Potencijal primene *Big Data* koncepta u domenu zvanične statistike

Tijana Čomić,
Aleksandar Đoković
Dragan Vukmirović
Fakultet organizacionih nauka, Univerzitet u Beogradu
Beograd, Republika Srbija
tijana.comic@gmail.com
djokovic.aleksandar@fon.bg.ac.rs
vuk@fon.bg.ac.rs

Sažetak - **Centralni problem koji se razmatra u ovom radu predstavlja istraživanje mogućnosti unapređenja sistema zvanične statistike primenom Big Data tehnologija. Ubrzani razvoj IKT, ekspanzija Interneta i društvenih mreža doveo je do informacione eksplozije – pojave velike količine podataka koji su na raspolaganju praktično svima, što ukazuje na neophodnost uvođenja inovacija u procese proizvodnje zvaničnih statističkih podataka. Većina ovih podataka su rasprostranjeni po globalnoj mreži bez reda i strukture i pred istraživačima je izazov da ih prikupe i obrade na valjan način.**

Ključne riječi-zvanična statistika; istraživanje; podaci; informacije; Big Data.

I. UVOD

Savremene tehnologije, pre svega internet, mobilna telefonija i društveni mediji, dale su veliki podsticaj razvoju društva, kreirajući potrebe koje do skora nisu postojale ili su bile na znatno nižem nivou. U prethodnom periodu razvoj civilizacije, ekonomije i društva uopšte, bio je značajno sporiji a informacije nisu bile dostupne svima, pogotovo ne u isto vreme. Pristup informacijama dovodio je do segmentacije, dovodeći u bolji položaj sve oni koji su im imali pristup. Međutim, danas je pristup informacijama omogućen gotovo svima. Dovoljno je imati pristup internetu, znanje kako da se informacije pronađu a ostalo je pitanje tehnike i brzine dolaska do informacija. Razvoj informatičkog društva ide upravo u smeru što bržeg dobijanja informacija, uz istovremeno smanjivanje troškova da se do njih dođe.

Od mnoštva podataka koji su dostupni korisnicima, većina je data u nestrukturiranoj formi (tekst, slika, multimedijlni zapis i sl.). Da bi se ovako strukturirani podaci koristili moraju se prethodno obraditi. Obradom podataka se bavi statistika na tradicionalni način, držeći se važeće naučne metodologije, u skladu sa kojom se prikupljeni podaci izražavaju na različitim mernim skalama. Pri tome, najveći broj metoda statističke obrade i analize se obavlja nad numeričkim obeležjima (mereni na ordinalnim ili intervalnim skalama), koji su visoko strukturirani. Na istim naučnim postavkama bazira se i metodološki pristup na kojima se zasniva zvanična statistika.

II. ZVANIČNA STATISTIKA

Zvanična statistika jeste neophodan element u informacionom sistemu demokratskog društva koji snabdeva vladu, ekonomiju i javnost podacima o ekonomskoj, demografskoj i socijalnoj situaciji i stanju životne sredine. U tom cilju zvanične statističke agencije obezbeđuju i na nepristrasnoj osnovi čine dostupnom zvaničnu statistiku koja ispunjava zahtev praktične korisnosti, uvažavajući pravo građana na javnu informaciju [1].

Razvoj zvanične statistike se može podeliti u tri razdoblja:

- prvo, koje se bazira na tradicionalnim metodama prikupljanja podataka koristeći uzorak i teoriju velikih brojeva, metode ocenjivanja i sl.;
- drugo, u kome dolazi do korišćenja administrativnih izvora podataka i objedinjavanja podataka kojima raspolažu različiti proizvođači podataka u jednu (ili više baza) korišćenjem jedinstvenog ključa (matičnog broja lica, PIB-a firme i slično);
- treće, koje je tek u začetku, a inicira korišćenje podataka sa interneta ili podataka koji su produkt modernih tehnologija, uređaja, društvenih medija i sl.

