PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 176
Taksonomia 18. Klasyfikacja i analiza danych – teoria i zastosowania 2011

Katarzyna Kopczewska
Uniwersytet Warszawski

MODELOWANIE INTERAKCJI PRZESTRZENNYCH Z WYKORZYSTANIEM PROGRAMU R

Streszczenie: Modele interakcji przestrzennych pozwalają ocenić znaczenie odległości w przepływie dóbr, osób, procesów, wiedzy, innowacji etc. Wykorzystywane powszechnie funkcje, w których przepływ ten tłumaczony jest odległością, mogą mieć postać wykładniczą, potęgową lub wielomianową, ich estymacja zaś możliwa jest przy wykorzystaniu metod klasycznych lub przestrzennych, w których dobór macierzy wag przestrzennych może być dokonany według różnych kryteriów sąsiedztwa. Celem artykułu jest analiza porównawcza wspomnianych modeli na przykładzie danych dla gmin i powiatów, z uwzględnieniem oceny jakości dopasowania przez SRMSE czy zysk informacyjny, a także prezentacja metod aplikacji w programie R.

Słowa kluczowe: modelowanie interakcji przestrzennych, dyfuzja, model wielomianowy, modelowanie przestrzenne.

1. Wstęp

W modelowaniu interakcji przestrzennych istnieje kilka etapów decyzyjnych, których rozstrzygnięcie jest znaczące dla wyników. Jest to wybór postaci funkcyjnej modelu, metody estymacji czy poziomu agregacji danych. W artykule zaprezentowane zostaną teoretyczne i techniczne aspekty modelowania dyfuzji zjawisk ekonomicznych przy wykorzystaniu metod ekonometrii przestrzennej i programu R.

2. Modele interakcji przestrzennych

Modele interakcji przestrzennych (*spatial interactions models*) wykorzystywane są do badania słabnącego przepływu dóbr i usług pomiędzy lokalizacjami wraz ze wzrostem odległości. Mają one szerokie zastosowania − od problemów komunikacji miejskiej, przez ocenę migracji czy handlu, do badania dyfuzji polityki, wiedzy, innowacji itp. Modelowanie opiera się na funkcjach wygasających wraz z odległością (*distance decay*): wykładniczej, potęgowej czy wielomianowej.

W podstawowym modelu interakcji przestrzennych zakłada się istnienie macierzy T przepływów pomiędzy lokalizacjami od M do N (tzw. *orgin-destination*) oraz

macierzy *d* odległości pomiędzy lokalizacjami. Wykorzystując określoną formę funkcyjną modelu, wyznacza się macierz *T'* wartości teoretycznych.

$$T = \{T_{ij}\}_{i,j=1}^{M.N}, \quad d = \{d_{ij}\}_{i,j=1}^{M.N}, \quad T' = \{T_{ij}^{'}\}_{i,j=1}^{M,N}.$$

Założenie o dwukierunkowych przepływach w parach jest ważne w modelach handlowych czy migracyjnych. W przypadku przepływów bodźców rozwojowych czy innowacji (nieobserwowanych) konieczne jest często przyjęcie przepływów jednokierunkowych, szczególnie z rdzenia na peryferie, gdzie jeden rdzeń ma wiele peryferii, a przepływ odbywa się w kierunku odśrodkowym. Macierz odległości i przepływów uprasza się wtedy do wektora.

Dobór funkcji wygasającej jest szeroko dyskutowany w literaturze. Z jednej strony istnieją klasyfikacje [Goux 1962; Fotheringham, O'Kelly 1989] wskazujące adekwatność postaci funkcyjnych do problemu badawczego, z drugiej zaś można znaleźć długą listę wad i zalet doboru funkcji, np. model potęgowy okazuje się lepszy niż wykładniczy, gdy konieczne jest zagwarantowanie porównywalności parametrów między badaniami niezależnie od skali pomiaru czy też gdy addytywna zmiana koszów transportu jednej pary *od-do* zmienia koszty innych par, lecz niestety okazuje się gorszy, gdyż zawyża niskie odległości i koszty (problem zbliżania się do 0). Funkcje wielomianowe są bardziej elastyczne w dopasowaniu do danych, jednak ich wadą jest zmiana kierunku na rosnący, możliwość prognozowania ujemnych interakcji, gdy funkcja spada poniżej osi *x*, oraz niejednoznaczna interpretacja wyrazu wolnego [Taylor 1975]. W nowszej literaturze można znaleźć zaawansowane modele przepływów dwukierunkowych [LeSage, Pace 2009].

