E/B/R ECONOMIC AND BUSINESS REVIEW

Volume 21
Issue 4 *Vol 21, Posebna številka*

Article 9

7-16-2019

Novejši pristopi v analizi podatkov o smrtnosti

Simona Korenjak-Černe

Aleša Lotrič Dolinar

Recommended Citation
Korenjak-Černe, S., & Lotrič Dolinar, A. (2019). Novejši pristopi v analizi podatkov o smrtnosti. *Economic and Business Review, 21*(4). https://doi.org/10.15458/2335-4216.1062

NOVEJŠI PRISTOPI V ANALIZI PODATKOV O SMRTNOSTI

SIMONA KORENJAK-ČERNE[1]
ALEŠA LOTRIČ DOLINAR[2]

POVZETEK: *Pri uporabi statističnih metod in modelov in tudi pri ostalih metodah analize podatkov se srečujemo s številnimi omejitvami, ki narekujejo potrebo po iskanju novih pristopov. Namen tega prispevka je prikazati razloge za iskanje novih pristopov v analizi pri podatkih o smrtnosti v evropskih državah, ki jo opazujemo razčlenjeno po spolu, starosti in vzroku smrti, na kratko predstaviti nekaj rezultatov dosedanjega dela in nakazati raziskovalno delo v povezavi s temi podatki, ki poteka v sodelovanju z drugimi raziskovalci.*

Ključne besede: *smrtnost po spolu, starosti in vzroku smrti, simbolna analiza podatkov, razvrščanje v skupine, program clamix*

1 UVOD

Statistične metode in modeli temeljijo na določenih predpostavkah, ki jih je pri uporabi treba upoštevati. Dodatne omejitve uporabe klasičnih statističnih metod in tistih metod analize podatkov, ki večinoma temeljijo na računalniških metodah strojnega učenja, predstavlja tudi tako imenovana klasična predstavitev podatkov, ki jo te metode navadno predpostavljajo. Pri taki predstavitvi je namreč vsaka spremenljivka predstavljena z eno samo vrednostjo, kar pa ne dopušča vključitve vsebinskih povezav med njimi.

S potrebo po iskanju drugih, neklasičnih pristopov v analizi podatkov smo se srečali pri obdelavi podatkov o smrtnosti, saj nas je zanimala podrobnejša slika, tj. smrtnost po spolu, starosti in vzroku smrti v evropskih državah. Analizirali smo podatke za 32 evropskih držav iz leta 2014, dosegljive na spletni strani Eurostata (http://ec.europa.eu/eurostat/data/database). Za ustrezno primerljivost podatkov med državami smo smrtnost v posameznih državah, gledano razčlenjeno po spolu, starosti in vzroku smrti, ustrezno preračunali na smrtnost po teh spremenljivkah glede na standardno populacijo (Lotrič Dolinar, Sambt, Korenjak-Černe, 2017).

Na podlagi podobnosti porazdelitev smrtnosti po vzrokih v nekaterih petletnih starostnih razredih, ki so na voljo v originalnih podatkih Eurostata, smo število starostnih razredov nekoliko zmanjšali na naslednjih sedem starostnih razredov: 0–14, 15–34, 35–54, 55–64, 65–74, 75–84, 85+. Vsako državo smo tako na klasičen način predstavili s 56

1 Univerza v Ljubljani, Ekonomska fakulteta, Ljubljana, Slovenija, e-pošta: simona.cerne@ef.uni-lj.si

2 Univerza v Ljubljani, Ekonomska fakulteta, Ljubljana, Slovenija, e-pošta: alesa.lotric.dolinar@ef.uni-lj.si

spremenljivkami (2 spol x 7 starost x 4 vzrok), ki v vsaki kombinaciji spol-starost-vzrok predstavljajo število smrti na 100.000 prebivalcev. Pri tem smo upoštevali tri najpogostejše vzroke smrti (bolezni obtočil, neoplazme, bolezni dihal) in vse ostale vzroke vključili v kategorijo "drugo".

Raziskovanje poteka v sodelovanju s prof. Jožetom Sambtom, ki je opozoril na dostopnost tovrstnih podatkov, pomagal pri pripravi njihove primerljivosti med državami in izpostavil dvodimenzionalnost podatkov: stopnje (ravni) smrtnosti na eni strani in njene strukture po različnih vzrokih na drugi (Lotrič Dolinar, Sambt, Korenjak-Černe, 2019).

