Modellierung räumlicher Interaktion mit Neuronalen Netzen

Marion Czeranka und Adrian Trapletti

Zusammenfassung

Räumliche Interaktion wird in klassischen Gravitationsmodellen als multiplikativer Zusammenhang zwischen den Modellgrößen abgebildet. Allerdings ist davon auszugehen, daß diese Annahme vereinfachend ist und daher mit der Realität nicht optimal übereinstimmt. Aufgrund der mittlerweile erheblich größeren Rechenleistung im PC-Bereich und aufbauend auf neuen Arbeiten im Bereich der Neuronalen Netze (NN) lassen sich Modelle entwickeln, welche die realweltlichen Zusammenhänge besser abbilden und bessere Prognosen zulassen, als die klassischen Modelle. Allerdings ist die Bestimmung des adäquaten Modells sowie die Schätzung der Modellparameter nicht trivial. In dieser Arbeit wird ein solcher NN-Modellansatz für die Interaktionsmodellierung vorgestellt. Untersuchungsgegenstand sind die Pendlerbeziehungen zwischen österreichischen Gemeinden (Auszug aus der Berufspendlererhebung, Volkszählung 1991). Es kann gezeigt werden, daß der tatsächlich beobachtete Zusammenhang bei den Pendlerdaten signifikant von dem mit dem klassischen Modell beschreibbaren Zusammenhang abweicht. Abschließend werden Möglichkeiten zur inhaltlichen Interpretation der Modellergebnisse vorgestellt.

1 Einleitung

Der Begriff "Räumliche Interaktion" beschreibt Wechselbeziehungen zwischen Objekten im Raum. Das Instrumentarium zur quantitativen Modellierung von Wechselbeziehungen wird von räumlichen Interaktionsmodellen zur Verfügung gestellt. Mit diesen Modellen lassen sich Verkehrsflüsse verschiedenster Art beschreiben: die Anzahl von Nachrichtenübertragungen kann grundsätzlich ebenso modelliert werden, wie Verkehrsströme von PKWs oder von Frachtgütern (vgl. ORTÚZAR und WILLUMSEN, 1990). Motivationen für die Erstellung von Interaktionsmodellen sind vielfältig; als Stichworte seien hier Verkehrsplanung, Infrastrukturausbau, Wirtschaftsförderung und Umweltschutz genannt. Unabdingbar notwendig ist allerdings die Verfügbarkeit entsprechenden empirischen Datenmaterials: zur Modellierung des Straßenverkehrs stammen Daten entweder aus Verkehrszählungen an fixen Zählstellen oder aus Mobilitätserhebungen (Erfassung der dynamischen Beziehungen zwischen Quell- und Zielorten). Nur diese letzteren Daten sind in Interaktionsmodellen verwendbar, da nur hier Quellen und Ziele der Interaktion festgehalten sind.
Klassische Interaktionsmodelle gehen von einer multiplikativen Verknüpfung der Modellvariablen aus (FOTHERINGHAM und O'KELLY, 1989). Im Gegensatz dazu wird in dieser Arbeit ein Ansatz vorgestellt, welcher nicht von einem a priori angenommenen Zusammenhang ausgeht, sondern welcher sich einzig vom empirischen Datenmaterial leiten läßt: Neuronale Netze (NN) dienen in dieser Arbeit als Tool, den tatsächlichen Zusammenhang der Daten zu

modellieren. Eine Möglichkeit zum Einsatz von NN zur Interaktionsmodellierung wird z.B. bei FISCHER, GOPAL (1994) beschrieben.
Im Kapitel 2 werden die beobachteten Pendlerdaten charakterisiert und deren statistische Kennwerte vorgestellt. Im Kapitel 3 wird das klassische Interaktionsmodell als Benchmark beschrieben, während im Kapitel 4 das NN-Modell hergeleitet wird. In beiden Kapiteln wird jeweils auf Methodologie und Ergebnisse eingegangen. Im Kapitel 5 werden die Modellergebnisse miteinander verglichen und darüber hinaus werden Analyseansätze zur Interpretation der NN-Ergebnisse gegeben. Kapitel 6 schließt mit Weiterführungsmöglichkeiten der hier gezeigten Arbeiten.

