Analiza ritmičnosti štev nih podatkov z uporabo modela cosinor

Nina Velikajne
Faculty of Computer and Information Science,
University of Ljubljana,
Večna pot 113, SI-1000 Ljubljana, Slovenia
nv6920@student.uni-lj.si

Miha Moškon
Faculty of Computer and Information Science,
University of Ljubljana,
Večna pot 113, SI-1000 Ljubljana, Slovenia
miha.moskon@fri.uni-lj.si

Povzetek

Analiza ritmičnosti štev nih podatkov je postala pomembna v mnogih vidikih znanosti, inženirstva in celo ekonomije. Obstajajo metode z namenom detekcije ritmičnosti zveznih podatkov, ki pa večinoma niso primerne za analizo štev nih podatkov. V prispevku predstavimo metodologijo, ki omogoča analizo ritmičnosti v štev nih podatkih. Metoda združuje metodo cosinor z uporabo različnih računskih regresijskih modelov, ki so primerni za analizo štev nih podatkov. Omogoča tako detekcijo ritma kot tudi ocenitev parametrov ritma, primerjavo zgrajenih modelov in iskanje optimalnega števila komponent za metodo cosinor ter iskanje najbolj ustreznega tipa štev nega modela. Vzpostavljena metoda omogoča primerjavo zaznanega ritma v odvisnosti od različnih parametrov ritmičnosti in izračun njihovih intervalov zaupanja. Celotno metodologijo smo testirali na tedenski periodičnosti realnih podatkov COVID-19 obolenj v Sloveniji.

Ključne besede metoda cosinor, analiza ritmičnosti, štev ni podatki, pojavnost dogodkov, regresija.

1 Uvod

Detekcija in analiza ritmičnih vzorcev v štev nih podatkih ima pomembno vlogo pri mnogih vidikih znanosti. Periodični podatki so podatki, v katerih se vzorci ponavljajo z določeno periodo. Zelo pogost tip periodičnih procesov so procesi, ki odražajo cirkadiano nihanje – procesi s periodo 24-ur [3]. Regulira jih Zemljina rotacija in izmenjavanje dneva in noči, ki vpliva na vse organizme in posledično na njihovo obnašanje ter gibanje (glej [20]). Poseben tip podatkov predstavljajo štev ni podatki, ki opisujejo pojavnost izbranih dogodkov [3].

Štev ni podatki se pogosto pojavljajo periodično. V naravi in naši okolici so tovrstni podatki vseprisotni. Z njihovo analizo lahko pripomoremo k razumevanju različnih dogodkov in posredno k razumevanju delovanja organizmov in družbenih sistemov. Na primer, analiza števila porodov glede na uro v dnevu lahko pomaga pri organizaciji medicinskega in babiškega osebja v porodnišnici [14]. Analiza dnevnih vzorcev v prometu in njihovo spreminjanje skozi čas nam lahko pove veliko o gibanju in obnašanju populacije, posebej v času epidemije (glej [4]).

Za analizo štev nih periodičnih podatkov potrebujemo posebne računske metode, ki upoštevajo in ohranjajo odnose med podatki. Metode morajo upoštevati diskretno porazdelitev, ki je omejena le na nenegativne cele vrednosti. Pri uporabi navadne linearne regresije so lahko napovedane vrednosti negativne, kar je teoretično nemogoče [8]. Za zaznavanje in analizo ritmičnosti v zveznih podatkih obstaja kar nekaj neparametričnih metod (glej [18, 12]). V primerjavi z omenjenimi metodami (glej [16]) nam uporaba trigonometričnih regresijskih metod v povezavi z različnimi cosinor modeli predstavlja številne prednosti, npr. ocenitev parametrov ritma [18]. Izkaže se tudi, da alternativne metode v določenih primerih odpovejo zaradi velikega števila osamelcev (angl. *outliers*), same velikosti podatkov, neuravnoteženosti podatkov in zbiranja podatkov brez ponovitev (glej [17]).

