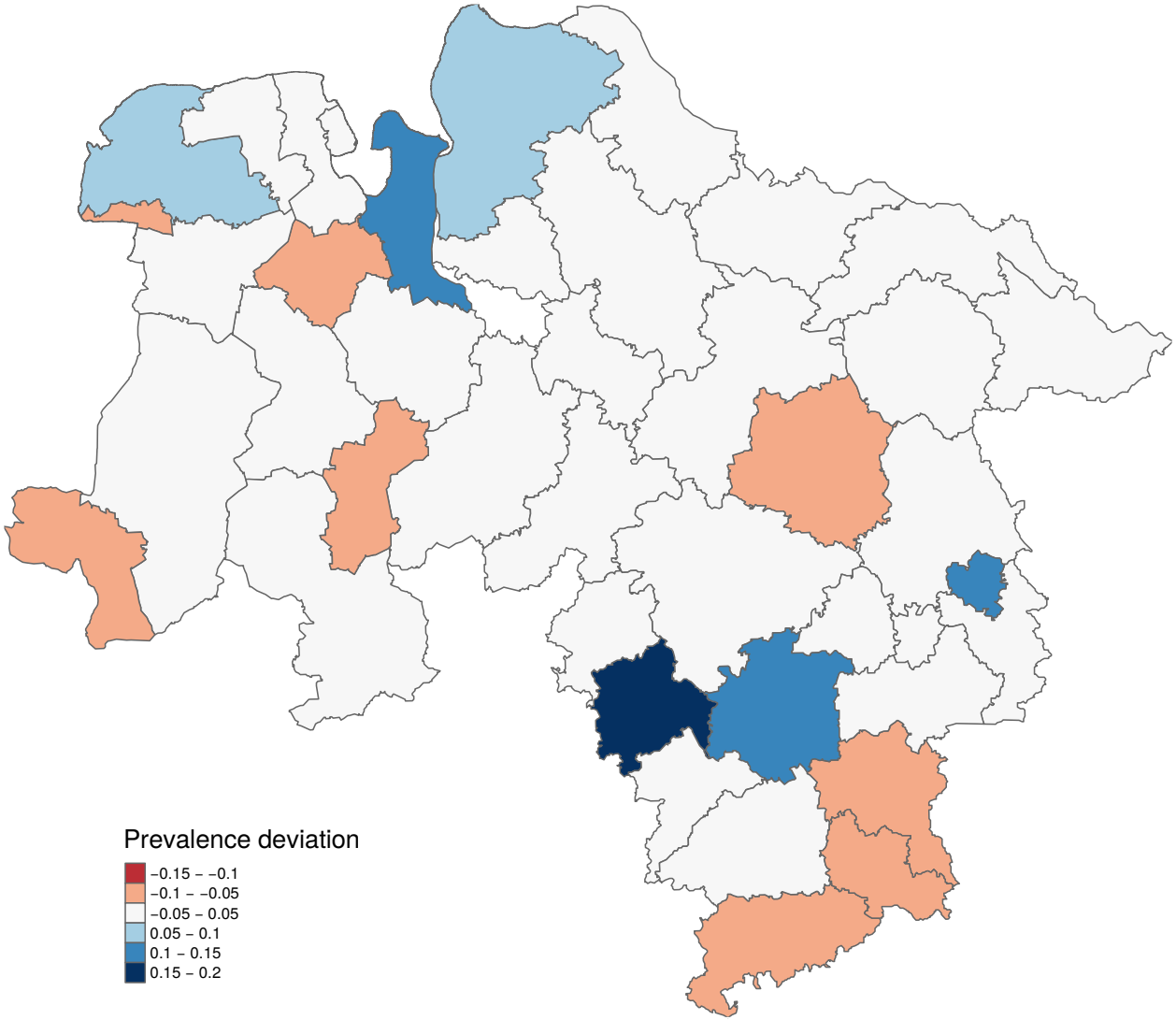
Using these predicted prevalences, we calculate the difference between the observed and predicted values for each model. For this, we use the mean absolute error (MAE):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \tilde{y}_i|$$

