

Zur Messung räumlicher Autokorrelation

PETER M. SCHULZE

Institut für Statistik und Ökonometrie
Johannes Gutenberg - Universität Mainz

Zusammenfassung/Summary

- 1. Einführung**
 - 2. Gewichtungsfaktoren**
 - 3. RAK-Koeffizienten**
 - 3.1 Allgemeine Analyse**
 - 3.2 Beispiel**
 - 4. RAK im Regressionsmodell**
 - 5. Zusammenfassung**
- Literaturverzeichnis**

Zusammenfassung

Daten aus räumlich benachbarten Regionen können miteinander korreliert sein. Dabei ist das Beziehungsgefüge zwischen den Raumeinheiten in aller Regel vielfältiger als bei der traditionellen Betrachtung der Zusammenhänge zwischen Zeitreihen- oder Querschnittsdaten.

Der vorliegende Beitrag gibt einen Überblick über die statistisch-ökonomische Analyse räumlicher Autokorrelation. Zunächst wird dabei der Frage nachgegangen, welche Probleme bei der Festlegung der Gewichtungsfaktoren, die etwas über die Art und/oder die Intensität der Beziehungen der Raumeinheiten aussagen, bestehen. Es werden verschiedene Maße vorgestellt. In der eigentlichen Analyse wird unterschieden, ob es sich um die Messung der Korrelation zwischen verschiedenen Raumeinheiten für ein und dieselbe Variable handelt oder aber um die Prüfung und Berücksichtigung von Autokorrelation im Rahmen eines Regressionsmodells.

Für die Analyse des ersten Sachverhalts hat sich weitgehend der Koeffizient von Moran durchgesetzt. Hierfür wird ein Illustrationsbeispiel dargestellt. Danach wird der Frage nachgegangen, wie räumliche Autokorrelation in den Residuen eines Regressionsmodells nachgewiesen werden kann und welche Schätzmöglichkeiten bei deren Vorliegen bestehen.

Summary

It is the purpose of the paper to give a survey of the phenomenon of spatial autocorrelation. First, the problem of the spatial weight matrices is considered, and different measures are presented.

The next section shows several coefficients discussed in the literature to measure spatial autocorrelation. Here the Moran-coefficient is mostly preferred because of several useful properties. The use of this measure is illustrated by an example.

Finally the paper deals with the treatment of spatial correlation in the disturbances of the linear regression model. It is shown how to test in this case and which estimation procedures may be used in the presence of spatial autocorrelation.

1. EINFÜHRUNG

Werte aus Zeitreihendaten sind sehr oft untereinander korreliert [Zeitliche (Auto-) Korrelation]. Genauso können auch Daten miteinander korreliert sein, die aus räumlich benachbarten Raumeinheiten stammen [Räumliche (Auto-) Korrelation: RAK]. Dies kann daraus resultieren, daß räumlich nahe beieinander liegende Regionen sich im Hinblick auf bestimmte sozio-ökonomische Merkmale stärker gleichen können als weiter voneinander entfernt liegende.

Im Bereich der quantitativen Regionalforschung hat es in den letzten 40 Jahren wichtige methodische Fortschritte gegeben, die sich aus der besonderen Natur räumlich verteilter Daten ergeben. So erzeugen Verflechtungen aufgrund des interregionalen Beziehungsgefüges "Nachbarschaftseffekte". Das gleiche gilt für Attraktionswirkungen bestimmter Regionen auf andere Raumeinheiten. Diese Effekte können gleichgerichtete, systematische Auswirkungen auf die gleichen Variablen in verschiedenen Regionen haben. Allerdings haben diese räumlichen Besonderheiten in der angewandten Regressionsanalyse, z.B. in Form von verschiedenen Tests und Schätzmethoden, vergleichsweise geringe Berücksichtigung gefunden.

Nehmen wir an, wir unterteilen einen (Wirtschafts-)Raum in R verschiedene, sich nicht überschneidende Raumeinheiten (Regionen), die insgesamt gerade den Gesamttraum ausfüllen. Die interessierende Variable sei mit Y bezeichnet. Betrachten wir jeweils zwei Regionen r und s ($r, s = 1, 2, \dots, R$), so sprechen wir davon, daß keine RAK vorhanden ist, wenn Y_r und Y_s nicht korreliert sind. Andererseits liegt RAK vor, wenn die Datenwerte nicht alle paarweise unkorreliert sind.

Zwischen zeitlicher und räumlicher Autokorrelation besteht ein wesentlicher Unterschied. Bei zeitlichen Autokorrelationen gibt es eine natürliche Unterscheidung zwischen Vergangenheit und Gegenwart. In einem räumlichen Prozess kann es dagegen zu Verbindungen bzw. Abhängigkeiten in allen Richtungen kommen; die Beziehungen sind meist bilateral, bei Zeitreihendaten unilateral.

Zwei Fragen können hierbei von Interesse sein:

1. Wie läßt sich prüfen, ob RAK zwischen den Originaldaten vorliegt?
2. Wie lassen sich Parameter in einem Regressionsmodell bei vorhandener RAK schätzen, d.h. wenn RAK zwischen den Regressionsresiduen vorliegt?

Beide Fälle sollen im folgenden behandelt werden. Zuvor ist jedoch auf Fragen der Gewichtungsfaktoren einzugehen.

2. GEWICHTUNGSFAKTOREN

Die Gewichtungsfaktoren sollen über das Beziehungsgefüge der Regionen untereinander Auskunft geben. Die Wahl dieser Gewichtungsfaktoren, die in die Berechnung eines RAK-Koeffizienten eingehen soll, ist sehr viel subjektiver als im Fall von Zeitreihendaten. Die Ansätze von Moran (1950) und Geary (1954) [vgl. unten Gleichungen (5) und (8)] basieren auf binären Nachbarschaftsfeststellungen: Haben zwei räumliche Einheiten eine gemeinsame Grenze, so werden sie als benachbart angesehen und dem Gewichtungsfaktor w dieser Beziehung der Wert 1 zugeordnet, sonst eine Null:

$$(1) \quad w_{rs} = \begin{cases} 1 & \text{wenn } r \text{ und } s \text{ benachbart} \\ 0 & \text{sonst} \end{cases}$$

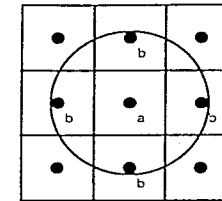
Diese Festlegung der Nachbarschaft erfordert eine Karte, in der die Grenzen erkannt werden können. Für eine naturräumliche Anordnung der Regionen ist dies meist einfach. Wenn die Raumeinheiten sich jedoch z.B. auf ein reguläres Gitter beziehen, so ist die Festlegung der Nachbarschaft nicht eindeutig, etwa bei der Frage, ob Eckpunkte eine Nachbarschaft etablieren sollen oder nicht [In Abb. 1 z.B. Region A mit den Regionen C].