Do oceny jakości modelu[1] można przyjąć kilka miar: R^2 w przypadku kalibracji przez MNK, zysk informacyjny *I* (*Information Gain*) dla kalibracji przez MNW oraz *SRMSE* (*Standarized Root Mean Square Error*), niezależnie od przyjętej postaci funkcyjnej i metody estymacji.

$$I = \sum_I \sum_j T_{ij} \ln\left(T_{ij} / T_{ij}^{'}\right), \quad SRMSE = \frac{\sqrt{\frac{\sum_{i,j=1}^{M,N}(T_{ij} - T_{ij}^{'})^2}{M \cdot N}}}{\frac{\sum_{i,j=1}^{M,N} T_{ij}}{M \cdot N}}.$$

W interpretacji wykorzystuje się zgodnie z regułą kciuka następujące progi: *SRMSE* < 0,5 bardzo dobre dopasowanie; *SRMSE* ~ 0,75 umiarkowane dopasowanie uwzględniające główne trendy*; SRMSE* > 1 słabe dopasowanie, obserwacje po-

[1] Użyteczną metodą jest diagnostyka wizualna modelowanej relacji, co umożliwia *ex ante* wybór funkcji odzwierciedlającej badany związek i pozwala zapobiec wykorzystaniu funkcji liniowej do predykcji relacji nieliniowej i *vice versa*.

za skalą [Andersson i in. 2008]. Zysk informacyjny może przyjmować wartości od 0 (idealne dopasowanie) do ∞.

Ważnym elementem jest definicja odległości – przestrzennej separacji. W najprostszym podejściu przyjmuje się odległość euklidesowską[2], która obarczona jest wieloma niedoskonałościami – nie uwzględnia ukształtowania geofizycznego i barier naturalnych (góry, jeziora, lasy), sieci dróg – ich gęstości, przepustowości, jakości itp. Z tego względu w bardziej wrażliwych badaniach wykorzystuje się wyrafinowane miary odległości, jak odległość drogowa w kilometrach, czas podróży czy koszt podróży. Jednak największy problem stanowi pozyskanie tych informacji, zwłaszcza w przypadku zbiorów dla wielu jednostek terytorialnych (np. gmin w Polsce). Istnieje więc wyraźny *trade-off* między ceną a jakością informacji o odległości.

3. Wykorzystanie metod przestrzennych w estymacji postaci funkcyjnych

Funkcje interakcji przestrzennych estymowane są najczęściej w sposób klasyczny, metodą najmniejszych kwadratów (MNK) lub największej wiarygodności (MNW). Ciekawą modyfikacją jest wykorzystanie metod przestrzennych uwzględniających strukturę sąsiedztwa. Można podejrzewać, że jeżeli badana jest relacja między lokalizacjami *A* i *B* oraz *A* i *C* oraz *B* i *C* są sąsiadami, *A* zaś jest odlegle, to wystąpić może autokorelacja błędów dla obu par ze względu na podobieństwo interakcji. W modelach interakcji przestrzennych główną (a najczęściej jedyną) zmienną objaśniającą jest odległość. Jeśli odległość między *A* i *B* jest taka, jak między *A* i *C*, to relacja może być zbliżona. Konieczne staje się zatem odfiltrowanie zależności pomiędzy *B* i *C*, by modelować relację *A* i *C* oraz *A* i *B*. Pomocne jest tu wykorzystanie macierzy wag przestrzennych, będącej przekształceniem macierzy sąsiedztwa. Powszechnie wykorzystuje się kilka typów macierzy wag przestrzennych − zależnie od kryterium sąsiedztwa. W wypadku modeli tej klasy nie należy wykorzystywać macierzy odwrotnej odległości, gdyż powiela ona informację zmiennej objaśniającej. Do wykorzystania pozostaje więc macierz **K** najbliższych sąsiadów[3] oraz macierz oparta na kryterium wspólnej granicy. Ta ostatnia jest najbardziej rozpowszechniona, a jej wykorzystanie i interpretacja nie budzą wątpliwości. Można a także zastosować macierz wyższego, niż pierwszy, rzędu. Dobór macierzy wag następuje zazwyczaj *a priori* i jest testowany *ex post*[4].

[2] W praktyce mierzona jako dystans między środkami ciężkości figur geometrycznych (wieloboków, *polygons*) reprezentujących badane jednostki terytorialne.