V tem prispevku predstavimo metodologijo, ki z združevanjem metode cosinor skupaj z različnimi štev nimi računskimi modeli upošteva vse omejitve štev nih periodičnih podatkov in tako omogoča tudi ocenitev parametrov ritma. Metoda omogoča izračun intervalov zaupanja za posamezen parameter ritma s pomočjo samovzorčenja (angl. *bootstrapping*). S F testom določimo optimalno število komponent za metodo cosinor, za iskanje najbolj ustreznega tipa štev nega modela pa uporabimo Vuongov test.

Članek je razdeljen na pet poglavij. V drugem in tretjem poglavju so predstavljeni metoda cosinor za analizo periodičnih podatkov in pet računskih regresijskih modelov za delo s štev nimi podatki. Sledi poglavje, ki opisuje postopek izbire najbolj ustreznega računskega modela. V poglavju Rezultati so predstavljeni rezultati testiranja vzpostavljene metode na realnih podatkih. V zadnjem poglavju so zajete ključne ugotovitve in postopki celotne analize.

2 Metoda cosinor

Pri periodičnih podatkih se opazovani vzorci ponovijo z določeno periodo. Njihovo analizo lahko naslovimo kot regresijski problem, pri katerem pa je potrebno upoštevati tudi karakteristike ritma, npr. fazo in amplitudo nihanja. Metoda cosinor se uporablja za analizo časovnih vrst in se posveča tako detekciji ritma kot tudi oceni parametrov ritma. Model v ozadju metode lahko opišemo z Enačbo 1, kjer je N število komponent, M srednja vrednost ritma (MESOR, angl. *Midline Estimating Statistic of Rhythm*), A amplituda, P perioda in $e(t)$ funkcija napake. Spremenljivka t označuje čas, i pa iterira po številu

DOI https://doi.org/10.18690/978-961-286-516-0.14
ISBN 978-961-286-516-0

komponent – od 1 do N [1, 15, 16].

$$Y(t) = M + \sum_{i=1}^{N} \left(A_{i,1} \cdot \sin\left(2\pi \frac{t}{P/i}\right) + A_{i,2} \cdot \cos\left(2\pi \frac{t}{P/i}\right) \right) + e(t), \quad (1)$$

Če je perioda ritma znana vnaprej, lahko enačbo metode cosinor poenostavimo v model linearne regresije:

$$Y(t) = M + \sum_{i=1}^{N} (A_{i,1} \cdot X_{i,1}(t) + A_{i,2} \cdot X_{i,2}(t)) + e(t), \quad (2)$$

kjer je $X_{i,1}(t) = \sin\left(2\pi \frac{t}{P/i}\right)$ in $X_{i,2} = \cos\left(2\pi \frac{t}{P/i}\right)$. V kolikor perioda ni znana, jo lahko ocenimo s pomočjo uporabe periodogramov (angl. *periodograms*) [1, 15, 16].

3 Analiza števnih podatkov z metodo cosinor

Za analizo in detekcijo ritma na izvornih očiščenih podatkih naprej uporabimo metodo cosinor. V primeru znane periode uporabimo Enačbo 2, ki podatke razdeli na poljubno število komponent in jih transformira do regresijske oblike. Na transformirane podatke nato apliciramo regresijske računske modele. Regresijski modeli omogočajo identifikacijo in karakterizacijo odnosov med mnogimi faktorji. Zaradi dela s števnimi podatki moramo izbrati ustrezne regresijske računske modele, ki upoštevajo vse lastnosti tovrstnih podatkov.

Podatki so diskretni, omejeni na nenegativne cele vrednosti in velikokrat tudi razpršeni (angl. *dispersed*). Srečamo se s pojmom povečane (angl. *overdispersion*) ali pa zmanjšane razpršitve (angl. *underdispersion*) [8]. Pri povečani razpršitvi imajo podatki večjo varianco, kot bi jo sicer pričakovali. Če je varianca večja kot povprečje podatkov, gre za povečano razpršitev. Obratno velja za zmanjšano razpršitev [7]. Z uporabo navadne linearne regresije bi dobili nepravilne rezultate, saj tak računski model ne bi upošteval omenjenih lastnosti [8].