Abb. 1: Nachbarschaft mit Eckpunkten

C	B	C
B	A	B
C	B	C

Bei räumlichen Einheiten, die Punkte (Knoten) enthalten - so wie z.B. Städte über den Raum verteilt sind - kann Nachbarschaft auch gemessen werden als die kürzeste Entfernung zwischen diesen Punkten [vgl. Abb. 2].

Abb. 2: Nachbarschaft mit Knoten



In einem solchen Netzwerk lassen sich Knoten als benachbart ansehen, wenn sie innerhalb einer gegebenen Distanz liegen [Die in Abb. 2 mit b bezeichneten Punkte liegen innerhalb eines Kreises mit einem bestimmten Radius und können als benachbart im Rahmen dieser Distanzbetrachtung angesehen werden, vgl. Anselin (1986), S. 18].

Darüber hinaus lassen sich verschiedene Ordnungen der Nachbarschaft betrachten: Stoßen die Raumeinheiten unmittelbar aneinander, so spricht man von Nachbarschaft 1. Ordnung, von Nachbarschaft 2. Ordnung, wenn eine dazwischen liegende Region vorhanden ist, usw.. In Abb. 1 haben die Region A und B die Nachbarschaft 1. Ordnung und A und C eine Nachbarschaft 2. Ordnung [wenn man die Eckpunkte als Berührungspunkte außer acht läßt]. Eine solche einfache Nachbarschaftsordnung ist relativ grob - wenngleich sie bei Anwendungen aus Einfachheitsgründen oft benutzt wird.

Erweiterungen des einfachen Konzepts einer binären Nachbarschaftsmatrix durch Maße, die die Interaktion zwischen zwei räumlichen Einheiten messen, sind möglich [Cliff/Ord (1973) S. 11ff.]. Dies läßt sich durch eine Gewichtungsmatrix W erreichen. Die Festlegung der Elemente kann allerdings im Einzelfall schwierig sein. So wurde vorgeschlagen [Cliff/Ord (1969) S. 32f.], eine Kombination von Abstandsmaßen und relativer Länge der gemeinsamen Grenzen zwischen zwei Raumeinheiten zu benutzen:

$$(2) \quad w_{rs} = d_{rs}^{-a} \beta_{rs}^b$$

Die relative Länge β_{rs} läßt sich dabei angeben als Anteil der gemeinsamen Grenze zwischen r und s an der gesamten Grenzlänge von r . Hier gilt i.a. nicht $w_{rs} = w_{sr}$, es sei denn, die beiden betrachteten Raumeinheiten haben die

gleiche Gesamtgrenzlänge. d_{rs} ist die Distanz zwischen Raumeinheit r und s ; a und b sind Parameter. Dacey [(1968) S. 490] schlägt Gewichtungsfaktoren vor, die darüberhinaus den relativen Flächen-Anteil α_r der Einheit r an der Gesamtfläche aller Raumeinheiten berücksichtigen:

$$(3) \quad w_{rs} = d_{rs} \alpha_r \beta_s$$

Bei der Berechnung dieser Gewichte geht die geographische Eigenart der betrachteten Raumeinheiten ein. Wie die binären Nachbarschaftskoeffizienten sind sie dann weniger nützlich, wenn bei der zugrundeliegenden Analyse etwa Punkte [vgl. Abb. 2] interessieren, denn dann verlieren Grenzlänge und Flächenmaße an Bedeutung. Sie sind auch dann weniger aussagekräftig, wenn das zu analysierende Merkmal durch rein ökonomische Faktoren und weniger durch die räumliche Konfiguration der Grenzen bestimmt wird. Deshalb führen z.B. Bodson/Peeters [(1975) S. 465 ff.] Gewichte ein, die das untersuchte Phänomen dadurch direkt einbeziehen, daß in einer logistischen Funktion der Einfluß verschiedener Kommunikationsverbindungen zwischen den Regionen, wie z.B. Straßen, Eisenbahnstrecken, Flußverbindungen, berücksichtigt wird:

$$(4) \quad w_{ij} = \sum_{j=1}^m k_j \frac{a}{1 + b^{(c_j d_{ij})}}$$

Es gibt m ($j = 1, \dots, m$) Verbindungen, gemessen durch die Distanz d_{ij} ; k_j gibt die relative Bedeutung der jeweiligen Kommunikationsverbindung an. a , b und c_j sind Parameter, die geschätzt werden müssen.

Ein nicht zu unterschätzendes Problem stellen die Parameter dar, die zur Bestimmung der w_{rs} dienen. Sie werden entweder vorgegeben oder aber durch eine separate Analyse bestimmt. Die Gültigkeit der Schätzungen kann davon abhängen, wie genau die Gewichte die räumliche Struktur abbilden. Dies kann aber bedeuten, daß die RAK, die durch die Analyse aufgedeckt werden soll, bereits vor der eigentlichen Datenanalyse vorweggenommen werden muß. Insgesamt ist zu sagen, daß es keine allgemeine Übereinstimmung darüber gibt, welche Gewichtungsfaktoren zur Berechnung der räumlichen AK benutzt werden sollen. Die Spezifikation der Gewichtungsmatrix W ist abhängig zu machen vom konkret benutzten Modell [Hordijk (1979) S. 111]. Eine Unterscheidung läßt sich treffen, je nachdem, ob W bei der Berechnung von Nachbarschaftskoeffizienten zur Feststellung räumlicher AK zwischen den beobachteten Variablen Y [s. Kap. 3] oder im Rahmen von Regressionsmodellen dienen soll [s. Kap. 4].

Kritisch bleibt aber für alle Spezifizierungsversuche von W festzuhalten, daß die Bedeutung der Regionen, ihr naturräumliches Muster und die relative Stärke der Verbindungen zwischen den Raumeinheiten kaum berücksichtigt wird. Außerdem ist W räumlich invariant, d.h. ein und dieselbe Gewichtungsmatrix kann zu unterschiedlichen räumlichen Anordnungen der Regionen gehören [vgl. Cliff/Ord (1969) S. 29f.].

