Kurzfassung

Optimierung von Algorithmen zur Schätzung von robusten Spatial Small Area Modellen

Masterarbeit

Philip Rosenthal

31. März 2015

Betreuer und Gutachter: Prof. Dr. Ralf Münnich
Betreuer und Gutachter: Prof. Dr. Ekkehard Sachs

In Politik, Wissenschaft und Wirtschaft steigt der Bedarf an regional oder inhaltlich differenzierten Daten. So benötigen zum Beispiel politische Entscheidungsträger Informationen über Einkommen, Arbeitslosigkeit und Mietpreisentwicklung auf Kreisebene, um geeignete Maßnahmen zu treffen und finanzielle Mittel effektiv und effizient allozieren zu können. Aus Kostengründen werden solche Daten jedoch in der Regel auf aggregierter Ebene durch geeignete Stichprobenziehungen erhoben (Beispiel: Mikrozensus). Die Betrachtung von fein untergliederten Variablen und die damit einhergehende Disaggregation der Daten kann zur Folge haben, dass die Substichprobengrößen von einigen Kreisen zu klein sind, um mit klassischen statistischen Methoden Schätzwerte mit angemessener Präzision erhalten zu können. Spätestens im Extremfall einer Substichprobengröße von Null können herkömmliche Verfahren nicht mehr verwendet werden.

Sind geeignete Hilfsvariablen vorhanden, können modellbasierte Small Area Verfahren eingesetzt werden, um den effektiven Stichprobenumfang in solch kleinen Regionen (engl. „small areas“) zu erhöhen und so die Schätzung zu verbessern. In vielen Small Area Modellen wird von einer Normalverteilung der abhängigen Variable ausgegangen. Das ist insbesondere bei der Betrachtung von Wirtschaftsdaten problematisch, da diese oft eine eher rechtsschiefe Struktur aufweisen. Wenn außerdem Ausreißer in den Daten vorhanden sind, kann dies zu zusätzlichen Verzerrungen der Small Area Schätzer führen. Von einer Anwendung robuster Schätzmethoden ist in solchen Fällen eine deutliche Verbesserung der Ergebnisse zu erwarten.

Wird zusätzlich von räumlichen Abhängigkeiten in den Daten ausgegangen, wie beispielsweise bei Mietspiegeln nahe Ballungsgebieten, können diese Zusammenhänge in der Korrelationsstruktur des Small Area Modells berücksichtigt werden, selbst wenn keine geeigneten Hilfsvariablen vorhanden sind.

Schmid (2011)[1] fasste beide Ideen, sowohl die Robustifizierung gegenüber Ausreißern als auch die Berücksichtigung räumlicher Korrelationsstrukturen, in einem Modell, dem *robusten Spatial Small Area Modell*, zusammen.

Für eine Anwendung in der Praxis ist jedoch nicht nur ein gutes Modell, sondern auch ein guter Algorithmus notwendig, der in möglichst vielen Fällen brauchbare Ergebnisse berechnen kann. Bei der Anwendung der komplizierten Modellgleichungen des robusten Spatial Small Area Modells auf synthetische Daten mit Ausreißern sowie auf reale Wirtschafstsdaten stellte sich jedoch heraus, dass das von Schmid verwendete gewöhnliche Newton-Verfahren an seine Grenzen stieß und in einigen Fällen nicht gegen eine Lösung konvergierte.

[1] Schmid, T. (2011): *Spatial Robust Small Area Estimation applied on Business Data.* Doktorarbeit.

Das Ziel dieser interdisziplinären Masterarbeit, die als Gemeinschaftsprojekt der angewandten Statistik und numerischen Mathematik der Universität Trier unter Betreuung von Prof. Dr. Ralf Münnich und Prof. Dr. Ekkehard Sachs geschrieben wurde, ist neben der ausführlichen Beschreibung und Herleitung des robusten Spatial Small Area Modells auch die Entwicklung eines geeigneten numerischen Algorithmus, der die komplizierten Gleichungen schnell und vor allem verlässlich lösen kann.

Viele Small Area Modelle, wie z. B. das berühmte Fay-Herriot- oder das Battese-Harter-Fuller-Modell, lassen sich auf das *Generelle Lineare Gemischte Modell*

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X\beta} + \boldsymbol{Zv} + \boldsymbol{e}, \\ \boldsymbol{v} &\sim \boldsymbol{N(0, G)}, \\ \boldsymbol{e} &\sim \boldsymbol{N(0, R)}, \end{aligned} \tag{1}$$

zurückführen. Salvati (2004)[2] verwendete u.a. einen simultanen autoregressiven Prozess, um eine Nachbarschaftsmatrix, die Informationen über räumliche Distanzen der abhängigen Variable enthält, in die Kovarianzmatrix G des sogenannten *Zufallseffekts* oder *Area-Effekts* v einzubauen.