Interesovanje istraživača za nove izvore podataka, bazirane prvenstveno na *Big Data* tehnologijama, počelo je paralelno sa razvojem internet tehnologija, mobilne telefonije i društvenih medija. Termin *Big Data* se prvi put pojavljuje 1997. u radu naučnika iz *NASA* [2], gde oni opisuju probleme sa kojima se suočavaju prilikom vizualizacije podataka koje poseduju, s obzirom na to da su te baze podataka prilično velike, što stvara problem sa memorijom računara pa čak i na udaljenim, spoljnim diskovima. Te podatke, koje ne mogu da smeste na memoriju koja im je na raspolaganju nazvali su Big Data. Od tada se u krugovima koji se bave pitanjem baza podataka koje su toliko velike i kompleksne da ih je

nemoguće obraditi tradicionalnim statističkim softverima ustalio termin „Big Data“ [3].

U međuvremenu, korisnici zvaničnih statističkih podataka, istraživači, kreatori politika i ostali, sve više koriste mogućnosti za upotrebu mnoštva podataka koji ne dolaze iz izvora zvanične statistike i pretvaraju ih u njima korisne informacije. Osim dostupnosti, ove izvore podataka karakteriše i brzina diseminacije. Bez obzira na jasno odsustvo metodoloških postavki, koje mogu dovesti (i dovode) u pitanje kvalitet dobijenih podataka, korisnici navode da tako dobijeni podaci mogu biti od koristi za brzo identifikovanje problema, potreba, pružanje usluga, ali i za predviđanje i sprečavanje kriza, a radi dobrobiti stanovništva.

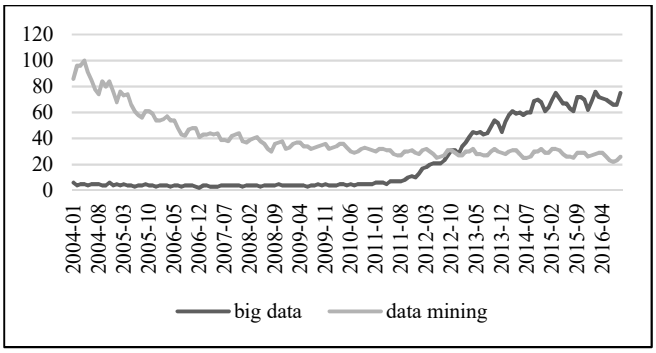
Iz navedenog jasno sledi neophodnost pune implementacije Big Data u sistem zvanične statistike, čime bi se obezbedila primena metodološkog okvira za upotrebu novih izvora podataka. Da bismo došli u fazu implementacije Big Data u sistem zvanične statistike, ne samo kao novog izvora podataka, neophodno je obezbediti da ti podaci ispunjavaju određene kriterijume kvaliteta [4]: relevantnost, nepristrasnost, dostupnost, poverljivost i sl.

Zvanična statistička istraživanja karakteriše vrlo čvrst zakonodavni i metodološki okvir koji mora biti ispoštovan o ma kojim izvorima podataka da je reč.

III. BIG DATA KONCEPT

Imajući u vidu da je koncept *Big Data* relativno skorijeg datuma, teško je pričati o „istorijatu“ razvoja *Big Data*. Koncept *Big Data* je tek u povoju i u većini slučajeva još uvek je na konceptualnom nivou, dok se prava implementacija Big Data tek očekuje kada se na svim nivoima ispune pretpostavke za njihovo korišćenje. Iako podaci kao pojam nisu ništa novo, kao ni činjenica da ih je vremenom sve više i više, kada je termin „*Big Data*“ prvi put upotrebljen i od strane koga, ne može se sa sigurnošću reći. Ipak, neki izvori navode da je vrlo moguće da je to bio Džon Maši (*John Mashey*), sredinom 1990-ih godina prošlog veka, koji je u tom periodu bio vodeći stručnjak u kompaniji *Silicon Graphics, Inc*. iz SAD [5]. Skoro deceniju i po kasnije, od 2010-2012. godine, *Big Data*, kao pojam i uopšte tema, postaje jedna od najaktuelnijih u IT svetu, a interesovanje stručnjaka (pa i onih koji to nisu), kao i bavljenje istom, u konstantnom je porastu.

