Pareto-Einkommensverteilung

GERHARD KOCKLÄUNER, KIEL

Zusammenfassung: *Nachfolgend wird die Einkommensverteilung in der Bundesrepublik Deutschland über eine Pareto-Verteilung modelliert. Die Pareto-Verteilung wird in Theorie und Empirie präsentiert. Es zeigen sich einfache Darstellungen von zugehöriger Lorenz-Funktion und davon abhängigem Gini-Koeffizienten. Die Modellanpassung erweist sich als gut, die Einkommenskonzentration kann im Vergleich zur Vermögenskonzentration noch als moderat beschrieben werden.*

1 Einleitung

Das Einkommen, speziell das Haushaltsnettoeinkommen, ist in allen Ländern ungleich verteilt, wobei sich die Einkommenskonzentration durchaus von Land zu Land unterscheidet (vgl. z. B. UNDP 2010, S. 186 ff). Einen kondensierten Überblick hinsichtlich des jeweiligen Ausmaßes an Konzentration bzw. Ungleichheit liefern, theoretisch wie empirisch, die Lorenz-Funktion und der Gini-Koeffizient.

Die Modellierung von Einkommensverteilungen erfolgt traditionell durch eine Pareto-Verteilung oder eine logarithmische Normalverteilung. Eine solche Modellierung soll nachfolgend mit bundesdeutschen Daten für das Jahr 2010 vorgenommen werden. Nach der Vorstellung der betreffenden Einkommensdaten wird, weil nur durch einen einzigen Parameter gekennzeichnet, die Pareto-Verteilung mit ihren speziellen Darstellungen für die Lorenz-Funktion und den Gini-Koeffizienten in Theorie und Empirie präsentiert. Das Ergebnis zeigt neben einem im Vergleich zur Vermögenskonzentration noch moderaten Ausmaß an Einkommenskonzentration eine gute Modellanpassung.

2 Daten

Die Bundeszentrale für politische Bildung (bpb) veröffentlichte am 27.9.2013 die Angaben der sich auf Haushaltsnettoeinkommen beziehenden Tabelle 1. Haushaltsnettoeinkommen ergeben sich aus den Haushaltsbruttoeinkommen, d. h. den gesamten Einnahmen aller Mitglieder eines Haushaltes aus Erwerbstätigkeit, Vermögen und eventuellen Transferzahlungen, indem davon sämtliche Steuern sowie die Pflichtbeiträge zur Sozialversicherung abgezogen werden.

Rundungsbedingt ergeben die in Tabelle 1 aufgeführten Dezilanteile in der Summe nicht exakt 100 %. Die 10 % Haushalte mit den niedrigsten Einkommen erhalten nach Tabelle 1 nur einen Anteil von 3,7 % an der Summe aller Haushaltsnettoeinkommen, die 10 % Haushalte mit den höchsten Einkommen verfügen dagegen über 23,1 % dieser Summe.

Die vorgenommene Anteilsberechnung beruht auf Ergebnissen der 28. Befragung (Welle) des sozioökonomischen Panels (SOEPv28, Personen in Privathaushalten), durchgeführt vom Deutschen Institut für Wirtschaftsforschung (DIW). Das sozioökonomische Panel ist eine jährlich erfolgende repräsentative Wiederholungsbefragung von über 12000 Privathaushalten. Das weite Themenspektrum reicht dabei von der Demographie über Einkommen und gegebenenfalls auch Vermögen bis hin zur Bildung. Die hier betrachteten einzelnen Haushaltsnettoeinkommen sind dabei gemäß OECD-Vorgaben äquivalenzgewichtet, d. h. in jedem Haushalt bekommt der erste Erwachsene das Gewicht 1, weitere Erwachsene sowie Kinder ab 14 Jahren das Gewicht 0,5, Kinder unter 14 Jahren das Gewicht 0,3. So ergibt sich für einen 4-Personen-Haushalt bei zwei Erwachsenen und zwei Kindern, beide unter 14 Jahren, die Gewichtssumme 2,1. Liegt in diesem Haushalt nun das monatliche Nettoeinkommen bei z. B. 2100 €, wird dieser Haushalt so bewertet, als ob alle Mitglieder über ein monatliches Nettoeinkommen von jeweils 1000 € verfügten.