[3] Macierz nie ma jednoznacznej interpretacji, najczęściej jest wykorzystywana jako etap pośredni do uzyskania macierzy odwrotnej odległości.

[4] Decyzja o przyjęciu macierzy podejmowana jest na podstawie istotności współczynnika przestrzennego z modelu regresji, a także kryteriów informacyjnych – procedura testowa jest zbliżona do doboru zmiennych w modelu.

W klasie modeli przestrzennych najczęściej dokonuje się wyboru pomiędzy dwoma podstawowymi modelami (więcej w: [Anselin 1988; Kopczewska 2006; Kopczewska i in. 2009]):

opóźnienia przestrzennego (*spatial lag model*) $y_i = \rho W y_i + \beta X + \varepsilon_i$

i błędu przestrzennego (*spatial error model*) $y_i = \beta X + \varepsilon_i$,

gdzie $\varepsilon_i = \lambda W \varepsilon_i + u_i$. Wybór może być uzasadniony wynikami testów LM lub wynikać z teoretycznych przesłanek. Gdy celem estymacji nie jest otrzymanie parametrów strukturalnych, lecz prognozowanie w oparciu o skalibrowany model, wtedy specyfikacja opóźnienia przestrzennego wymaga użycia opóźnienia czasowego, co jest często niemożliwe w modelach interakcji przestrzennych ze względu na niedostępność zmiennej objaśnianej dla innego momentu czasowego. Z tego względu ta forma modelowania wymaga wykorzystania modeli błędu przestrzennego, w których informacja o strukturze przestrzennej zawarta jest na poziomie błędu.

Ważnym elementem modelowania jest dobór poziomu agregacji danych. W obszarze danych statystycznych najczęściej jest to wybór pomiędzy kolejnymi szczeblami sprawozdawczości i statystyki publicznej. W praktyce wybór zależy od kształtu badanego obszaru i liczby regionów – zbyt mała liczba regionów lub duży odsetek leżących na krawędzi obszaru i graniczących z „próżnią" prowadzą do błędów w estymacji, m.in. ze względu na „efekt krawędzi". W polskich warunkach uzasadniony jest wybór między poziomem NUTS3, NUTS4 i NUTS5, który oznacza przejście od 66 do 379 czy blisko 2500 obserwacji.

4. Wykorzystanie R w estymacji

Program R ma rozbudowane narzędzia do modelowania przestrzennego[5]. Do najważniejszych pakietów w tym obszarze należą: `spdep` do modelowania zależności przestrzennych na danych regionalnych oraz powiązane z nim `maptools` do wykonywania operacji na mapach i `sp` definiujący klasy obiektów przestrzennych. Jedną z najważniejszych cech R jest kompatybilność wielu pakietów przestrzennych, zarówno do analiz na punktach, jak i regionach. Jest to możliwe dzięki dedykowanej klasie obiektów – pakiet `sp` definiuje klasy dla wszystkich typów danych przestrzennych, kompatybilne zaś z `sp` pakiety działają w ramach tych klas lub umożliwiają dwukierunkowa konwersję danych wejściowych i wyników (więcej w: [Bivand i in. 2008]).

[5] Strona internetowa programu R oferuje dokładny opis dostępnych narzędzi przestrzennych – zebrane zostało to w kategorii SPATIAL w TaskViews.

Wczytanie mapy konturowej w formacie *shapefile* jest możliwe komendą `readShapePoly()`, gdzie wczytywana jest tylko część mapy o rozszerzeniu **.shp*, powiązane zaś elementy (**.shx*, **.dbf*) o tej samej nazwie, znajdujące się w tym samym katalogu, wczytywane są automatycznie. Dane do modelowania mogą być zawarte w powiązanej z mapą bazie danych **.dbf*, ale w praktyce najczęściej wczytuje się zewnętrzne dane w formacie np. **.csv*.

```
> pow<-readShapePoly("powiaty_mapa.shp", IDvar="ID_POW",
proj4string=CRS("+proj=longlat+ellps=clrk80"))
> dane<-read.csv("powiaty_dane.csv", header=T, sep=";", dec="." )
```