V vzpostavljeni metodologiji smo se odločili za uporabo petih različnih računskih modelov, ki se najpogosteje uporabljajo za analizo števnih podatkov. Uporabili smo Poissonov model (angl. *Poisson model*), generaliziran Poissonov model (angl. *generalised Poisson model*), Poissonov model z inflacijo ničel (angl. *zero-inflated Poisson model*), negativen binomski model (angl. *negative binomial model*) in negativen binomski model z inflacijo ničel (angl. *zero-inflated negative binomial model*).

Poissonov model predpostavlja, da so podatki porazdeljeni s Poissonovo porazdelitvijo:

$$P(y = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad (3)$$

kjer je λ povprečna pričakovana vrednost oz. število dogodkov na enoto časa. Povprečje podatkov μ je pri Poissonovi porazdelitvi enako povprečni pričakovani vrednosti λ. Varianca σ^2 je enaka povprečju, poenostavljeno velja, da je $\sigma^2{=}\lambda$. Model ne upošteva povečane ali pa zmanjšane razpršitve podatkov [9].

Generaliziran Poissonov model izhaja iz navadnega Poissonovega modela. Bistvena razlika tega modela v primerjavi s Poissonovim modelom je ta, da je generaliziran Poissonov model primeren tudi za podatke, ki imajo povečano ali zmanjšano razpršitev. Vpeljemo dodaten parameter α, ki opisuje stopnjo disperzije. Model torej ne zahteva, da je povprečje μ enako varianci σ^2. Obstajata dve različici generaliziranega Poissonovega modela – GP-1 in GP-2 [6]. V vzpostavljeni metodologiji smo uporabili različico GP-1.

Poissonov model z inflacijo ničel je razširitev navadnega Poissonovega modela. Tovrstni model za razliko od navadnega in generaliziranega Poissonovega modela upošteva, da so ničelne vrednost bolj pogoste kot ostale vrednosti. Izhaja iz tega, da obstajata dva dejavnika, ki vplivata na izid, ali je vrednost ničelna ali neničelna [13].

Negativen binomski model predpostavlja, da so podatki porazdeljeni z negativno binomsko porazdelitvijo. Uporabljata se dve verziji negativnega binomskega modela, tj. NB-1 in NB-2. Različica NB-1 se je izkazala kot bolj primerna, prilagojena krivulja se je namreč vidno lepše prilegala izvornim podatkom, zato smo v metodologiji uporabili verzijo NB-1. Varianca takega modela je definirana kot $\sigma^2 = \mu + \alpha \cdot \mu$, kjer je α parameter disperzije in μ povprečje. Povprečje je enako povprečni pričakovani vrednosti λ. Model je zato primeren tudi za podatke s povečano ali pa zmanjšano razpršitvijo [2, 11]. Ob večanju parametra α varianca konvergira k povprečju in negativna binomska porazdelitev postane Poissonova [2].

Negativen binomski model z inflacijo ničel je razširitev negativnega binomskega modela. Podobno kot Poissonov model z inflacijo ničel upošteva, da so ničelne vrednosti bolj pogoste kot ostale (neničelne) vrednosti. Ključna razlika v primerjavi s Poissonovim modelom z inflacijo ničel je, da ta model temelji na negativni binomski porazdelitvi. Model je zato primeren tudi za podatke, ki imajo povečano ali pa zmanjšano razpršitev [10, 11].

4 Izbira najustreznejšega modela

Izbiro najbolj ustreznega računskega modela razdelimo na dva nivoja. Na prvem nivoju iščemo optimalno število komponent za metodo cosinor. Na tem mestu smo izhajali iz tega, da so modeli gnezdeni. Dva modela sta gnezdena, če lahko prvi model izrazimo z drugim oz. če drugi model poleg vsaj enega dodatnega parametra vsebuje enake parametre kot prvi [5]. Implementirali smo F test. Na podlagi dveh zgrajenih modelov izračunamo F vrednost. F test temelji na razliki vsote kvadratov (angl. *sum of squares*) dveh modelov in upošteva število parametrov modela [5].

Na drugem nivoju smo vrednotili tip računskega modela. Uporabili smo Vuongov test, ki je primeren tako za gnezdene modele kot tudi za ne gnezdene in prekrivajoče se (angl. *overlapping*) modele. Vuongov test omogoča izračun Z vrednosti na podlagi logaritma največjega verjetja (angl. *maximum log-likelihood*) dveh modelov. Tudi ta test upošteva število parametrov modelov [19].