Koristeći jedan od najznačajnijih izvora *Big Data* – internet i njegov najznačajniji brend (*Google*) možemo zaključiti da se termin *Big Data* uočljivo nameće 2011 godine, da bi samo za godinu dana nadmašio “*data mining*” kao do tada vodeći naučni postupak za naprednu analizu podataka i nastavio sa eksponencijalnim rastom (ostavljajući “*data mining*” u blagom linearnom padu) (Sl.1).



Slika 1. Pretraživanje termina na Google-u

Izvor: https://www.google.rs/trends (stranica posećena 1.10.2016)

A. *Osnovne definicije*

S obzirom da ne postoji jedna, opšte prihvaćena, definicija pojma *Big Data*, u ovom radu navodimo nekoliko odabranih, radi stvaranja opšte i što kompletnije slike o tome šta *Big Data* zapravo predstavlja. Većina ovih definicija su u osnovi slične.

Big Data je termin koji se koristi kada se govori o kompilaciji podataka i informacijama, koje su toliko velike i kompleksne da je teško vršiti njihovu obradu pomoću standardnih, trenutno dostupnih metoda i alata za uređivanje podataka. Pod teškoće spadaju prikupljanje, ažuriranje, skladištenje, pretraživanje, grafički prikaz, utvrđivanje raspodele, i naročito analiza podataka. Trend korišćenja velikih kompilacija podataka i međusobno povezanih informacija se nastavlja zbog pojave novih, dodatnih informacija koje se dobijaju njihovom obradom, u poređenju sa obradom manjih, odvojenih kompilacija koje ukupno sadrže istu količinu podataka. Kroz obradu većih skupova moguće je uočiti poslovne trendove, ustanoviti kvalitet tržišnih ili drugih istraživanja, sprečiti bolesti, boriti se protiv kriminala, ustanoviti saobraćajne uslove u realnom vremenu, itd. [6]

Big Data je termin koji se koristi da opiše eksponencijalni rast i dostupnost podataka, kako struktuiranih, tako i nestruktuiranih. Big Data je bitan u poslovanju, kao i u društvu, u istoj meri kao internet, jer više podataka omogućava preciznije sprovođenje analiza podataka u raznim oblastima [7].

Big Data predstavlja veliku količinu podataka koji se proizvode iz mnogih izvora, alarmantnom brzinom, obimom i raznovrsnošću a kao rezultat različitih digitalnih procesa i društvenih medija. Kako bi se značajne i vredne informacije izvukle iz *Big Data*, neophodno je imati optimalnu moć obrade, analitičke mogućnosti i veštine [8].

Ono što je zajedničko za većinu definicija Big Data je upotreba u nekom obliku tri “V“, koja predstavljaju početna slova od engleskih reči: *Volume* (obim), *Variety* (raznovrsnost) i *Velocity* (brzina). Ovu definiciju uveo je još 2001, analitičar iz IT industrije Dag Lani (*Doug Laney*),

analitičar zaposlen u *Gartner*-u, da bi danas bila opšte prihvaćena u IT industriji [9].

Obim (*Volume*) podataka. Do povećanja obima podataka uglavnom dolazi usled ubrzanog razvoja IKT.

Generalno, izvori Big Data se mogu klasifikovati na sledeći način [10]:

a) Administrativni izvori
b) Transakcioni (podaci koji su rezultat transakcija između dva entiteta, bilo da je reč o licima ili poslovnim subjektima), npr. transakcije sa kreditnim karticama, on-line transakcije (uključujući i one koje se obavljaju putem mobilnih telefona), itd.
c) Podaci sa spoljnih senzora, npr., senzori na putevima, senzori za klimatske uslove, itd.
d) Satelitski snimci, uključujući hidrometeorološke podatke
e) Sa uređaja za praćenje, npr. podaci o kretanju sa mobilnih telefona, GPS, itd.
f) Bihevioralni, npr. on-line pretrage (o proizvodima, uslugama ili nekim drugim podacima), pregledi on-line stranica itd.
g) Stavovi, npr. komentari na socijalnim medijima (facebook, tweeter, youtube) itd.
h) Internet inteligentnih uređaja (IoT)

