Z. geol. Wiss., Berlin 340 (2012) 6: 381 – 391, 6 Abb., 5 Tab.

Global Estimates of Clustered Samples

Globale Schätzer für geclusterte Daten

HEINZ BURGER & MICHAEL FIETZ, Berlin

Key words: Clustered samples, global error estimates, confidence limits, Voronoi-tesselation.

Abstract

Clustered samples can lead to biased estimates and unreliable confidence limits for the mean value of spatially distributed variables. Spatial autocorrelation has the effect that samples within a cluster are more similar than samples from different clusters. Classical statistical methods require independent samples for calculating mean values and errors. This paper presents two case studies which require different statistical treatments for calculating statistical parameters. The first example is a case study in urban ecology: For a long-term study infrared aerial photographs have been used for the evaluation of the vitality of individual trees along traffic roads in the central area of Berlin, Germany. The sample pattern of these data is a set of 180 clearly separated clusters. It is the main goal of this study to detect changes of tree vitality in consecutive sampling campaigns, differentiated by species, urban districts and age classes of the trees. The detection of significant changes of the vitality of urban trees requires the calculation of the standard deviation of the mean proportion of healthy or damaged trees. Two methods are applied for estimating the mean value and standard deviation of these highly clustered samples: 1) an approximation formula which handles the clustering effect by weighting each proportion with respect to the number of trees in the test site and 2) the simulation of random sampling selecting a single sample tree from each test area and subsequent calculation of the statistics of this sample set.

The second example presents an exploration data set with a typical sample pattern: dense sampling in traverses and various clusters in other areas of the deposit. The variogram of the data reveals spatial autocorrelation so that declustering methods (cell declustering, polygonal method) are necessary for calculating local and global estimates of the mean. Block kriging can be used to estimate the total amount of reserves but the estimation of confidence intervals is more complicated because kriging errors depend on the size of the blocks and the errors aren't independent. We use a generalization of JOURNEL & HUIJBREGT's method "combination of independent elementary errors" for estimating the standard deviation of the global mean. This method is based on the Voronoi-tesselation of the sample domain and calculation of the extension variance of the generated polygons.

Zusammenfassung

Geowissenschaftliche Daten weisen immer eine räumliche Korrelation auf, d.h. benachbarte Probenwerte sind ähnlicher als weit entfernte. Dieser Effekt bewirkt, dass bei unterschiedlicher Datendichte und besonders beim Auftreten von Datenclustern eine einfache Berechnung von statistischen Parametern wie Mittelwert und Streuung von ortsabhängigen Variablen systematische Fehler auftreten können. Es gibt eine Reihe von Verfahren, die den Clustereffekt bei der Schätzung von lokalen Erwartungswerten berücksichtigen (z. B. durch geeignete Wichtung von Daten) und somit verzerrungsfreie Schätzer liefern, die zur Berechnung des globalen Erwartungswertes kombiniert werden können. Die berechneten Schätzfehler (z. B. bei den verschiedenen Krigingverfahren) können jedoch nicht in einfacher Weise zur Berechnung von Vertrauens-

intervallen für den globalen Erwartungswert verwendet werden, weil sie nicht unabhängig sind. Die Praxis erfordert jedoch Angaben über die Aussagesicherheit solcher Schätzwerte, z. B. die ökonomische Bewertung von Vorräten in einer Lagerstätte, der Nachweis zeitlicher Veränderungen von Klimadaten oder ökologischen Parametern.

In der vorliegenden Arbeit werden zwei Muster von geclusterten Daten genauer untersucht: Beim ersten Beispiel sind die Datencluster klar voneinander getrennt (s. Abb. 1), beim zweiten sind Cluster mit unterschiedlicher Form und Dichte in ein irreguläres Probenahmemuster eingestreut (Abb. 5 und 6).

In einer Langzeitstudie wurde mit Hilfe von Infrarotphotos die Vitalität von Straßenbäumen in den zentralen Bezirken von Berlin untersucht. Hierfür wurde die Vitalität von etwa 6000 Straßenbäumen an 180 Standorten anhand eines Photoschlüssels klassifiziert und die Ergebnisse nach den vier Gattungen (Linde, Ahorn, Platane, Kastanie) sowie nach drei Altersklassen und den Stadtbezirken differenziert. Das Ziel dieser Studie war festzustellen, ob signifikante Unterschiede der Straßenbaumvitalitäten in aufeinander folgenden Befliegungskampagnen nachgewiesen werden können, wobei der Anteil der ungeschädigten Bäume, getrennt nach Baumgattungen (Linde, Ahorn, Platane, Kastanie) sowie nach drei Altersklassen und den Stadtbezirken, bestimmt wurden. Zum Nachweis von Unterschieden der Straßenbaumvitalitäten wurden jeweils die Anteile der Bäume in einer gewissen Klasse und die Streuung dieses Parameters, der den Clustereffekt berücksichtigt (Abschn. 2, Tab. 2 bis 4) berücksichtigt. Es wurden zwei Methoden eingesetzt, um die Streuung für Clusterdaten zu erhalten: 1) eine Approximationsformel von Cochran (Abschn. 2.1), bei der die Anteile der Bäume einer Vitalitätsklasse im jeweiligen Cluster durch die Gesamtzahl der Bäume in diesem Cluster gewichtet werden; und 2) ein Simulationsansatz, bei dem aus jedem Cluster genau eine Zufallsprobe gewählt wird und anschließend der Anteil der Bäume in den verschiedenen Vitalitätsklassen bestimmt wird (Abschn. 2.4, Abb. 2, Tab. 5). Diese Zufallsprobe hat den Umfang der Anzahl der Testgebiete und kann als unabhängige Stichprobe betrachtet werden und die Streuung der simulierten Werte ist ein Schätzwert für die Standardabweichung des Mittelwerts.

Für stetige, ortsabhängige Variable wird die Polygonmethode vorgestellt, um unverzerrte Schätzer des globalen Mittelwerts der Variablen sowie seiner Streuung zu erhalten. Dieses Verfahren basiert auf der Voronoi-Zerlegung des Untersuchungsgebietes und der Berechnung von lokalen Schätzfehlern für jedes Polygon: Jedem Polygon wird der Datenwert des erzeugenden Probenpunktes zugeordnet; der Fehler, der dabei gemacht wird heißt Ausdehnungsfehler (geostatistical extension variance, s. Abb. 5). Jede Probe wird genau auf ihren Einflussbereich beschränkt und die Ausdehnungsfehler werden als unabhängig betrachtet. Die globale Streuung wird dann nach JOURNEL & HUIJBREGTS (1978) durch die Kombination dieser Elementarfehler bestimmt (Abschn. 3).