Bayesian generalized least square regression

Alexander Juschitz

09-03-2009

Inhaltsverzeichnis

1 Einleitung

Die Entwicklung eines rigorosen und effizenten Verfahrens zu Schätzung der regionalen Schiefe von Parametern in einem Modell und damit auch die Genaugkeit für eine log-Pearson typ 3 (LP3) Verteilung ist ein zentraler Bestandteil eines Berichtes, welcher in den USA bezüglich Richtlinien zu Ergreifung von Maßnahmen gegen Flut und ähnliche Umweltkatastrophen herausgegeben wurde. Dafür wird jedoch der Schätzer für die Schiefe und dessen Varianz benötigt. Hierbei liegt das besondere Augenmerk an einer möglichst niedrigen Varianz. Die üblichen Verfahren zur Schätzung dieses Parameters lassen jedoch an Schärfe fehlen [Tasker and Stedinger, 1986]. Das Bayes'sche verallgemeinerte Regressionsmodell (GLS), wie es hier vorgestellt wird, geht auf diese Probleme ein. Es biete ein Verfahren, das auf regionale Probleme eingeht, Quantile für Flut, und Statistiken für Regen und Strömung erstellt [Tasker and Stedinger, 1989].

1.1 Hintergrund

[Cunnane, 1988] und eine Gruppe von Wissenschafter haben unterschiedliche Methoden nur Nutzbarmachung von regionalen Daten zur Schätzng von Wasserstatistiken untersucht. Eine Möglichkeit war Strömung und den physiographischen Eigenschaften einer Region durch Regression in Verbindung zu setzen. Solche Verahren wurden oft genutzt um Wasserstatistiken zu schätzen in Gegenden ohne einer Meßstelle und um die Genauigkeit für Gegenden mit unterschiedlicher Datenlänge zu verbessern.

Die Regressionsmodelle für regionale Daten zielen darauf ab die Wasserstatistik für örtliche Erscheinungen anhand ihrer physiographischen Parameter, wie zum Beispiel Abflussgebiet, Gefälle zum Hauptfluss, prozentuelle Waldbedeckung,

herzuleiten. Über viele Jahre hinweg ist dies durch das klassische lineare Regressionsmodell (OLS) geschehen, welches die Fehler homogen verteilt und unabhängig annimmt [Riggs, 1973]. Gegen diese Annahmen wird jedoch häufig verstossen, da die Schätzer für Statistiken an unterschiedlichen Orten häufig unterschiedliche Datenlänge haben und voneinander abhängig sind, da der Flussstand oft räumlich korrelliert.

[Tasker, 1980] schlug darauf hin vor das verallgemeinerte lineare Regressionsmodell zu verwenden (WLS), welches unterschiedlich die Datenlänge gewichtet. Tasker kam zum Schluß, dass das WLS zu niedrigeren Fehlerquadraten in der Paramterschätzung führt als es das OLS Verfahren tut und ähnlich effizient ist.

[Kuczera, 1983] behandelte inwiefern kurze Datenlänge und die räumliche Korrelation sich auf die Parameterschätzung auswirken, wenn regionale und örtliche Daten empirisch in Verbindung gesetz werden. Er entwickelte ein Modell welches sowohl zeitliche, als auch räumliche Abhängigkeit miteinbezieht. [Stedinger and Tasker, 1985] entwickelten das GLS-Modell weiter um auch Quantile für Flut zu berechnen. Sie lösten auch das Problem, auf welches [Kuczera, 1983] und [Tasker, 1980] hinwiesen, wie man die Fehlervarianzmatrix schätzen kann.

Die geschätze Kovarianzmatrix sollte dabei unabhängig sein von den Quantilen um unbayse Parameter zu erhalten. Das führt zu einem Problem, da die Varianz der Fehler abhängig ist von der örtliche Varianz. [Stedinger and Tasker, 1985] haben ein geglättete Kovarianzmatrix verwendet, die zu regionalen Annahmen der örtlichen Varianz führt. Ihre Monte Carlo Simulation vergleicht die Effizienz von OLS, WLS und GLS bei der Schätzung von Quantilen von Flut, der Berechnung des Maximum Likelihood Funktion. Ihr Ergebnis zeigt das die Parameterschätzung durch WLS und GLS effizienter ist als durch OLS. GLS

führt zudem zu einem besseren Ergebnis als WLS wenn die Kreuzkorrelation zwischen den Flussströmungen hoch und die der Fehler gering ist.

1.2 Schätzung des Modellfehlers

[Stedinger and Tasker, 1986] untersucht die Genauigkeit der verallgemeinerten Methode der Momente und des Maximum Likelihoodschätzers. Dies wird vor allem zur Berechnung der Schiefe benötigt. Desweiteren, kann sogar die Momentenmethode und der Maximum Likelihoodschätzer den Wert Null annehmen wenn die Varianz der Modellfehler klein ist im Vergleich zum Fehler der Messung. Dies kommt oft zum tragen beim Versuch die Geländeparameter auf eine Region umzulegen.