Dezil	1	2	3	4	5	6	7	8	9	10
Anteil	3,7	5,4	6,5	7,4	8,3	9,3	10,4	11,9	14,2	23,1

Tab. 1: Dezilanteile für das Haushaltsnettoeinkommen der Bundesrepublik Deutschland im Jahr 2010
www.bpb.de/nachschlagen/zahlen-und-fakten/soziale-situation-in-deutschland/61769/einkommensverteilung

3 Pareto-Verteilung: Theorie

Die Pareto-Verteilung ist eine Wahrscheinlichkeitsverteilung für stetige Zufallsvariablen und wird traditionell zur Modellierung von Einkommensverteilungen genutzt. Für Y als Haushaltsnettoeinkommen und y_0 als kleinstes (positives) Haushaltsnettoeinkommen ist die Verteilungsfunktion F der Pareto-Verteilung durch

$$F(y) = 1 - \left(\frac{y}{y_0}\right)^{-k} \text{ für } k > 2 \text{ und } y \geq y_0 \tag{1}$$

definiert, die Dichtefunktion f als erste Ableitung von F damit durch

$$f(y) = \frac{k}{y_0}\left(\frac{y}{y_0}\right)^{-k-1} \text{ für } k > 2 \text{ und } y \geq y_0 \tag{2}$$

(Mood et al. 1974, S. 118). Gleichung (1) kennzeichnet Wahrscheinlichkeiten $P(Y \leq y)$ für Einkommen Y von höchstens y. Da $P(y - 1 \leq Y \leq y)$ mit steigendem y sinkt, zeigen die Verteilungsfunktion F und die Dichtefunktion f ein für Einkommensverteilungen typisches Bild. So treten empirisch höhere Einkommen seltener als niedrige auf, was sich nach Gleichung (2) in einem monoton fallenden Verlauf der Dichtefunktion f widerspiegelt.

Bei $k > 2$ als Vorgabe für den konstanten Parameter k existieren der Erwartungswert $E(Y)$ und die Varianz $V(Y)$, liegen diese doch bei

$$E(Y) = \int_{y_0}^{\infty} x f(x)dx = \frac{k y_0}{k-1} \text{ und}$$

$$V(Y) = \frac{k y_0^2}{k-2} - E(Y)^2 \tag{3}$$

(Mood et al. 1974, S. 118). Der Erwartungswert aus Gleichung (3) ist Bestandteil der nach Lorenz (1905) benannten Funktion L, die eine Veranschaulichung vorhandener Einkommenskonzentration ermöglicht. Die Funktionswerte von L sind allgemein als – mit dem minimalen Einkommen beginnend – kumulierte Anteile von $E(Y)$ (vgl. Gleichung (3)) definiert. D. h. für die Stelle $F(y)$, dass

$$L(F(y)) = \frac{1}{E(Y)} \int_{y_0}^{y} x f(x)dx \text{ für } y \geq y_0 \tag{4}$$

(Lambert 2001, S. 32).

Mit $f(y)$ aus Gleichung (2) lässt sich $L(F(y))$ für eine Pareto-Verteilung konkretisieren. Speziell ergibt sich gemäß Gleichung (3) und Gleichung (4), aber analog auch für zur Pareto-Verteilung alternative Verteilungen, $L(F(y_0)) = 0$ und $L(F(\infty)) = 1$. Jede Lorenz-Funktion L ist damit selbst eine Verteilungsfunktion.

Bei $F(y) = p$, F^{-1} als Umkehrfunktion von F und somit $F^{-1}(p) = y$ kann Gleichung (4) aber nach Substitution auch als

$$L(p) = \frac{1}{E(Y)} \int_0^p F^{-1}(q)dq \tag{5}$$

geschrieben werden.