Führt man einen RAK-Test durch, so sind dessen Ergebnisse immer als bedingt in dem Sinn zu betrachten, als sie von der Art der gewählten w_{rs} abhängen. Eine Standardinterpretation eines Tests bei einem bestimmten Signifikanzniveau kann deshalb in diesem Zusammenhang irreführend sein, wenn nicht die Unsicherheit berücksichtigt wird, die durch die vorherige Wahl von W entsteht - wie bei der Bayes-Statistik [vgl. Anselin (1988a) S. 288].

3. RAK - KOEFFIZIENTEN

3.1 Allgemeine Analyse

Betrachten wir die RAK zwischen ein und derselben Variablen Y in verschiedenen räumlichen Einheiten. Die zugehörigen Koeffizienten sollen mit $C^{(k)}$ (entspr. Contiguity) bezeichnet werden, wobei k für die Ordnung der Nachbarschaft steht. Zunächst ist das von Moran [(1950) S. 21ff.] entwickelte Maß $C_M^{(k)}$ zu nennen:

$$(5) \quad C_M^{(k)} = \frac{R \sum_{r=1}^R \sum_{s=1}^R (y_r - \bar{y})(y_s - \bar{y}) w_{rs}^{(k)}}{R_k \sum_{r=1}^R (y_r - \bar{y})^2} \quad \text{mit } w_{rr}^{(k)} = 0$$

R bedeutet wie bisher die Zahl der betrachteten Regionen; R_k ist die Zahl der Regionen, die aufgrund der Ordnung k benachbart sind. $w_{rs}^{(k)}$ bezeichnet binäre

Gewichtungskoeffizienten der Ordnung k nach (1) und $\bar{y} = \frac{1}{R} \sum_{r=1}^R y_r$.

$C_M^{(k)}$ weist formale Ähnlichkeit zum Bravais-Pearson-Korrelationskoeffizienten auf. Ist keine RAK vorhanden, so wird $\sum (y_r - \bar{y})(y_s - \bar{y}) w_{rs}^{(k)} = 0$ und damit ist $C_M^{(k)} = 0$. Positive RAK wird demnach durch $C_M^{(k)} > 0$ und negative RAK durch $C_M^{(k)} < 0$ gekennzeichnet [vgl. Hordijk (1974) S. 124].

Unter der Voraussetzung normalverteilter und unabhängiger $y_1, \dots, y_R$ haben Cliff und Ord [(1973) S. 30f.] gezeigt, daß sich Erwartungswert und Varianz von $C_M^{(k)}$ wie folgt ableiten lassen:

$$(6) \quad E[C_M^{(k)}] = -\frac{1}{R-1}$$

$$(7) \quad \text{Var}(C_M^{(k)}) = \frac{R^2 S_1 - RS_2 + 3R_k^2}{(R^2 - 1)R_k^2}$$

$$\text{mit } S_1 = \frac{1}{2} \sum_{r=1}^R \sum_{s=1}^R (w_{rs} + w_{sr})^2$$

$$S_2 = \sum_{r=1}^R (w_r + w_{.r})^2$$

$$\text{und } w_{.r} = \sum_{s=1}^R w_{rs} \text{ bzw. } w_{.r} = \sum_{s=1}^R w_{sr}.$$

Ein weiterer Koeffizient geht auf Geary [(1954) S. 116] zurück und lautet:

$$(8) \quad C_G^{(k)} = \frac{(R-1) \sum_{r=1}^R \sum_{s=1}^R (y_r - y_s)^2 w_{rs}^{(k)}}{2R_k \sum_{r=1}^R (y_r - \bar{y})^2} \quad \text{mit } w_{\pi}^{(k)} = 0$$

Auch hier bezeichnet R_k die Zahl der Regionen, die aufgrund der Ordnung k miteinander verbunden sind. $C_G^{(k)}$ weist eine formale Ähnlichkeit zu dem von J. v. Neumann aufgestellten Koeffizienten zur Prüfung zeitlicher Autokorrelation auf. Interpretieren läßt sich $C_G^{(k)}$ wie folgt: Wenn keine RAK der Ordnung k vorliegt, dann ist $C_G^{(k)} = 1$. Wenn $C_G^{(k)} > 1$, so spricht man von negativer RAK, während bei $C_G^{(k)} < 1$ positive RAK vorliegt [vgl. Hordijk (1974) S. 125]. Anstatt $C_G^{(k)}$ zu benutzen, läßt sich auch ein komplementärer Koeffizient $g^{(k)}$ mit

$$(9) \quad g^{(k)} = 1 - C_G^{(k)}$$

definieren, was sich als Koeffizient der räumlichen Beeinflussung bezeichnen läßt [Paelinck/Nijkamp (um 1975) S. 227]. Wenn $g^{(k)} = 0$, dann gibt es keine

Beeinflussung der Ordnung k zwischen den Regionen, wenn $g^{(k)} > 0$ oder $g^{(k)} < 0$, dann liegen positive bzw. negative Nachbarschaftseffekte vor.

Cliff und Ord [(1969) S. 32] haben die Ansätze von Moran und Geary weiterentwickelt und die binären Gewichte (1) durch solche in (2) ersetzt. Es resultiert eine Art gewichteter Moran-Koeffizient $C_{CO}^{(k)}$ mit

$$(10) \quad C_{CO}^{(k)} = \frac{R \sum_{r=1}^R \sum_{s=1}^R (y_r - \bar{y})(y_s - \bar{y}) w_{rs}^{(k)}}{W \sum_{r=1}^R (y_r - \bar{y})^2} \quad \text{mit } w_{\pi}^{(k)} = 0$$

wobei $W = \sum_{r=1}^R \sum_{s=1}^R w_{rs}^{(k)}$.

Dieser Koeffizient hat sich allerdings nicht durchgesetzt, sicherlich nicht zuletzt deshalb, weil die Bestimmung der Gewichte nach Gleichung (2) sehr aufwendig ist. Vielmehr wird oft der Moran-Koeffizient (5) benutzt, auch deshalb, weil $C_M^{(k)}$ asymptotisch normalverteilt ist, wenn keine RAK vorliegt [vgl. Cliff/Ord (1973) Chap. 2,5]. Außerdem entspricht die Interpretation von $C_M^{(k)}$ derjenigen der üblichen Korrelationskoeffizienten; darüber hinaus spricht für ihn die relativ leichte Berechenbarkeit.

