Механика Транспорт Комуникации

Научно списание http://www.mtc-aj.com

ISSN 1312-3823 (print)
ISSN 2367-6620 (online)
том 13, брой 3/1, 2015 г.
статия № 1150

НОВ ПОДХОД И АЛГОРИТЪМ ЗА ДИНАМИЧНО СЕГМЕНТИРАНЕ ПРИ КАЛИБРИРАНЕ НА АГРЕГИРАНИ ЛОГИТ МОДЕЛИ

Тодор Размов, Любомир Клямбарски
t.razmov@gmail.com, klqmba@mail.bg

Висше транспортно училище „Тодор Каблешков,
ул.“Гео Милев“ № 158, София 1574,
БЪЛГАРИЯ

Ключови думи: *модален сплит, агрегирани данни, сегментиране, полезност, калибриране, нормиране*

Резюме: *Предложен е нов подход и алгоритъм за динамично сегментиране при калибриране на агрегирани логит модели. При този подход O-D двойките се разпределят в различни класове в зависимост от разстоянието, но ширината на класовете се определя динамично в процеса на калибриране (динамична сегментация).*

Предложеният нов подход и алгоритъм са тествани при разпределението на междуградските пътувания между конкуриращите се железопътен и автобусен обществен транспорт (модален сплит). Използваните данни за пътуванията са агрегирани на зоново ниво. Всяка зона отговаря на област от приетото административното деление на страната.

Направен е анализ и оценка на резултатите от предложения подход.

ВЪВЕДЕНИЕ

Логит моделите са едни от най-простите модели за математическо описване на модалния сплит. В основата на прилагането им в областта на транспортното моделиране стои икономическата теория за полезността. Рационалния избор в случая се свежда до избор на алтернативата осигуряваща най-голяма полезност [1]. Полезността най-често се представя в следния вид:

$V_i = \sum_k \beta_{ki}.X_{ki}$, където:

V_i е полезността на алтернатива i,

β_{ki} е теглови коефициент, а

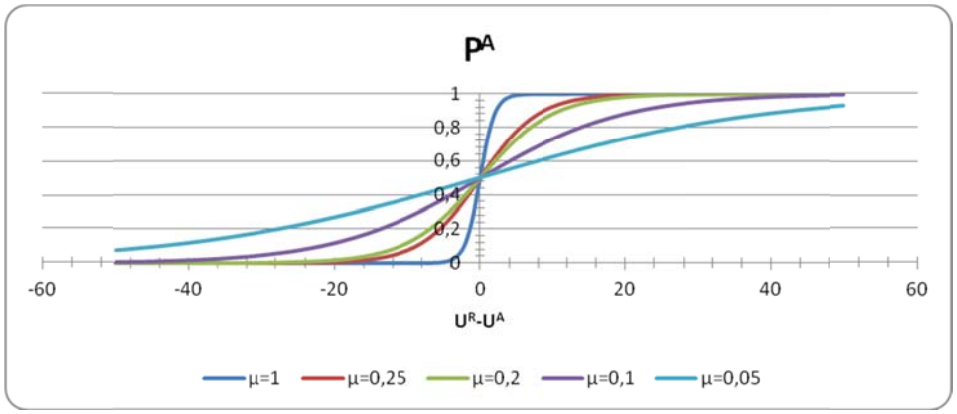
X_{ki} е параметър, влияещ на полезността.

Разглеждайки две алтернативи, изборът се определя от разликата между двете полезности. Алтернатива 1 може да бъде избрана, ако $V_1 - V_2 > 0$ и обратно. Такова строго определяне обаче предполага всички индивиди да избират алтернативата с по-голяма полезност и никой да не избира по-малко полезната, а също така и полезността да бъде оценена по един и същ начин от всеки отделен индивид. Оттук се определя

нуждата от използване на вероятностен подход при описването на избора между алтернативи.

В множество трудове, изследващи избора на вид транспорт [2] е доказано, че логистичните модели добре описват вероятността за този избор. Вероятността за избор на една от две алтернативи се представя в следния вид:

$$(1)\quad P^A = \frac{\exp(-\mu U^A)}{\exp(-\mu U^A) + \exp(-\mu U^R)} = \frac{1}{1 + \exp[-\mu(U^R - U^A)]}$$



Фиг. 1 Логистична функция

На фиг.1 е представен характера на логистичната функция в зависимост от скалиращия параметър μ. При по-големи стойности на μ зависимостта клони към реализация на подхода „всичко или нищо“.

От формула 2 личи, че изборът между алтернативите зависи не от стойностите, а от разликата между полезностите. Основният недостатък от това свойство е, че една и съща разлика се оценява по един и същ начин както при високи, така и при ниски стойности на полезността. С други думи, например разлика от един лев има по-висока стойност при пътуване, струващо 5 лева, отколкото при такова струващо 50. В литературата се предлага подход за сегментиране на разглежданите алтернативи и определянето на подходящи скалиращи параметри и теглови коефициенти за всеки сегмент.