Metodološko se zdi najbolj smiselno najprej uporabiti metode faktorske analize, a te metode zaradi majhnega števila enot, tj. evropskih držav, ki so v središču našega zanimanja, v primerjavi z velikim številom spremenljivk ne moremo uporabiti. Čeprav smo število spremenljivk nekoliko zmanjšali z združitvijo nekaterih petletnih starostnih razredov in omejili število kategorij za vzroke smrti, smo za to, da pri zmanjševanju števila spremenljivk ne bi izgubili preveč informacij, še vedno ohranili skupno 56 spremenljivk, kar je za možnost uporabe metode faktorske analize bistveno preveč.

Takšno krčenje podatkov nam torej zaradi prevelike izgube informacij ne omogoča zmanjšanja števila spremenljivk v tolikšni meri, da bi lahko uporabili klasično faktorsko analizo. V statistiki se za zmanjšanje števila spremenljivk sicer najpogosteje uporablja metoda glavnih komponent, kjer pa ne moremo ohraniti vsebinske povezanosti med spremenljivkami, zato smo se lotili problema s pomočjo novejših pristopov v analizi podatkov.

Smiselnost upoštevanja vseh omenjenih treh kategorij smrtnosti (spola, starosti in vzroka smrti) smo ovrednotili na podlagi primerjave teh rezultatov z rezultati analize brez upoštevanja vzrokov smrti, ki smo jo napravili na osnovi pričakovanega trajanja življenja ob rojstvu (e_0) kot tradicionalnega kazalnika v tovrstnih raziskavah (OECD, 2017). Poleg tega smo identificirali tiste spremenljivke, ki najizraziteje ločujejo nastale skupine držav.

2 UPORABA KLASIČNIH METOD RAZVRŠČANJA

Najprej smo podatke analizirali s kombinacijo Wardove hierarhične metode in metode *k*-središč (Ward, 1963, Anderberg, 1973, Hartigan, 1975, Kaufman in Rousseeuw, 1990). Za merjenje različnosti med državami smo uporabili kvadrat evklidske razdalje, saj sta pri tej izbiri različnosti metodi usklajeni, posledica izbire te različnosti pa je minimiziranje notranje variabilnosti skupin in maksimiranje razlik med skupinami (Podani, 1989, Murtagh, 2014). Z uporabo klasičnih metod razvrščanja je vsaka od kombinacij upoštevana kot samostojna spremenljivka in tako ne vključuje nobene informacije o medsebojni povezanosti spremenljivk. A vse te spremenljivke med seboj nikakor niso neodvisne niti vsebinsko (pri vsaki kombinaciji spola in starosti podatki vsebinsko predstavljajo strukturo

smrtnosti po vzrokih smrti) niti statistično (kar npr. lahko preverimo s korelacijsko matriko), zato smo želeli raziskati še druge možnosti.

3 NOVEJŠI PRISTOPI V ANALIZI PODATKOV: SIMBOLNA ANALIZA PODATKOV

Eno od možnih rešitev za upoštevanje medsebojne povezanosti spremenljivk v podatkih ponujajo metode simbolne analize podatkov. Ta veja analize podatkov se je začela razvijati v 80. letih prejšnjega stoletja na pobudo francoskega profesorja Edwina Didaya. Osnovna ideja temelji na neklasičnih, kompleksnejših predstavitvah podatkov, ki omogočajo ohranjanje notranje strukture, pri čemer to kompleksnost podatkov upoštevamo v prilagojenih metodah (Billard in Diday, 2006, Noirhomme-Fraiture in Brito, 2011, Diday, 2016). V razvoj metod simbolne analize podatkov smo predvsem pod mentorstvom prof. Vladimirja Batagelja že od začetka aktivno vključeni tudi slovenski raziskovalci.

3.1 Simbolna tabela podatkov

V tabeli 1 je prikazan izsek simbolne tabele naših podatkov o smrtnosti v evropskih državah, predstavljenih tako s stopnjo (ravnjo) kot z relativno strukturo po vzroku za izbrane štiri kategorije vzrokov smrti glede na spol in starost. Pri tovrstni predstavitvi podatkov zlahka opazimo, da med spoloma in med starostnimi razredi obstajajo razlike tako v stopnji smrtnosti (predstavljena je s številčno vrednostjo v vsakem polju) kot v njeni porazdelitvi po vzrokih smrti. Opazimo tudi, da sta si Avstrija in Belgija bolj podobni (tako po stopnjah kot večinoma tudi po strukturah), Bolgarija pa se od obeh, še posebno pa od Belgije, zelo loči. Ker pa je takšen pregled pri večjem številu držav težko opraviti, potrebujemo za to ustrezne metode, ki bi kar najbolje vključile obe dimenziji naših podatkov.