Utworzenie macierzy sąsiedztwa klasy *nb* możliwe jest dzięki komendzie `poly2nb()`, natomiast jej konwersja do macierzy wag przestrzennych dzięki komendzie `nb2listw()`. Tak przygotowany obiekt klasy *listw* wykorzystany zostanie w regresji przestrzennej. Poniższy przykład dotyczy macierzy sąsiedztwa według kryterium wspólnej granicy. Tworzenie innych typów macierzy można znaleźć w pracach Kopczewskiej k [2006, 2009]. Współrzędne oraz odległości między wielobokami wyznaczone zostały przez `coordinates()` oraz `spDistsN1()`z pakietu `sp`. Modele oszacowano za pomocą funkcji `glm()` dla modeli a-przestrzennych oraz `errorsarlm()` z pakietu `spdep` dla modeli przestrzennych.

```
> cont.nb.pow<-poly2nb(as(pow, "SpatialPolygons"))
> cont.listw.pow<-nb2listw(cont.nb.pow, style="W")
> crds<-coordinates(pow)
> for(i in 1:dim(pow)[1]){
> km[i,]<-spDistsN1(crds, crds[i,], longlat=TRUE)}
> model.gls<-glm(dane$y~dane$dist+I(dane$dist^2)+I(dane$dist^3)+
I(dane$dist^4))
> model.error<-
   errorsarlm(dane$y~dane$dist+I(dane$dist^2)+I(dane$dist^3)+
I(dane$dist^4), data=dane, cont.listw.pow, tol.solve=2e-40)
> SRMSE<-sqrt(sum((model$fitted.values-dane$y)^2)/
   dim(dane)[1])/
(mean(dane$y))
```

5. Przykład empiryczny estymacji

W przykładzie empirycznym porównana została jakość oszacowania modelu, w którym liczba firm w powiecie na 1000 mieszkańców w ujęciu[6] Polska = 100% tłumaczona jest odległością powiatu od jego miasta wojewódzkiego. Zjawisko ma charakter wygasający wraz z odległością. Im dalej położona jednostka terytorialna,

[6] Wszystkie obserwacje podzielone przez ich średnią wartość. Otrzymane wartości powyżej 1 oznaczają poziom powyżej średniej.

tym słabsza gospodarczo, co przejawia się m.in. w spadającej liczbie podmiotów gospodarczych. Siłą sprawczą jest kombinacja nieobserwowanych sił dośrodkowych i odśrodkowych. W ujęciu teoretycznym odzwierciedla to dyfuzję bodźców rozwojowych z rdzenia na peryferie. Pozwala to na ocenę, jak daleko od centrum można obserwować wpływ dużego ośrodka. Zmienną objaśniającą jest odległość euklidesowa w ramach województwa (NUTS2) pomiędzy miastem wojewódzkim a powiatami (NUTS4).

Wykorzystane zostały trzy postacie funkcyjne modelu:

1) wielomianowa (*multinomial*)

$$x_1 = \beta_0 + \beta_1 x_0 + \phi_1 D^1 + \phi_2 D^2 + \phi_3 D^3 + \phi_4 D^4 + ... + e,$$

2) potęgowa (*power*)

$$\ln x_1 = \beta_0 + \beta_1 \ln D + e,$$

3) wykładnicza (*exponential*)