КАЛИБРИРАНЕ НА БИНАРЕН ЛОГИТ МОДЕЛ

В процесът на калибриране на логит моделите целта е да се определят такива теглови коефициенти и скалиращи параметри, с помощта на които моделът да описва достатъчно точно разглежданите данни.

Логит моделите могат да бъдат калибрирани вземайки се предвид всички наблюдавани *O-D* двойки. Тогава се получава един общ комплект коефициенти за определяне на функциите на полезност, а оттам и на пазарните дялове на видовете транспорт. Поради сложността при изчисление за голям обхват на модела се прилага подход на агрегиране на данните, който предполага приобщаване на *O-D* двойките в по-големи класове в зависимост от даден критерий. Идеята при този подход е да се определи един комплект стойности на параметрите на модела, но на базата на по-малко, усреднени за класа данни. При този подход се моделира дела на вида транспорт в класа. В литературата се предлага този критерий за класификация да бъде разликата между функциите на полезност на разглежданите видове транспорт. При този подход най-често се разглеждат класове с еднаква ширина, като най-подходяща е най-ниската възможна ширина, при която няма празни интервали, т.е. във всеки интервал има поне

един елемент (*O-D* двойка). Друга възможност за определяне на интервалите е предварително да се определи минимален брой пътувания в даден клас. В този случай интервалите са разнородни, но отново са предварително определени. Основният недостатък при моделите, използващи един комплект стойности на параметрите, е че не се отчитат тежестите в разликата между полезностите. За това се предлага сегментиране на разглежданите данни и определяне на коефициенти за всички класове. Подбирането на класовете се извършва, спазвайки се принципа за липса на празни класове. Възможно е подбраните интервали да са лесно интерпретируеми (от 0 до 20 км, 21-40 км., 41-60 км. и т.н.). В [3] са изследвани два от подходите, а именно такъв с един комплект коефициенти и такъв със коефициенти за всеки клас. В изследването се наблюдава, че при по-населените класове (тези с по-голям брой *O-D* двойки, съответно и пътници) се наблюдава по-голяма неточност в изчислените стойности на пазарните дялове. Оттук може да се заключи, че в модела не са включени всички влияещи фактори и характерни особености при избора на вид транспорт. В настоящия доклад е предложен подход за динамично определяне ширината на класа, което само по себе си би включило всички неразгледани фактори, определящи вида на модалния сплит. При зададена достатъчно висока стойност на точността, този подход гарантира пълното възпроизвеждане на наблюдавания модален сплит.

АЛГОРИТЪМ НА ПРЕДЛОЖЕНИЯ ПОДХОД ЗА КАЛИБРИРАНЕ

1. O-D двойките се подреждат във възходящ ред според средното разстояние за пътуване между тях
2. Първата двойка се присъединява към клас 1. Във всеки клас трябва да има поне толкова на брой елементи, колкото са параметрите на модела, за да могат тези параметри първоначално да бъдат определени еднозначно. Добавят се необходимия брой двойки в класа.
3. Чрез линейна регресия се определят стойностите на коефициентите и се определя грешката.
4. Ако сумарната грешка за класа преминава приет праг на точност, новите коефициенти не се запазват, а последната двойка се изважда от класа и се създава нов, в който тя става първа.
5. Проверява се дали всички двойки са присъединени към даден клас. Ако са присъединени калибрирането приключва. Ако не са се пристъпва към стъпка 2.

ПРИЛОЖЕНИЕ НА ПОДХОДА

Подходът е приложен за моделиране на пазарните дялове на автобусния и железопътния транспорт при междуградските пътувания в България за 2011 г., като доразвитие на разработения модел в докладът на авторите „Моделиране на търсенето на транспортни услуги за превоз на пътници с обществен транспорт в Република България“ [3]. Разгледано е административното деление на страната (28 области), като форма на агрегиране на данните. Двойките, включени в изследването са 410 между които е налична информация за извършени пътувания и с двата вида транспорт. Съгласно преобразуванията в [2] моделът може да се представи в линейна форма, както следва:

$$(2)\quad \ln\left(\frac{P^A}{P^R}\right) = \mu\left(U^R - U^A\right) + \mu\delta$$

Полезностите на двата вида транспорт включват компонентите скорост на пътуване (v), цена на пътуване (C) и честота на предоставяната услуга (F).

