Notwendige Datenbankabfragen zum statistischen Schließen mit räumlich und zeitlich korrelierten Daten

Karl Gerald van den Boogaart*

5. März 2003

Zusammenfassung

Die in Umweltdatenbanken und Geoinformationsystemen gespeicherten Daten repräsentieren tatsächliche Vorgänge in Raum und Zeit, die in komplizierten Wechselwirkungen untereinander stehen. Typischerweise weisen Meßwerte von benachbarten Orten ähnliche Meßwerte auf und nahe gelegene Gebiete werden durch die gleichen Ursachen beeinflußt. Daher können diese Informationen nicht als stochastisch unabhängige Realisationen von Zufallsvariablen angesehen werden, was aber die zentrale Voraussetzung zur Anwendung der klassischen Statistischen Methoden ist. Dabei sind die Abhängigkeit sind gewöhnlich stärker je näher sich zwei Dinge räumlich oder zeitlich sind und verschwindet für große Distanz gewöhnlich ganz.

Im Beitrag werden grundlegende Methoden aufgezeigt, um statistische Verfahren auch auf stochastisch abhängige Daten übertragen zu können. Dabei wird die spezielle Struktur der räumlichen und zeitlichen Abhängigkeit ausgenutzt. Die zentrale Rolle spielt dabei eine neue Methode Bestimmung der Schätzgenauigkeit von Mittelwerten räumlich korrelierter Daten. Mittels dieses Ansatzes lassen sich klassische statistische Verfahren, wie Varianzanalyse und Regression auf räumlich und zeitlich stochastisch abhängige Zufallsvariablen übertrage und somit auf Umweltinformationen anwenden.

Im Zusammenhang mit der Anwendung dieser Verfahren auf Umweltdatenbanken und Geoinformationssystemen erweist sich allerdings, daß die Schätzformeln für die Varianz im Falle unabhängiger Beobachtungen zwar in SQL fest implementiert ist, die neue Formel allerdings eine Summe über Produkte vieler Datenpaare benötigt und somit im Prinzip n^2 Zeit benötigen würde. Ein geschicktes umorganisieren der Formel könnte allerdings zusammen mit geeigneten Abfragemöglichkeiten, die in einer räumliche Datenbank implementiert sein sollten, genutzt werden um das Verfahren mit wenige schnellen Datenbankabfragen durchführen zu können.

*TU Bergakademie Freiberg, Institut für Geologie

1 Räumliche Abhängigkeit

Die in Umweltdatenbanken und Geoinformationssystemen gespeicherten Daten beschreiben physikalische Phänomen im Raum, deren physikalischen Wechselwirkungen sich nicht an die durch das Datenmodell vorgegebene Grenzen halten, sondern durch gemeinsame Ursachen und gegenseitige Beeinflussung miteinander interagieren. Dadurch wird die Voraussetzung der stochastischen Unabhängigkeit, die als Generalvoraussetzung der Statistik gesehen werden kann, verletzt. Dabei gilt der Erfahrungsgrundsatz, daß räumlich nahe benachbart liegende Objekte aufweisen, als weiter entfernt liegende Objekte. Dieser Zusammenhang wird in der Geostatistik oft über das sogenannte Semivariogramm γ oder die Korrelationsfunktion quantifiziert[Cressie 1993].

$$\gamma(\mathbf{x},\mathbf{y}) := \frac{1}{2}E[(Z(\mathbf{x})-Z(\mathbf{y}))^2] \text{ Semivariogramm} \quad (1)$$

$$c(\mathbf{x},\mathbf{y}) := \text{cov}(Z(\mathbf{x}),Z(\mathbf{y})) \text{ Kovarianzfunktion} \quad (2)$$

Wobei $Z(\mathbf{x})$ die Zufallsgröße des am Ort $\mathbf{x}$ gemessenen Wertes, $E[]$ den Erwartungswert einer Zufallsgröße und cov die Kovarianz zweier Zufallsgrößen bezeichnet. Zwischen diesen beiden Funktionen gilt die Beziehung:

$$\gamma(\mathbf{x},\mathbf{y}) = \frac{1}{2}c(\mathbf{x},\mathbf{x}) + \frac{1}{2}c(\mathbf{y},\mathbf{y}) - c(\mathbf{x},\mathbf{y})$$

Die räumliche Abhängigkeit drückt sich darin, daß die Kovarianz nahen, aber verschiedenen Orten gemessener Z-Werte nicht 0 ist oder gleichbedeutend darin, daß das Semivariogramm $\gamma(\mathbf{x},\mathbf{y})$ für nahe Orte mbx,$\mathbf{y}$ kleinere Werte (damit erwartete quadratische Unterschiede) aufweißt, als für Orte mit großem Abstand.