Wird nun nach Gleichung (1)

$$F^{-1}(p) = y_0(1-p)^{-\frac{1}{k}} \text{ für } 0 \leq p \leq 1 \tag{6}$$

bestimmt und Gleichung (6) in Gleichung (5) eingesetzt, ergibt eine einfache Integralrechnung für $L(0) = 0$ und $L(1) = 1$ die von y_0 unabhängige Lorenz-Funktionsgleichung der Pareto-Verteilung

$$L(p) = 1 - (1-p)^{1-\frac{1}{k}} \text{ für } 0 \leq p \leq 1. \tag{7}$$

Gleichung (7) zeigt die Lorenz-Funktion der Pareto-Verteilung als einfach strukturierte Verteilungsfunktion. Gleichung (7) zeigt auch die allgemein für Lorenz-Funktionen gültige Ungleichung $L(p) \leq p$. D. h. eine Lorenz-Funktion kann mit ihren Funktionswerten die Werte auf einer Winkelhalbierenden nicht überschreiten. $L(p) = p$ für alle p findet sich im für Einkommensverteilungen unrealistischen Fall einer Ein-Punkt-Verteilung der Variable Y. Die Ungleichung $L(p) < p$ ergibt sich für $0 < p < 1$, wenn die Variable Y wie im Falle realer Einkommensverteilungen unterschiedliche Werte annehmen kann. Ungleichheit stellt damit eine Voraussetzung für vorhandene Konzentration dar.

Die Funktion L aus Gleichung (7) weist zudem bei L' bzw. L'' als erster bzw. zweiter Ableitung wegen $L'(p) \geq 0$ sowie $L''(p) \geq 0$ den für alle Lorenz-Funktionen vorhandenen konvexen Verlauf auf. Für die Pareto-Verteilung gilt speziell

$$L'(p) = \frac{\left(1 - \frac{1}{k}\right)(1 - L(p))}{1-p} \text{ für } p < 1. \tag{8}$$

Aus Gleichung (8) ergibt sich folgende Interpretation des Parameters k der Pareto-Verteilung: Bei $F(y) = p$ ist $1 - L(p)$ der Anteil von $E(Y)$, der auf den Anteil $1 - p$ von Einkommen größer als y entfällt. Das für diesen Anteil zu erwartende Einkommen liegt also für alle $p < 1$ bei $\frac{(1 - L(p))}{1-p}$. Nach Gleichung (8) beträgt die Steigung der Lorenz-Funktion an der Stelle p nun gerade das $\left(1 - \frac{1}{k}\right)$-fache dieses zu erwartenden Einkommens (vgl. eine ausführliche Diskussion der (Differenzial)gleichung (8) bei Kämpke et al. 2003).

Mit Hilfe der Lorenz-Funktion kann nun das Ausmaß relativer Konzentration bzw. Ungleichheit für die Verteilung von Y, also hier speziell des Haushaltsnettoeinkommens, auch quantitativ bestimmt werden. Der für die Konzentrations- bzw. Ungleichheitsmessung üblicherweise genutzte Gini-Koeffizient G ist allgemein als

$$G = 2\int_0^1 (p - L(p))dp = 1 - 2\int_0^1 L(p)dp \qquad (9)$$

definiert (Lambert 2001, S. 33). G erfasst als geometrisches Konzentrationsmaß in einem (p, L)-Koordinatensystem das Zweifache des Flächeninhalts zwischen $L(p) = p$, d. h. dem Verlauf der Lorenz-Funktion im Fall ohne Konzentration, und der in der Regel vorhandene Konzentration ausweisenden Lorenz-Funktion L (vgl. Gleichung (5)).