2 Regression Problem

In diesem Abschnitt werden wird das verallgemeinerte Regressionsmodell (GLS) untersuchen.

Das GLS Modell nimmt an, dass die Variable y_i, die durch den Ort i gegeben ist, als lineare Funktion von physiographischen Eigenschaften erklärt werden kann. Dabei nennen wir y die zu erklärende Variable, und die physiographischen Eigenschaften die erklärenden Variablen. Zudem wird bei dem Modell noch ein Fehlerterm ϵ hinzugefügt. So lässt sich das Model in Matrixnotation anschreiben als

$$y = X\beta + \epsilon, \tag{1}$$

wobei X eine Matrix der Dimension $(n \times k + 1)$ ist mit Einsern in der ersten Spalte und Werten in den restlichen k Spalten, welche die erklärenden Variablen darstellen.

In der Literatur können wir folgenden Annahmen bezüglich des Modells finden:

1. X nicht stochastisch
2. $X^T X$ nicht singulär
3. $$\mathbb{E}[\epsilon] = 0 \tag{2}$$
4. $\mathbb{E}\left[\epsilon\epsilon^T\right] = \delta^2\Omega$ mit $\Omega^T = \Omega, \Omega > 0$, $\delta^2 \in \mathbb{R}^+$ unbekannt, $\Omega \in \mathbb{R}^{n\times n}$
5. keine a-priori Information bezüglich $\left(\beta, \delta^2\right) \in \mathbb{R}^{k+1} \times \mathbb{R}^+$

Durch eine unterschiedliche Wahl der Matrix Ω kann man auf verschiedene Eigenschaften für das Modell eingehen. Ist Ω eine Einheitsmatrix, so bedeutet das, dass eine Homoskedastie für die Fehler vorliegt, und das verallgemeinerte

Regressionsmodell sich auf das klassische reduziert (von GLS zu OLS). Sollten die Fehler unkorreliert sein, aber eine unterschiedlich Varianz haben, drückt sich das in der Matrix Ω so aus, dass sich unterschiedliche Wert in der Spur vorfinden. Alle restlichen Elemente sind dabei 0. In diesem Fall ändert sich das GLS Model in (2) zu WLS. Im allgemeinen Fall, hat Ω das Ziel korrelierte heteroskädastische Fehler zu repräsentieren. Nach dem Gauss-Markov-Aitken Theorem für bekanntes Ω, erhält man den unbayes'schen Schätzer für β mit kleinster Varianz

$$\hat{\beta} = \left(X^T \Omega^{-1} X\right)^{-1} X^T \Omega^{-1} y \tag{3}$$

mit

$$\Sigma \left[\hat{\beta}\right] = \delta^2 \left(X^T \Omega^{-1} X\right) - 1 \tag{4}$$

als Varianz für den Schätzer.

Der unbayes'sche Schätzer der Varianz für den unbayes'sche Fehler ist

$$\hat{\delta} = \frac{\left(y - X\hat{\beta}\right)^T \Omega^{-1} \left(y - X\hat{\beta}\right)}{n - k - 1} \tag{5}$$

In der Praxis ist Ω selten bekannt und muss geschätzt werden, was wiederum zu einem Verlust von Effizienz für die Schätzer von β und δ^2 führt.

2.1 Stedinger-Tasker Modell

[Stedinger and Tasker, 1985] und [Tasker and Stedinger, 1986] entwickelten eine leicht abgewandelte Form des GLS Modells für die Analyse eines regionales Wasserkreislaufs. Der Unterschied liegt in der Aufteilung der Kovarianzmatrix der Fehler. Ihr Modell nimmt an, dass der gesamte Fehler Resultat zweier ist

1. Modellfehler ϵ_i, die unabhängig identisch verteilt sind mit $\mathbb{E}[\epsilon_i] = 0$ und einer gemeinsamen Varianz δ^2,

2. Messfehler, der dadurch auftaucht, dass man nicht die genauen Quantile von y_i kennt.

Somit lässt sich die Gleichung (1) umformen, und wir erhalten

$$\hat{y} = X\beta + \omega + \epsilon = X\beta + \eta, \tag{6}$$

wobei ω der Messfehler ist. Somit sehen wir, dass der Regressionsfehler η_i eine Summe ist von

1. Zeitmessfehler im Schätzer $\hat{y}_i$ von y_i und
2. einem grundlegendem Modellfehler ϵ_i

Für den gesamten Fehler η können wir folgende Eigenschaften anführen

$$\mathbb{E}[\eta] = 0 \tag{7}$$

$$\mathbb{E}\left[\eta\eta^{\mathsf{T}}\right] = \Lambda\left(\delta^2\right) = \delta^2 I + \Sigma\left(\hat{y}\right), \tag{8}$$

wobei Σ die Kovarianzmatrix der Meßfehler in Schätzer ist. Die Fehler durch die Meßfrequenz in den y_i sind meist korreliert auf Grund der örtlichen Nähe. Die Schätzung der Kovarianzmatrix im GLS Verfahren ist von großer Bedeutung, auf die wir später noch eingehen werden.