(3) $U_{kij}^{R} = \alpha_k.v_{ij}^{R} + \beta_k.C_{ij}^{R} + \gamma_k.F_{ij}^{R}$

(4) $U_{kij}^{A} = \alpha_k.v_{ij}^{A} + \beta_k.C_{ij}^{A} + \gamma_k.F_{ij}^{A}$, където:

α, β и *γ* са тегловите коефициенти за наблюдавания сегмент *k*. Данните са за пътуване между пунктове *i* и *j*. Тъй като теглата са неизвестни, то скалиращия параметър μ може да се интегрира в тегловите коефициенти. Тогава формула 3 се преобразува до:

(5) $$\ln\left(\frac{P^A}{P^R}\right) = \left(U^R - U^A\right) + \delta = \alpha_k.v_{ij}^{R} + \beta_k.C_{ij}^{R} + \gamma_k.F_{ij}^{R} - \alpha_k.v_{ij}^{A} - \beta_k.C_{ij}^{A} + \gamma_k.F_{ij}^{A} + \delta =$$
$$\alpha_k.\left(v_{ij}^{R} - v_{ij}^{A}\right) + \beta_k.\left(C_{ij}^{R} - C_{ij}^{A}\right) + \gamma_k.\left(F_{ij}^{R} - F_{ij}^{A}\right) + \delta$$

Последното уравнение е удобно за изследване чрез многофакторен регресионен анализ. Тъй като изследваните променливи в модела са от различни порядъци (скоростите - десетичен и стотичен порядък; цените – стотици и хиляди евроцента; честотите – за някои направления стигат и до десетохиляден в годишно изражение) е предприето нормиране на данните от 0 до 1 спрямо горни допустими граници (120 км/ч за средната скорост, около 60 лв или 3100 евроцента за пътуване и 600 услуги дневно – стойността е висока, но тя се определя честотата на най-обвързаните *O-D* двойки, като напр. София-град – София-област).

РЕЗУЛТАТИ ОТ МОДЕЛА

Методът за калибриране е тестван с различни стойности на допустимата грешка (разликата между реалните и моделираните данни на квадрат). В таблица 1 са представени резултатите от тестването.

Таблица 1

	Точност	Брой класове	R^2 дялове	R^2 пътници	Грешка
1	20,00	38	0,6302739	0,9674664	509,9584
2	10,00	49	0,8646952	0,9871275	216,5495
3	5,00	60	0,9142814	0,9876856	111,8409
4	3,33	66	0,9416643	0,9871726	73,43948
5	2,50	67	0,9479683	0,9871277	49,86884
6	2,00	67	0,9542667	0,9932255	43,5167
7	1,67	69	0,9619063	0,9730882	40,35413
8	1,43	72	0,9731348	0,9863800	32,34762
9	1,25	73	0,9745112	0,9880879	28,01147
10	1,11	73	0,9816021	0,9964640	24,66999
11	1,00	74	0,9801448	0,9965237	24,06543
12	0,50	84	0,9944319	0,9998115	7,618476
13	0,33	85	0,9958155	0,9998390	5,262318
14	0,25	88	0,9973355	0,9998705	2,960742
15	0,20	90	0,9970767	0,9998764	1,850786
16	0,17	90	0,9972164	0,9998831	1,834497
17	0,14	91	0,9983548	0,9998869	1,173735
18	0,13	91	0,9983548	0,9998869	1,173735
19	0,11	92	0,9984125	0,9998871	1,005705
20	0,10	92	0,9984125	0,9998871	1,005705

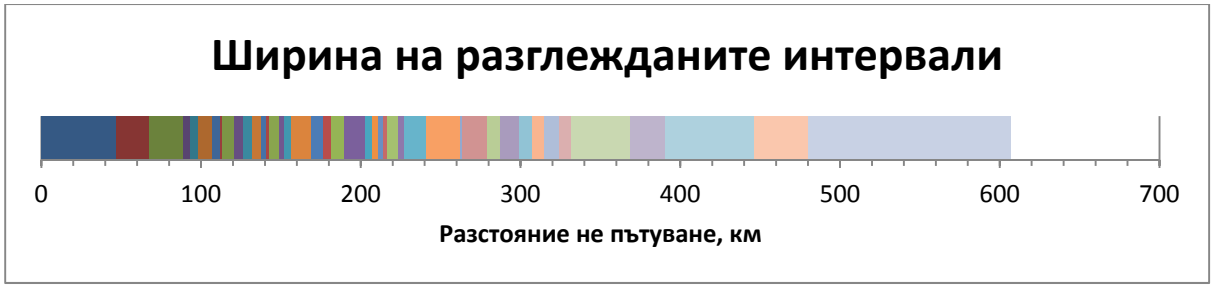
Високата точност определя големия брой класове необходим за точното възпроизвеждане на наблюдавания модален сплит. В този случай моделът се стреми да опише възможно най-точно всеки един наблюдаван дял, т.е. моделът не прави разлика дали по *O-D* направлението са извършени 10 или 10 хил. пътувания. Аналитично погледнато, при двойките с по-малко извършени пътувания е допустимо грешката в моделираните дялове да е по-голяма. За извършването на такова приоритизиране е определена тежестта на грешката, като е определен делът на пътуванията по *O-D* двойката от всички разглеждани пътувания. В таблица 2 са представени резултатите от извършените тестове чрез използване на такъв тип приоритизиране.

