

Car/Griesebner/Strobl (Eds.)

# **Geospatial Crossroads @ GI\_Forum '09**

Proceedings of the  
Geoinformatics Forum Salzburg



Wichmann

# Statistical Confidence on Maps by Bootstrapping

Adam WALANUS and Dorota NALEPKA

## 1 Statistical Confidence (Significance) of Isolines

Isolines (isarithms) are probably the most frequent form of space data visualization and are lines on maps which represent points of equal (exact) value. However, since measurements will always have limited precision, it follows that the isolines which are based on a number of points on the map, are also not absolutely precise.

So, the assessment of its precision is necessary, and here we propose to use confidence intervals as the form of uncertainty presentation. The simple method of bootstrapping is proposed for creating confidence intervals for isolines. The method may be deemed more primitive, because it is possible to apply it without any special tool, in contrast to more classical methods of error estimation (JAGER & KING 2004; HEUVELINK & BURROUGH 2002). It is worth mentioning that the probabilistic approach is similar only to that applied within the theory of possibility (BORDOGNA & CHIESA 2003).

Presenting the statistical confidence interval instead of a single (estimated) value is the good way of the reliability description. The confidence interval should be based on a standard 95 % confidence level, which means that in one random case per 20, the interval will not comprise the true value. If only one “exact” value is presented, one could *almost* be sure that it is not the true value. It is also *almost* certain that the true value is within the confidence interval. However, the first “*almost*” means, in fact, something like 99.99 % (where the number of 9s’ is strictly related to the number of digits in the given “exact” number) whilst the second “*almost*” means precisely 95 %. The aforementioned single estimated value would be the value presented on the map, for the given position (x, y).

However, while it is easy to plot point values on the map, especially using isolines, it would be difficult to plot intervals, instead. One possible solution is to present only one isoline on the map.

## 2 Bootstrapping

Bootstrapping (CHERNICK 1999; EFRON & TIBSHIRANI 1993; MANLY 1997) is a metaphor for a self-sustaining process that proceeds without external help. It is a computer-intensive, general purpose approach to statistical inference. It is the practice of estimating properties of an estimator, which is here, rather unusually, applied to the isolines. The method is based on the resampling from the, so called, empirical distribution of the observed data. Here, a thousand resamples have been used.

The simplest application of bootstrapping to the map (MARIGNANI et al. 2007) is to remove randomly some data (points), and apply the standard procedure of map creation to the new, slightly reduced set of data. The random selection of data is essential for bootstrapping, and

is not problematic as even common spreadsheets have random number generators. The typical resampling process with a replacement procedure is not recommended in the case of maps, because of an unavoidable doubling of some sites, which is unnatural, while to have less data is completely normal.

The idea behind bootstrapping is connected with the fact, that data points are frequently collected in a random way (in some sense, at least).

### 3 Exemplary Data

The map presented in Figure 1, an exemplary one, is called the isopollen. In fact, it illustrates the amount of grass pollen flying over an area (Poland) 10,000 years ago is. It is very difficult to estimate the precision of the palaeo-pollen data.

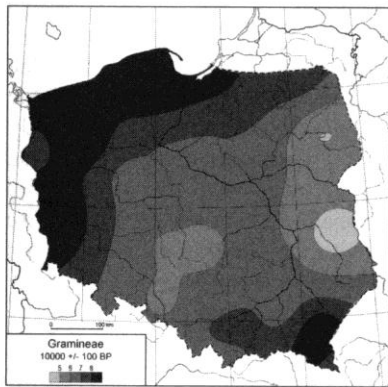


**Fig. 1:**

Isopollen map of grass cover of Poland 10,000 years ago (HARMATA et al. 2004). The number of data points is 97; however, the sites are neither strictly randomly distributed, nor equally, throughout Poland. The algorithm, used to plot isolines, effectively smooths the raw data.

The method of choice of sites for the project (about 100) is essential for a bootstrap idea. By no means, are sites related to the grass cover in the beginning of the Holocene. The site places possessing the most advantageous conditions for coring, should be lakes or peat-bogs. Therefore, the geographical position of the original sites, as well as their number, is virtually random, in some sense. We would be confronted with a map comprising a total number of only 75 % of the sites. The 125 % of data is equally well imaginable (in the future).

The grass cover image, which is estimated from the grass pollen stored in the sediment archive, should by no means depend on the random site selection, as it in fact does. Thus, the random selection of data by bootstrap is deeply rational.

The map constructed (by the method treated as standard for the given problem) with the random 25 % of data removed, differs from the original map (comprising 100 % of data). The inspection of a few maps, which are produced for independent random selections, gives an idea of the original map precision (reliability). When evaluating a certain feature of an isoline (certain shape of it) appreciated as important, and noting, at the same time, that this feature is absent from one of bootstrapping maps, must create some doubts in the observers mind (Fig. 2).