Raznovrsnost (*Variety*). Podaci se u današnje vreme javljaju u raznim oblicima Razvoj informacionih tehnologija, pre svih doveo je do pojave različita vrsta podataka koji se mogu prikupiti. Jedna od osnovnih podela je na struktuirane, polustruktuirane i nestruktuirane podatke.

*a) **Struktuirani podaci** su postojeći podaci, smešteni u bazama podataka, po svim pravilima skladištenja podataka.*
*b) **Polu-strukturiranih podaci** se koriste za opisivanje strukturiranih podatka koji se ne uklapa u formalnu strukturu modela podataka. Ovi podaci ne sadrže oznake koje razdvajaju semantičke elemente, a koji poseduju sposobnost sprovođenja hijerarhije unutar podataka.*
*c) **Nestrukturirani podaci** su u osnovi informacije koje nemaju unapred definisani model podataka (meta podatke) i/ili se dobro ne uklapaju bazu podataka. Nestrukturirani podaci su obično tekstualni ili multimedijalni podaci, ali mogu biti i numerički – kao što su datumi, brojevi, i sl.*

Konkretno, pod nestrukturiranim podacima se podrazumevaju tekst, audio, video, slike, geoprostorni podaci, internet podaci, *click streams*, *log*-ovi i sl.

Brzina (*Velocity*). Vreme koje je neophodno da se dobije krajnji rezultat istraživanja u značajnoj meri definiše Big Data koncept. Pod brzinom se u Big Data konceptu pretpostavlja realno ili približno realno vreme za dostavljanje (praćenje) rezultata za razliku od tradicionalnog istraživanja gde se na konačan rezultat čeka danima, neretko mesecima, čak i godinama (npr. zvanična statistika – BDP obračun)

IV. PRIMENA BIG DATA U ZVANIČNOJ STATISTICI

Imajući u vidu sve dobrobiti informatičke revolucije, zvanična statistika će, iako po mnogima rigidan sistem, što pre morati da se prilagodi promenama i iskoristi blagodeti novih tehnologija kako bi išla u korak sa potrebama donosioca odluka i kako bi uz smanjenje troškova, povećanje kvaliteta i pravovremeno mogla da odgovori potrebama donosioca odluka ali i drugih korisnika podataka. Zvanična statistika ulazi u treću fazu svog razvoja u kojoj će se pored popisa, istraživanja na uzorku i korišćenja administrativnih izvora podataka koristiti i podaci koji su dostupni, kao sekundarni podaci, na internetu [11].

Pad stope odgovora prilikom anketnih istraživanja, bilo da se radi o istraživanjima koja se vrše na domaćinstvima ili istraživanjima u okviru poslovnog sektora, sve je izraženiji. U tom smislu korišćenje Big Data donosiocima odluka može doneti informacije u realnom vremenu naročito u oblastima kao što su statistika cena, zaposlenosti, industrijske proizvodnje, demografije i sl. [12] Big Data imaju potencijal pružanja relevantnijih i blagovremenije podatke nego što je to slučaj sa tradicionalnim metodama kao što su ankete i administrativni izvori. Međutim, većina izvora Big Data je u vlasništvu privatnog sektora te je neophodno zakonodavstvo koje bi omogućilo korišćenje ovih podataka u zvaničnoj statistici [13]. Nepostojanje zakonodavstva je jedan od glavnih razlog zašto Big Data podaci nisu još uvek u široj primeni u zvaničnoj statistici.