Oba testa sledita podobnemu postopku. Za dva modela, tj. model A in model B, izračunamo F oz. Z vrednost.

Model A zavržemo, če je izračunana vrednosti manjša od vnaprej določene meje, tj. statistične signifikance (angl. *statistical significance*) [5, 19].

5 Rezultati

Vzpostavljeno metodologijo smo preizkusili na realnih podatkih. Podatke smo pridobili s strani COVID-19 sledilnik[1] in analizirali število pozitivnih testov glede na dan v tednu. Upoštevali smo vse teste, tj. seštevek PCR (angl. *polymerase chain rection*) in hitrih antigenskih (HAGT) testov. Metodo smo izvedli na treh podatkovnih zbirkah. Prva zbirka beleži število pozitivnih testov od 20. oktobra 2020 – razglasitev 2. vala epidemije v Sloveniji, do 10. februarja 2021. Povprečje pozitivnih testov je 1340, varianca pa 276278. Druga zbirka vsebuje primere od 11. februarja 2021 do 26. aprila 2021. Povprečje podatkov je 796, varianca pa 90328. Tretja podatkovna zbirka združuje časovni obdobji prve in druge podatkovne zbirke. Povprečje zadnje zbirke je 1082, varianca pa 232224. Metodo smo preizkusili tudi za časovno obdobje 1. epidemije v Slovenji – pomlad 2020, vendar je teh podatkov premalo za smiselno analizo.

V podatkovni zbirki smo najprej odstranili osamelce (angl. *outliers*), tako da smo za posamezno uro odstranili vnose, kjer so bile vrednosti števila pozitivnih testov večje ali manjše od 0,15 kvantila. Nato smo izvedli metodo cosinor za posamezno število komponent – od 1 do 4, in zgradili posamezen tip računskega modela (glej Poglavje 1). Pri številu komponent 4 smo se ustavili, ker se računski modeli zaradi prevelikega števila komponent niso več prilagajali izvornim podatkom. Zgrajene modele smo nato ovrednotili, najprej smo poiskali optimalno število komponent na podlagi F testa in nato še najbolj ustrezen tip modela s pomočjo Voungovega testa. Opisan postopek se ponovi za posamezno podatkovno zbirko. Vse podatkovne zbirke imajo večjo varianco kot povprečje kar pomeni, da imajo podatki povečano razpršitev. Na podlagi porazdelitve izvornih podatkov (glej Sliko 1) lahko ugotovimo, da podatki nimajo ničelnih vrednosti.

Težave se pojavijo pri vseh modelih, v kolikor je število komponent pri metodi cosinor večje od 3. Opazimo, da se izvornim podatkom najbolje prilegata negativen binomski model in generaliziran Poissonov model. Oba modela, sta namreč primerna za podatke s povečano razpršitvijo. Za vse podatkovne zbirke smo dobili enak rezultat. Optimalno število komponent je 3, najbolj ustrezen tip modela pa je generaliziran Poissonov model (glej Sliko 1). Na podlagi izvedene analize smo lahko ovrednotili parametre ritmičnosti (glej Tabelo 1) in njihove intervale zaupanja (glej Tabelo 2). Celoten postopek analize z vsemi vmesnimi rezultati je dostopen v repozitoriju GitHub[2].

6 Diskusija in zaključek

Vzpostavljena in implementirana metodologija omogoča analizo številnih periodičnih podatkov. Metodologija je se-