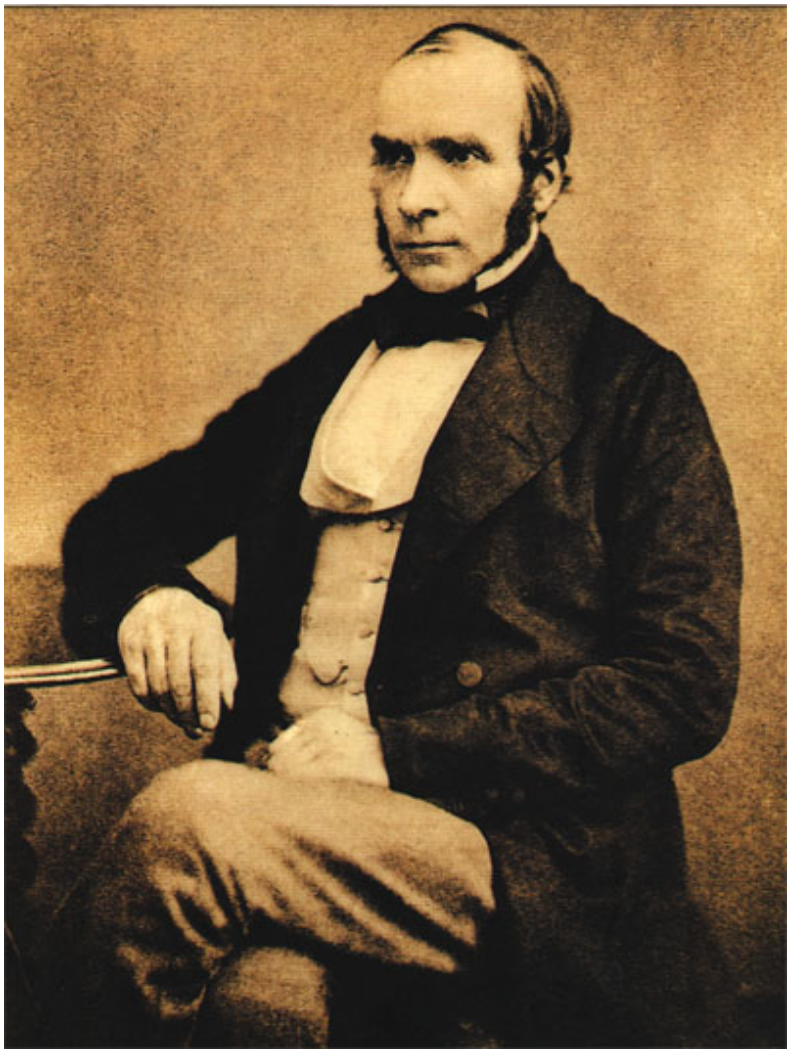
The model that resulted in the predicted prevalence with the lowest MAE value is considered as the best.



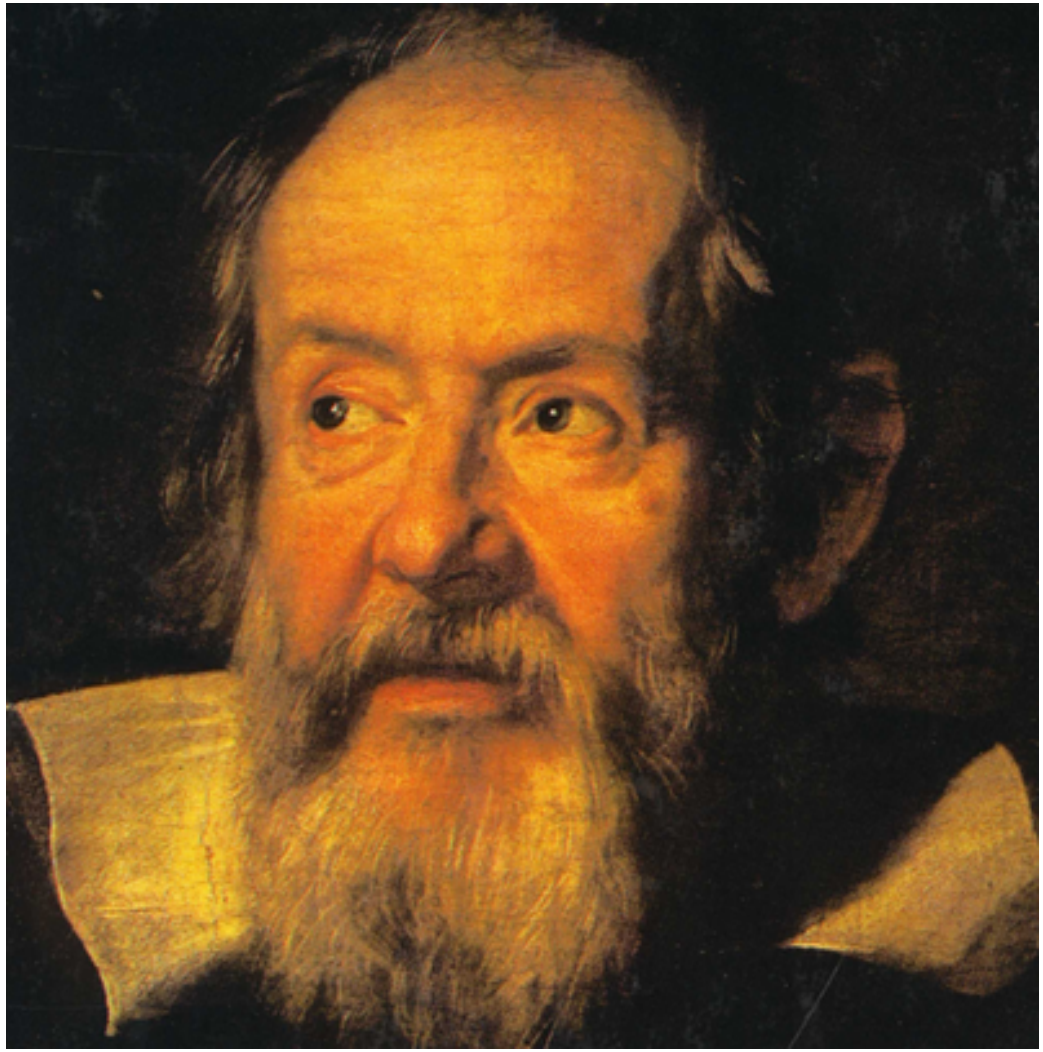
Prevalence per district predicted based on the model with the lowest MAE value.