Tabela 1: *Izsek simbolne tabele podatkov: Struktura smrtnosti v evropskih državah v letu 2014 po vzroku smrti in ravni glede na spol in starost, preračunana na standardno populacijo velikosti 100 000 (vir: Eurostat in lastni izračuni)*

		Starost (v dopolnjenih letih)						
Država	Spol	0–14	15–34	35–54	55–64	65–74	75–84	85 +
Avstrija (AT)	M	2,7	7,4	34,6	54,9	95,3	141,0	117,0
	Ž	2,2	3,0	18,3	29,9	59,2	134,7	237,2
Belgija (BE)	M	3,2	8,4	36,2	57,5	94,5	151,9	122,8
	Ž	2,3	3,2	22,5	33,9	60,1	140,1	227,4
Bolgarija (BG)	M	6,8	12,1	72,7	125,1	176,5	246,1	155,6
	Ž	4,9	4,9	33,4	52,0	98,1	253,8	311,0
…								

3.2 Prilagojeni metodi razvrščanja s programom clamix

Za bolj avtomatiziran pregled podobnosti med državami smo izbrali prilagojeno Wardovo in prilagojeno metodo *k*-središč (Batagelj, Kejžar in Korenjak-Černe, 2015). Metodi sta implementirani v R-paketu clamix (Batagelj in Kejžar, 2012). Izbrani metodi razvrščanja sta kombinacija kvalitativne in kvantitativne analize podatkov, saj po eni strani ohranjata strukturo smrtnosti po vzroku smrti, po drugi strani pa rešujeta problem razvrščanja kot optimizacijski problem.

Vsako državo smo tako predstavili s 14 simbolnimi spremenljivkami (kombinacijami spola in starosti), predstavljenimi z relativnimi strukturami smrtnosti po 4 vzrokih, kot utež pa za vsako od 14 spremenljivk upoštevali ustrezno stopnjo (raven) smrtnosti. Najbolj očitno je razbitje držav v dve skupini, ki sta predstavljeni v tabeli 2. Omenjeni skupini držav sta zelo izrazito ločeni, saj dobimo enako razbitje v dve skupini tudi s klasično metodo razvrščanja in z razvrščanjem držav glede na podatek o pričakovani življenjski dobi ob rojstvu (e_0).

V tabeli 2 takoj opazimo, da med skupinama pri obeh dimenzijah (stopnji in strukturi smrtnosti po vzroku) obstajajo razlike v vseh opazovanih kategorijah: spolu, starosti in vzroku smrti. V skupini vzhodnih držav med vzroki smrti pri moških prevladujejo bolezni obtočil že od srednjih let dalje, medtem ko sta pri ženskah in v zahodnih državah bolj opazna tudi deleža neoplazem in ostalih vzrokov. Med skupinama so velike razlike tudi v povprečnih stopnjah smrtnosti na državo, saj so te v skupini vzhodnih držav večinoma od enainpolkrat do dvakrat tolikšne kot v skupini zahodnih držav.

Še pomembnejšo vlogo pri odločanju ima podrobnejše razbitje na več skupin. Pri razbitjih v več skupin pa so med rezultati, dobljenimi z novejšimi pristopi, in med tistimi na osnovi klasičnih pristopov bistvene razlike, zato je pomembno, da izbrana metoda res vključi obe dimenziji, tako stopnjo kot tudi razčlenjenost po vzroku.

3.3 Simbolna analiza podatkov s programskim paketom SYR

Dodatne analize nam ponujajo tudi druge prilagojene metode simbolne analize podatkov, zato smo se na pobudo prof. Didaya povezali s podjetjem SYMBAD – Le Symbolic Data Lab, ki ima v svojem programskem paketu SYR razvite in implementirane nekatere dodatne metode simbolne analize.

V naši aplikaciji smo uporabili metodo glavnih komponent, prilagojeno simbolni predstavitvi podatkov, razvrščanje v skupine in spremljanje vzorcev poti spreminjanja vzrokov po državah s pomočjo prestrukturiranja v t. i. metabine (Diday, 2013). Pri tem še dodatno razvijamo in prilagajamo metode v skladu s simbolno predstavitvijo obravnavanih podatkov, ki bi nam omogočile dodaten vpogled v podatke in povezave med njimi.