2 Pendlerbeziehungen zwischen den Gemeinden Österreichs

Diese Arbeit verfolgt das Ziel, die Pendlerbeziehungen zwischen den österreichischen Gemeinden zu modellieren. Allerdings geht es dabei nicht in erster Linie um eine fachliche Interpretation der Pendlerdaten in Bezug auf Wirtschaftsfaktoren, Infrastrukturbeschaffenheiten, Umweltschutzaspekte oder Raumplanung, sondern es geht hier vor allem darum, eine Methode vorzustellen, die für Prognosen und Szenarien gute Ergebnisse bereitstellt.
Die hier verwendeten Interaktionsdaten sind ein Auszug aus der österreichischen Volkszählung von 1991: im folgenden wird die Berufspendlermatrix zwischen 2377 Quellen (2354 Gemeinden und die 23 Bezirke Wiens) und 222 ausgewählten Zielen betrachtet. Letztere Ziele sind alle Gemeinden sowie die Bezirke Wiens mit einer Bevölkerungsanzahl größer als 5000 Einwohner bzw. mit einem Einpendleraufkommen von mindestens 1000. Zur Charakterisierung der Quellen und Ziele wird pragmatisch die jeweilige Bevölkerungsanzahl herangezogen. Als Merkmal der räumlichen Separation werden die Luftlinienentfernungen verwendet.
Die Ziele der Pendlermatrix sind in Abbildung 1 dargestellt: ersichtlich wird, daß diese Ziele unregelmäßig verteilt liegen, teils gehäuft auftreten und durch die räumlichen Eigenschaften Österreichs vielfach periphere Lagen einnehmen. Dies läßt ungleichmäßige Pendlerflüsse für die Gesamtheit der Gemeinden erwarten. Außerdem treten natürliche Barrieren auf (einige Höhenlinien sind skizziert), welche die Pendlerbewegungen und –mengen sicherlich ebenfalls beeinflussen. Es existieren sehr viele Gemeinden, von denen nur wenig bis gar keine Flüsse ausgehen: die Hälfte aller Pendlerbewegungen pro Gemeinde liegt bei nur 1 bis maximal 3 Personen (vgl. Tabelle 1). Auch finden sehr viele kurze Pendlerbewegungen statt: 25 % der gesamten Pendlerbewegungen gehen über maximal 20 km Luftlinienentfernung; 50 % gehen bis maximal 35 km. Die maximale Distanz, über welche ein Pendlerfluß stattfindet, liegt bei 191 km Luftlinienentfernung (vgl. Tabelle 1 sowie Abbildung 1).
Insgesamt weist die Pendlermatrix (2.377 x 222) - 2.377 = 525.317 interregionale Pendlerflüsse auf, von denen allerdings ein Großteil Null beträgt (d.h.: kein einziger Pendler zwischen jeweiliger Quelle und Ziel). Somit verbleiben positive 37.880 Pendlerflüsse, die als Beobachtungen dem Modell zur Parameterschätzung zur Verfügung gestellt werden.