Bez obzira na velike mogućnosti Big Data, ne treba razmišljati o ovim podacima kao o zameni za tradicionalne metode prikupljanja podataka, već više kao mogućeg izvora koji bi upotpunio statistički sistem. I pored pokušaja da Big Data pruže što iscrpnije podatke, pre svega kad se radi o podacima koji dolaze sa socijalnih medija, oni su, po svojoj prirodi uvek parcijalna, sa različitim prazninama, pristrasnostima i neizvesnostima. Big Data su proizvedeni od strane sistema koji su dizajnirani i testirani u određenim naučnim okvirima i okruženi mnoštvom različitih konteksta i interesa [14].

Treba imati na umu da Big Data nikada ne govore sami za sebe i neophodno ih je sistematizovati i na pravi način izvući odgovarajuće informacije iz njih za tačno definisan deo populacije.

Dakle, da bi se Big Data koristili kao dodatni izvor podataka u zvaničnoj statistici, neophodno je da se ispoštuju sve relevantne faze GSBPM i da se ispune svi principi koda prakse.

UN i *Eurostat* kroz osnivanje radnih grupa visokog nivou u velikoj meri već rade na istraživanjima mogućnosti

korišćenja ovog koncepta za potrebe zvanične statistike [15]. Modernizacija statistike se odvija u smeru podrške praćenju indikatora održivog razvoja i koncepta „*Data Revolution* za održivi razvoj“ [16].

A. Statistika saobraćaja i statistika transporta

Primer koji se ovde navodi za slučaj Holandije u okviru statistike saobraćaja je, kasnije rezultirao u objavljivanju prvih ikada zvaničnih statističkih podataka koji su dobijeni na bazi Big Data u Holandiji [17]. U Holandiji postoji preko 60,000 senzora (*road sensors*) od kojih je 20,000 na autoputevima koji služe za brojanje vozila različitih veličina, svakog minuta. Prilikom obrade ovih podataka pokazalo se da sam kvalitet podataka značajno varira, kako iz minuta u minut tako i iz dana u dan. Kako bi se ovaj problem prevazišao razvijeni su određeni filteri koji su podešeni na stohastičko ponašanje prilaska vozila u okvir senzora. Korigovanjem podataka i kombinovanjem dnevnog „profila“ koje daje senzor na jednom putu, pokrivenost i kvalitet podataka su unapređeni. Na osnovu ovoga moguće je izvesti indekse stanja na holandskim putevima na regionalnom nivou [18].

B. Statistika društvenih medija

Ovaj primer je, kao i prethodni, na podacima iz Holandije. U Holandiji se dnevno objavi preko tri miliona poruka na javnim socijalnim medijima [19]. Na društvenim (socijalnim) medijima pojedinci razmenjuju informacije, učestvuju u diskusijama i komuniciraju sa svojim prijateljima i porodicom. Kako bi istražili koje polje zvanične statistike bi moglo biti pokriveno iz ovog izvora podataka, sve objave su sagledavane iz dva ugla: sadržaj i oblasti. Najčešće korišćen medij u Holandiji je *Twitter* te su i korišćeni podaci sa ovog medija. Analiza je pokazala da je oko 50% „besmisleno brbljanje“, dok je ostatak poruka bio u vezi aktivnosti u slobodno vreme (10%), posao (7%), mediji (5%) i politika (3%). Kasnije studije su pokazale da je raspoloženje na holandskim društvenim medijima visoko korelisano sa poverenjem potrošača. Posmatrano raspoloženje je bilo stabilno kada se posmatra na mesečnom ili nedeljnom nivou ali ne i kada se posmatra na dnevnom nivou.

Kada je reč o društvenim medijima kao izvorima podatak za Big Data najveći problem je pokrivenost populacije. Uzmimo za primer Tweeter. Ma koliko on bio globalno rasprostranjen i mesto gde pojedinci izražavaju svoje stavove o mnoštvu tema, korisnici Tweeter su samo jedan, specifičan deo društva. Sa druge strane, veliki broj korisnika Tweeter-a nisu aktivni već pasivni korisnici, odnosno ne objavljuju svoje „Tweet-ove“ (stavove) već samo prate druge. Prema studiji *An Exhaustive Study of Twitter Users Across the World* [20] koja je sprovedena na 36 miliona korisnika, 25% korisnika Tweeter-a, nikada nije *Tweet*-ovalo. Pored toga, 74% korisnika je starosti od 15 do 25 godina starosti, a samo 6% spada u kategoriju starih 46 i više godina. Otuda, preko Tweeter-a dostupni su nam samo podaci o specifičnom delu društva, ne ulazeći ovde u specifični profil aktivnih korisnika.