In allgemeiner Form lassen sich alle drei Koeffizienten als Nachbarschaftskoeffizienten c subsumieren unter

$$(11) \quad c = a \frac{y' W y}{y' y}$$

wobei a eine Konstante ist; y enthält mittelwertbereinigte Größen. W stellt eine entsprechende $r \times r$ Gewichtungsmatrix [s. Kap. 2] dar und sollte Restriktionen derart erfüllen, daß das Maß im Wertebereich zwischen +1 und -1 liegt [letzteres gilt z.B. nicht für den Geary-Koeffizienten (8)].

Ein Signifikanztest - mit H_0 : Keine RAK - läßt sich dann mit der asymptotisch standardnormalverteilten Größe

$$(12) \quad r_c = \frac{c - E(c|H_0)}{\sqrt{\text{Var}(c|H_0)}}$$

durchführen. Für $R < 50$ wird man beim Testen kaum brauchbare Annäherungen mit (12) erhalten. In solchen Fällen lassen sich bei bestimmten Annahmen nur über Simulationen Anhaltspunkte gewinnen [vgl. Cliff/Ord (1973) S. 37].

3.2 Beispiel

Zur Illustration der vorangegangenen Überlegungen soll der Koeffizient der räumlichen Autokorrelation nach Moran [Gleichung (5)] für die Variable Veränderung des Volkseinkommens je Einwohner in % in den Jahren 1971 und 1983 und die Bevölkerungsdichte 1970 und 1987 in den alten Bundesländern der Bundesrepublik Deutschland ohne Westberlin dienen [vgl. Gemeinschaftsveröffentlichung (1986), Heft 15, S. 292 und Statistisches Bundesamt (1990) S. 44]. Wegen der geringen Zahl von Regionen läßt sich sinnvoll kein Signifikanztest durchführen, vielmehr wird hier $C_M^{(k)}$ als deskriptives Korrelationsmaß betrachtet.

Abb. 3: Bundesländer als Regionen (Alte Bundesländer ohne Westberlin)

1. Schleswig-Holstein SLH
2. Niedersachsen NS
3. Nordrhein-Westfalen NRW
4. Hessen HE
5. Rheinland-Pfalz RPL
6. Baden-Württemberg BW
7. Bayern BY
8. Saarland SL
9. Hamburg HH
10. Bremen HB

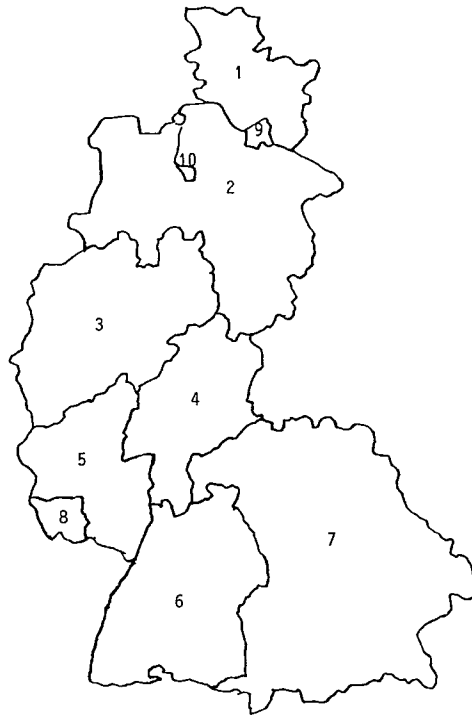


Tabelle 1 zeigt zunächst die aufgrund von Abb. 3 erstellte Nachbarschaftsmatrix.

Tabelle 1: Nachbarschaftsmatrix

Land	1	2	3	4	5	6	7	8	9	10
SLH 1	-	+	0	0	x	x	x	*	+	0
NS 2	+	-	+	+	0	0	0	x	+	+
NRW3	0	+	-	+	+	0	0	0	0	0
HE 4	0	+	+	-	+	+	+	0	0	0
RPL 5	x	0	+	+	-	+	0	+	x	x
BW 6	x	0	0	+	+	-	+	0	x	x
BY 7	x	0	0	+	0	+	-	x	x	x
SL 8	*	x	0	0	+	0	x	-	*	*
HH 9	+	+	0	0	x	x	x	*	-	0
HB 10	0	+	0	0	x	x	x	*	0	-

Es bedeutet:

- +
 - 0
 - x
 - *
- : Nachbarschaft 1.Ordnung
 : Nachbarschaft 2.Ordnung
 : Nachbarschaft 3.Ordnung
 : Nachbarschaft 4.Ordnung