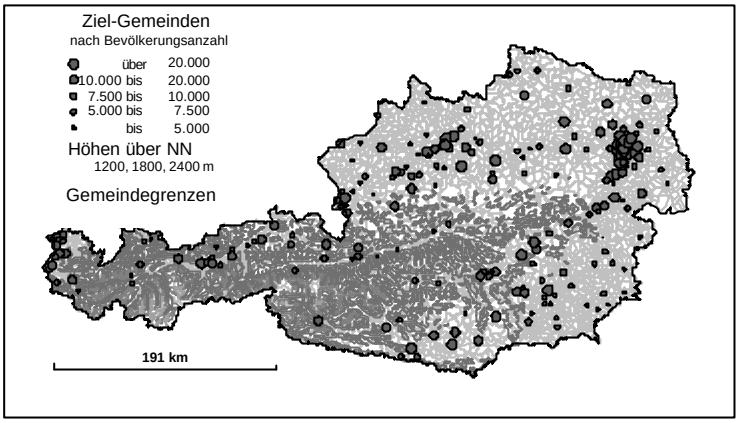

Abb. 1: Übersicht über die Ziele der verwendeten Pendlerdaten

Tatsächlich wäre es durchaus interessant, ebenfalls die Pendlerflüsse von Null zu analysieren. Mit dem klassischen Modell ist dies allerdings gar nicht möglich, da dort die Parameterschätzung über die logarithmierte Gleichung geschieht, log 0 aber nicht definiert ist. Da in dieser Arbeit der direkte Vergleich mit dem klassischen Modell durchgeführt wird, werden auch im NN-Modell die Flüsse von 0 nicht zur Modellierung verwendet. Allerdings sei herausgestellt, daß ein NN-Modell grundsätzlich sehr wohl in der Lage ist, die Null-Beziehungen als Information zu verwenden und zu modellieren.

Tab. 1: Charakterisierung der verwendeten Pendlerdaten

statistische Kennwerte	Bev. der Quellen Q_i	Bev. der Ziele Z_j	Distanz in km Luftlinie D_{ij}	beobachtete Pendleranzahl P_{ij}
Minimum	50	1547	1.00	1.0
1. Quantil	1230	6691	20.00	1.0
Median	1950	15450	35.00	3.0
3. Quantil	3387	66880	58.00	12.0
Mittelwert	5731	40150	43.79	34.4
Maximum	237800	237800	191.00	6785.0

3 Benchmark: das klassische Interaktionsmodell

Generell wird bei den klassischen Interaktionsmodellen davon ausgegangen, daß der Umfang der Interaktion proportional mit zunehmender Distanz zwischen Quell- und Zielort abnimmt. Je nach Informationen über das Interaktionssystem stehen verschiedene Modelltypen zur Verfügung; in Formel (1) ist das in dieser Arbeit verwendete Benchmark-Modell - das klassische unbeschränkte Gravitationsmodell - dargestellt.
Die Stärke des Interaktionsflusses hängt von den ausgewählten Attributen der beteiligten Objekte ab: zur Charakterisierung des Quellortes und zur Beschreibung der Attraktivität des

Zielortes werden daher aussagekräftige Maßzahlen benötigt. Üblicherweise dienen hierzu Wirtschaftskennzahlen bzw. sozio-ökonomische Kennzahlen, da von einer hohen Korrelation zwischen Wirtschaftsfaktoren und Verkehrsaufkommen ausgegangen werden kann. Beispielsweise wäre aber auch denkbar, die Anzahl vorhandener Arbeitsplätze zur Kennzeichnung des Zielortes zu verwenden. In dieser Arbeit wird vereinfachend die Bevölkerungsanzahl für die Charakterisierung des Quellortes sowie des Zielortes benutzt (Q und Z in Formel 1). Desweiteren wird eine Maßzahl zur Spezifizierung der räumlichen Separation zwischen den jeweiligen Ortspaaren benötigt. Üblicherweise wird hier die Luftlinienentfernung herangezogen; Straßendistanzen oder Reisezeiten stellen sinnvolle Alternativen dar.

$$P_{ij} = a_0 \, Q_i^{a_1} \, Z_j^{a_2} \, D_{ij}^{a_3} \, e_{ij} \tag{1}$$

mit P_{ij}: Pendlerfluß von Quelle i nach Ziel j; Q_i: Bevölkerungsanzahl der Quelle i;

Z_j: Bevölkerungsanzahl des Ziels j; D_{ij}: Distanz zwischen Quelle i und Ziel j;

a_0, a_1, a_2, a_3: Parameter des Modells;

e_{ij}: Fehler: identisch, unabhängig, lognormalverteilt.

Zur Berechnung der gesuchten Parameter dieses multiplikativen Modells wird das gesamte Modell logarithmiert, wodurch sich ein additiver Zusammenhang ergibt. Mittels der Methode der kleinsten Fehlerquadrate lassen sich sodann die Modellparameter bestimmen. Diese so geschätzten Modellparameter sind in Tabelle 2 dargestellt: es zeigt sich, daß alle Werte hoch signifikant sind. Außerdem wird aus den Werten der Schätzer deutlich, daß wie erwartet mit zunehmender Distanz der Pendlerfluß abnimmt. Ebenso wird erkenntlich, daß mit zunehmender Bevölkerungsanzahl der Quelle wie auch des Ziels der Pendlerfluß zunimmt.