C. Statistika cena

Primer koji se ovde navodi je korišćenje *web scraping* tehnike za automatsko prikupljanje podataka o cenama sa interneta. *Billion Prices Project* je projekat Instituta za tehnologiju u Masačusetsu (*MIT*) kroz koji se korist cene sakupljene od stotina on line prodavnica na dnevnom nivou kako bi se sprovela ekonomska istraživanja [21].

D. Statistika turizma

Ideja ove studije izvodljivosti bila je da se istraži koliko podaci o poziciji mobilnih telefona mogu biti od koristi u statistici turizma kao i ocena njihovih prednosti i nedostataka [22]. Glavni zaključci ove studije bili su [23]:

a) Podaci o poziciji mobilnih telefona su veoma limitirani pre svega zbog zakonodavnih ograničenja.

b) Neophodni su longitudinalni podaci kako bi se prikupili što precizniji podaci o kretanju pretplatnika.

c) Podatke mobilne telefonije je bolje koristiti kao dodatni nego kao jedinstven izvor podataka za obračun indikatora iz oblasti turizma.

d) Korišćenjem ovog izvora kao dodatnog, poboljšava se pravovremenost, omogućen je pristup podacima koji su ranije bili nedostupni (čime je omogućen obračun novih indikatora), poboljšavaju se mogućnosti za kalibraciju podataka, bolja „rezolucija“ (podaci na nižem nivou) i veća preciznost u prostoru i vremenu.

e) I druge statističke oblasti mogu imati korist od ovog izvora podataka.

E. Korišćenje informaciono komunikacionih tehnologija (IKT)

Ideja ove studije izvodljivosti je da se testiraju mogućnosti korišćenja podataka sa interneta za izradu statistika iz oblasti korišćenja informaciono komunikacionih tehnologija.

F. Podaci sa pametnih brojila

Pametna brojila su elektronska brojila koja omogućavaju automatsko prikupljanje podataka o potrošnji električne energije u domaćinstvima i malim poslovnim subjektima. Ova pametna brojila omogućavaju uvid u potrošnju električne energije u svakom trenutku [24].

Primena brojača bi trebalo da bude šire razmatrana imajući u vidu prednosti IoT i činjenice da u današnje vreme, mnogi kućni aparati komuniciraju što međusobno, što sa nekim centralnim serverom. *People metre*-i su odavno u primeni u marketinške svrha ali bi se mogla razmotri njihova primena i za potrebe društvenih statistika u okviru zvanične statistike.

G. Podaci o otvorenim radni mestima

Podaci o otvorenim radnim mestima najčešće su dostupni na veb stranicama za zapošljavanje, ko što su stranice nacionalnih zavoda za zapošljavanja ili privatnih veb portala koji se bave ovim pitanjima. Ideja eksperimenta je bila da se koristeći web scraping tehniku, scrap-uju podaci sa najvećih sajtova za zapošljavanje. Za implementaciju ovog koncepta neophodno je ostvariti partnerstva sa vlasnicima portala jer je time omogućeno prikupljanje detaljnijih i tačnijih podataka [25].

H. Pregledi na on lajn enciklopediji Wikipedia

Wikipedia je najposećenija on-lajn enciklopedija, a ovaj projekat se fokusirao na preglede stranica posvećenih mestima koji pripadaju UNESCO svetskoj zaostavštini. Ovim eksperimentom dobijaju se podaci koji su relevantni za statistiku kulture i regionalne statistike. Ovi podaci pokrivaju segmente koji ranije nisu bili pokriveni statistikom kulture.