Tabela 2: *Simbolna tabela skupin: Struktura smrtnosti v dveh osnovnih skupinah evropskih držav v letu 2014 po vzroku smrti in povprečni ravni na državo v skupini glede na spol in starost, preračunana na standardno populacijo velikosti 100 000 (vir: Eurostat in lastni izračuni)*

		Starost (v dopolnjenih letih)						
Skupina držav	**Spol**	**0–14**	**15–34**	**35–54**	**55–64**	**65–74**	**75–84**	**85+**
Vzhod (BU, CZ, EE, HR, HU, LT, LV, PL, RO, RS, SK)	M	4,4	13,5	73,6	114,2	166,8	222,6	138,5
	Ž	3,4	4,4	29,6	48,1	90,6	214,1	270,1
Zahod (AT, BE, CH, CY, DE, DK, EL, ES, FI, FR, IE, IT, LI, LU, MT, NL, NO, PT, SE, SI, UK)	M	2,6	6,9	32,1	50,6	85,3	142,4	120,9
	Ž	2,1	2,9	17,5	29,0	54,9	136,6	229,1

4 ZAKLJUČEK

V prispevku smo izpostavili, zakaj nam klasične metode ne ponujajo zadovoljivih rešitev v primeru analize podatkov o smrtnosti v evropskih državah, ki jih želimo proučevati razčlenjene po spolu, starosti in vzroku smrti. Nakazali smo nekatere možne rešitve, ki nam jih ponujajo novejši pristopi v analizi podatkov s področja simbolne analize podatkov. Glavna prednost teh pristopov je v bogatejši predstavitvi podatkov in ustrezno prilagojenih metodah, ki omogočajo ohranjanje več informacij (v našem konkretnem primeru strukture smrtnosti po vzrokih). Predstavljeno delo poteka v sodelovanju s prof. Jožetom Sambtom z Ekonomske fakultete Univerze v Ljubljani, prof. Edwinom Didayem, zaslužnim profesorjem s francoske univerze Dauphine Université Paris, in dr. Filipejem Afonsom, podatkovnim analitikom podjetja SYMBAD – Le Symbolic Data Lab.

REFERENCE

Anderberg, M. (1973). *Cluster analysis for applications*. New York: Academic Press.

Batagelj, V. & Kejžar, N. (2012). clamix – Clustering Symbolic Objects, R package, https://r-forge.r-project.org/projects/clamix/, 2012.

Batagelj, V., Kejžar, N. & Korenjak-Černe, S. (2015). Clustering of Modal Valued Symbolic Data. *ArXiv e-prints*, 1507.06683.

Billard, L. & Diday, E. (2006). *Symbolic Data Analysis: Conceptual Statistics and Data Mining.* Chichester: John Wiley.

Diday, E. (2013). Principal Component Analysis for Bar Charts and Metabins Tables. *Statistical Analysis and Data Mining, 6*(5), 403–430.

Diday, E. (2016). Thinking by classes in data science; the symbolic data analysis paradigm. *WIREs Computational Statistics, 8*, 172–205.

Eurostat (2013). *Revision of the European Standard Population, Report of Eurostat's Task Force*, Eurostat Methodologies and Working Papers, 2013 edition. Eurostat, European Commission.

Hartigan, J. (1975). *Clustering algorithms.* New York: Wiley-Interscience.

Kaufman, L. & Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: Wiley.

Lotrič Dolinar, A., Sambt, J. & Korenjak-Černe, S. (2017). Mortality by causes of death in European countries. V: Malačič, J. & Gams, M. (ur.), *Soočanje z demografskimi izzivi: zbornik 20. mednarodne multikonference Informacijska družba - IS 2017*, 9.–13. oktober 2017, Ljubljana, Slovenija (str. 56–59). Ljubljana: Institut Jožef Stefan.

Lotrič Dolinar, A., Sambt, J. & Korenjak-Černe, S. (2019). Clustering EU Countries by Causes of Death. *Population Research and Policy Review, 38*(1), 157–172.

Murtagh, F. (2014). Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? *Journal of Classification, 31*, 274–295.

Noirhomme-Fraiture, M. & Brito, P. (2011). Far Beyond the Classical Data Models: Symbolic Data Analysis. *Statistical Analysis and Data Mining, 4*(2), 157–170.

OECD (2017). *Life expectancy at birth (indicator)*. (pridobljeno 2. 10. 2017).

Podani, J. (1989). New combinatorial clustering methods. *Vegetatio, 81*, 61–77.

Ward, J. H. Jr. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association, 58*(301), 236–244.