Tab. 2: Geschätzte Parameter des klassischen Modells

Parameter	Schätzer	Standardfehler	t-Wert
$\log a_0$	-1.484823	0.059151	-25.1
a_1	0.446703	0.005218	85.6
a_2	0.526656	0.004570	115.3
a_3	-1.605160	0.007232	-221.9
σ^2	1.061787		

4 Modellierung mittels Neuronaler Netze

Neuronale Netze bieten grundsätzlich einen sehr flexiblen Modellierungsansatz, da die Modellzusammenhänge nicht explizit festgelegt werden, sondern einzig durch die zugrundeliegenden empirischen Daten bestimmt werden. So können weitaus komplexere realweltliche Zusammenhänge abgebildet werden, als dies mit expliziten Verknüpfungsvorschriften möglich ist. Allerdings ist die Bestimmung der NN-Modellarchitektur sowie der Modellparameter

und die Vermeidung des Overfitting nicht trivial (vgl. zu diesem Themenkreis die Einführung CZERANKA, TRAPLETTI, 1998, inkl. dortiger Literaturhinweise).
Die hier vorgestellte Pendlermodellierung ist ein Regressionsproblem: damit kann das Feedforward NN verwendet werden (s. CZERANKA, TRAPLETTI, 1998, Abbildung 3). Das zu schätzende NN-Modell besitzt die gleichen Einflußvariabeln wie das Gravitationsmodell, s. Formel (1).
Zur Modellselektion wird eine Kreuzvalidierung durchgeführt (Crossvalidation, s. EFRON, TIBSHIRANI, 1993, Kapitel 17): so wird die beste Modellarchitektur ausgewählt und gleichzeitig ein Overfitting vermieden, da die Güte der verschiedenen Modelle nur out-of-sample miteinander verglichen wird. Die vollständige Kreuzvalidierung würde darin bestehen, jeweils alle Daten mit Ausnahme eines Datensatzes zur Modellschätzung zu verwenden, sodann das Modell mit dem jeweils ausgelassenen Datensatz zu testen (leave-one-out cross-validation) und dieses Verfahren zu wiederholen, bis für alle Datensätze ein Testwert zur Verfügung steht. Da dies sehr rechenaufwendig ist, wurde bei diesem Beispiel das gesamte Datenset (37.880 Beobachtungen) mittels Zufallsauswahl in 10 gleich große Teile aufgeteilt (10-fach Kreuzvalidierung). Von diesen 10 Teilen werden jeweils 9 Teile für die Schätzung der Parameter verwendet (in-sample) und 1 Teil wird zur out-of-sample Vorhersage benutzt. Eine endgültige Maßzahl ergibt sich dann durch Mittelung dieser Testwerte.
Grundsätzlich wird die beste Modellarchitektur für die verwendeten Daten gesucht. Daher wird zunächst eine Reihe von Modellen, angefangen bei 0 Hidden Units, mit sukzessiver Steigerung der Anzahl der Hidden Units um 1, geschätzt. Erst nachfolgend wird das beste Modell anhand des besten R^2 ausgewählt. R^2 ist ein aus der Summe der Fehlerquadrate hergeleitetes normiertes Fehlermaß: es ist das Verhältnis von der mit dem jeweiligen Modell erklärten Variation des Pendlerflusses zur unerklärten Variation. R^2 liegt normalerweise zwischen 0 und 1: $R^2=0$ würde besagen, daß das betrachtete Modell nicht besser ist, als eine konstante Prognose, die den Mittelwert verwendet. $R^2=1$ würde bei in-sample-Berechnung ein vollständiges Overfitting bzw. bei out-of-sample Berechnung die perfekte Vorhersage bedeuten. Die verschiedenen Arbeitsschritte zur kombinierten Modellselektion und Parameterschätzung sind im folgenden aufgeschlüsselt:

1) Parameterschätzung jeweils 10 mal pro Modellarchitektur. Die verwendete Methode für die Parameterschätzung ist "Quasi Newton" (PRESS et al., 1995).
2) Berechnung des Bestimmtheitsmaßes R^2 (coefficient of determination) pro Modellarchitektur, gemittelt über die jeweils zehn geschätzten Modelle.
3) Systematische Schätzung der Parameter und Berechnung von R^2 gemäß obiger Schritte 1) und 2) für die verschiedenen Modellarchitekturen: angefangen bei 0 Hidden Units, was exakt der Schätzung des klassischen Modells entspricht (womit automatisch das Benchmark-Modell geschätzt ist), bis hin zu in diesem Falle 12 Hidden Units.
4) Vergleich der Modellgüten über die out-of-sample Bestimmtheitsmaße und Selektion jenes Modells, welches zur tatsächlichen Modellschätzung verwendet werden soll. Die Entwicklung des Bestimmtheitsmaßes R^2 für die verschiedenen Modellarchitekturen ist in Abbildung 2 dargestellt. Aufgrund der dort dargestellten Entwicklung von R^2 und unter der Prämisse, daß die Modellkomplexität nicht zu groß werden soll, werden 9 Hidden Units ausgewählt. Wie ersichtlich wird, bildet das NN-Modell die Daten erheblich besser ab, als das klassische Modell (aufgrund der out-of-sample Schätzung ist ein Datenfitting ausgeschlossen).

5) Schätzung des letztlich zur Verwendung kommenden Modells mit 9 Hidden Units anhand des vollen Datensets.

Die gesamten für das Modell zu schätzenden Parameter belaufen sich also auf 49. Diese setzen sich zusammen aus. 3x9 Gewichten zwischen Input und Hidden Layer, 9 Gewichten zwischen Hidden und Output Layer und 10 Gewichten der beiden Bias Units sowie 3 weiteren Gewichten, da Shortcuts zwischen den Input- und Output-Units zugelassen sind (die geschätzten Parameter sind in der Online-Fassung dieses Beitrages aufgelistet).
R^2 zeigt nun an, daß das NN-Modell eine bessere Schätzung darstellt, als das Benchmark. Wo sich nun die Abweichungen zwischen dem klassischen Modell und dem NN-Modell befinden bzw. wie sich die genauere Modellierung der Daten auswirkt, wird im folgenden Kapitel ansatzweise aufgezeigt.

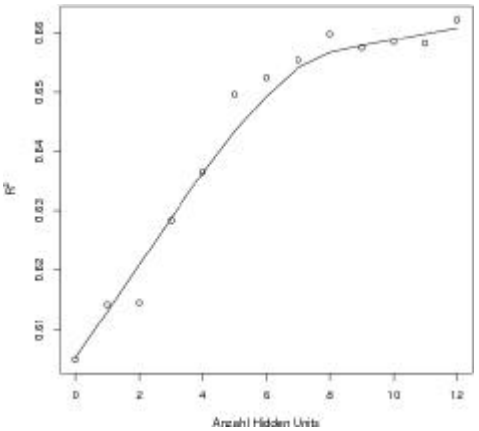

Abb. 2: Entwicklung der Modellgüte bei steigender Modellkomplexität

5 Analyse der Modellergebnisse

Das NN-Modell läßt sich aufgrund der Anzahl seiner Parameter und der komplexen Beziehungen nicht mehr interpretieren. Dies ist allerdings auch gar nicht nötig: letztlich interessieren Prognoseergebnisse unter ganz bestimmten Prämissen. Die Abbildungen 3 bis 6 zeigen einige Ansätze, wie die Analyse der Modellergebnisse unter bestimmten Gesichtspunkten geschehen könnte. Aufgezeichnet sind jeweils der bedingte Erwartungswert des Pendlerflusses in Abhängigkeit einer Variable. Die verbleibenden zwei Variablen wurden grundsätzlich auf ihren Median gesetzt (vgl. hierzu Tabelle 1). Einzig in Abbildung 5 wurde zudem die Entfernung wie angegeben variiert.
Die Abbildungen 3 und 4 zeigen den Anstieg der Pendlerzahlen bei steigender Bevölkerungsgröße zum einen des Ziels und zum anderen der Quelle. Deutlich wird z.B., daß vor allem bei hohen Bevölkerungsanzahlen starke Abweichungen vom klassischen Modell auftreten. Das klassische Modell tendiert dazu, vor allem bei großen Zielen weitaus kleinere Pendlerflüsse zu prognostizieren. In der Abbildung 6 wird deutlich, daß das klassische Modell eine Besonderheit des Datenmaterials gar nicht in der Lage ist, wiederzugeben: speziell bei Entfernungen zwischen 1 und 10 km weichen die beiden Modelle sehr stark voneinander ab. Das NN zeigt bei sehr kleinen Entfernungen sogar völlig gegenläufige Bewegungen auf. Damit wird deutlich, daß die Daten durchaus statistische Eigenschaften aufweisen, die vom klassischen Modell nicht erfaßt werden.