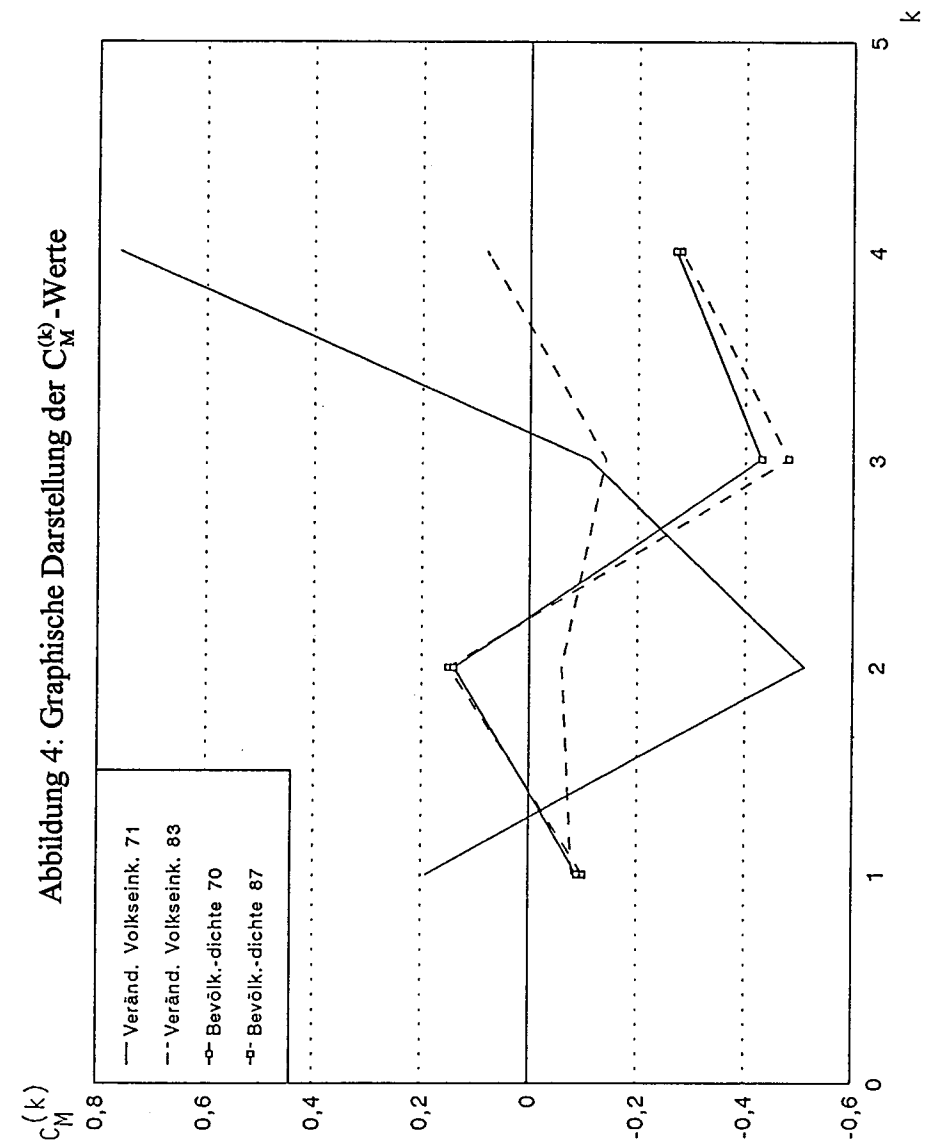
Es können Nachbarschaftskoeffizienten bis zur 4. Ordnung berechnet werden. In Tabelle 2 sind die R_k -Werte, die zur Berechnung von $C_M^{(k)}$ benötigt werden, angegeben. Diese $C_M^{(k)}$ -Werte zeigt Tabelle 3.

Tabelle 2: R_k -Werte

Land	1. Ordnung	2. Ordnung	3. Ordnung	4. Ordnung
1	2	3	3	1
2	5	3	1	-
3	3	6	-	-
4	5	4	-	-
5	4	2	3	-
6	3	3	3	-
7	2	3	4	-
8	1	3	2	3
9	2	3	3	1
10	1	4	3	1
$\Sigma = R_k$	$R_1 = 28$	$R_2 = 34$	$R_3 = 22$	$R_4 = 6$

Tabelle 3: $C_M^{(k)}$ -Werte

		$C_M^{(1)}$	$C_M^{(2)}$	$C_M^{(3)}$	$C_M^{(4)}$
Veränderungen des Volkseinkommens	1971	0,19	-0,51	-0,11	0,76
	1983	-0,08	-0,06	-0,14	0,08
Bevölkerungsdichte	1970	-0,09	0,14	-0,43	-0,27
	1987	-0,10	0,15	-0,48	-0,28



Die Interpretation der Werte in Tabelle 3 bzw. ihr Verlauf in der Grafik - vgl. Abb. 4 - läßt sich in dreierlei Hinsicht vornehmen:

1. Vergleich der RAK-Werte für die einzelnen Variablen mit zunehmendem k . Bei der Bevölkerungsdichte zeigt sich in beiden betrachteten Jahren z. B. eine gewisse positive Autokorrelation bei der Nachbarschaft 2. Ordnung, weil hier überwiegend Bundesländer mit gleichen Bevölkerungsdichten korreliert sind. Dagegen ist die RAK bei $k = 1$ nur gering ausgeprägt, da hier Bundesländer mit hohen und niedrigen Bevölkerungsdichten in gemischter Form benachbart sind. Bezüglich der Nachbarschaft 3. Ordnung gibt es dagegen die (negative) Korrelation von Regionen mit hoher Bevölkerungsdichte (Stadtstaaten) mit solchen niedriger Bevölkerungsdichte (Flächenstaaten). Bei der Änderung des Volkseinkommens ergibt sich 1971 bei $k = 2$ eine negative RAK, da hier Länder mit in diesem Jahr hohen Änderungsraten des Volkseinkommens (z.B. HB) mit solchen mit niedriger Änderungsraten korreliert sind (z. B. NRW, HE). Bei $k = 4$ ist eine ausgeprägte positive AK festzustellen, da hier vorwiegend Regionen mit hohen Änderungsraten miteinander verknüpft sind (SLH, SL, HB, HH).
2. Vergleich der RAK-Werte verschiedener Variablen bei einem bestimmten k . Werden mehrere Variablen verglichen, so kann man erkennen, ob sie ein ähnliches räumliches Muster für verschiedene k zeigen oder nicht. Bei den hier beispielhaft ausgewählten beiden Variablen ist eine Ähnlichkeit im räumlichen Muster nicht zu erkennen.
3. Intertemporaler Vergleich der RAK-Werte ein und derselben Variablen. Betrachtet man die Bevölkerungsdichte, so hat sich die RAK zwischen 1971 und 1987 praktisch nicht verändert, sie ist also bezüglich der beiden betrachteten Jahre zeitlich stabil. Bei der Änderung des Volkseinkommens zeigen sich in den Jahren 1971 und 1983 jedoch deutliche Unterschiede. Während sich 1971 der unter 1. interpretierte Verlauf zeigt, ergibt sich für 1983 praktisch bei keinem k eine auffallende RAK: Die Änderungsraten des Volkseinkommens haben sich in fast allen Regionen stark nivelliert.

Insgesamt auffallend ist, daß die RAK mit größerem k nicht zwangsläufig abnehmen, sondern teilweise größer werden. Dies kann daher rühren, daß R_k die Zahl der Regionen, die aufgrund der Ordnung k benachbart sind - oft sehr viel kleiner ist als R_k bei einem k niedrigerer Ordnung. Dadurch wird der Nenner von (5) kleiner und $C_M^{(k)}$ größer. Andererseits sollten die Ergebnisse als illustrierende Beispiele angesichts des geringen Beobachtungsumfangs nicht überbewertet werden.