$$\ln x_1 = \beta_0 + \beta_1 D + e.$$

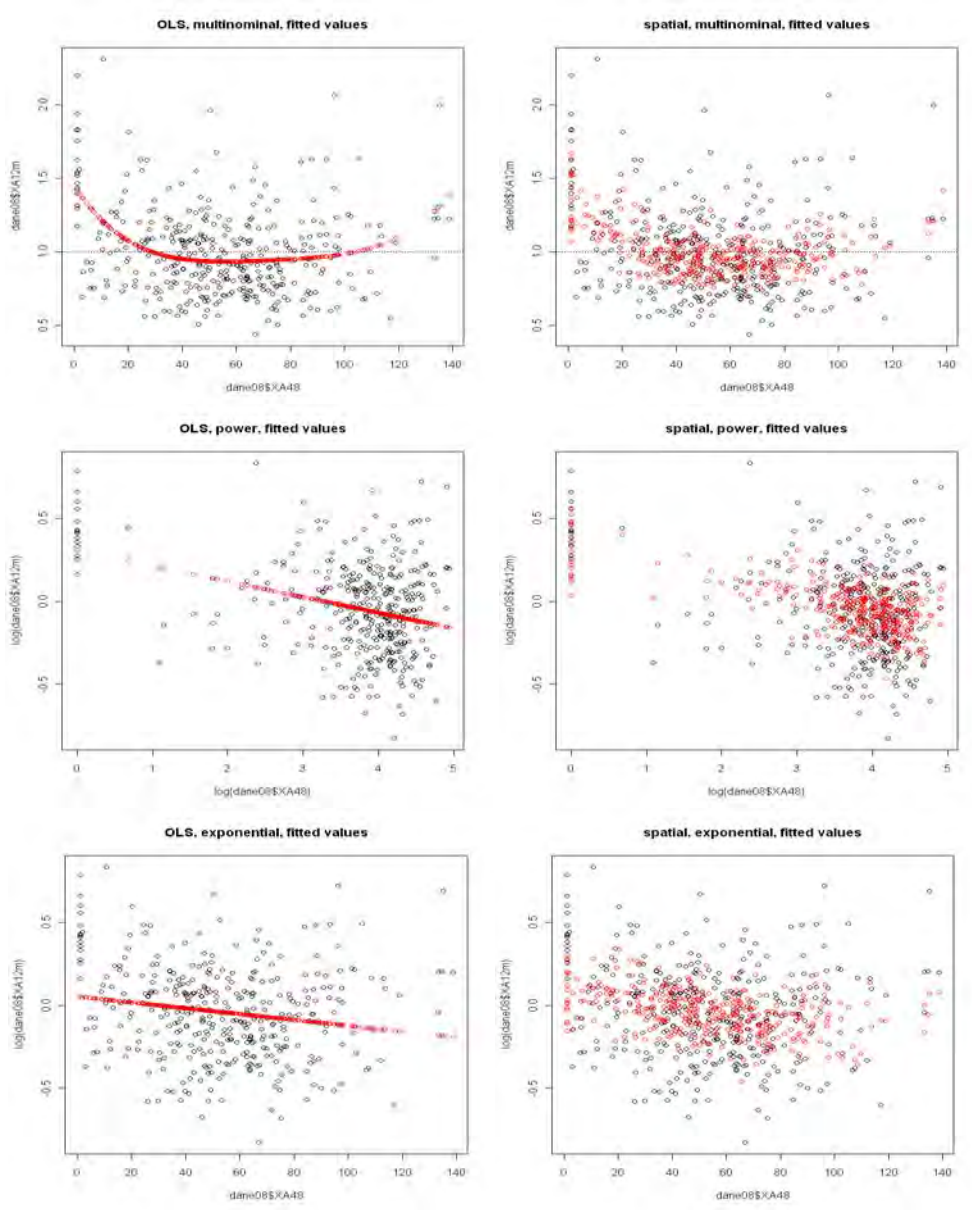
Każda z form funkcyjnych modelu oszacowana została na dwa sposoby: klasyczny oraz modelem błędu przestrzennego, przy wykorzystaniu macierzy sąsiedztwa według kryterium wspólnej granicy. Wyniki zbiorcze przedstawiono w tabeli 1 i na rysunku 1.

Tabela 1. Wyniki estymacji

Model	Wielomianowy		Potęgowy		Wykładniczy	
	$y \sim f(x)$		$\log(y) \sim f(x)$		$\log(y) \sim f(x)$	
Zmienne	klasyczny	przestrzenny	klasyczny	przestrzenny	klasyczny	przestrzenny
Stała	1,448***	1,482***	0,310***	0,3975***	0,0538*	0,1242***
Log(Dist)	---	---	-0,094***	-0,1000***	---	---
Dist	-0,028***	-0,0027***	---	---	-0,0018***	-0,00174**
$Dist^2$	0,0005***	0,0005***	---	---	---	---
$Dist^3$	-0,000005**	-0,000005**	---	---	---	---
$Dist^4$	0,00000001*	0,00000001*	---	---	---	---
SRMSE	0,2710	0,2322	1,0767	1,0665	1,08084	1,06959
AIC	97,947	7,0344	76,18	-29,609	109,06	3,92
Lambda	---	0,54***	---	0,55***	---	0,56***

Źródło: opracowanie własne.

Jak pokazują wyniki, specyfikacja wielomianowa jest lepsza niż potęgowa czy wykładnicza. Każdorazowo modele przestrzenne danej specyfikacji okazywały się lepsze niż modele klasyczne, czego dowodem jest niższe AIC oraz istotna lambda.



Rys. 1. Wizualizacja dopasowania modeli

Źródło: opracowanie własne.