Stedinger und Tasker entwickelten einen Schätzer für β, der gegeben ist durch

$$\beta = \left[X^T\Lambda\left(\delta^2\right)^{-1}X\right]^{-1}X^T\Lambda\left(\delta^2\right)^{-1}\hat{y}. \tag{9}$$

mit der Kovarianzmatrix für gegebenes δ^2

$$\Sigma\left[\hat{\beta}\right] = \left[X^T\Lambda\left(\delta^2\right)^{-1}X\right]^{-1}. \tag{10}$$

Die Varianz der Fehler δ^2 kann entweder durch die Momentenmethode oder durch die Maximum-Likelihoodfunktion geschätzt werden ([Stedinger and Tasker, 1985] und [Tasker and Stedinger, 1986]).

Den Schätzer durch die Momentenmethode (MM) erhält man indem man Gleichung (9) iterativ löst gemeinsam mit der Kleinsten-Quadrat-Gleichung

$$\left(\hat{y} - X\hat{\beta}\right)^T \left[\delta^2 I + \Sigma\right]^{-1} \left(\hat{y} - X\hat{\beta}\right) = n - (k+1) \tag{11}$$

für n Meßstellen und $k+1$ Parameter. Dies ist eine Verallgemeinerung der Gleichung (5). Manchmal kann die Kovarianzmatrix der Meßfehler die Unterschiedlichkeit der gemessenen Daten komplett erklären. In diesem Fall ist die linke Seite der Gleichung (11) kleiner als $n - (k+1)$ sogar wenn δ^2 gleich Null ist. Man setzt dann δ^2 als Null an [Stedinger and Tasker, 1985].

Der Maximum Likelhiood Schätzer wird hergeitet unter der Annahme, dass die Fehler normal verteilt sind mit Erwartungswert Null und der Kovarianzmatrix wie in Gleichung (8). Dann kann β und δ^2 gemeinsam geschätzt werden mit Nebenbedingung $\delta^2 \geq 0$ indem man folgende Gleichung minimiert

$$\frac{1}{2} \ln\left[\det\left(\delta^2 I + \Sigma\right)\right] + \frac{1}{2} \left(\hat{y} - X\beta\right)^T \left[\delta^2 I + \Sigma\right]^{-1} \left(\hat{y} - X\beta\right) \tag{12}$$

Klarerweise ist der Maximum Likelihood Schätzer für β genau der gleiche wie in Gleichung (9), mit dem alleinigen Unterschied, dass sich δ^2 verändert hat. Da β und δ^2 assymptotisch unabhängig sind [Rencher, 2000], kann die Varianz von β dargestellt werden durch die Inverse der Fishermatrix [Bickel and Doksum, 1977],

$$\Sigma\left[\hat{\beta}\right] = \left\{ \mathbb{E}_{\hat{y}} \left[\frac{\partial^2 \ln f\left(\hat{y}|\hat{\beta}\right)}{\partial \beta^2} \right] \right\}^{-1} \tag{13}$$

was genau der Gleichung (10) entspricht.

2.2 Schätzer für die Fehlervarianz

Da die Verwendung einer geschätzten und geglätteten Kovarianzmatrix für die Messfehler zu einem Verlust von Effizienz in der Schätzung von β führt, kann

man nicht notwendigerweise davon ausgehen, dass die Schätzung durch GLS effizienter ist als jene durch OLS. Jedoch haben unterschiedliche Studien [[Stedinger and Tasker, 1985], [Moss and Tasker, 1991], [Kroll and Stedinger, 1998]] gezeigt, dass das verallgemeinerte Regressionsmodel zu besseren Resultaten führt als dies WLS oder OLS tun, vor allem in Wasserstatistiken.

Anhand von Monte Carlo Simulationen, sind [Stedinger and Tasker, 1986] zum Entschluss gekommen, dass die Momenten Methode zur Schätzung der Fehlervarianz schneller und robuster ist. Die Robustheit begründet sich darauf, dass keine Annahmen für die Verteilung der Fehler gemacht werden. Zudem kommt, dass das Ergebnis weniger verzerrt ist wenn die tatsächliche Fehlervarianz groß ist. Ihre Resultate haben gezeigt, dass bei kleiner Fehlervarianz die Maximum Likelihood Methode zu bevorzugen ist bei der Schätzung der Schiefe. Eine Analyse mittels bayes'scher Statistik empfiehlt sich hier ebenfalls, da sie ja auf der Likelihood Methode aufbaut.