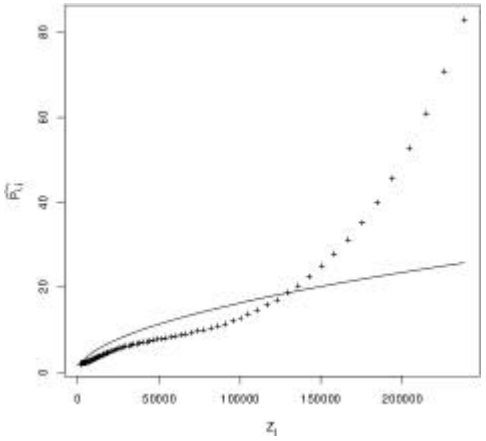

Abb. 3: Prognose des Pendlerflusses in Abhängigkeit der Bevölkerungsgröße des Ziels;
Kreuze: NN-Modell
Linie: klassisches Modell

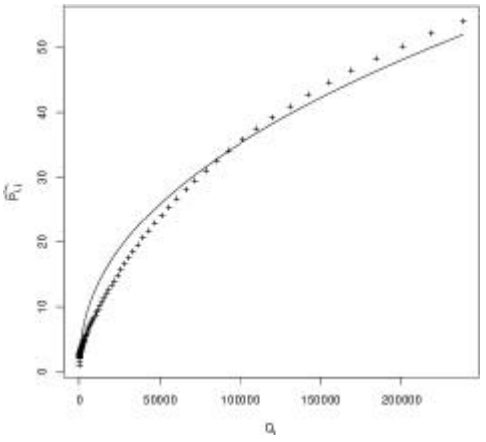

Abb. 4: Prognose des Pendlerflusses in Abhängigkeit der Bevölkerungsgröße der Quelle;
Kreuze: NN-Modell
Linie: klassisches Modell

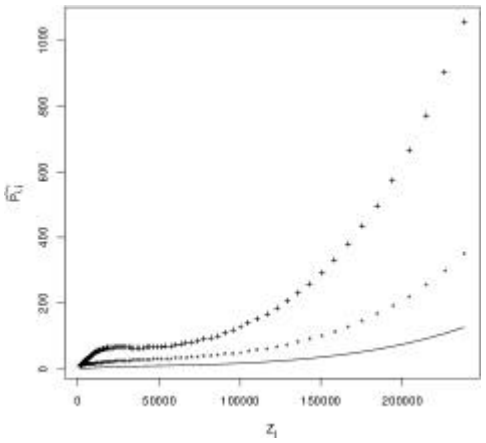

Abb. 5: Prognose des Pendlerflusses für zunehmende Bevölkerungsgröße des Ziels (alles: NN-Modell);
Kreuze: D = 10 km
Sterne: D = 20 km
Linie: D = 30 km

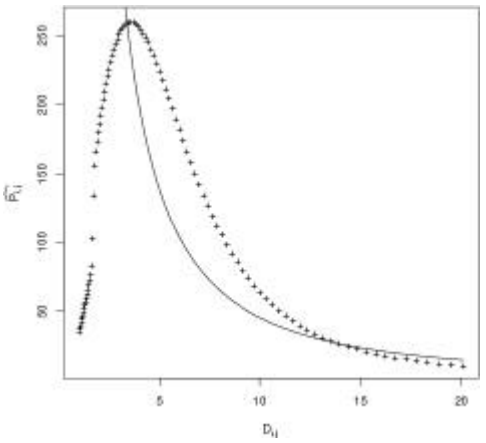

Abb. 6: Prognose des Pendlerflusses für zunehmende Distanz;
Kreuze: NN-Modell
Linie: klassisches Modell

