



SEMINAR IN STATISTICS

Wednesday, May 7, 2008

1:00 p.m.

223 Weber Building

Constructing Confidence Regions for Level Curves

Joshua French

PhD Candidate

Department of Statistics

Colorado State University

Abstract

The display of three-dimensional data is an important component of any spatial data analysis. One of the most common display tools used to represent spatial data is the contour plot. Informally, a contour plot is created by taking a “slice” of a three-dimensional surface at a certain level of the response variable and projecting the slice onto the two-dimensional coordinate-plane. The “slice” at each level is known as a level curve.

Suppose we have a continuous Gaussian random field $\{Z(s) : s \in \mathcal{D}\}$ where s is the location in the continuous two-dimensional region of interest $\mathcal{D} \subset \mathbb{R}^2$. The level curve for the process Z at level u is defined to be $I_u = \{s : Z(s) = u\}$. Based on the observed data $z(s_1), z(s_2), \dots, z(s_N)$ from a single realization of Z , our goal is to construct a confidence region containing I_u with high probability. From the sample data one can predict $\hat{z}(s)$ for any location $s \in \mathcal{D}$ using kriging, and from this create level curves $\hat{I}_u = \{s : \hat{z}(s) = u\}$ as an approximation of I_u . Typically, researchers analyzing spatial data focus only on “vertical” error, which is the the distance between the true response $Z(s)$ and the predicted response $\hat{z}(s)$. However, in some applications, one may be interested in the error in the horizontal direction. In this setting, horizontal error refers to the disparity between I_u and its approximation \hat{I}_u , i.e., the location of the true level curve compared to the location of the estimated level curve. This research will focus on accounting for horizontal error instead of vertical error.

Two main approaches have been used in the statistical literature to account for the horizontal error associated with finding I_u . One method uses a specific probability to assess the quality of a contour plot. The predicted value of $Z(s)$ depending on $z(s_1), \dots, z(s_N)$ for $s \in \mathcal{D}$ is used to create level curves for various levels. In the actual contour plot, s will lie between two level curves that represent the levels a and b , implying that $a < Z(s) < b$. The quality of the contour plot is assessed by calculating $p(s) = \mathbb{P}(a < Z(s) < b)$ for every $s \in \mathcal{D}$. If this probability tends to be high, then \hat{I}_u is a reasonable approximation of I_u . The other approach presents a method to add confidence limits to level curves. Starting at a particular location $s_0 \in \hat{I}_u$, one considers the process $Z(s)$ given $z(s_1), \dots, z(s_N)$ along a transect $T(s_0)$ perpendicular to \hat{I}_u at s_0 . Using theory related to crossing intensity, a confidence interval is constructed for the distance from s_0 to the first location along $T(s_0)$

where the process $Z(s)$ given $z(s_1), \dots, z(s_N)$ crosses the level u . This results in a confidence interval in the direction of $T(s_0)$ that contains a point along I_u with guaranteed probability. One drawback of this method is that each confidence interval is constructed independently and does not account for the problem of multiple inference. The method we propose will attempt to correct this deficiency.

To construct a confidence region for the true level curve, we would like to find a region S such that $\mathbb{P}(I_u \subseteq S) \geq 1 - \alpha$. Instead of finding S directly, we will adopt a different tack and try to find a set E which *does not* intersect I_u with guaranteed confidence, i.e., a set E such that $\mathbb{P}(\{I_u \cap E\} = \emptyset) \geq 1 - \alpha$. Consequently, the set $S = E^c$ since $\mathbb{P}(I_u \subseteq E^c) \geq 1 - \alpha$. The set E will be constructed by testing $H_0 : Z(s) = u$ versus $H_a : Z(s) \neq u$. If there is sufficient evidence to conclude that $Z(s)$ is different than u , then that location will be included in E . In order to make the scale of the problem more manageable, we will approach this problem by dividing \mathcal{D} into m rows and n columns of equal sized “pixels”. Let R denote $\{1, \dots, m\} \times \{1, \dots, n\}$ and $P_{i,j}$ represent the pixel in row i , column j . If the pixels are small enough, the center of each pixel should be a good representative of the process over that pixel. Let $Z_{i,j}$ represent $Z(s_0)$ where s_0 is the centerpoint of pixel $P_{i,j}$. Consider $\hat{Z}_{i,j} = \mathbb{E}(Z_{i,j} | Z(s_1), \dots, Z(s_N))$ and $\hat{\sigma}_{i,j}^2 = \mathbb{E}(Z_{i,j} - \hat{Z}_{i,j})^2$. Then $Z'_{i,j} = (Z_{i,j} - \hat{Z}_{i,j})/\hat{\sigma}_{i,j}$ has a standard normal distribution. Assuming that the Gaussian random field has mean zero and known covariance function, then $\hat{Z}_{i,j}$ corresponds to the usual kriging predictor (though of course in practice one would need to estimate the mean function and/or parameters of the covariance function). We seek a cutoff value C_α such that $\mathbb{P}\left(\max_{(i,j) \in R} |Z'_{i,j}| \leq C_\alpha\right) = 1 - \alpha$. Note that C_α does not depend on the observed data because $Z'_{i,j}$ is independent of the observed data $Z(s_1), \dots, Z(s_N)$. Finally, setting $E = \left\{ \bigcup P_{i,j} : |(\hat{Z}_{i,j} - u)/\hat{\sigma}_{i,j}| > C_\alpha, (i,j) \in R \right\}$ and $S = E^c$, we have the desired property $\mathbb{P}(I_u \subseteq S) \geq 1 - \alpha$.

A key component of this research is to find a good approximation to C_α . One approach is to consider the asymptotic behavior of the maximum of a triangular array of possibly non-stationary normal random variables. A second approach employs simulation to approximate the distribution of the maximum of $Z'_{i,j}$ for all $(i,j) \in R$.

Other possible approaches that we have considered, and will be explored further, include construction of confidence regions along transects perpendicular to $\hat{I}(u)$, construction of the set E directly through simulation of $Z(s)$ given $z(s_1), \dots, z(s_N)$, and the adaptation of methods based on distinct error criteria such as False Detection Rate (FDR).

Advisory Committee

Richard Davis, Adviser
 Jay Breidt, Committee Member
 Dan Cooley, Committee Member
 Robin Reich (Forest, Rangeland, and Watershed Stewardship)