2.3 Vorhersagen der Varianz

Es sei Y_0 eine Statistik, die es durch die Werte $x_0\hat{\beta}$ in einem neuen Gebiet mit physiographischen Eigenschaften x_0 zu schätzen gilt. Dann ist die Varianz gegeben durch

$$\mathbb{E}\left[\left(Y_0 - x_0\hat{\beta}\right)^2\right] = \delta^2 + x_0\left(X^T\Lambda^{-1}X\right)^{-1}x_0^T \tag{14}$$

wobei die zwei Terme auf der rechten Seite der Gleichung die regionale Fehlervarianz und die Varianz der Meßfehler durch $x_0\hat{\beta}$ sind. Die Varianz wie sie in Gleichung (14) definiert ist, kann auch als Kriterium zur Auswahl des Modells für einen bestimmten Ort verwendet werden. Da man jedoch hauptsächlich an der Regression für eine ganze Region interessiert ist, und nicht nur für einen

bestimmten Ort, haben [Stedinger and Tasker, 1986] den AVP_{new} (average variance of prediction for new sites) entwickelt, welcher anzeigt, wie gut ein Model sich auf ein Gebiet anwenden lässt

$$AVP_{new} = \delta^2 + \frac{1}{n}\sum_{i=1}^{n} x_i \left(X^T \Lambda^{-1} X\right)^{-1} x_i^T = \delta^2 + ASV \tag{15}$$

mit ASV als die durchschnittliche Stichprobenvarianz.

Manchmal kann versucht werden das Regressionsmodell für eine Region durch Information von einem bestimmten Ort zu verbessern [Kuczera, 1983]. In diesem Fall, muss der AVP_{new} auch die Korrelation zwischen $x_0\hat{\beta}$ und den Modellfehler berücksichtigen

$$AVP_{old} = \delta^2 + \frac{1}{n}\sum_{i=1}^{n} \left[x_i \left(X^T \Lambda^{-1} X\right)^{-1} x_i^T - 2\delta^2 x_i \left(X^T \Lambda^{-1} X\right)^{-1} X^T \Lambda^{-1} e_i\right] \tag{16}$$

mit e_i als Spaltenvektor mit einer 1 in der i-ten Reihe und 0 sonst [Reis, 2005].

3 Bayes'sche Ansatz

3.1 Hintergrund

Die derzeitige Analyse zum GLS Modell, wie sie in [Tasker and Stedinger, 1989] beschrieben wird, liefert keine Auskunft über die Unsicherheit in der geschätzten Fehlervarianz. Die Analyse beinhaltet auch keinen Schätzer der Varianz von β. Dazu kommt, dass in den Fällen, wo die Fehlervarianz gering ist, wie bei der Regionalisierung des Schätzers zur Schiefe, sowohl durch die Momentenmethode, als auch durch die Maximum Likelihoodmethode, der Schätzer für die regionale Fehlervarianz gleich Null ist. Das lässt zu dem Schluss kommen, dass das Modell perfekt ist, was wiederum dazu führt, dass es zu einer Unterschätzung bei der Unsicherheit in den Quantilen zur Flut und zu einer zu hohen Relevanz des regionales Modells kommt. Die Bayes'sche Statistik ist in diesem Fall ein nützliches Werkzeug, weil sie sowohl für die Fehlervarianz als auch für die Parameter eine posteriori Verteilung liefert. Zudem erklärt sie die Fehler besser wenn die Maximum Likelihood Schätzer Null ist. Bayes'sche Analyse ist nachzulesen in [Kitanidis, 1986], [Guadard et al., 1999], [Best et al., 2000], [Schmidt and Gelfand, 2003], [Wikle and Anderson, 2003], [Wikle, 2003], und [Richardson and Best, 2003]. Zum Beispiel verwendete [Holland et al., 2000] Bayes'sche Statistik um die regionalen Verbreitung von Schwefeldioxid in den östlichen Vereinigten Staaten zu schätzen. Ähnlich unserem GLS Modell, hatten sie auch eine Kovarianzmatrix mit zwei additiven Termen, einen für das eigentliche Fehlermodell und einen anderen Term für die Messfehler mit einer festgelegten Kovarianzmatrix. Sie verwendeten den Markov Chain Monte Carlo Algorithmus um die regionale Entwicklung, den Schätzfehler, als auch die posteriori Verteilung des

Fehlervarianz für das Modell zu erhalten.