4. RAK IM REGRESSIONSMODELL

Im vorigen Abschnitt wurde die RAK im Zusammenhang mit Stichprobenvariablen betrachtet. Es ist aber auch in vielen Fällen, in denen die Regressionsanalyse auf Querschnittsdaten verschiedener Raumeinheiten angewendet wird, mit Abhängigkeiten in der latenten Variablen e zu rechnen. Im Vergleich zur zeitlichen AK wird die RAK relativ selten untersucht, obwohl sie oft ernste Konsequenzen besitzt. Die RAK kann im Regressionsmodell zu verzerrten Schätzungen der Residualvarianz und ineffizienten Schätzungen der Regressions-Koeffizienten bei Verwendung der einfachen Kleinst-Quadrat-Methode führen. Deswegen sollte auf das Vorhandensein von RAK geprüft werden.

Die in Kap. 3 dargestellten Maße lassen sich dabei in analoger Form jetzt mit den Restgrößen e berechnen [vgl. Sen/Srivastava (1990) S. 143].

Haben sich Hinweise auf RAK ergeben, dann kann dies (1) ein Hinweis auf eine oder mehrere vernachlässigte erklärende Variablen in der Regressionsbeziehung sein, (2) eine Fehlspezifikation der Funktion oder (3) eine räumlich autoregressive Struktur vorliegen.

Gehen wir vom linearen Modell aus

$$(13) \quad y = x\beta + \varepsilon$$

mit y als $(n \times 1)$ -Vektor der abhängigen Variablen, x als $(n \times m)$ -Matrix der (nichtstochastischen) Regressoren vom Rang m , β als der $(m \times 1)$ -Vektor der zu schätzenden Regressionsparameter und e als $(n \times 1)$ -Vektor der latenten Variablen. Im übrigen sollen die üblichen Annahmen des klassischen Regressionsmodells gelten. Ein räumliches autoregressives Modell 1. Ordnung für den Vektor der latenten Variablen e ist

$$(14) \quad e = r W e + u$$

wobei $u \sim NV(0, \sigma^2 I)$.

Wie bei den RAK-Maßen in Kap. 3 bereits erwähnt, wird für W meist ein binäres Nachbarschaftsmuster unterstellt. Es ist aber zu bedenken, daß dies keinesfalls die einzige Spezifikationsmöglichkeit ist. W wird oft deshalb in der binären Form vorgegeben, weil es unmöglich ist, $n\beta$ -Koeffizienten, $n(n-1)$ Gewichtungsfaktoren w_{rs} , σ_e^2 und ρ aus n Beobachtungsdaten zu schätzen.

Die Gewichtungsmatrix $\mathbf{W}$ hat in diesem einfachen Fall die Form

$$\mathbf{W} = \begin{bmatrix} 0 & 1 & 0 & . & . & . & 0 \\ 1 & 0 & 1 & . & . & . & 0 \\ 0 & 1 & 0 & . & . & . & 0 \\ 0 & 0 & 1 & . & . & . & 0 \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ 0 & 0 & 0 & . & . & 1 & 0 \end{bmatrix}$$

Ist $\mathbf{I} - \rho \mathbf{W}$ nichtsingulär, so läßt sich (14) schreiben als

$$(15) \quad \varepsilon = (\mathbf{I} - \rho \mathbf{W})^{-1} \mathbf{u}$$

und daraus

$$(16) \quad \Omega = E(\varepsilon \varepsilon') = \sigma^2 (\mathbf{I} - \rho \mathbf{W})^{-1} [(\mathbf{I} - \rho \mathbf{W})^{-1}]'$$

Man sieht, daß in der Struktur von ε die Probleme begründet liegen, nämlich zum einen in der Schätzung von ρ und zum anderen in der Formulierung von $\mathbf{W}$.

Zunächst wird man untersuchen, ob sich ρ signifikant von Null unterscheidet. Um dies zu testen, benutzt man die Teststatistik (12) mit $H_0: \rho = 0$ und $H_1: \rho \neq 0$. Ausdruck (11) läßt sich jetzt schreiben als

$$(17) \quad c = \frac{R}{S_0} \frac{\mathbf{e}' \mathbf{W} \mathbf{e}}{\mathbf{e}' \mathbf{e}}$$

wobei $\mathbf{e}$ der Vektor der Residuen aus dem linearen Regressionsmodell (13) und

$$S_0 = \sum_{r=1}^R \sum_{s=1}^R w_{rs} \text{ ist; } a \text{ in (12) entspricht hier } R/S_0.$$

Für eine Gewichtungsmatrix, die so normiert ist, daß die Zeilenelemente in $\mathbf{W}$ sich zu Eins summieren, vereinfacht sich (17) zu

$$(18) \quad c = \frac{\mathbf{e}' \mathbf{W} \mathbf{e}}{\mathbf{e}' \mathbf{e}}.$$

Cliff/Ord [(1981) S. 201ff.] haben für (17) den Erwartungswert und die Varianz von c unter der Bedingung, daß H_0 : "Keine räumliche Abhängigkeit" gilt und $\varepsilon \sim NV(0, \sigma^2 \mathbf{I})$, abgeleitet. Es ist

$$(19) \quad E(c_M) = \frac{R \operatorname{sp}(\mathbf{A})}{(R - m) S_0}$$

wobei m die Zahl der erklärenden Variablen, $\operatorname{sp}(\mathbf{A})$ die Spur der Matrix $\mathbf{A}$, d.h. die Summe ihrer Elemente in der Diagonalen, und $\mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \mathbf{W} \mathbf{X}$. Die Varianz läßt sich angeben mit

$$(20) \quad \operatorname{Var}(c_M) = \frac{R^2}{S_0^2 (R - m)(R - m + 2)} \left\{ S_1 + 2 \operatorname{sp}(\mathbf{A}^2) - \operatorname{sp}(\mathbf{B}) - \frac{2 [\operatorname{sp}(\mathbf{A})]^2}{R - m} \right\}$$

mit der $m \times m$ -Matrix

$$\mathbf{B} = 4 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \mathbf{W}^2 \mathbf{X}.$$

c in Gleichung (17) ist asymptotisch normalverteilt. Dies gilt sowohl für die KQ-Residuen als auch für die - noch zu erwähnenden - BLUS-Residuen [vgl. Brandsma/Ketellapper (1979) S. 119ff.]. Wenn keine Untergruppe von Raumeinheiten die Struktur des Gesamttraumes dominiert, so stimmen die Enden der Dichtefunktion der Normalverteilung und von (17) im wesentlichen bereits ab $R > 10$ überein [Cliff/Ord (1972) S. 279]. [Ein Likelihood-Ratio-Test, der sich auf die RAK in den Residuen bezieht, findet sich bei Brandsma/Ketellapper (1979) S. 122ff. und Anselin (1988b) S. 103f.] Als Prüfgröße läßt sich dann das standardnormalverteilte Maß (12) benutzen, in das jetzt die Werte von (17) bzw. (18), (19) und (20) eingehen.