I. Mobilnosti lica

Rikato i dr. (2015) su u studiji koja se bavi merenjem gustine naseljenosti na bazi upotrebe mobilnih telefona [26] opisali tehnologiju na bazi koje se prikupljaju podaci o korišćenju mobilne telefonije. Naime, infrastruktura operatera mobilnih mreža (OMM) sastoji se iz velikog broja „ćelija“ različitih veličina, koje pokrivaju prostor koji može biti veličine od 10 metara do nekoliko kilometara. Ove ćelije emituju signal koji primaju mobilni telefoni tako da nakon bilo kog „događaja“ – primanje ili upućivanje poziva, primanje ili slanje SMS-ova – mobilni telefon otkriva u okviru koje ćelije se nalazi i ova informacija se zauvek čuva u takozvanoj *Call Detail Record* (CDR) bazi podataka kako bi „događaj“ bio naplaćen. Pored ovoga mreža beleži i prelazak iz jedne u drugu ćeliju.

J. Prognoza stopa nezaposlenosti

U ovom radu [27] korišćen je *Google trends* radi poboljšavanja predikcije stope nezaposlenosti, polazeći od jednostavnog autoregresionog modela koji uključuje jednu docnju stope nezaposlenosti. Podaci koji su korišćeni su na mesečnom nivou. Početni autoregresivni model je oblika:

$$\log(y_t) = a + b * \log(y_{t-1}) + e_t \qquad (1)$$

gde je y_t nezaposlenost u mesecu t, a i b koeficijenti, a y_{t-1} nezaposlenost u mesecu *t-1*. e_t je greška modela.

Nakon ocene ovakvog modela, u model su uključene dodatne varijable koje prestavljaju indeks pretraživanja tri različita termina na *Google trends*, a dovode se u vezu sa stopom nezaposlenosti. To su termini:

a) *„pole emploi“ – francuska agencija za zapošljavanje*
b) *„etre au chomage” – biti nezaposlen*
c) *“indemnité” – džeparac*

Kao zaključak ocenjivanja modela je da uključivanje dodatnih prediktora poboljšava kvalitet modela.

ZAKLJUČAK

Tradicionalna metodologija statističkog istraživanja koja se primenjuje u sistemima zvanične statistike bazira se na prikupljanju podataka upotrebom anketnih istraživanja, korišćenju administrativnih izvora podataka, registara, popisa i sl. Nedostatak ovih metoda je u tome što prikupljanje podataka, njihova obrada i objavljivanje rezultata istraživanja traje duže nego što bi korisnici to želeli.

U radu je analiziran potencijal korišćenja velike količine podataka (Big Data) u sistemu zvanične statistike. Ukazano je na glavne izazove u implementaciji Big Data koncepta i predložene su osnovne odrednice budućeg pravca razvoja modela istraživanja korišćenjem Big Data koncepta u zvaničnoj statistici.

Razmatrana je veza između tradicionalnog koncepta statističkog istraživanja i primene Big Data koncepta i sagledana uloga IKT u savremenom pristupu koji podrazumeva internet kao medij. Osnovni nedostaci koje je neophodno otkloniti za dalji razvoj statistike u ovom pravcu navodi se: pravni osnov za korišćenje Big Data, metodološki okvir i nedostatak stručnog kadra („naučnici za podatke“).

Kroz analizu i pregled postojećih pilot-studija i rešenja, baziranih u oblasti primene Big Data resursa, ukazano je na trenutno stanje u oblasti zvanične statistike na međunarodnom nivou.

LITERATURA

[1] Osnovni principi zvanične statistike, Izvod iz izveštaja Statističke komisije Ujedinjenih nacija sa njene Specijalne sednice, održane u Njujorku, 11–14. aprila 1994. Zvanični izveštaj Ekonomskog i socijalnog saveta, 1994, Prilog br. 9.

[2] National Aeronautics and Space Administration - www.nasa.gov

[3] B. Brown, J. Sikes & P. Willmott, „Bullish On Digital“, McKinsley Global Survey Results, 2013.