Auch Sinha und Rao (2009)[3] gingen von Modell (1) aus. Allerdings führten sie die Robustifizierung des Modells mit Hilfe der sogenannten *Normalengleichungen*, die zur der Herleitung des *Besten Linearen Unverzerrten Prädiktors* (BLUP) verwendet werden können, durch. Um den Abstand zwischen beobachteten und geschätzten Werten zu beschränken, wendeten sie die Huber-Funktion Ψ_k auf die Residuen an (siehe Abbildung 1).

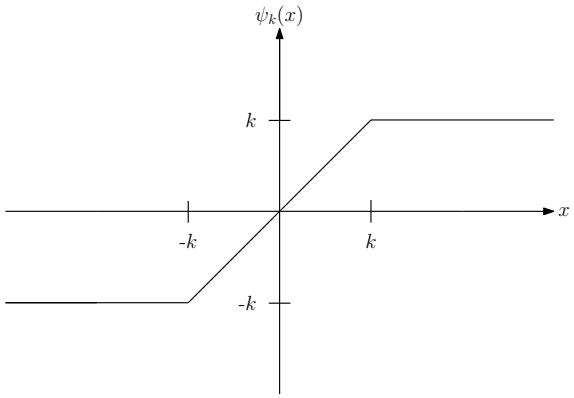

Abbildung 1: Huber-Funktion $\psi_k(x)$

Fügt man diese beiden Ansätze, wie bei Schmid (2009) beschrieben, zusammen, erhält man folgendes Gleichungssystem:

[2]Salvati, N. (2004): *Small Area Estimation by Spatial Models: The Spatial Empirical Best Linear Unbiased Prediction (Spatial EBLUP).*

[3]Sinha, S. K. und Rao, J. N. K. (2009): *Robust Small Area Estimation.*

$$
\begin{aligned}
\alpha(\sigma_e^2, \sigma_u^2, \rho, \beta) :=& X^T V^{-1} U^{\frac{1}{2}} \Psi_k(r) && \stackrel{!}{=} 0 \\
\Phi(\sigma_e^2, \sigma_u^2, \rho, \beta) :=& \Psi_k^T(r) U^{\frac{1}{2}} V^{-1} \frac{\partial V}{\partial \sigma_u^2} V^{-1} U^{\frac{1}{2}} \Psi_k(r) - \text{spur}(V^{-1} \frac{\partial V}{\partial \theta_l} K) && \stackrel{!}{=} 0 \\
\Gamma(\sigma_e^2, \sigma_u^2, \rho, \beta) :=& \Psi_k^T(r) U^{\frac{1}{2}} V^{-1} V^{-1} U^{\frac{1}{2}} \Psi_k(r) - \text{spur}(V^{-1} K) && \stackrel{!}{=} 0 \\
\Omega(\sigma_e^2, \sigma_u^2, \rho, \beta) :=& \Psi_k^T(r) U^{\frac{1}{2}} V^{-1} \frac{\partial V}{\partial \rho} U^{\frac{1}{2}} \Psi_k(r) - \text{spur}(V^{-1} \frac{\partial V}{\partial \theta_l} K) && \stackrel{!}{=} 0
\end{aligned}
\tag{2}
$$

Durch die Lösung von (2) erhält man Werte für $\sigma_e^2, \sigma_u^2, \rho$ und β, die letztendlich für die Schätzung von Statistiken wie Mittel- oder Totalwerten benötigt werden.

Nach mathematischer Analyse des Gleichungssystems stellte sich schließlich heraus, dass anders als in (2) *alle* Gleichungen von *allen* Parametern abhängen. Um aufwändige Ableitungsberechnungen zu vermeiden, wurde zunächst eine Version des numerisch effizienten *Newton-GMRES-Algorithmus* angewendet, der zur Klasse der matrix- und ableitungsfreien Verfahren gehört. Da sowohl das bisher verwendete gewöhnliche als auch das Newton-GMRES-Verfahren sensitiv auf Startwertveränderungen reagieren, die hier mit der *Henderson-Methode-III* berechnet wurden, wurde der Algorithmus um einen einfachen Globalisierungsansatz ergänzt.

Wider erwarten funktionierte die neue Methode zunächst *schlechter* als die bislang verwendete. Das Aufsplitten des Gleichungssystems und die Verwendung eines neu entwickelten *Hybridalgorithmus*, der nur noch einen Teil des Gleichungssystems mit Newton-GMRES und den anderen mit einem Fixpunkt-Verfahren löst, führte jedoch zu deutlich höheren Erfolgsraten als die bisherige Verwendung des gewöhnlichen Newton-Verfahrens. Die Praxistauglichkeit des Hybridalgorithmus wurde schließlich durch eine modellbasierte Simulationsstudie belegt, in der unterschiedliche Ausreißerszenarien betrachtet wurden.

Die weitere Erforschung solcher Hybridalgorithmen, für die bislang kaum mathematische Theorie existiert, könnte langfristig dazu führen, dass komplizierte statistische Small Area Modelle anwendungsfreundlich und effizient in praxisrelevante Softwareanwendungen implementiert werden können.