Difference between observed and predicted prevalence. The overpredicted districts are coloured with the red scale, and the underpredicted ones are coloured with the blue scale.



John Snow (1813–1858)



Galileo Galilei (1564–1642)

A 21st century clinician who cannot critically read a study is as unprepared as one who cannot take a blood pressure or examine the cardiovascular system (Glasziou et al., 2008).

Anselin, L. (1992). Local indicators of spatial association – LISA. *Geogr. Anal. 27*(2), 93–115.

Barratt, A., P. C. Wyer, R. Hatala, T. McGinn, A. L. Dans, S. Keitz, V. Moyer, et al. (2004). Tips for learners of evidence-based medicine: 1. relative risk reduction, absolute risk reduction and number needed to treat. *Canadian Medical Association Journal 171*(4), 353–358.

Berke, O. (2001). Choropleth mapping of regional count data of Echinococcus multilocularis among red foxes in Lower Saxony, Germany. *Preventive veterinary medicine 52*(3), 119–131.

Besag, J., J. York, and A. Mollié (1991). Bayesian image restoration with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics 43*(1), 1–20.

CDC (1990). Guidelines for investigating clusters of health events. Technical Report RR-11.

Dubecz, A., N. Solymosi, M. Schweigert, R. J. Stadlhuber, J. H. Peters, D. Oefner, and H. J. Stein (2013). Time-Trends and Disparities in Lymphadenectomy for Gastrointestinal Cancer in the United States: A Population-Based Analysis of 342,792 Patients. *Journal of Gastrointestinal Surgery 17*(4), 611–619.

Glasziou, P., A. Burls, and R. Gilbert (2008). Evidence based medicine and the medical curriculum. *BMJ (Clinical Research Ed.) 337*(7672), 704–705.

Kulldorff, M. and N. Nagarwalla (1995). Spatial disease clusters: Detection and inference. *Statistics in Medicine 14*, 799–810.

Lawson, A. B. (2013a). *Bayesian disease mapping: hierarchical modeling in spatial epidemiology*. CRC press.

Lawson, A. B. (2013b). *Statistical methods in spatial epidemiology*. John Wiley & Sons.

Lawson, A. B., S. Banerjee, R. P. Haining, and M. D. Ugarte (2016). *Handbook of spatial epidemiology*. CRC Press.

Monmonier, M. S. (1973). Maximum-difference barriers: An alternative numerical regionalization method. *Geogr. Anal. 3*, 245–261.

Noordhuizen, J. P. T. M., K. Frankena, M. Thrusfield, and E. A. M. Graat (2001). *Application of Quantitative Methods in Veterinary Epidemiology*. Wageningen, The Netherland: Wageningen Pers.

Solymosi, N., J. Reiczigel, A. Harnos, J. Mészáros, L. D. Molnár, and F. Rubel (2005). Finding spatial barriers by Monmonier's algorithm. In *International Society for Clinical Biostatistics Conference*, Szeged, Hungary.

Stevenson, M. (2012). An introduction to veterinary epidemiology. EpiCentre, IVABS, Massey University, Palmerston North, New Zealand.

Stevenson, M. (2019). *epiR: Tools for the Analysis of Epidemiological Data*. R package version 1.0-4.

Thrusfield, M., R. Christley, H. Brown, P. J. Diggle, N. French, K. Howe, L. Kelly, A. O'Connor, J. Sargeant, and H. Wood (2018). *Veterinary Epidemiology* (4th ed.). Oxford, UK: Wiley.

Ward, M. P. and T. E. Carpenter (2000a). Analysis of time-space clustering in veterinary epidemiology. *Prev. Vet. Med. 43*(4), 225–237.

Ward, M. P. and T. E. Carpenter (2000b). Techniques for analysis of disease clustering in space and in time in veterinary epidemiology. *Prev. Vet. Med. 45*, 257–284.