Außerdem sind die Streuungen, der an verschiedenen Stellen im Raum oft unter unterschiedlichen Bedingungen entstehenden Werte, in vielen Fällen nicht konstant sondern von Ort zu Ort unterschiedlich. Das wiederspricht der in der Statistik oft angewendeten Annahmen der identischen Verteilung. Wir müssen also alle statistischen Verfahren in deren Annahme die „unabhängige identische Verteilung“vorkommt auf ihre Gültigkeit im Zusammenhang mit Umweltdaten überprüfen.

Insbesondere die räumliche Abhängigkeit, aber auch die nicht gleichmäßige Varianz, haben starke Auswirkungen auf statistische Schlüsse aus räumlich abhängigen Daten. Dies soll im Beispiel an einer fiktiven Problemstellung erläutert werden, in der Änderung der Belastung eines Gebiets mit einem Luftschadstoff (z.B Schwefeldioxid) überprüft, der an mehreren Stellen des Gebiets regelmäßig gemessen wird. Ähnliche Probleme treten aber mit praktisch allen räumlich erfaßten Daten auf. Wir nehmen also an, wir verfügen für den Zeitraum von zwei Sommern (2002,2003) über tägliche Messungen $Z(\mathbf{x}_i,j,t)$, $i=1,\ldots,32$ (Meßstationen), $j=2002,2003$ (Jahre),$t=210,\ldots,315$ (Sommertage) des Luftschadstoffes. Unsere Aufgabe ist nun festzustellen, ob sich die Schwefeldioxidkonzentration von einem Sommer zum nächsten im Durchschnitt geändert hat und wenn

ja in welche Richtung. Ein übliches stochastisches Modell würde nun annehmen, daß der Mittelwert der Messungen von der Meßstation und dem Tag des Jahres abhängt, so daß unsere Fragestellung im wesentlichen auf darauf hinausläuft zu überprüfen, ob die Werte an der gleichen Station am gleichen Tag des nächsten Jahres im Schnitt angewachsen sind, gefallen sind oder ob kein Unterschied nachweisbar ist. Geht man nun von der unrealistischen Annahme aus, daß abgesehen von dieser Mittelwertsabhängigkeit die Messungen stochastisch unabhängig sind so wird diese Frage von einem klassischen Wilcoxon-Vorzeichen-Rank-Test bzw. unter Annahme gleicher Streuung und Normalverteilung von einem gepaarten t-Test für die gepaarten Datensätze $Z(\mathbf{x}_i, 2002, t)$ und $Z(\mathbf{x}_i, 2002, t)$ beantwortet (s. z.B. [Rinne 1997]).

Die zur Berechnung der Tests nötigen Statistiken in einer relationalen Datenbank zu berechnen ist nicht gerade trivial, aber dennoch effizient machbar. Wir machen das am Beispiel des einfacheren t-Tests, der auf der sogenannten t-Statistik aufbaut:

$$\begin{aligned}
t &:= \frac{\bar{Z_D}}{\sqrt{\frac{1}{n}\hat{\sigma}^2_{Z_D}}} \\
Z_D(\mathbf{x}_i, t) &:= Z(\mathbf{x}_i, 2003, t) - Z(\mathbf{x}_i, 2002, t) \\
\bar{Z}_D &= \frac{1}{n}\sum_t \sum iZ_D(\mathbf{x}_i, t) \\
\hat{\sigma}^2_{Z_D} &:= \frac{1}{n-1}\sum_t \sum i(Z_D(\mathbf{x}_i, t) - \bar{Z}_D) \\
n &= \sum_t \sum_i 1
\end{aligned}$$

Die t-Statistik ist also der Quotient aus dem Mittelwert der Änderungen geteilt durch seine geschätzte Varianz. Sind diese Datensätze in einer relationalen Datenbank gespeichert, so genügt es durch die durch jahr=2002 und jahr=2003 definierte Selektionen, über Station und Tag mit einem inneren Join zu verknüpfen. Damit lassen sich die Z_D berechnen und die Abfrage von deren Mittelwert und Varianz genügen, um die t-Statistik direkt zu berechnen. In einer Datenbank mit einem Index für Station und Tag lassen sich diese Abfragen in $o(n)$ Zeit direkt in SQL durchführen. Bei dem Wilcoxon Vorzeichnen Rank Test wird das deutlich zeitaufwendiger, da zunächst die Sortierreihenfolge der Z_D berechnet werden muß.