Einsetzen von Gleichung (7) in Gleichung (9) und wiederum einfache Integralrechnung liefern für die Pareto-Verteilung

$$G = \frac{1}{2k - 1}. \qquad (10)$$

Nach Gleichung (10) hängt die Einkommenskonzentration gemäß Pareto ausschließlich vom Verteilungsparameter k ab.

4 Pareto-Verteilung: Empirie

Ein Vergleich von Kapitel 2 mit Kapitel 3 zeigt, dass in Tabelle 1 die für eine empirische Bestimmung des Parameters k einer Pareto-Verteilung erforderlichen Daten vorliegen. Insbesondere sind die dort angegebenen Dezilanteile, in kumulierter Form, als kumulierte Anteile an der Summe aller Haushaltsnettoeinkommen auch kumulierte Anteile am arithmetischen Mittel der betrachteten Haushaltsnettoeinkommen. Sie bilden damit das empirische Gegenstück zu den Funktionswerten einer Lorenz-Funktion.

Konkret bietet Tabelle 1 mit (p_i, L_i), $i = 1, \ldots, l = 10$ und $p_i = 0{,}1i$ sowie L_i als kumulierten Dezilanteilen, also speziell $L_1 = 0{,}037$, $L_2 = 0{,}037 + 0{,}054 = 0{,}091$ usw. und rundungsbedingt $L_{10} = 1{,}002$ die Grundlage für einen Datensatz zur regressionsanalytischen Berechnung des Verteilungsparameters k. Ergänzt um den Punkt $(p_0 = 0, L_0 = 0)$, ist dieser Datensatz in der Abbildung 1 dargestellt. Diese zeigt mit den Variablenbezeichnungen p und l einen für vorhandene Konzentration bzw. Ungleichheit charakteristischen konvexen Verlauf. Werden die einzelnen Punkte aus Abbildung 1 linear verbunden, ergibt sich der stückweise lineare Verlauf einer empirischen Lorenz-Funktion. Dieser findet sich als unterer Funktionsverlauf in Abbildung 3 unten.

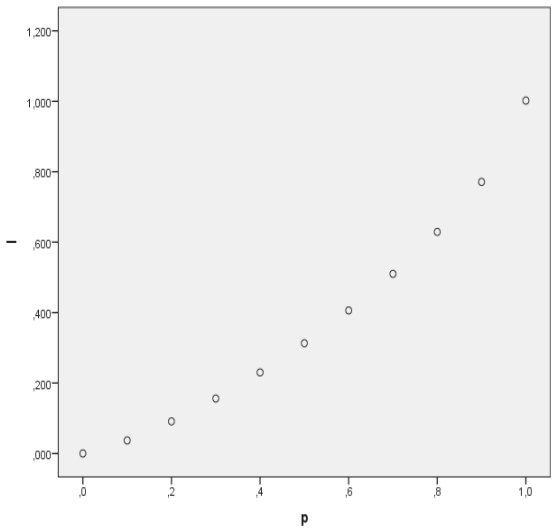

Abb. 1: Kumulierte Dezilanteile (Daten)

Im Rahmen einer Regressionsanalyse ist nun speziell die im Parameter k nichtlineare Lorenz-Funktion L aus Gleichung (7) mit dem beschriebenen Datensatz zu konfrontieren. Eine nichtlineare Regression wird dafür aber nicht benötigt. Gleichung (7) kann linearisiert werden und zeigt sich in linearisierter Form als

$$\ln(1 - L(p)) = \left(1 - \frac{1}{k}\right)\ln(1 - p). \qquad (11)$$

Gleichung (11) weist $\ln(1 - p)$ als unabhängige und $\ln(1 - L(p))$ als abhängige Variable aus. Dem entsprechen auf der Datenebene die Wertepaare $(\ln(1 - p_i), \ln(1 - L_i))$, $i = 1, \ldots, n = 9$. Der Fall $l = 10$, also $(p_{10} = 1, L_{10} = 1{,}002)$, muss offensichtlich ausgeschlossen werden. Ergänzt um den weiteren Punkt $(\ln(1 - p_0), \ln(1 - L_0))$ sind diese Wertepaare in Abbildung 2 grafisch dargestellt.