3.2 Bayes'sche Statistik

Die Bayes'sche Statistik ist eine Alternative zur klassischen Statistik und dessen Betrachtungsweise eines Problems. Dabei wird das Wissen über eine Modellparameter durch eine Wahrscheinlichkeitsverteilung beschrieben. Das ermöglicht das apriori Wissen über einen Parameter mit den zur Verfügung stehend Daten zu verbinden [Zellner, 1971] indem man Bayes Theorem

$$p\left(\theta|\hat{y}\right) = \frac{p\left(\hat{y}|\theta\right)\xi\left(\theta\right)}{\int p\left(\hat{y}|\theta\right)\xi\left(\theta\right)d\theta} \tag{17}$$

verwendet. Hier stellt $p(\theta|\hat{y})$ die a-posteriori Verteilung des Parametervektors θ dar bei gegebenen Daten $\hat{y}$. $p(\hat{y}|\theta)$ ist die Maximum Likelihood Funktion für Daten und $\xi(\theta)$ ist die a-priori Verteilung von θ. Der Nenner dient zur Normalisierung der a-posteriori Dichte.

Der große Vorteil der Bayes'schen Statistik im Gegensatz zur Klassischen ist die Verfügbarkeit der a-posteriori Verteilung der Parameter um die Annahme über asymptotische Normalverteilung bei Unsicherheit treffen zu können [Bickel and Doksum, 1977]. Durch die Verfügbarkeit der vollständigen a-posteriori Verteilung der Parameter sind keine numerische Approximation zur Bestimmung der Unsicherheit mehr von Nöten. Vielmehr können nun Parameterintervalle viel leichter angegeben und interpretiert werden als bei Konfidenzintervallen der klassischen Statistik [Congdon, 2001].

3.3 Bayes'sche GLS Model

3.3.1 A-priori Verteilung

Hier wird die a-priori Information bezüglich der β Parameter dargestellt durch eine multivariate Normalverteilung mit Erwartungswert β_p und der Matrix P, welche die Inverse der Kovarianzmatrix von β_p ist. Wenn keine a-priori Information vorliegt, kann eine a-priori Verteilung mit geringer Aussagekraft und hoher Varianz ($P \approx 0$) gewählt werden. Eine multivariater normalverteilter Prior ist

$$\xi(\beta) = \frac{|P|^{1/2}}{(2\pi)^{(k+1)/2}} \exp\left[-0.5\,(\beta - \beta_p)^T P\,(\beta - \beta_p)\right] \tag{18}$$

wobei β die Dimension $k+1$ hat.

Wenn keine Information zur Fehlervarianz verfügbar ist, kann der Prior durch den Kehrwert der Varianz dargestellt werden [Zellner, 1971]. Der Prior ist zwar unpassend, jedoch führt das zu einer passenden a-posteriori Verteilung, wenn man ihn mit der Likelihood Funktion für die klassische Regression verbindet [Zellner, 1971]. Ein andere Weg ist eine $Gamma(\alpha, \beta)$ Verteilung zu benutzen [Congdon, 2001]. Wenn keine vorangehende Information verfügbar ist, sollten die Parameter α und β so gewählt werden, dass die Dichtefunktion relativ flach für die zu behandelnde Region ist.

3.3.2 Likelihood Funktion

Die Likelihood Funktion für Daten ist die multivariate Normalverteilung mit

$$L\left(\hat{y}|\beta, \delta^2\right) = (2\pi)^{-\pi/2} \frac{1}{|\Lambda|^{1/2}} \exp\left[-0.5\,(\hat{y} - X\beta)^T \Lambda^{-1}\,(\hat{y} - X\beta)\right] \tag{19}$$

wobei die Kovarianzmatrix die Form annimmt, wie sie in Gleichung (8) beschrieben wird, n die Anzahl der Meßstellen ist, $\hat{y}$ der Vektor mit den hydrologischen Meßdaten und X die Matrix der erklärenden Variablen ist.

3.3.3 Quasi-analytische Lösung

Die numerische Berechnung der Konstanten in Gleichung (17) kann sehr rechenintensiv sein, abhängig von der Größe des Problems. Ein möglicher Ansatz zur Lösung des Problems liegt in der Verwendung des Markov-Chain-Monte-Carlo-Verfahrens (MCMC) [Gilkis et al., 1996], welches eine korrelierte Stichprobe von Parameter des Modells von der a-posteriori Verteilung (17) generiert ohne die eigentliche normalisierende Konstante zu kennen. Diese Stichprobe wird dann verwendet um die a-posteriori Verteilung der Varianz des Modellfehlers und der β Parameter zu erhalten. [Reis et al., 2005] entdeckte, dass das Problem viel leichter gelöst werde kann, indem man die quasi-analytische Approximation der a-posteriori Grenzverteilung der Fehlervarianz verwendet.