[1] `https://covid-19.sledilnik.org/sl/data`

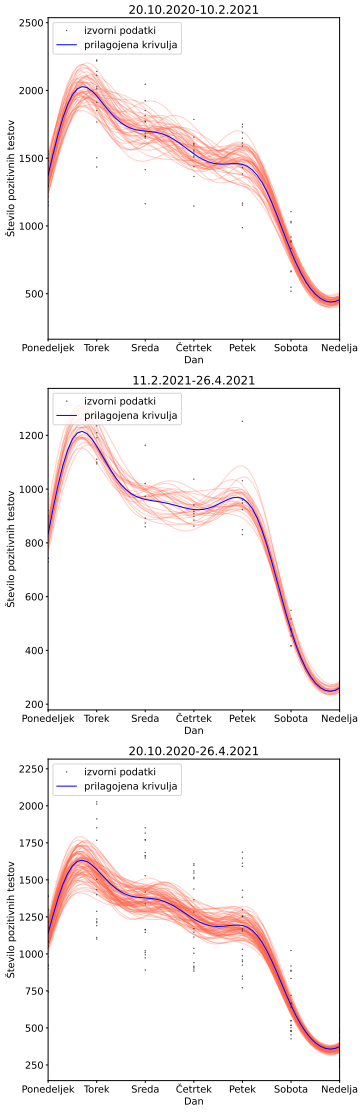
[2] `https://github.com/ninavelikajne/Ritmicnosti-stevnih-podatkov`

Tabela 1: Ocenjeni parametri ritmičnosti za posamezno podatkovno zbirko.

pod. zbirka	tip modela	št. komponent	amplituda	mesor	vrhovi	št. poz. testov
1.	gen_poisson	3.0	794.62	1233.67	0.7	2028.29
2.	gen_poisson	3.0	483.37	731.32	0.7 3.9	1214.69 969.59
3.	gen_poisson	3.0	637.66	995.12	0.7	1632.77

Tabela 2: Intervali zaupanja parametrov ritmičnosti za posamezno podatkovno zbirko.

pod. zbirka	amplituda	mesor	vrhovi	št. poz. testov
1.	[727.44 863.81]	[1164.72 1299.48]	[0.63 0.85]	[1897.05 2158.4]
2.	[452.11 514.36]	[701.36 757.49]	[0.56 0.83] [1.96 4.82]	[1154.14 1271.19] [902.71 1040.37]
3.	[596.96 691.66]	[955.36 1051.93]	[0.59 0.86]	[1556.13 1739.78]



Slika 1: Zmagovalni modeli za posamezno podatkovno zbirko. V vseh primerih je število komponent enako 3 in tip modela generaliziran Poissonov model. Oranže črte označujejo intervale zaupanja modela.

stavljena iz dveh delov. Prvi del predstavlja analizo periodičnih podatkov. Implementirana je metoda cosinor, ki ji lahko uporabnik nastavlja poljubno število komponent. Drugi del zajema analizo številnih podatkov. Uporabljeni so različni računski regresijski modeli, ki so primerni za delo s številnimi podatki. Implementirali smo pet tovrstnih modelov, tj. Poissonov model, generaliziran Poissonov model, Poissonov model z inflacijo ničel, negativen binomski model in negativen binomski model z inflacijo ničel. Vzpostavljena metodologija omogoča vrednotenje zgrajenih modelov. Tudi vrednotenje se tako kot grajenje modelov deli na dva dela. V prvem delu se osredotočimo na iskanje optimalnega števila komponent za metodo cosinor. Izhajamo iz dejstva, da so modeli gnezdeni in jih zato lahko ovrednotimo s F testom. V drugem delu iščemo najbolj ustrezen tip računskega modela. Uporabimo Vuongov test, ki je primeren tako za gnezdene, negnezdene in tudi prekrivajoče se modele.

Računsko metodo smo preizkusili na tedenski periodičnosti števila obolenj z boleznijo COVID-19 v Sloveniji. Podatke smo razdelili na 3 podatkovne zbirke, vsaka opisuje različno obdobje. Za vse podatkovne zbirke se kot najboljši tip modela izkaže generaliziran Poissonov model, optimalno število komponent pa je število 3 (glej Sliko 1). V podatkih se ritmičnost podatkov lepo izraža. Ocenili smo parametre ritma (glej Tabelo 1) in njihove intervale zaupanja (glej Tabelo 2). Z uporabo predstavljene metode na dobljenih rezultatih razberemo, da je število pozitivnih testov največje ob torkih nato pa skozi teden upada. Kot pričakovano je število pozitivnih testov najmanjše ob vikendih. Oblika zaznanega ritma je za vse podatkovne zbirke podobna (glej Sliko 1).

Celotna metoda je implementirana kot modul v jeziku Python in je prosto dostopna. Omogoča širok spekter funkcionalnosti za analizo tovrstnih podatkov. Potencialni uporabnik lahko direktno spreminja in prilagaja funkcionalnosti glede na svoje potrebe.