6 Ausblick

Im Sinne einer explorativen Mobilitätsanalyse (vgl. KOLLARITS, HEUEGGER, USCHNIGG, 1998) sind weitere geographische Gesichtspunkte jederzeit in die NN-Modellierung integrier-

bar. Im Rahmen dieses Aufsatzes wurde zwar nur eine Variablenkonstellation analysiert, es ist jedoch sicherlich interessant, den spezifischen Einfluß der Straßendistanzen oder von Reisezeiten zu untersuchen. Diese Aspekte lassen sich relativ einfach integrieren; auf diese Weise fände die geographische Komponente bei der Modellentwicklung bessere Berücksichtigung und darauf aufsetzende Planungen oder Prognosen könnten ein weiteres Spektrum als bisher abdecken. Weiterhin lassen sich raum-zeitliche Entwicklungen mittels des (für die Zeitreihenanalyse leicht abgewandelten) NN-Ansatzes modellieren: wöchentliche oder monatliche Pendlerdaten bzw. Verkehrsflüsse lassen sich ebenso abbilden, wie der hier analysierte statische Zustand zum Zeitpunkt der Volkszählung. Diesbezüglich wird es allerdings schwierig sein, die passenden empirischen Daten zu beschaffen (vgl. SPIEGEL, 1998).

Verwendete Daten und Software

- Pendler-Daten: Auszug aus der Pendlererhebung der Volkszählung 1991; Berufspendler in Zielgemeinden >5000 Einwohner oder >1000 Berufseinpendler. Quelle: ÖSTAT, ÖIR
- Bevölkerungszahlen auf Gemeindeebene. Quelle: ÖSTAT; Volkszählung 1991
- Gemeindegrenzen: dig. Datenaufbereitung durch AGIS GmbH, Wien; Quelle: ÖSTAT
- GIS: MapInfo™
- Statistiksoftware: R : http://www.ci.tuwien.ac.at/R/contents.html
- NN-Software: ffnet - Zusatzmodul zu R; anforderbar per email beim Zweitautor: adrian.trapletti@wu-wien.ac.at

Literatur

Czeranka, M. & Trapletti, A. (1998): *Statistische Modellierung mit Neuronalen Netzen und Anwendungen in der geographischen Informationsverarbeitung*. In: Strobl, J. und Dollinger. F. (Hr.): Angewandte Geographische Informationsverarbeitung. Wichmann, Heidelberg, S. 39-50.

Efron, B. & R.J. Tibshirani, (1993): *An Introduction to the Bootstrap*. Chapman&Hall, NY.

Fischer, M.M. und Gopal, S. (1994): *Artifical Neural Networks: A New Approach to Modeling Interregional Telecommunication Flows*. In: Journal of Regional Science Vol. 34, No. 4, S. 503-527.

Fotheringham, S.A. & M.E. O'Kelly, (1989): *Spatial Interaction Models: Formulations and Applications.* Kluwer, Dordrecht.

Kollarits, S., Heuegger, M. & M. Uschnigg, (1998): *Explorative Mobilitätsanalyse*. CORP 98: http://osiris.iemar.tuwien.ac.at/~corp/html/kollarits.htm.

Ortúzar, J.deD. & L.G. Willumsen, (1990): *Modelling Transport*. John Wiley, Chichester.

Press, W.H., Teukolsky, S.A., Vetterling, W.T. & B.P. Flannery, (1995): *Numerical Recipes in C. The Art of Scientific Computing*. Cambridge Univ. Press.

Spiegel, T. (1998): *Überregionale Mobilitätserhebungen: Organisations- und Finanzierungsformen*. CORP 98: http://osiris.iemar.tuwien.ac.at/~corp/html/spiegel.htm.