Diese einfache Berechenbarkeit liegt in dem Aufbau der zu berechnenden Statistiken aus Mittelwerten und empirischen Varianzen von Attributen von innerhalb der relationalen Algebra berechenbaren Relationen.

2 Varianz räumlicher zeitlicher Mittelwerte

Im Fall der Schwefeldioxidkonzentrationen ist die Annahme räumlich und zeitlich unabhängiger Beobachtungen allerdings nicht realistisch. Nahe Stationen werden nicht nur ähnliche Meßwerte aufweisen, sondern werden sich auch von Jahr zu Jahr gleichsinnig verändern, da die Verschmutzungen ja auch gleiche Ursachen haben wird, die sich eigentlich verändern. Auch die Abhängigkeit vom Tag im Jahr ist wohl nur eine mittelbare, da der Sonnenstand das Wetter zwar beeinflußt, aber nicht bestimmt. Es ist also davon auszugehen, daß die Meßwerte naher Stationen eine positive Kovarianz aufweisen. Ebenso sollten Messungen mit geringem zeitlichen Abstand als abhängig angesehen werden, da sie aus der gleichen Wetterlage entstehen. Somit sind fast alle Beobachtungspaare als wechselseitig stochastisch abhängig anzusehen. [Boogaart 2002] zeigt, daß in Fällen abhängiger Daten $\hat{\sigma}^2_{Z_D}$ die Varianz von $\bar{Z}_D$ stark unterschätzt. Damit würde die t-Statistik auch im Falle im Mittel gleich gebliebener Schadstoffbelastungen große Werte bekommen und zur fälschlichen Ablehnung der Hypothese unveränderter Schadstoffbelastungen führen. In [Boogaart 2002] wird ein alternativer erwartungstreuer Schätzer $\hat{\sigma}^2_{s,Z_D}$ für die Varianz des Mittelwertes angeboten, der auch im Falle abhängiger Daten die wahre Varianz im Mittel trifft bleibt. Das s steht für spatial dependence (räumliche Abhängigkeit). Er funktioniert aber auch für zeitliche Abhängigkeiten und basiert auf der Kenntnis einer Menge $N \subset \{((i,t),(j,s)) : i,j = 1,\ldots,32, t,s = 210,\ldots,315\}$ (für Nachbarn) der stochastisch unabhängigen Beobachtungspaare $Z_D(\mathbf{x}_i,t), Z_D(\mathbf{x}_j,s)$:

$$\hat{\sigma}^2_{s,Z_D} = \bar{Z}^2_D - \frac{1}{|N^c|} \sum_{((i,t),(j,s))\in N^c} Z_D(\mathbf{x}_i,t) Z_D(\mathbf{x}_j,s)$$

Die Erwartungstreue dieses Schätzers unter der Annahme, das sich nichts verändert hat (also $E[Z_D(\mathbf{x}_i,t)] = 0 = E[\bar{Z}_D]$. schnell gezeigt:

$$\begin{aligned}
E[\hat{\sigma}^2_{s,Z_D}] &= E[\bar{Z}^2_D] - \frac{1}{|N^c|} \sum_{((i,t),(j,s))\in N^c} E[Z_D(\mathbf{x}_i,t) Z_D(\mathbf{x}_j,s)] && (3)\\
&= E[\bar{Z}^2_D] - \frac{1}{|N^c|} \sum_{((i,t),(j,s))\in N^c} E[Z_D(\mathbf{x}_i,t)] E[Z_D(\mathbf{x}_j,s)] && (4)\\
&= E[\bar{Z}^2_D] - \frac{1}{|N^c|} \sum_{((i,t),(j,s))\in N^c} E[\bar{Z}_D] E[\bar{Z}] && (5)\\
&= E[\bar{Z}^2_D] - E[\bar{Z}_D]^2 = \mathrm{var}(\bar{Z}_D) && (6)
\end{aligned}$$

In Zeile 4 nutzt man hier die Unabhängigkeit von $Z_D(\mathbf{x}_i,t)$ und $Z_D(\mathbf{x}_j,s)$. Nähere Untersuchungen finden sich in [Boogaart 2002]. Wie man klar sehen kann benötigt man jetzt eine viel größere Summe, nämlich über alle unkorrelierten Paare. Ob ein Paar unkorreliert ist entscheidet man aufgrund der räumlichen und zeitlichen

Entfernung. Um die Reichweite der räumliche und zeitliche Abhängigkeit zu bestimmen kann man zum Beispiel Methoden der Variographie benutzen, wie sie in z.B. in [Cressie 1993] erläutert wird oder auf fachwissenschaftliche Überlegungen zurückgreifen. Dieser Beweis ist außerdem völlig unabhängig von irgendwelchen Überlegungen der identischen Verteilung der Z_D und erlaubt somit verschiedene Streuungen an verschiedenen Stationen.