Eine allererste Schätzmöglichkeit im Regressionsmodell bestünde darin, auf (13) die KQ-Methode anzuwenden und dann ρ mit Hilfe der Residuen $\mathbf{e}$ in

(14) zu schätzen. Damit könnte $(I - \rho W)$ berechnet und β geschätzt werden. Ein Problem bei der Analyse der räumlichen Abhängigkeit in den Regressionsresiduen e ergibt sich allerdings daraus, daß diese nur unvollkommene Schätzungen der latenten Variablen ε sind. Führt der Test zur Annahme von H_0 , so könnte man mit der KQ-Methode befriedigende Ergebnisse erzielen. In aller Regel sind die KQ-Residuen aber korreliert. Sie besitzen nämlich keine skalare Kovarianz-Matrix, weshalb die Momente von c abhängig von der Matrix der Regressoren sind. Aus diesem Grund ist die KQ-Methode nur bedingt brauchbar, und es stellt sich die Frage nach alternativen Schätzprozeduren.

Es liegt z.B. nahe, Schätzungen mit einer skalaren Kovarianz-Matrix zu benutzen. Als Beispiel sei die BLUS [Best Linear Unbiased Scalar]-Schätzung genannt [vgl. Bartels/Hordijk (1977) S. 88f.; Cliff/Ord (1981) S. 203ff.].

Weiterhin könnte man die Verallgemeinerte Kleinst-Quadrat-Methode (VKQ) heranziehen, wenn die Kovarianz-Matrix (16) bekannt wäre. Dies würde bedeuten, daß die Struktur der räumlichen Abhängigkeit - also W - und der Koeffizient ρ numerisch bekannt sein müßte. Da dies i.d.R. nicht der Fall ist, erhält man einen Geschätzten-Verallgemeinerten-Kleinst-Quadrat-(GVKQ)-Koeffizienten-Vektor aus der Anwendung des VKQ-Prinzips mit konsistenten Schätzungen für die Parameter in (16) mit

$$(21) \quad b_{GVK} = [X' (I - \hat{\rho} W)^2 X]^{-1} X' (I - \rho W)^2 y$$

[vgl. Anselin (1988b) S. 109]. Die formalen Eigenschaften der GVKQ-Methode sind asymptotischer Natur. Bei kleinen Stichprobenumfängen sind die Eigenschaften nicht gesichert. Außerdem reagieren diese Schätzungen sensitiv im Hinblick auf die korrekte Spezifikation der Kovarianz-Matrix (16), was im räumlichen Modell in erster Linie die Wahl der Gewichte W betrifft [vgl. Anselin (1988b) S. 109] [Für den Sonderfall gleichkorrelierter Residuen läßt sich die Gleichheit der VKQ-mit der KQ-Methode nachweisen, vgl. Krämer/Donninger (1987) S. 577ff.].

Darüber hinaus werden Suchtechniken benutzt. Es läßt sich etwa daran denken, ρ mit der KQ-Methode in (14) zu schätzen, nachdem die Residuen aus (17) vorliegen, und dann - wie bei der Orcutt-Cochrane-Prozedur - die Originaldaten zu transformieren. Erneute Schätzung von (13) führt zu neuen Residuen. Eine Iteration bringt dann letztendlich Schätzwerte, die - nach dem Test - keine RAK mehr enthalten. Ein solches Verfahren wurde im räumlichen Kontext erstmals von Hordijk [(1974) S. 131ff.] vorgeschlagen. Diese KQ-Schätzung führt allerdings zu verzerrten und inkonsistenten Parameter-

schätzungen [vgl. Bodson/Peeters (1975) S. 460; die Autoren bevorzugen ein Vorgehen, das verschiedene Werte von $\hat{\rho}$ benutzt und dann denjenigen Wert wählt, der die Summe der quadrierten Residuen minimiert]. Weiterhin lassen sich verschiedene VKQ-Schätzungen für unterschiedliche ρ (z.B. zwischen 0 und 0,25) vornehmen. Dann wird diejenige Schätzung ausgewählt, die den kleinsten Wert von

$$(y - Xb\rho)' (I - \rho W)^{-2} (y - Xb\rho) - 2 \lg \{ \det [(I - \rho W)] \}$$

ergibt. Dies beruht auf dem ML-Prinzip [vgl. Ord (1975) S. 121]. Ebenso wird bei Hordijk/Paelinck [(1976) S. 180ff.] im Rahmen eines ML-Ansatzes eine Suchtechnik benutzt. Auch Hepple (1976) und Case (1991) verwenden ML-Ansätze, um ρ und β simultan zu bestimmen. Diese Maximum Likelihood-Schätzungen haben die üblichen asymptotischen Eigenschaften [Konsistenz, Effizienz]. Für kleine Stichproben gibt es keine gesicherten Ergebnisse.

Eine letzte Möglichkeit sei erwähnt: Aus (13) läßt sich schreiben $\varepsilon = y - x\beta$, was zusammen mit (14) zu

$$(22) \quad y = x\beta + \rho W y - \rho W x\beta + u$$

führt. Wendet man hierauf die KQ-Methode an, so sind die Schätzungen verzerrt und inkonsistent [vgl. Case (1991) S. 956]. Ein ML-Ansatz führt dagegen zu konsistenten Schätzungen [vgl. Anselin (1988b) S. 111].

Simulationsstudien von Bartels/Hordijk [(1977), S. 91ff.] und Brandsma/Ketellapper [(1979) S. 125ff.] vergleichen die Eigenschaften von Moran's Koeffizienten bei der Benutzung von Residuen aus verschiedenen Schätzmethoden, u.a. KQ- und ML-Schätzungen. Dabei hatte zwar der C_M -Test aufgrund der KQ-Residuen z.B. gegenüber solchen aus BLUS die größte Macht, ansonsten lassen sich aber kaum allgemeingültige Schlußfolgerungen ziehen.