In diesem Abschnitt wird die quasi-analytische Approximation der a-posteriori Verteilung der Fehlervarianz hergeleitet mit Hilfe der selben Faktorisierung, wie sie von [Kitanidis, 1986] und [Zellner, 1971] verwendet wurde. Die numerische Integration des eindimensonalen Integrals liefert den Erwartungswert und die Varianz der regionalen Fehlervarianz. Wenn man die a-posteriori Verteilung der Fehlervarianz kennt, kann man die posteriori Momente der β Parameter durch ein weiteres eindimensionales Integral numerisch berechnen.

Mit der Likelihood Funktion aus Gleichung (19) und der a-priori Verteilung für β aus Gleichung (18), kann man die geminsame a-posteriori Verteilung von δ^2 und β durch die möglichen Werte für β berechnen um numerisch die a-posteriori Grenzverteilung von δ^2 zu erhalten. Daher gilt :

$$f\left(\delta^2|\hat{y}\right) = \int f\left(\beta, \delta^2|\hat{y}\right) d\beta \propto \int f\left(\hat{y}|\beta, \delta^2\right) \xi\left(\beta, \delta^2\right) d\beta \tag{20}$$

wobei $f(\beta, \delta^2|\hat{y})$ die gemeinsame a-posteriori Verteilung der Parameter ist.

$f(\hat{y}|\beta, \delta^2)$ ist die Likelihood Funktion, und $\xi(\beta, \delta^2)$ ist der gemeinsame Prior für δ^2 und β. Wenn man die gemeinsame Verteilung von β aus Gleichung (18) verwendet, so ist die Verteilung des gemeinsamen Priors gegeben als

$$\xi\left(\beta, \delta^2\right) \propto \xi\left(\delta^2\right) \exp\left[-0.5\left(\beta - \beta_p\right)^T P\left(\beta - \beta_p\right)\right] \tag{21}$$

wobei $\xi(\delta^2)$ der Prior für die Fehlervarianz ist. Das Integral in Gleichung (20) ergibt die a-posteriori Grenzverteilung von δ^2, welche gegeben ist durch

$$\begin{aligned} f\left(\delta^2|\hat{y}\right) \propto \xi\left(\delta^2\right) \left|\Lambda^{-1/2}\right| \int \exp\Big[-0.5\left(\hat{y} - X\beta\right)^T \Lambda^{-1}\left(\hat{y} - X\beta\right) \\ -0.5\left(\beta - \beta_p\right)^T P\left(\beta - \beta_p\right)\Big]. \end{aligned} \tag{22}$$

Mit der Substitution

$$\begin{aligned} \left(\hat{y} - X\beta\right)^T \Lambda^{-1}\left(\hat{y} - X\beta\right) = \left(\hat{y} - X\hat{\beta}\right)^T \Lambda^{-1}\left(\hat{y} - X\hat{\beta}\right) \\ + \left(\beta - \hat{\beta}\right)^T X^T \Lambda^{-1} X\left(\beta - \hat{\beta}\right) \end{aligned} \tag{23}$$

kann das Integral aus Gleichung (20) neu angeschrieben werden

$$\begin{aligned} \exp\left\{-0.5\left[\left(\hat{y} - X\hat{\beta}\right)^T \Lambda^{-1}\left(\hat{y} - X\hat{\beta}\right) + \beta_p^T P\beta_p + \hat{\beta}^T X^T \Lambda^{-1} X\hat{\beta} - \beta_o^T C\beta_o\right]\right\} \\ \cdot \int \exp\left\{-0.5\left[\left(\beta - \beta_o\right)^T C\left(\beta - \beta_o\right)\right]\right\} d\beta \end{aligned} \tag{24}$$

wobei $\beta_o = C^{-1}[P\beta_p + X^T\Lambda^{-1}X\hat{\beta}]$ und $C = P + X^T\Lambda^{-1}X$ gilt. Daher ergibt sich für die a-posteriori Grenzverteilung von δ^2

$$\begin{aligned} f\left(\delta^2|\hat{y}\right) \propto \left[\left|\Lambda^{-1}\right| |C|\right]^{-1/2} \exp\left\{-0.5\left[\left(\hat{y} - X\hat{\beta}\right)^T \Lambda^{-1}\left(\hat{y} - X\hat{\beta}\right)\right.\right. \\ \left.\left. + \beta_p^T P\beta_p + \hat{\beta}^T X^T \Lambda^{-1} X\hat{\beta} - \beta_o^T C\beta_o\right]\right\} \xi\left(\delta^2\right) \end{aligned} \tag{25}$$