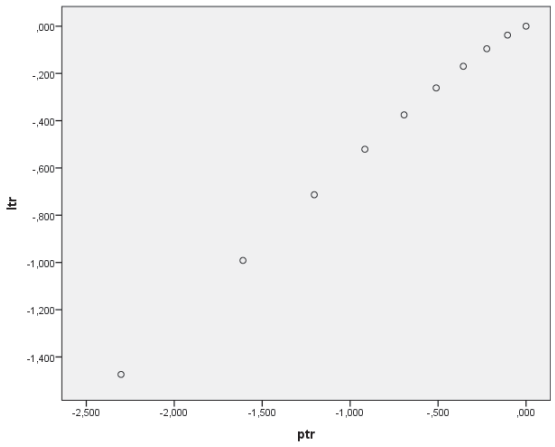

Abb. 2: Transformierte kumulierte Dezilanteile (Daten)

In Abbildung 2 stehen die Bezeichnungen ltr und ptr für die transformierten Variablen $\ln(1 - L(p))$ bzw. $\ln(1 - p)$. Im Gegensatz zu Abbildung 1 bietet Abbil-

dung 2 aber nun ein Punkt-Streudiagramm, zu dessen modellbezogener Anpassung ein linearer Ansatz für weite Bereiche (bis auf Punkte in der Nähe des Nullpunkts) angemessen erscheint (vgl. die Anpassungsdiskussion unten).

So kann für Gleichung (11) eine homogene lineare Regressionsanalyse mit $1 - \frac{1}{k}$ als einzigem Regressionskoeffizienten erfolgen. Um k zu bestimmen, ist für die gegebenen Daten die Summe von Quadraten der Residuen $\ln(1 - L_i) - \left(1 - \frac{1}{k}\right)\ln(1 - p_i)$ $i = 1, \ldots, n = 9$ bezüglich $1 - \frac{1}{k}$ zu minimieren.

Wird eine solche Minimierung, z. B. mit einem Standard-Softwarepaket wie Excel oder SPSS durchgeführt, ergibt sich für die beschriebenen Daten $1 - \frac{1}{k} = 0{,}613$ und damit $k = 2{,}584$.

Einsetzen des ermittelten k-Wertes in Gleichung (10) führt auf $G = 0{,}240 \leq 0{,}3$ und damit ein im Vergleich zu anderen Nationen und zur Vermögenskonzentration noch moderates Ausmaß an Einkommenskonzentration (vgl. unten). Die Korrelation zwischen gegebenen und mit dem berechneten Wert von k gemäß Gleichung (11) modellierten Werten der Variable $\ln(1 - L(p))$, also zwischen $\ln(1 - L_i)$ und $0{,}613\ln(1 - p_i)$ für $i = 1, \ldots, n = 9$, liegt zudem bei 0,999 und deutet auf eine hervorragende Modellanpassung hin. Entsprechend sollten auch die über Gleichung (7) modellierten kumulierten Dezilanteile nahe bei denjenigen aus dem Datensatz liegen. Die folgende Abbildung 3 zeigt – jeweils mit linearer Verbindung und ergänzt um Randwerte – den Vergleich dieser Anteile.

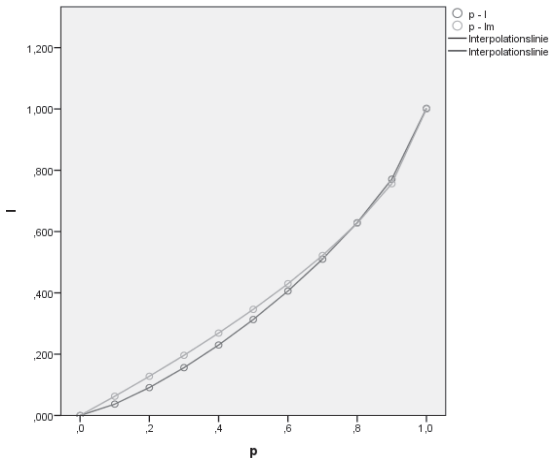

Abb. 3: Kumulierte Dezilanteile (Daten und Modellierung)