Najlepszą kombinacją jest model wielomianowy z uwzględnieniem czynnika przestrzennego. Jest to szczególnie widoczne przy *SRMSE*, które w modelach wielo-

mianowych okazało się znacząco niższe niż w modelach potęgowych czy wykładniczych oraz w modelach przestrzennych niższe niż w modelach tradycyjnych. Co więcej, w modelach wielomianowych *SRMSE* wykazuje dobre dopasowanie, przeciwnie do pozostałych modeli. W warstwie ekonomicznej model wyraźne pokazuje, że w odległości ok. 25 km od miasta wojewódzkiego poziom badanej zmiennej spada poniżej 1, co oznacza, że zlokalizowane tam jednostki terytorialne są słabsze niż średnie. Podnoszenie się funkcji powyżej $y = 1$ poza odległością 100 km wynika z właściwości funkcji wielomianowej oraz małej próby[7]. Oznacza to, że wpływ dużych miast na promocję postaw przedsiębiorczości i podnoszenie atrakcyjności inwestycyjnej wygasa w powiatach, które są sąsiadami już drugiego rzędu dla miasta wojewódzkiego. Dalsze zwiększanie odległości od rdzenia nie ma znaczenia dla procesów dyfuzji – są one zbliżone w całym obszarze „peryferyjnym".

6. Podsumowanie

Celem artykułu było pokazanie kilku aspektów modelowania nieobserwowanych zjawisk dyfuzji polityki rozwojowej. Zmodyfikowany jednokierunkowy model interakcji przestrzennych, w którym wykorzystano specyfikację wielomianową oraz uwzględniono autokorelację przestrzenną, okazał się dobry w sensie *SRMSE* oraz znacząco lepszy niż powszechnie wykorzystywane modele wykładnicze i potęgowe, przestrzenne i a-przestrzenne. Widoczna jest dyfuzja od rdzenia w kierunku peryferii w odległości ok. 25 km, czyli sąsiedztwa pierwszego rzędu. Zaprezentowane podejście jest więc efektywnym narzędziem modelowania polityki regionalnej.

Literatura

Andersson J., Joernsten K., Pettersen Strandenes S., *Modeling Freight Markets for Coal*, NHH Discussion Paper, 2008.

Anselin L., *Spatial Econometrics: Method and Models*, Kluwer Academic Publishers, Dordrecht 1988.

Bivand R., Pebesma E., Gomez-Rubio V., *Applied Spatial Data Analysis with R*, Springer, New York 2008.

Fotheringham A.S, O'Kelly M.E., *Spatial Interaction Models: Formulations and Applications*, Kluwer Academic Publishers, New York 1989.

Goux J.M., *Structure de l'espace et migration*, [w:] J. Sutter (red.), *Human Displacements*, Entretiens de Monaco en Sciences Humaines: Premiere Session, Paris 1962.

Kopczewska K., *Ekonometria i statystyka przestrzenna*, CeDeWu, Warszawa 2006.

Kopczewska K., *Przestrzenne modelowanie zmian stopy bezrobocia*, „Wiadomości Statystyczne" 2010, 5.

[7] W zbiorze znajduje się niewiele obserwacji – jednostek terytorialnych zlokalizowanych w tej odległości.

Kopczewska K., Kopczewski T., Wójcik P., (red), *Metody ilościowe w R. Aplikacje ekonomiczne i finansowe*, CeDeWu, Warszawa 2009.

LeSage J., Pace R.K., *Introdution to Spatial Econometrics*, CRC Press Taylor & Francis Group, London 2009.

Taylor P., *Distance decay models in spatial interactions*, CATMOG, „Concepts and Techniques in Modern Geography" 1975, no. 2.

Vries J., Nijkamp P., Rietveld P., *Expotential or power distance-decay for commuting*? An alternative scpecification, Paper presented at 45th Congress of the European Regional Science Association, Amsterdam 2005.

SPATIAL INTERACTION MODELING USING R PROGRAM

Summary: Spatial interaction models allow to assess the importance of distance in the flow of goods, people, processes, knowledge, innovation, etc. Commonly used functions in which this flow is explained by the distance may take the exponential, exponential or polynomial form, and their estimation is possible using classical or spatial methods, in which the choice of spatial weights matrix can be according to various criteria neighborhood. The purpose of this article is to compare these models estimated on data for counties (NUTS4), including the assessment of quality matching or profit by SRMSE information and the presentation of application methods in R software.