Таблица 2

	Точност	Брой класове	R^2 дялове	R^2 пътници	Грешка
1	1,0000	43	0,6959268	0,9674620	24,24255
2	0,5000	51	0,8682068	0,9911379	12,76784
3	0,3333	58	0,8829959	0,9873862	8,043049
4	0,2500	63	0,9203621	0,9859790	5,349707
5	0,2000	67	0,9454640	0,9871481	4,233168
6	0,1667	67	0,9484356	0,9871578	3,361971
7	0,1429	67	0,9589868	0,9876392	2,871934
8	0,1250	67	0,9542667	0,9932255	2,701113
9	0,1111	69	0,9637680	0,9915636	2,384094
10	0,1000	71	0,9677211	0,9863796	2,061070
11	0,0500	78	0,9860979	0,9965003	1,076277
12	0,0333	83	0,9926102	0,9997703	0,541948
13	0,0250	84	0,9929019	0,9998098	0,422331
14	0,0200	85	0,9958014	0,9998390	0,319712
15	0,0167	88	0,9977944	0,9998747	0,197883
16	0,0143	90	0,9970767	0,9998764	0,114880
17	0,0125	90	0,9970767	0,9998764	0,114880
18	0,0111	90	0,9972164	0,9998831	0,113869
19	0,0100	91	0,9981352	0,9998861	0,080548

Използването на приоритизиране намалява необходимия брой класове, за сметка на увеличаващата се сумарна грешка. Въпреки това моделът описва достатъчно точно наблюдавания модален сплит.

На фигура 2 е показана ширината на всеки от класовете, определени чрез подхода за калибриране, при 43 класа и използване на приоритизиране (таблица 2 ред 1).



Фиг. 2 Ширина на разглежданите интервали

ЗАКЛЮЧЕНИЕ

Предложеният подход представлява инструмент, с който всеки един модален сплит може да се възпроизведе с необходимата точност. Този подход е приложим при наличие на малко на брой изследвани фактори, влияещи върху модалния сплит. Подходът е теоретически издържан, тъй като разглеждания брой променливи е значително по-малък от наблюдавания брой данни (при 43 класа с четирите променливи се получават 170 коефициента за целия модел). Другите известни в теорията подходи използват по-малко на брой променливи, но при тях има предварително сегментиране (по цели на пътуване, характеристики на пътуващите, а също и дължина на пътуването), след което реално за моделирането на всички пътувания са използвани значителен брой променливи.

На авторите не е известен подобен подход за калибриране на модален сплит.

ЛИТЕРАТУРА:

[1] Карагьозов К., Р. В., „Прогнозиране и сегментиране на пазара на пътническите превози“, в *Транспорт 2004*, с-ци 59–64.

[2] Ortúzar, J. de D. и L. G. Willumsen, *MODELLING TRANSPORT*, 4th изд. UK: John Wiley & Sons, Ltd, 2011.

[3] Размов, Т. и Л. Клямбарски, „МОДЕЛИРАНЕ НА ТЪРСЕНЕТО НА ТРАНСПОРТНИ УСЛУГИ ЗА ПРЕВОЗ НА ПЪТНИЦИ С ОБЩЕСТВЕН ТРАНСПОРТ В РЕПУБЛИКА БЪЛГАРИЯ“, *Механика Транспорт Комуникации*, том 12, бр 3, с-ци I–38/I–46, 2014.

NEW APPROACH AND ALGORITHM FOR DYNAMIC SEGMENTATION IN THE PROCESS OF CALIBRATION OF AGGREGATED LOGIT MODELS

Todor Razmov, Lyubomir Klyambarski
t.razmov@gmail.com, klqmba@mail.bg

Todor Kableshkov Univercity of Transport,
158 Geo Milev Steet., Sofia 1574,
BULGARIA

Key words: *modal split, aggregated data, segmentation, utility, calibration, standardization,*

Abstract: *A new approach and algorithm for dynamic segmentation in the process of calibration of aggregated logit models is proposed in this article. This approach consists of segmenting the travel demand (the O-D pairs) by trip length, but the segment size is dynamically defined in the process of calibration (dynamic segmentation).*

These new approach and algorithm are tested in modeling the modal split between the two competing public transport modes in interurban passenger travels in Bulgaria – rail and bus. The modal split is calibrated using zonal data according the first level administrative subdivision of the country.

The results from the proposed approach are analyzed and evaluated.