Für eine nicht-informative a-priori Verteilung für die Parameter β, sodass in Gleichung (9) $P \to 0$, dann $C \to X^T\Lambda^{-1}X$ und $\beta_o \to \hat{\beta}$ gilt, vereinfacht sich die a-posteriori Verteilung von δ^2 zu einen Ausdruck, wie er schon früher in [Reis

et al., 2003] erwähnt wurde

$$\begin{aligned} & f\left(\delta^2|\hat{y}\right)\left[|\Lambda|\left|X^T\Lambda^{-1}X\right|\right]^{-1/2} \\ \cdot \exp & \left[-0.5\left(\hat{y}-X\hat{\beta}\right)^T\Lambda^{-1}\left(\hat{y}-X\hat{\beta}\right)\right]\xi\left(\delta^2\right) \end{aligned} \tag{26}$$

Man kann nun entweder Gleichung (25) oder (26) zur numerischen Berechnung der a-posteriori Dichte des Modellfehlers verwenden, und damit auch für die Berechnung des Erwartungswertes und der Varianz. Die Dichtefunktion der Varianz des Modellfehlers kann später verwendet werden um die a-posteriori Verteilung der Parameter β zu erhalten indem man

$$f\left(\beta|\hat{y}\right) = \int f\left(\beta|\delta^2,\hat{y}\right) f\left(\delta^2|\hat{y}\right) d\delta^2 \tag{27}$$

numerische löst, wobei $f(\beta|\delta^2,\hat{y})$ die eigentliche multivariate Normalverteilung für jeden Wert von δ^2 entsprechend $(\beta|\delta^2,\hat{y}) \sim N[\beta_o, C^{-1}]$ ist.

Die a-posteriori Momente von β können nun mehr numerisch bestimmt werden durch die Gleichungen

$$\begin{aligned} \mu_\beta = \mathbb{E}\left(\beta|\hat{y}\right) &= \int \beta f\left(\beta|\hat{y}\right) d\beta = \int \mathbb{E}\left(\beta|\delta^2,\hat{y}\right) f\left(\delta^2|\hat{y}\right) \\ &= \int \beta_o f\left(\delta^2|\hat{y}\right) d\delta^2 \end{aligned} \tag{28}$$

$$\begin{aligned} Var\left(\beta|\hat{y}\right) &= \iint \left(\beta-\mu_\beta\right)^2 f\left(\beta|\delta^2,\hat{y}\right) f\left(\delta^2|\hat{y}\right) d\beta d\delta^2 \\ &= \int \left\{\left[\beta_o\left(\delta^2\right)-\mu_\beta\right]+C^{-1}\right\} f\left(\delta^2|\hat{y}\right) d\delta^2 \end{aligned} \tag{29}$$

Dieses Ergebnis ist weit einfacher als das GLS Verfahren angeführt in [Stedinger and Tasker, 1985]. Mit geeigneten numerischen Verfahren können die Integrale in Gleichung (28) und (29) gelöst werden.

4 Conclusio

In diesem Bericht geht es darum, einen bayes'schen Ansatz zur Analyse eines GLS Modells zur Regionalisierung von hydrologischen Daten vorzustellen. Die posteriori Verteilung der Modellfehlervarianz, mit Ausnahme der Konstanten zur Normalisierung, zur numerischen Berechnung der Dichtefunktion wird hergeleitet, wie auch Erwartungswert, Standardfehler der regionalen Modellfehlervarianz. Die Parameter β werden unter Verwendung von einfachen eindimensionalen numerischen Integralen hergeleitet. Die quasi-analytische bayes'sche Lösung liefert eine vollständige posteriori Verteilung der regionalen Modellparameter und die Modellfehlervarianz. Bisher war das traditionelle GLS Verfahren, welches durch [Tasker and Stedinger, 1989] vorangetrieben wurde, unvollständig in Bezug auf die Beschreibung der Genauigkeit der computererrechneten Modellfehlervarianz. Diese Lücke wurde somit geschlossen. In manchen Fällen, wo die Stichprobenfehler $Var[y_i]$ größer als der regionale Modellfehler δ^2 sind, führte das bisherige GLS Verfahren zu einer Modellfehlervarianz von Wert Null. Dies kommt durch Überbestimmung der Genauigkeit der regionalen Schätzung zustande, was wiederum zu einer Unterschätzung der Unsicherheit in den Flutquantilen und zu hoher Abhängigkeit zum regionalen Modell führt. Der Lösungsansatz durch bayes'sche Statistik mit einen geeigneten Prior führt zu einer sinnvollen Werten für die Modellfehlervarianz in solchen Fällen. Die quasi-analytische Lösung liefert eine einfache und notwendige Erweiterung zum GLS Verfahren, wie in [Stedinger and Tasker, 1985] und [Tasker and Stedinger, 1989] beschrieben, und zum WLS Verfahren [Stedinger and Tasker, 1986]. Man kann damit aber auch die Schiefe, die regionale Schiefe, das regionale κ für GEV Modelle schätzen.