3 Die zur Berechnung notwendigen Datenbankabfragen

Man erkennt leicht, daß man in dieser Formel Summen über sehr viele Paare benötigt. Die Anzahl von Paaren liegt hier in der Größenordnung n^2. Bei einem Zugriff auf eine Datenbank kann das über ein äußeres Join realisiert werden, von dem dann ein großer Teil aufgrund der aus den Koordinatenwerten berechneten Entfernungen selektiert wird und dann die Summe über die Produkte gebildet wird. In Bezug auf die Rechenzeit scheint es lediglich sehr besorgniserregend, daß dieses äußere Join tatsächlich zwar nicht im Speicher, aber doch in Rechenzeit realisiert werden muß, da tatsächlich jedes Entfernung berechnet und jedes in der Summe vorhandenen Produkt berechnet werden muß. Daher kann die nicht im Rahmen der relationale Algebra automatische vereinfacht werden. Während wir im Falle unabhängiger Daten die Varianz praktisch durch eine einzige schnelle SQL Abfrage geliefert bekommen muß für die Abfrage der Varianz bei abhängigen Größen eine extrem zeitaufwendige Abfrage gestartet werden.

4 Algorithmen zur schnellen Berechnung

Um diese Abfragen zu beschleunigen benötigen wir spezielle Algorithmen. Die schnelle Aufzählung naher Ort in der hat wohl schon Aufnahme in die Geoinformatik und Computergeometrie gefunden, so daß Ortsmengen, die sicher nahe oder sicher ferne sind z.B.Über Quadtreezerlegungen schnell gefunden werden können. Um das Ausnutzen zu können benötigt man eine einfache Identität, die für jede Paarmenge $M \subset A \times B$ gilt:

$$\sum_{(i,j)\in M} x_i y_i + \sum_{(i,j)\in M^C} x_i x_j = \left(\sum_{i\in A} x_i\right)\left(\sum_{j\in B} x_j\right)$$

Die rechte Seite kann dabei immer in o(—A—+—B—) Zeit berechnet werden, während auf der anderen Seite des Gleichheitszeichen eine naive Implementierung zumindest eine der beiden Summen immer $o(|A||B|)$ benötigt. Falls es eine der beiden Summen schnell in o(—A—+—B—) berechnet werden kann, sollte diese verwendet werden und die andere daraus berechnet werden. Wie aber kann man mit Fällen umgehen in den beide Summen $o(|A||B|)$ viele Summanden besitzen?

5 Zusammenfassung

Räumliche statistische Abhängigkeit ist ein zentrales Problem in der statistischen Auswertung von Daten, wie sie typischerweise in Raumbezogenen Datenbanksystemen gespeichert werden. Selbst einfache Abfragen, wie etwa das schätzen der Stichprobenvarianz oder der Varianz des Mittelwertes erfordern Aufwendige Abfragen der Datenbank, die über die statistische Standardfunktionalität von SQL hinausreichen. Darüber hinaus ist eine schnelle Algorithmische Abarbeitung der notwendigen Abfragen nicht garantiert und erfordert selbst dann vom Nutzer die Entscheidung, welche Abfrage er verwenden sollte um die Antwort in nicht quadratischer Zeit zu erhalten. Ein effizienter Einsatz von Statistik in Umweltdatenbanken ist nur möglich, wenn sich Statistiken von fundamentaler Bedeutung auch in einfacher Weise und vernünftigen Rechenzeiten berechnen lassen.

Literatur

[Boogaart 2002] Boogaart, K. G. v.d. (2002): Estimating the variance of the spatial mean, *Mathematical Geology*, submitted

[Boogaart 2001] Boogaart, K. G. v.d. (2002): *Statistics of crystallographic individual orientation measurements*, Dissertationsschrift, Shaker Verlag, Aachen

[Cressie 1993] Cressie, N. (1993): *Statistics for spatial data*, revised edition, Wiley, New York, 900 p.

[Rinne 1997] Rinne, H. (1997): *Taschenbuch der Statistik*, 2., überarb. und erw. Auflage, Harry Deutsch, Thun