ZUSAMMENFASSUNG

Bei der Benutzung von Daten aus verschiedenen Regionen im Rahmen von quantitativen Analysen ist immer mit RAK zu rechnen. Es ist deshalb sinnvoll, auf solche räumlichen Zusammenhänge zu prüfen. Trotz der Möglichkeit, dabei die unterschiedlichsten Gewichtungen zu berücksichtigen, werden meist

binäre Faktoren benutzt und der Koeffizient von Moran bevorzugt. Ein exakter Signifikanztest ist nur bei einer großen Zahl von Raumeinheiten möglich. Bei einer kleinen Zahl kann jedoch die Berechnung von C_M als deskriptives Maß auch sinnvoll sein, um Anhaltspunkte über die RAK zu erhalten.

Auch bei der Regressionsanalyse sollte zunächst auf mögliche RAK untersucht werden. Ist diese vernachlässigbar, dann kann das klassische Regressionsmodell mit der KQ-Methode als erste Annäherung benutzt werden. Ist RAK nachweisbar, so ist nach anderen Modellspezifikationen bzw. Schätzprozeduren zu suchen. Probleme bestehen hierbei in der Festlegung der Elemente der Gewichtungsmatrix W und der Schätzung von ρ . In der Literatur sind etliche Prozeduren vorgeschlagen worden, von denen einige skizziert wurden. Hierbei sind oft nur die asymptotischen Eigenschaften bekannt; für kleine Stichprobenumfänge liegen teilweise nur Simulationsergebnisse vor. Trotz fehlender theoretischer Eigenschaften wird oft der KQ-Methode der Vorzug gegeben; zum einen wegen des vertretbaren Rechenaufwandes und brauchbarer Ergebnisse - auch bei kleinen Stichproben, zum anderen weil keine der bisher entwickelten alternativen Spezifikationen und Schätzansätze in allen Fällen befriedigende Ergebnisse gezeigt haben. Deshalb wurde auch auf eine Beispielrechnung in diesem Überblick verzichtet.

Festzuhalten ist, daß eine Vernachlässigung der räumlichen Effekte in Form von RAK durch die ernsthaften Konsequenzen bei den Schätzungen zu Fehleinschätzungen im Regressionsmodell führen kann.

LITERATURVERZEICHNIS

- ANSELIN, L. (1980), Estimation Methods for Spatial Autoregressive Structures, A Study in Spatial Econometrics, Regional Science Dissertation & Monograph Nr. 8, Cornell University, Ithaca.
- ANSELIN, L. (1988a), Model Validation in Spatial Econometrics: A Review and Evaluation of Alternative Approaches, *International Regional Science Review* 11, S. 279 - 316.
- ANSELIN, L. (1988b), *Spatial Econometrics: Methods and Models*, Dordrecht usw.
- BARTELS, C.P.A./HORDIK, L. (1977), On the Power of the Generalized Moran Contiguity Coefficient in Testing for Spatial Autocorrelation among Regression Disturbances, *Regional Science and Urban Economics* 7, S. 83 - 101.
- BODSON, P./PEETERS, D. (1975), Estimation of the Coefficients of a Linear Regression in the Presence of Spatial Autocorrelation. An Application to a Belgian Labour-Demand Function, *Environment and Planning A* 7, S. 455 - 472.
- BRANDSMA, A.S./KETELLAPPER, R.H. (1979), Further Evidence on Alternative Procedures for Testing of Spatial Autocorrelation among Regression Residuals, in: Bartels, C.P.A., Ketellapper (eds.) *Exploratory and Explanatory Statistical Analysis of Spatial Data*, Boston usw., S. 113 - 135.
- CASE, A.C. (1991), Spatial Patterns in Household Demand, *Econometrica* 59, S. 953 - 965.
- CLIFF, A.D./ORD, J.K. (1969), The Problem of Spatial Autocorrelation, in: Scott, A.J. (ed.) *Studies in Regional Science*, London, S. 25 - 55.
- CLIFF, A.D./ORD, J.K. (1972), Testing for Spatial Autocorrelation among Regression Residuals, *Geographical Analysis* 4, S. 267 - 284.
- CLIFF, A.D./ORD, J.K. (1973), *Spatial Autocorrelation*, London.
- CLIFF, A.D./ORD, J.K. (1981), *Spatial Processes. Models & Applications*, London.
- DACEY, M.F. (1968), A Review on Measures of Contiguity for Two and K-Color Maps, in: BERRY, B.J.L., MARBLE, D.F. (eds.) *A Reader in Statistical Geography*, Englewood Cliffs, S. 479 - 495.
- GEARY, R.C. (1954), The Contiguity Ratio and Statistical Mapping, *The Incorporated Statistician* 5, S. 115 - 145.
- GEMEINSCHAFTSVERÖFFENTLICHUNG DER STATISTISCHEN LANDESÄMTER (Hrsg.) (1986), *Volkswirtschaftliche Gesamtrechnungen der Länder*, Heft 15, Stuttgart.
- HEPPLE, L.W. (1976), A Maximum Likelihood Model for Econometric Estimation with Spatial Series, in: MASSER, I. (ed.) *Theory and Practice in Regional Science* London, S. 90 - 104.
- HORDIK, L. (1974), Spatial Correlation in the Disturbances of a Linear Interregional Model, *Regional and Urban Economics* 4, S. 117 - 140.
- HORDIK, L. (1979), Problems in Estimating Econometric Relations in Space, *Papers of the Regional Science Association* 42, S. 99 - 115.
- HORDIK, L./PAELINCK, J. (1976), Some Principles and Results in Spatial Econometrics, *Recherches Economiques de Louvain* 42, S. 175 - 197.
- KRÄMER, W./DONNINGER, CHR. (1987), Spatial Autocorrelation Among Errors and the Relative Efficiency of OLS in the Linear Regression Model, *Journal of the American Statistical Association* 82, S. 577 - 579.
- MORAN, P.A.P. (1950), Notes on Continuous Stochastic Phenomena, *Biometrika* 3, S. 17 - 23.

- ORD, J.K. (1975), Estimation Methods for Models of Spatial Interaction, Journal of the American Statistical Association 70, S. 120 - 126.
- PAELINCK, J.H.P./NIJKAMP, P. (um 1975), Operational Theory and Method in Regional Economics, Westmead, Farnborough/ Lexington (Mass.) o.J.
- SEN, A./SRIVASTAVA, M. (1990), Regression Analysis. Theory, Methods and Application, New York usw.
- STATISTISCHES BUNDESAMT (Hrsg.) (1990), Statistisches Jahrbuch für die Bundesrepublik Deutschland, Wiesbaden.