Literatura

[1] Bingham, C., Arbogast, B., Guillaume, G. C., Lee, J. K., and Halberg, F. Inferential statistical methods for estimating and comparing cosinor parameters. *Chronobiologia 9*, 4 (1982), 397–439.

[2] Cameron, A. C., and Trivedi, P. K. Econometric models based on count data: Comparisons and applications of some estimators and tests. *Journal of Applied Econometrics 1*, 1 (1986), 29–53.

[3] Cameron, A. C., and Trivedi, P. K. *Regression analysis of count data*, vol. 53. Cambridge University Press, 2013.

[4] Chang, S., Pierson, E., Koh, P. W., Gerardin, J., Redbird, B., Grusky, D., and Leskovec, J. Mobility network models of covid-19 explain inequities and inform reopening. *Nature 589*, 7840 (2021), 82–87.

[5] Clark, T. E., and McCracken, M. W. Tests of equal forecast accuracy and encompassing for nested models. *Journal of econometrics 105*, 1 (2001), 85–110.

[6] Consul, P., and Famoye, F. Generalized Poisson regression model. *Communications in Statistics - Theory and Methods 21*, 1 (1992), 89–109.

[7] Coxe, S., West, S. G., and Aiken, L. S. The analysis of count data: A gentle introduction to Poisson regression and its alternatives. *Journal of personality assessment 91*, 2 (2009), 121–136.

[8] Gardner, W., Mulvey, E., and Shaw, E. Regression analyses of counts and rates: Poisson, overdispersed Poisson, and negative binomial models. *Psychological bulletin 118* (12 1995), 392–404.

[9] Gardner, W., Mulvey, E., and Shaw, E. Regression analyses of counts and rates: Poisson, overdispersed Poisson, and negative binomial models. *Psychological bulletin 118* (12 1995), 392–404.

[10] Greene, W. H. *Accounting for excess zeros and sample selection in Poisson and negative binomial regression models.* Working Paper EC-94-10. Leonard N. Stern School of Business, New York University, 1994.

[11] Hilbe, J. M. *Negative binomial regression.* Cambridge University Press, 2011.

[12] Hutchison, A. L., Maienschein-Cline, M., Chiang, A. H., Tabei, S. A., Gudjonson, H., Bahroos, N., Allada, R., and Dinner, A. R. Improved statistical methods enable greater sensitivity in rhythm detection for genome-wide data. *PLoS Comput Biol 11*, 3 (2015), e1004094.

[13] Lambert, D. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics 34*, 1 (1992), 1–14.

[14] Martin, P., Cortina-Borja, M., Newburn, M., Harper, G., Gibson, R., Dodwell, M., Dattani, N., and Macfarlane, A. Timing of singleton births by onset of labour and mode of birth in nhs maternity units in england, 2005–2014: A study of linked birth registration, birth notification, and hospital episode data. *PloS One 13*, 6 (2018).

[15] Nelson, W., LEE, J. K., et al. Methods for cosinor-rhythmometry. *Chronobiologia 6*, 4 (1979), 305–323.

[16] Refinetti, R., Cornélissen, G., and Halberg, F. Procedures for numerical analysis of circadian rhythms. *Biological rhythm research 38*, 4 (2007), 275–325.

[17] Ruben, M. D., Francey, L. J., Guo, Y., Wu, G., Cooper, E. B., Shah, A. S., Hogenesch, J. B., and Smith, D. F. A large-scale study reveals 24-h operational rhythms in hospital treatment. *Proceedings of the National Academy of Sciences 116*, 42 (2019), 20953–20958.

[18] Thaben, P. F., and Westermark, P. O. Detecting rhythms in time series with RAIN. *Journal of biological rhythms 29*, 6 (2014), 391–400.

[19] Vuong, Q. H. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica 57*, 2 (1989), 307–333.

[20] Zhdanova, I. Melatonin. In *Encyclopedia of the Neurological Sciences (Second Edition)*, M. J. Aminoff and R. B. Daroff, Eds., second edition ed. Academic Press, Oxford, 2014, pp. 1030 – 1033.