Wie Abbildung 3 aber verdeutlicht, liegen die modellierten kumulierten Dezilanteile (vgl. die Bezeichnung (p, lm) mit dem Buchstaben m zur Modellkennzeichnung) in weiten Bereichen oberhalb der sich aus den Daten ergebenden (vgl. (p, l)). Der Eindruck einer ausgezeichneten Modellanpassung wird dadurch relativiert. Er wird entsprechend auch durch einen Vergleich des nach Gleichung (10) berechneten Wertes von G mit dem Wert $G = 0{,}271$ getrübt, der sich mit den beschriebenen Daten für das empirische Gegenstück zu Gleichung (9), d. h. für

$$G = 1 - \sum_1^l (L_i + L_{i-1})(p_i - p_{i-1}) \qquad (12)$$

ergibt (vgl. Lambert 2001, S. 27). In Gleichung (12) ist für $i - 1 = 0$ ($p_0 = 0$, $L_0 = 0$) zu setzen, im Beispiel daneben rundungsbedingt $L_{10} = 1{,}002$. Wie Gleichung (12) zeigt, ergibt sich der empirische Gini-Koeffizient, indem von Eins das Zweifache des Flächeninhalts zwischen der empirischen Lorenz-Kurve und der p-Achse subtrahiert wird (vgl. die empirische Lorenz-Funktion in Abbildung 3).

$G = 0{,}271$ liegt nahe bei dem in UNDP (2010, S. 186) mit von Tabelle 1 abweichenden Daten für 2010 und Deutschland ausgewiesenen Wert von $G = 0{,}283$; vergleiche dagegen ebendort für 2010 und Großbritannien $G = 0{,}360$ sowie für die USA im selben Jahr $G = 0{,}408$. Wie Spannagel & Seils (2014, S. 622) dokumentieren, hat sich das Ausmaß der Konzentration der Haushaltsnettoeinkommen in Deutschland im Zeitablauf verändert. Es ist von 1991, damals lag der Gini-Koeffizient bei $G = 0{,}25$, bis 2004 mehr oder weniger kontinuierlich auf das Niveau von $G = 0{,}29$ angestiegen, danach auf das hier dokumentierte Niveau gesunken. Der in diesem Beitrag untersuchten Einkommenskonzentration steht mit $G = 0{,}78$ für das Jahr 2012 in Deutschland aber ein entschieden größeres Ausmaß an Vermögenskonzentration bei den Nettovermögen gegenüber (vgl. Grabka & Westermeier 2014). Deutschland weist damit im internationalen Vergleich ein ausgesprochen hohes Maß an Vermögensungleichheit auf.

Die Abweichung zwischen den beiden hier für die Einkommenskonzentration in Deutschland gefundenen Werten des Gini-Koeffizienten erklärt sich nach einer Residuenanalyse wie folgt: Bei homogenen Regressionen muss die Residuensumme nicht notwendig Null betragen. So findet sich für die Residuen $\ln(1 - L_i) - 0{,}613\ln(1 - p_i)$, $i = 1, \ldots, n = 9$, der abhängigen Variable $\ln(1 - L(p))$ aus Gleichung (11): Ihre Mehrheit und auch die Summe ist positiv. Für die Residuen der abhängigen Variable $L(p)$ aus Gleichung (7) gilt stattdessen: Wie bereits Abbildung 3 zeigt, sind die dort ersichtlichen Residuen $L_i - (1 - (1 - p_i)^{0{,}613})$, $i = 1, \ldots, n = 9$ mehrheitlich und auch in der Summe negativ. Letzteres bedeutet,

dass die modellierten L_i-Werte die nach Tabelle 1 ermittelten tatsächlichen L_i-Werte in der Regel, im Beispiel für die ersten sieben Dezile, überschreiten. Damit fällt das über den Gini-Koeffizienten erfasste Ausmaß an Konzentration im gemäß Pareto modellierten Fall niedriger aus als im tatsächlichen. Ein genaueres Bild über die Residuenstruktur der Variable $L(p)$ liefert die folgende Abbildung 4.