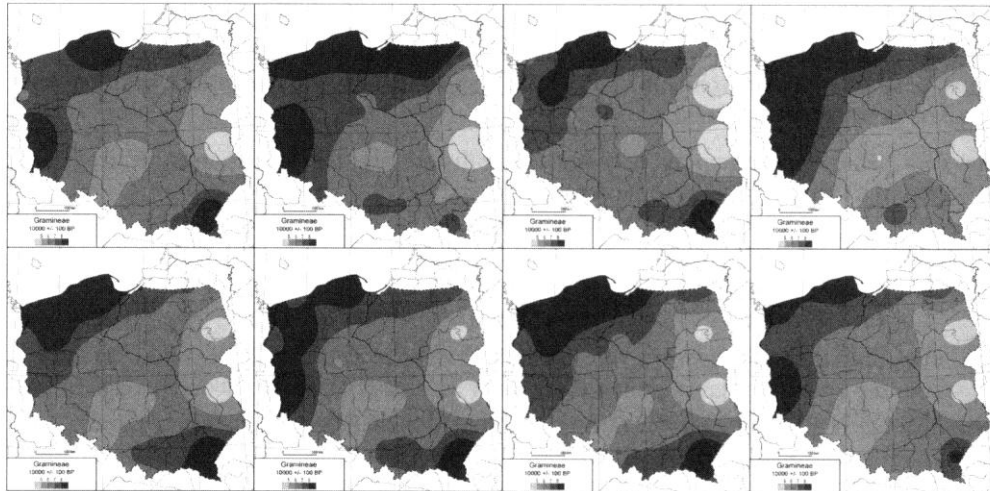


**Fig. 2:** Eight bootstrap maps, generated by random removal of approximately  $\frac{1}{4}$  of the original sites. Inspecting such versions of maps gives an idea of how reliable some features of the isolines shape are. For example, the dark area ( $>8$ ) evidently present on the main map in the SE corner (Fig. 1), disappears completely in the case of maps nr. 2 and 4, i.e. for  $2/8 = 25\%$  of maps. The quantity (25 %) represented a very rough estimation of the probability that the given feature ( $>8$  at SE) may be an artefact only, connected with the position of sites but not with the real grass cover.

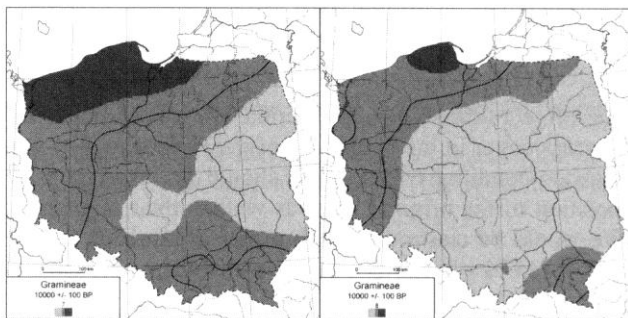


**Fig. 3:** Confidence intervals for isolines 7 and 8 (left and right map, correspondingly). The confidence level is 95 %. Isolines are presented as well. The three grey levels have another meaning from that of typical isoline maps (like those in Fig. 1 and 2). The darkest colour is for an area, where, with a statistical confidence of 97.5 %, the value of the Graminae percentage is  $>7$  (or  $>8$  for the right map). Similarly, for the lightest area, with the opposite sign:  $<7$ . The medium area of the medium darkness gives the confidence band, where the true isoline lies with the probability of 95 %. It is clear, that only the northern region had more than 7 of grasses (and even 8) for sure, where sure = 95 % (7, and 8 refer to the percent of grass pollen).

## 4 Creation of the Confidence Band for an Isoline by Bootstrap Resampling

As was already mentioned, the confidence band for an isoline may be plotted on the map, if only one isoline is considered.

The confidence interval estimation by bootstrapping is connected with the frequency estimation. The frequencies (quantiles) 0.025 and 0.975 produce a 95 % confidence interval. The map presented in Figure 3 has been obtained after 1000 resamplings. The palynological software POLPAL (NALEPKA & WALANUS 2003), with the bootstrap module was used.

## References

- BORDOGNA, G. & CHIESA, S. (2003), A fuzzy object-based data model for imperfect spatial information integrating exact objects and fields, *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 11, pp. 23-41.
- CHERNICK, M. R. (1999), *Bootstrap Methods, A practitioner's guide*, Wiley Series in Probability and Statistics.
- EFRON, B. & TIBSHIRANI, R. J. (1993), *An Introduction to the Bootstrap*. London, Chapman and Hall.
- HARMATA, K., MADEJA, J., OKUNIEWSKA-NOWACZYK, I. & NALEPKA, D. (2004), Gramineae Poaceae – Grass family. (Eds.): RALSKA-JASIEWICZOWA et al. (red.). Late Glacial and Holocene history of vegetation in Poland based on isopollen maps; W. Szafer Institute of Botany, Polish Academy of Sciences. Kraków, pp. 305-313.
- HEUVELINK, G. B. M. & BURROUGH, P. A. (2002), Developments in statistical approaches to spatial uncertainty and its propagation. *International Journal of Geographical Information Science*, 16 (2), pp. 111-113.
- JAGER, H. I. & KING, A. W. (2004), Spatial Uncertainty and Ecological Models, *Ecosystems*, 7, pp. 841-847.
- MANLY, B. F. J. (1997), *Randomization, bootstrap, and Monte Carlo Methods in biology*. 2. ed. London, Chapman & Hall.
- MARIGNANI, M., DEL VICO, E. & MACCHERINI, S. (2007), Spatial scale and sampling size affect the concordance between remotely sensed information and plant community discrimination in restoration monitoring. *Biodiversity and Conservation*, 16/13 pp. 3851-3861.
- NALEPKA, D. & WALANUS, A. (2003), Data processing in pollen analysis. *Acta Palaeobotanica*, 43 (1), pp. 125-134.
- NALEPKA, D. & WALANUS, A. (2004), Methods used for the construction of isopollen maps. (Eds.): RALSKA-JASIEWICZOWA et al. Late Glacial and Holocene history of vegetation in Poland based on isopollen maps. W. Szafer Institute of Botany, Polish Academy of Sciences. Kraków, pp. 21-23.