Literatur

[Best et al., 2000] Best, N. G., Ickstadt, K., and Wolpert, R. L. (2000). Spatial poisson regression for health and exposure data measured at disparate resolutions. *J. Am. Stat. Assoc., 95.*

[Bickel and Doksum, 1977] Bickel, P. J. and Doksum, K. A. (1977). *Mathematical Statistics: Basic Ideas and Selected Topics.* Boca Raton.

[Congdon, 2001] Congdon, P. (2001). *Bayesian Statistical Modelling.* Hoboken.

[Cunnane, 1988] Cunnane, C. (1988). Methods and merits of regional flood frequency analysis. *J. Hydrol., 100.*

[Gilkis et al., 1996] Gilkis, W. R., Richardson, S., and Spiegelhater, D. J. (1996). *Markov Chain Monte Carlo in Practice.* CRC Press.

[Guadard et al., 1999] Guadard, M., Karson, M., Linder, E., and Sinha, D. (1999). Bayesian spatial prediction. *Environ. Ecol. Stat., 6.*

[Holland et al., 2000] Holland, D. M., Oliveira, V., Cox, L. H., and Smith, R. L. (2000). Estimation of regional trends in sufur dioxide over the eastern united states. *Environmetrics, 11.*

[Kitanidis, 1986] Kitanidis, P. K. (1986). Parameter uncertainty in estimation of spatial functions: Bayesian analysis. *Water Resour. Res., 22(4).*

[Kroll and Stedinger, 1998] Kroll, C. N. and Stedinger, J. R. (1998). Regional hydrologic analysis: Ordinary and generalized least squares revisited. *Water Resour. Res., 34(1).*

[Kuczera, 1983] Kuczera, G. (1983). Effect of sampling uncertainty and spatial correlation on an empirical bayes procedure for combining site and regional information. *J. Hydrol., 65.*

[Moss and Tasker, 1991] Moss, M. and Tasker, G. D. (1991). An intercomparison of hydrological network-design technologies. *Hydrol. Sci. J.,36(3).*

[Reis, 2005] Reis, D. (2005). *Flood frequency analysis employing Bayesian regional regression and imperfect historical information.* PhD thesis, Sch. of Civ, and Environ. Eng., Cornell Univ., Ithaca, N.Y.

[Reis et al., 2003] Reis, D. S., Stedinger, J. R., and Martins, E. S. (2003). Bayesian gls regression with application to lp3 regional skew estimation. *in Proceedings World Water and Environmental Resources Congress 2003.*

[Reis et al., 2005] Reis, D. S., Stedinger, J. R., and Martins, E. S. (2005). Bayesian generalized least squares regression with application to log pearson type 3 regional skew estimation. *Water Resour. Research, 41.*

[Rencher, 2000] Rencher, A. C. (2000). *Linear Models in Statistics.* Hoboken.

[Richardson and Best, 2003] Richardson, S. and Best, N. G. (2003). Bayesian hierarchical models in ecological studies of health-environment effects. *Environmetrics, 14(2).*

[Riggs, 1973] Riggs, H. C. (1973). *Regional analyses of streamflow characteristics*. U.S. Geol. Surv. Tech. Water Resour. Invest.

[Schmidt and Gelfand, 2003] Schmidt, A. M. and Gelfand, A. E. (2003). A bayesian coregionalization approach for multivariate pollutant data. *J. Geophys. Res., 108(D24)*.

[Stedinger and Tasker, 1985] Stedinger, J. R. and Tasker, G. D. (1985). Regional hydrologic analysis: 1. ordinary, weighted and generalized least squares compared. *Water Resour. Res., 21(9)*.

[Stedinger and Tasker, 1986] Stedinger, J. R. and Tasker, G. D. (1986). Regional hydrologic analysis, 2: Model-error estimators, estimation of sigma and log-pearson type 3 distributions. *Water Res. Research, 22(10)*.

[Tasker, 1980] Tasker, G. D. (1980). Hydrologic regression with weighted least squares. *Water Resour. Res., 16(6)*.

[Tasker and Stedinger, 1986] Tasker, G. D. and Stedinger, J. R. (1986). Estimating generalized skew with weighted least squares regression. *Water Resour. Plann. Manage.,112(2)*.

[Tasker and Stedinger, 1989] Tasker, G. D. and Stedinger, J. R. (1989). An operational gls model for hydrologic regression. *Hydrol., 111(1-4)*.

[Wikle, 2003] Wikle, C. K. (2003). Hierarchical bayesian models for predicting the spread of ecological processes. *Ecology, 84(6)*.

[Wikle and Anderson, 2003] Wikle, C. K. and Anderson, J. (2003). Climatological analysis of tornado report counts using a hierarchical bayesian spatiotemporal model. *J. Geophys. Res., 108(D24)*.

[Zellner, 1971] Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*. Hoboken.