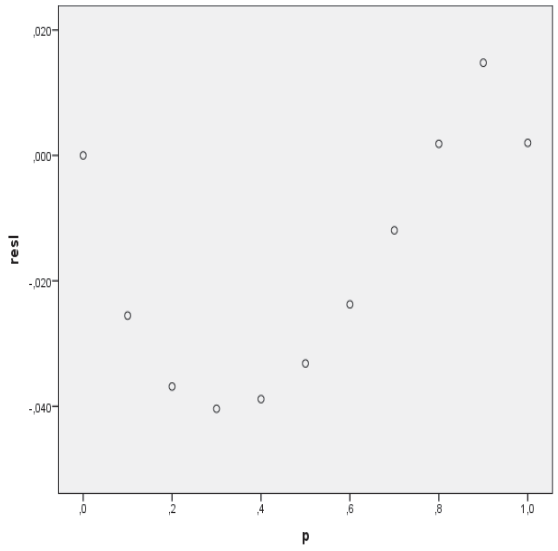

Abb. 4: Residuen bezüglich $L(p)$

In Abbildung 4 steht die Bezeichnung resl für die Residuenvariable bezüglich $L(p)$. Wie Abbildung 4 zeigt, weisen die dargestellten Residuen keine zufallsbehaftete, sondern eine deutlich nichtlineare Struktur auf. Derartige Strukturen gelten in der Regressionsanalyse als Indiz für eine Fehlspezifikation.

Soll also auch für den Gini-Koeffizienten eine bessere Modellanpassung erfolgen, sind Alternativen zur hier vorgestellten Pareto-Einkommensverteilung zu betrachten.

Einen Überblick bezüglich entsprechender Ansätze, speziell was mehrparametrige Modellierungen von Lorenz-Kurven angeht, gibt Chotikapanich (2008). Darunter findet sich dann auch der Ansatz von Singh & Maddala (1976), der eine dreiparametrige Erweiterung des Pareto-Ansatzes darstellend, gegenüber diesem insbesondere für die ersten Dezile bessere Anpassungen liefern kann.

Literatur

Chotikapanich, D. (Ed.) (2008): Modeling Income Distributions and Lorenz Curves. New York: Springer.

Grabka, M. M.; Westermeier, Ch. (2014): Anhaltend hohe Vermögensungleichheit in Deutschland. DIW-Wochenbericht Nr. 9.2014.

Kämpke, T.; Pestel, R.; Radermacher, F. J. (2003): A Computational Concept for Normative Equity. In: *European Journal of Law and Economics* Vol. 15, S. 129–163.

Lambert, P. J. (2001): The Distribution and Redistribution of Income. Manchester: Manchester University Press.

Mood, A. et. al. (1974): Introduction to the Theory of Statistics. Tokyo: McGraw-Hill.

Singh, S. K.; G. S. Maddala (1976): A Function für the Size Distribution of Incomes. In: *Econometrica* Vol. 44, S. 963–970.

Spannagel, D.; Seils, E. (2014): Armut in Deutschland wächst – Reichtum auch. WSI-Verteilungsbericht 2014. In: *WSI-Mitteilungen*, S. 620–627.

UNDP (2010): Bericht über die menschliche Entwicklung 2010. Berlin: Deutsche Gesellschaft für die Vereinten Nationen.

Anschrift des Verfassers

Gerhard Kockläuner
FB Wirtschaft
FH Kiel
Sokratesplatz 2
24149 Kiel
gerhard.kocklaeuner@fh-kiel.de