[4] ESS Standard for Quality Reports, Eurostat Metholodogies and working papers, 2009,

[5] Technology Trend Analysis, https://setandbma.wordpress.com/2013/02/04/who-coined-the-term-big-data/

[6] J.M. Cavanillas, E. Curry, W. Wahlster, Editors (2013), New Horizons for a Data-Driven Economy A Roadmap for Usage and Exploitation of Big Data in Europe, Springer Open, 2015

[7] http://www.sas.com/big-data/

[8] http://www.ibm.com/big-data/us/en/

[9] http://www.gartner.com/analyst/40872/Douglas-Laney

[10] What Does “Big Data” Mean For Official Statistics?, High-Level Group for the Modernisation of Official Statistics UNECE, 2013.

[11], [13] United Nations “Big data and modernization of statistical systems”, Report of the Secretary-General, Statistical Commission, Forty-fifth session, 4-7 March 2014

[12] Organization for Economic Cooperation and Development (OECD), “Exploring data-driven innovation as a new source of growth: mapping the policy issues raised by big data”, 2013.

[14] R. Kitchin, “Big data should complement small data, not replace them”,

[15], [16] http://www1.unece.org (stranica posećena 5.12.2016).

[17] “A first for Statistics Netherlands: launching statistics based on Big Data”, https://www.cbs.nl/NR/rdonlyres/4E3C7500-03EB-4C54-8A0A-753C017165F2/0/afirstforlaunchingstatisticsbasedonbigdata.pdf (stranica posećena 5.12.2016).

[18] P. J. H. Daas, M. Puts, M. Tenneks,and A. Priem, “Big Data as a Data Source for Official Statistics: experiences at Statistics Netherlands”, The Survey Statistician, Proceedings of Statistics Canada Symposium 2014: Beyond traditional survey taking: adapting to a changing world, доступно на:

[19] Daas, P. J. H., Puts, M., Tenneks, M. and Priem, A. “Big Data as a Data Source for Official Statistics: experiences at Statistics Netherlands”, The Survey Statistician, Proceedings of Statistics Canada Symposium 2014: Beyond traditional survey taking: adapting to a changing world, 2014.

[22] An Exhaustive Study of Twitter Users Across the World, http://www.beevolve.com/twitter-statistics/, (stranica posećena 1.10.2016).

[21] http://bpp.mit.edu/, (stranica posećena 7.12.2016).

[22] What Does “Big Data” Mean For Official Statistics?, 2013, http://www1.unece.org/stat/platform/pages/viewpage.action?pageId=77170622 (stranica posećena 5.12.2016).

[23] Feasibility Study on the Use of Mobile Positioning Data for Tourism Statistics Consolidated Report Eurostat Contract No 30501.2012.001-2012.452, 2014, Eurostat,

[24], [25], [26] http://www1.unece.org/stat/platform/display/bigdata/-Experiment+Reports, (stranica posećena 10.12.2016).

[27] F. Ricciato, P. Widhalm, M. Craglia and F. Pantisano, “Estimating Population Density Distribution from Network-based Mobile Phone Data”, dostupno na: https://ec.europa.eu/eurostat/cros/system/files/Final-%20jrc-AIT-MNO-study-compressed_1.pdf (stranica posećena 23.10.2016).

ABSTRACT

The central problem that is being addressed in the paper is exploring possibilities of improving the system of official statistics using Big Data. The rapid development of information and communication technologies, the expansion of the Internet and social networks has led to an explosion of information - the appearance of large amounts of data that are practically available to everyone, indicating the necessity of introducing innovations in the production processes of official statistics. Most of these data are dispersed on global network without order and structure, and the challenge that researchers are facing with is to collect and process those data in a proper way.

IMPROVEMENT OF OFFICIAL STATISTICS BY APPLYING THE CONCEPT OF BIG DATA

Tijana Čomić,
Aleksandar Đoković
Dragan Vukmirović